Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FACULTY
OF MATHEMATICS, INFORMATICS
AND NATURAL SCIENCES

Language Technology

DISSERTATION

On Strengths and Limitations of Multimodal Transformer Encoders and Decoders for Vision-Language Tasks

Florian Schneider

Language Technology

Department of Informatics

Faculty of Mathematics, Informatics and Natural Sciences

Universität Hamburg
Hamburg, Germany

A cumulative thesis submitted for the degree of

Doctor rerum naturalium (Dr. rer. nat.)

2025

On Strengths and Limitations of Multimodal Transformer Encoders and Decoders for Vision-Language Tasks

Dissertation submitted by: Florian Schneider

Date of Submission: 16.06.2025
Date of Disputation: 02.10.2025

Supervisor:

Prof. Dr. Chris Biemann, Universität Hamburg

Universität Hamburg, Hamburg, Germany
Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics

Language Technology

Affidavit

Eidesstattliche Versicherung:

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

Affidavit:

I hereby declare and affirm that this doctoral dissertation is my own work and that I have not used any aids and sources other than those indicated. If electronic resources based on generative artificial intelligence (gAI) were used in the course of writing this dissertation, I confirm that my own work was the main and value-adding contribution and that complete documentation of all resources used is available in accordance with good scientific practice. I am responsible for any erroneous or distorted content, incorrect references, violations of data protection and copyright law or plagiarism that may have been generated by the gAI.

Date

Signature
(Florian Schneider)

This thesis is dedicated to my mother for her unwavering lifelong support
and to my grandfather, who sparked my curiosity and verve for science and nature.

Acknowledgements

This thesis and my whole academic journey would not have been possible without the constant support of many people.

First and foremost, I would like to thank my supervisor, Prof. Dr. Chris Biemann. Only thanks to his initial and persistent support, guidance, and encouragement throughout from the very beginning to the end of my PhD and beyond, I started this journey and was finally able to succeed. Besides his professional support, I am also grateful for his emotional support, which encouraged me to believe in myself and reach out for goals that I would not have dared to pursue otherwise. I am further especially thankful for his support in pursuing my two internships at IBM Research in New York and Microsoft Research in Bangalore, which were both life-changing and inspiring experiences. Lastly, I deeply appreciate his commitment to resolving any administrative matters and ensuring me a smooth and pleasant PhD experience.

I would also like to extend my sincere gratitude to Prof. Dr. Anne Lauscher, who kindly agreed to collaborate with me and her team after I reached out, resulting in a fruitful collaboration that finally resulted in multiple joint publications central to this thesis. I am also very thankful for her invitations to several team events with her group, which made me feel very welcome and close to her amazing team.

Additionally, I want to thank my awesome colleagues at the Language Technology Group, who made my time there a truly enjoyable experience and helped me become a better Kicker player. I am especially grateful to my office mate and project partner, Tim Fischer, for the many memorable professional and personal moments we shared while developing and presenting the DATS. I also deeply appreciate taking on my workload during my two internships, which allowed me to dedicate myself fully to those experiences.

Also, I want to thank all my collaborators and co-authors who made my research journey a truly collaborative and enjoyable experience. Particularly, I want to thank Carolin Holtermann for her invaluable contributions to the GIMMICK project and for her professional and emotional support throughout the final years of my PhD.

Moreover, I would like to express my heartfelt gratitude to my mother, Anette, and my sister, Luzia, as well as my many friends for their unwavering support, patience, and encouragement throughout the many ups and downs of my life's journey that ultimately led me successfully to this point in fulfilling my PhD.

Finally, I want to express my most profound gratefulness to my girlfriend and partner-in-crime, Lisa, who has been my rock and constant source of joy and

strength throughout this journey. Her unwavering belief in me, inexhaustible patience during challenging times, and boundless love have been indispensable. I am deeply thankful for all the extra effort she put into making my life easier, especially during the final months of my PhD.

Third Party Software Acknowledgements

To write this thesis several third-party software tools were used, which I would like to acknowledge here. I wrote the thesis in LaTeX, using the VS Code IDE [1] and the `Visual Studio Code LaTeX Workshop Extension`[2]. The thesis is based on the `Universität Hamburg LT Thesis Template`[3]. To check the grammar and spelling, I used the premium version of `Grammarly`[4] through their web application. Further, for initial proof reading of single chapters, I used the premium version of `ChatGPT`[5] through the native application on my MacBook. Note that `ChatGPT` was only used to generate suggestions for grammar, spelling, or style corrections which was then incorporated manually by myself, but not to generate any content or text. For the figures, I used `draw.io`[6] through their native application on my MacBook. Lastly, I used `Google Translate`[7] to create a first draft for the German translation of the abstract, which I then manually corrected and polished.

1. https://code.visualstudio.com/
2. https://github.com/James-Yu/LaTeX-Workshop
3. https://github.com/uhh-lt/thesis-template-uhh-lt-latex/
4. https://www.grammarly.com/
5. https://www.openai.com/chatgpt
6. https://www.diagrams.net/
7. https://www.google.com/translate

Abstract

Transformer models, or their inherent attention mechanism in particular, have significantly advanced the fields of natural language processing (NLP) and computer vision (CV). Combining transformers for text processing and transformers for image processing led to multimodal transformers, enabling more natural and effective interactions between humans and Artificial Intelligence (AI), thus having great potential for real-world applications. Further, there are multimodal transformer models that can process even more modalities, such as audio, time series, graphs, or various kinds of digital sensor data. However, this thesis focuses on multimodal transformers for vision-language tasks (VLTs), i.e., models that jointly process textual and visual data from images or videos. These models can be grouped into two archetypes: Firstly introduced were Vision-Language Models (VLMs), which are based on two aligned transformer encoders, one for text and one for images, trained to compute semantically rich representation in the same multimodal embedding space. These VLMs can not generate text but are often used for discriminative tasks like image classification or cross-modal retrieval. Thanks to significant hardware improvements and the increased availability of immense computational resources, generative Large Vision-Language Models (LVLMs) were introduced after classical VLMs. LVLMs typically combine a transformer encoder for visual data processing with a transformer decoder, i.e., a pretrained Large Language Model (LLM) capturing extensive world knowledge and reasoning capabilities to enable coherent text generation conditioned on interleaved text-image inputs. These models demonstrated remarkable capabilities in various VLTs, such as visual question answering, image captioning, object detection, or optical character recognition, making them powerful tools for a wide range of real-world applications.

Despite their strengths, multimodal transformers exhibit notable limitations when faced with data significantly different from their training distribution, commonly referred to as "out-of-distribution" data. Performance degradation is particularly evident in scenarios involving diverse domains, low-resource languages, and non-Western cultural contexts. These limitations are especially problematic, considering the increasing global availability and reliance on AI systems, which require robust and equitable performance across diverse user groups and scenarios.

This dissertation investigates the strengths and limitations of multimodal transformer encoder and decoder models, i.e., VLMs and LVLMs. Through systematic experimentation, it demonstrates the practical applicability of such models in developing efficient and user-friendly cross-modal information retrieval systems for the interactive exploration of complex datasets. Conversely, analyses based on specially curated datasets designed explicitly to test robustness highlight

significant performance drops when models encounter complex data, multilingual inputs, or culturally diverse content. Further, experiments evaluating LVLMs through extensive multilingual benchmarks that span numerous low-resource languages highlight significant limitations in non-English languages across a wide range of tasks. Similarly, large-scale experiments assessing the cultural knowledge reveal pronounced biases toward Western cultures across LVLMs of all kinds and sizes.

Moreover, the thesis proposes effective strategies for addressing multilingual limitations, particularly through improved training regimens and carefully curated, massively multilingual datasets. By systematically analyzing various training configurations, including the optimal distribution and scale of multilingual training data, the dissertation successfully demonstrates strategies for creating massively multilingual LVLMs capable of maintaining high performance across various multilingual VLTs.

Overall, this thesis provides comprehensive insights into leveraging the strengths of multimodal transformer models, systematically investigates their inherent limitations, and proposes mitigation strategies to overcome restrictions in massively multilingual settings. The work presented aims to contribute towards developing more universally applicable, equitable, and robust multimodal AI systems for people from all over the world, irrespective of their language or cultural background.

Zusammenfassung

Transformer-Modelle, insbesondere deren “Attention” Mechanismus, haben die Bereiche der natürlichen Sprachverarbeitung (NLP) und des maschinellen Sehens (Computer Vision, CV) erheblich vorangebracht. Die Kombination von Transformern für die Textverarbeitung und für die Bildverarbeitung führte zu multimodalen Transformern, die natürlichere und effektivere Interaktionen zwischen Mensch und KI ermöglichen und somit großes Potenzial für reale Anwendungen bieten. Zudem existieren multimodale Transformer-Modelle, die noch weitere Modalitäten wie Audio, Zeitreihen, Graphen oder verschiedene Arten digitaler Sensordaten verarbeiten können. Diese Dissertation konzentriert sich jedoch auf multimodale Transformer für Vision-Language-Tasks (VLTs), d.h. Modelle, die textuelle und visuelle Daten aus Bildern oder Videos gemeinsam verarbeiten. Diese Modelle lassen sich in zwei Haupttypen einteilen: Zunächst wurden Vision-Language-Modelle (VLMs) erfunden, die auf zwei ausgerichteten Transformer-Encodern basieren: einer für Text und einer für Bilder. Diese Modelle sind darauf trainiert, semantisch reichhaltige Repräsentationen in einem gemeinsamen multimodalen Vektorraum zu erzeugen. VLMs können zwar keinen Text generieren, werden jedoch häufig für diskriminative Aufgaben wie Bildklassifikation oder kreuz-modale, also Text-Bild oder Bild-Text, Suche eingesetzt. Dank erheblichem Fortschritt der Hardware und der damit gesteigerten Verfügbarkeit enormer Rechenressourcen wurden nach den klassischen VLMs generative Large Vision-Language-Modelle (LVLMs) erfunden. LVLMs kombinieren typischerweise einen Transformer-Encoder zur Verarbeitung visueller Daten mit einem Transformer-Decoder, also einem großen vortrainiertem Sprachmodell, auch bekannt als Large Language Model (LLM), welches umfangreiches Wissen und logische Fähigkeiten kombiniert, um kohärente Textgenerierung auf Grundlage von Text- und Bildinputs zu ermöglichen. Diese Modelle zeigten bemerkenswerte Fähigkeiten bei verschiedenen VLTs wie der Beantwortung von Fragen auf Basis von Bildinhalten, Bildunterschriften, Objekterkennung oder optischer Zeichenerkennung und stellen somit leistungsstarke Werkzeuge für viele reale Anwendungen dar.

Trotz ihrer Stärken weisen multimodale Transformer erhebliche Einschränkungen auf, wenn sie mit Daten konfrontiert werden, die deutlich von ihrer Trainingsverteilung abweichen, also sogenannte „Out-of-Distribution“-Daten. Leistungseinbußen zeigen sich insbesondere deutlich bei Szenarien aus verschiedenen Domänen, Sprachen mit geringen Ressourcen sowie in nicht-westlichen kulturellen Kontexten. Diese Beschränkungen sind besonders problematisch angesichts der zunehmenden globalen Verfügbarkeit und Abhängigkeit von Systemen basierend auf Künstlicher Intelligenz, die eine robuste und gerechte Leistung über verschiedene Nutzergruppen und Szenarien hinweg erfordern.

Diese Dissertation untersucht Stärken und Schwächen multimodaler Transformer-Encoder- und Decoder-Modelle, also VLMs und LVLMs. Durch systematische und umfangreiche Experimente wird die praktische Anwendbarkeit solcher Modelle zur Entwicklung effizienter und benutzerfreundlicher Systeme zur kreuz-modalen Suche und interaktiven Exploration komplexer Datensätze demonstriert. Umgekehrt zeigen ausgedehnte Analysen anhand speziell kuratierter Datensätze, die explizit zur Überprüfung der Robustheit entwickelt wurden, deutliche Leistungseinbrüche, wenn die Modelle mit komplexen Daten, mehrsprachigen Eingaben oder kulturell vielfältigen Inhalten konfrontiert werden. Weiterhin verdeutlichen Experimente, welche LVLMs mit umfangreichen multilingualen Benchmarks evaluieren, erhebliche Einschränkungen bei nicht-englischen Sprachen in einer Vielzahl von Aufgaben. Ebenso zeigen großangelegte Experimente zur Bewertung des kulturellen Wissens ausgeprägte Vorurteile gegenüber westlichen Kulturen über LVLMs aller Arten und Größen hinweg.

Darüber hinaus schlägt die Dissertation effektive Strategien zur Überwindung multilingualer Einschränkungen vor, insbesondere durch verbesserte Trainingsverfahren und sorgfältig kuratierte, mehrsprachige Datensätze. Durch die systematische Analyse verschiedener Trainingskonfigurationen, einschließlich der optimalen Verteilung und des Umfangs multilingualer Trainingsdaten, demonstriert die Dissertation erfolgreich Strategien zur Entwicklung multilingualer LVLMs.

Insgesamt bietet diese Arbeit umfassende Erkenntnisse darüber, wie sich die Stärken multimodaler Transformer-Modelle nutzen lassen, untersucht systematisch ihre inhärenten Grenzen dieser Modelle und schlägt Strategien vor, um Einschränkungen in multilingualen Kontexten zu überwinden. Die vorgestellte Arbeit soll dazu beitragen, universell einsetzbare, gerechte und robuste multimodale Künstliche Intelligenz-Systeme für Menschen aus aller Welt zu entwickeln, unabhängig von ihrer Sprache oder ihrem kulturellen Hintergrund.

Contents

List of Figures ii

List of Tables iii

List of Abbreviations v

1 Introduction 1
- 1.1 Historical Background and Motivation . . . 2
- 1.2 Definitions . . . 5
- 1.3 Research Questions . . . 6
- 1.4 Publications . . . 7
  - 1.4.1 List of Foundational Papers . . . 7
  - 1.4.2 List of Related Papers . . . 8
  - 1.4.3 Degree of Authorship of Foundational Papers . . . 9
- 1.5 Contributions . . . 10
  - 1.5.1 Research Question 1: Strengths . . . 10
  - 1.5.2 Research Question 2: Limitations . . . 11
  - 1.5.3 Research Question 3: Mitigations . . . 12
- 1.6 Thesis Outline . . . 13

2 Theoretical Background 15
- 2.1 Introduction . . . 17
- 2.2 Traditional Neural Networks for Natural Language Processing . . . 17
  - 2.2.1 Language Modeling . . . 17
  - 2.2.2 Neural Networks Fundamentals . . . 19
  - 2.2.3 Neural Word Embeddings . . . 22
  - 2.2.4 Recurrent Neural Networks (RNNs) . . . 23
  - 2.2.5 Sequence-to-Sequence Models . . . 25
- 2.3 Transformer Models for Modern Natural Language Processing . . . 27
  - 2.3.1 The Transformer Architecture . . . 27
  - 2.3.2 Transformer Encoder Models . . . 31
  - 2.3.3 Transformer Decoder Models . . . 32
  - 2.3.4 Large Language Models (LLMs) . . . 33
- 2.4 Multimodal Transformer Models for Vision-Language Tasks . . . 38
  - 2.4.1 Vision Transformers (ViTs) . . . 38
  - 2.4.2 Encoder-based Vision-Language Models (VLMs) . . . 41
  - 2.4.3 Decoder-based Large Vision-Language Models (LVLMs) . . . 48
- 2.5 Summary . . . 56

3 Conclusion **57**
3.1 Summary . . . 58
3.1.1 Research Question 1: Strengths . . . 58
3.1.2 Research Question 2: Limitations . . . 59
3.1.3 Research Question 3: Mitigations . . . 60
3.2 Limitations . . . 61
3.3 Future Work . . . 62

References **65**

Papers **85**

I Golden Retriever:
A Real-Time Multi-Modal Text-Image Retrieval System with the Ability to Focus **87**

II CollEX:
A Multimodal Agentic RAG System Enabling Interactive Exploration of Scientific Collections **95**

III WISMIR3:
A Multimodal Dataset to Challenge Text-Image Retrieval Approaches **119**

IV M5 – A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks **127**

V GIMMICK:
Globally Inclusive Multimodal Multitask Cultural Knowledge Benchmarking **165**

VI Centurio:
On Drivers of Multilingual Ability of Large Vision-Language Model **231**

List of Figures

2.1 Schematic overview of a 3-layer Feed-Forward neural network architecture with bias terms and non-linearity omitted for better readability. . . . 20
2.2 A schematic overview of a Recurrent Neural Network (RNN) architecture. . . . 24
2.3 A schematic overview of a Sequence-to-Sequence (seq2seq) architecture. . . . 25
2.4 The original transformer architecture proposed as encoder-decoder seq2seq model for machine translation and its core attention mechanisms. Images taken from Vaswani et al. (2017). . . . 28
2.5 Left: An overview of the classical transformer encoder architecture. Center: Inner architecture of a transformer encoder layer with multi-head attention. Right: A random examplatory attention matrix. 32
2.6 Left: An overview of the classical transformer decoder architecture including sampling of the next token. Center: Inner architecture of a transformer decoder layer with causal multi-head attention. Right: A random examplatory causal attention matrix. . . . 33
2.7 Left: An overview of the vanilla Vision Transformer (ViT) architecture. Right: Inner architecture of a transformer encoder layer. Image taken from (Dosovitskiy et al., 2021). . . . 39
2.8 High-level overview of typical architectures for transformer encoder-based vision-language models (VLMs). The figures are inspired by (Shaikh et al., 2024). . . . 42
2.9 Architecture of the Transformer Encoder Reasoning and Alignment Network (TERAN) (Messina et al., 2021). Image taken from the original paper. . . . 44
2.10 An overview of CLIP models. Image taken from Radford et al. (2021). 46
2.11 A high-level overview of main LVLM architectures. . . . 51

3.1 Overview of the research questions and key findings of this thesis. 58

List of Tables

1.1 Contributions of the **foundational** papers to answer the research questions core to this thesis. ✓ denotes a strong contribution, (✓) denotes a moderate contribution. 10
1.2 Contributions of the **related** papers to answer the research questions core to this thesis. ✓ denotes a strong contribution, (✓) denotes a moderate contribution. 10

2.1 Frequently used symbols. Only the most common ones are listed here; chapter-specific notation is introduced locally as needed. . . 16

List of Abbreviations

AI . . . . . Artifical Intelligence
BERT . . . . . Bidirectional Encoder Representations from Transformers
CLIP . . . . . Contrastive Language-Image Pretraining
CNN . . . . . Convolutional Neural Network
CV . . . . . Computer Vision
FNN . . . . . Feedforward Neural Network
GPT . . . . . Generative Pre-trained Transformer
GRU . . . . . Gated Recurrent Unit
LLM . . . . . Large Language Model
LM . . . . . Language Model
LSTM . . . . . Long Short-Term Memory
LVLM . . . . . Large Vision-Language Model
ML . . . . . Machine Learning
MLP . . . . . Multilayer Perceptron
NER . . . . . Named Entity Recognition
NLP . . . . . Natural Language Processing
OCR . . . . . Optical Character Recognition
OOD . . . . . Out-of-Distribution
OVD . . . . . Open-Vocabulary Object Detection
QA . . . . . Question Answering
RAG . . . . . Retrieval-Augmented Generation
RLHF . . . . . Reinforcement Learning from Human Feedback
RNN . . . . . Recurrent Neural Network
RQ . . . . . Research Question
VGR . . . . . Visually-Grounded Reasoning
ViT . . . . . Vision Transformer
VLM . . . . . Vision-Language Model
VLT . . . . . Vision-Language Task
VQA . . . . . Visual Question Answering

1 Introduction

Contents

1.1 Historical Background and Motivation **2**

1.2 Definitions **5**

1.3 Research Questions **6**

1.4 Publications **7**

1.4.1 List of Foundational Papers 7

1.4.2 List of Related Papers 8

1.4.3 Degree of Authorship of Foundational Papers 9

1.5 Contributions **10**

1.5.1 Research Question 1: Strengths 10

1.5.2 Research Question 2: Limitations 11

1.5.3 Research Question 3: Mitigations 12

1.6 Thesis Outline **13**

1.1 Historical Background and Motivation

The idea of computer programs communicating with humans through natural language dates at least back to the mid-1960s when Joseph Weizenbaum introduced ELIZA (Weizenbaum, 1966). This pioneering chatbot simulated conversational interaction by utilizing pattern-matching algorithms and scripted responses. Today, about half a century later, recent advances in machine learning (ML), especially in natural language processing (NLP) and computer vision (CV), have drastically improved these systems and revolutionized how we use and interact with technology in our daily lives.

At the heart of this revolution are language models (LMs): Statistical models of natural languages that approximate a probability distribution to sample the next word in a sequence, given its preceding words. While first approaches were often pure statistical models that relied solely on word frequencies (Witten et al., 1991; Kneser et al., 1995; Gale et al., 1995; Katz, 2003), modern language models are neural networks (NNs) trained on massive amounts of textual data that "store" the information contained as abstract numerical representations within their parameters. Early neural LMs were based on simple feed-forward architectures (Bengio et al., 2003; Collobert et al., 2008) and later evolved to incorporate recurrent neural networks (RNNs) (Rumelhart et al., 1985) like Long Short-Term Memory networks (LSTMs) (Hochreiter et al., 1997) or Gated-Recurrent Units (GRUs) (Cho et al., 2014). These models slowly started to outperform count-based non-neural approaches but had severe limitations in terms of the amount of data they could process efficiently due to their sequential nature. The introduction of the "attention mechanism" for LSTMs by Bahdanau et al. (2015) allowed the models to focus on specific parts of an input sequence, enabling them to better capture long-range dependencies in the data and improving their performance on various NLP tasks. However, the architecture of recurrent LMs still forced models to process words sequentially and continued to hinder scalability, making it challenging to train on large-scale datasets required to solve complex tasks and acquire extensive knowledge.

Inspired by the "attention mechanism" for LSTMs, the pivotal moment happened in 2017 when researchers at Google introduced the Transformer architecture in their seminal paper "Attention is All You Need" (Vaswani et al., 2017). This new architecture replaced the sequential processing of RNNs with a mechanism based entirely on attention, allowing for parallel processing of all words in a sentence at once, thereby not only addressing the limitations of RNNs with long sequences but also enabling efficient large-scale training on modern hardware. Since then, the Transformer architecture has proven to be one of the most influential advancements in NLP. It profoundly reshaped the field and subsequently impacted other research areas, including computer vision (CV), speech signal processing, robotics, and ML in general.

Following the introduction of the transformers, significant milestones were achieved. In 2018, Google introduced BERT (Devlin et al., 2019), an encoder-only Transformer model based on self-attention, which rapidly became the state-of-the-art for a wide range of NLP tasks. However, encoder-only models, such as BERT, are not designed for text generation, and OpenAI introduced the Generative Pre-trained

Transformer (GPT) (Radford et al., 2018), a decoder-only transformer based on causal self-attention, that enabled next word prediction in the classical sense of language modelling.

Due to the general success of transformers, the architecture was also applied to modalities beyond text. Notably, the Vision Transformer (ViT) model (Dosovitskiy et al., 2021), introduced in 2021, successfully adapted transformer encoders to images by dividing an image into smaller patches treated as a sequence similar to words in a sentence. Furthermore, multimodal transformer encoders were developed to process data from multiple modalities, mainly text and images, simultaneously (Chen et al., 2020; Li et al., 2020). These vision-language models (VLMs) have proven highly effective for tasks like visual question answering (VQA), image captioning, object detection, or cross-modal retrieval. A particularly influential model in this space is CLIP (Radford et al., 2021), which was trained on large-scale text-image pairs to compute aligned vector embeddings of texts and images. Due to its broad applicability and strong performance in numerous vision-language tasks (VLTs), such as zero-shot image classification or cross-modal retrieval, CLIP has become a foundational model for multimodal scenarios.

The landscape of language models shifted again around 2022 with the emergence of instruction-tuned large language models (LLMs). Following the development of increasingly large models such as GPT-3 (Brown et al., 2020) trained on vast amounts of data, thereby having greater knowledge capacity, OpenAI's InstructGPT (Ouyang et al., 2022) paper introduced the concept of instruction following LLMs. By using reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2020), LLMs were trained to follow human instructions and align to human preferences, effectively becoming more general-purpose, task-agnostic models. Built upon these advancements, the public release of OpenAI's ChatGPT further popularized LLMs, leading to widespread interest and adoption by the general public.

In parallel, generative large vision-language models (LVLMs), which combine ViTs and LLMs, began to emerge (Alayrac et al., 2022; Chen et al., 2023b; Liu et al., 2023). These models are capable of jointly processing mixed multimodal content in the form of (interleaved) texts and images within a unified architecture, enabling comprehensive vision-language understanding and reasoning. This is achieved by resembling an architecture similar to the human visual system: The ViT acts analogously to the eye by capturing visual information, which the LLM, akin to the brain, contextualizes and integrates with existing knowledge. This close alignment with human perception makes LVLMs suitable for a broad range of VLTs and pushes the boundaries of multimodal AI applications.

However, despite the remarkable progress and performance of multimodal transformers, these neural networks are not without limitations. One key challenge is their performance and robustness on data different from the data they were trained on, often referred to as "out-of-distribution" (OOD) data. Such data can be characterized by different domains, e.g., medical images or legal documents, languages, e.g., English vs. German, or cultures, e.g., Western vs. African data. Models trained on data containing only a small set of domains, languages, or cultures reflect these inherent biases in their parameters and thus produce biased

results. As of today, the majority of datasets used to train the models behind modern, powerful AI systems are primarily in English and focus on Western cultures, which can lead to significant performance drops or even complete failures when the models are applied to data from other languages, cultures, or domains. This is particularly problematic in real-world scenarios involving a diverse group of users who come from different cultural backgrounds and speak different languages. Moreover, most users are unaware of the limitations and biases and expect the models to perform well when prompted in their languages to solve all kinds of tasks from various domains. Hence, it is crucial to identify and mitigate these shortcomings to ensure robust and reliable models required for globally equitable AI and thus guarantee that the benefits of this powerful technology are accessible to all people, regardless of their language or cultural background.

This dissertation explores the strengths and limitations of multimodal Transformer models—both encoder-only Vision-Language Models (VLMs) and generative encoder-decoder Large Vision-Language Models (LVLMs)—across a diverse set of vision-language tasks. Specifically, it investigates their performance degradation on out-of-distribution (OOD) data, evaluates the robustness across massively multilingual and global cultural contexts, and proposes effective strategies to mitigate multilingual limitations, thereby paving the way for more universally applicable multimodal AI systems for people all over the world.

1.2 Definitions

To investigate the strengths and limitations of transformer encoder and decoder models for vision-language tasks, or more specifically, vision-language models (VLMs) and large vision-language models (LVLMs), we first need to define how these terms are used in the scope of this thesis. More details on the architecture and training of VLMs and LVLMs can be found in Chapter 2.

Definition 1: Vision-Language Tasks (VLTs)

Vision-language tasks are problems that require the integration and joint understanding of visual and textual information. They evaluate a model's ability to learn cross-modal representations and effectively leverage them to compute solutions. Common tasks include visual question answering (VQA), visually grounded reasoning (VGR), open-vocabulary object detection (OVD), image captioning, or cross-modal retrieval.

Definition 2: Vision-Language Models (VLMs)

Vision-Language Models (VLMs) are neural networks based on transformer encoders that are trained on datasets containing aligned text-image pairs to compute joint cross-modal representations in a shared embedding space. Specifically, we focus on VLMs that consist of a vision encoder for processing visual data and a text encoder for processing textual data. The two encoders are trained jointly and compute aligned dense vector embeddings for both modalities usable for a wide range of vision-language tasks, especially for cross-modal retrieval.

Definition 3: Large Vision-Language Models (LVLMs)

Large Vision-Language Models (LVLMs) are generative neural networks that combine a transformer encoder for processing visual data and a transformer decoder, i.e., an LLM, for generating textual responses from visual, textual, or mixed multimodal input. Specifically, we focus on LVLMs that are trained on large-scale text-image datasets in multiple stages to understand multimodal data and follow human instructions. This makes LVLMs capable of performing a wide range of vision-language tasks that require textual solutions.

1.3 Research Questions

In recent years, rapid advances in NLP and CV research combined with ever-increasing computational resources have led to a growing interest in multimodal transformer-based models such as VLMs and LVLMs. Due to their ability to process not only text but also visual information, more closely resembling human perception, their potential field of use is manifold. However, deploying such models in real-world scenarios involving a diverse group of users from different cultural backgrounds, speaking different languages, and expecting robust performance for various inputs and tasks, it is crucial to understand the model's strengths and limitations in detail.

This gives rise to the following research questions that are core to this dissertation:

Research Question 1

What are the strengths and practical applications of VLMs and LVLMs, and how can they be employed in real-world scenarios?

Research Question 2

What are the inherent limitations of VLMs and LVLMs, and how do they affect their performance and robustness?

Research Question 3

How and to what degree can identified limitations in VLMs and LVLMs be mitigated?

1.4 Publications

In the following, accepted papers that address the research questions stated in this thesis are listed chronologically. The first list includes papers that form the foundation of this thesis, whereas the second list holds related to but rather peripheral work for this thesis. On all foundational papers, I am the first author or one of the joint first authors who equally contributed[1] to the respective work. Details on the degree of authorship of foundational papers are provided in §1.4.3.

1.4.1 List of Foundational Papers

Florian Schneider and Chris Biemann. 2022. Golden Retriever: A Real-Time Multi-Modal Text-Image Retrieval System with the Ability to Focus. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval,* 3245–3250. SIGIR '22. Madrid, Spain: Association for Computing Machinery

Florian Schneider and Chris Biemann. 2024a. WISMIR3: A Multi-Modal Dataset to Challenge Text-Image Retrieval Approaches. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR),* 1–6. Bangkok, Thailand: Association for Computational Linguistics

Florian Schneider and Sunayana Sitaram. 2024b. M5 – A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024,* 4309–4345. Miami, Florida, USA: Association for Computational Linguistics

Gregor Geigle*, **Florian Schneider***, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, and Goran Glavaš. 2025. Centurio: On Drivers of Multilingual Ability of Large Vision-Language Model. In *Proceedings of the Association for Computational Linguistics: ACL 2025,* in press. Vienna, Austria: Association for Computational Linguistics

Florian Schneider, Carolin Holtermann, Chris Biemann, and Anne Lauscher. 2025b. GIMMICK – Globally Inclusive Multimodal Multitask Cultural Knowledge Benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2025,* in press. Vienna, Austria: Association for Computational Linguistics

Florian Schneider, Narges Baba Ahmadi, Niloufar Baba Ahmadi, Iris Vogel, Martin Semmann, and Chris Biemann. 2025a. CollEX – A Multimodal Agentic RAG System Enabling Interactive Exploration of Scientific Collections. In *Proceedings of the 1st Workshop on Multimodal Augmented Generation via MultimodAl Retrieval (MAGMaR),* in press. Vienna, Austria: Association for Computational Linguistics

1. An equal contribution is indicated by an asterisk after the surname (*).

1.4.2 List of Related Papers

Florian Schneider, Özge Alaçam, Xintong Wang, and Chris Biemann. 2021. Towards Multi-Modal Text-Image Retrieval to Improve Human Reading. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop.* Online: Association for Computational Linguistics

Xintong Wang*, **Florian Schneider***, Özge Alacam, Prateek Chaudhury, and Chris Biemann. 2022. MOTIF: Contextualized Images for Complex Words to Improve Human Reading. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 2468–2477. Marseille, France: Association for Computational Linguistics

Anton Wiehe, **Florian Schneider**, Sebastian Blank, Xintong Wang, Hans-Peter Zorn, and Christian Biemann. 2022a. Language over Labels: Contrastive Language Supervision Exceeds Purely Label-Supervised Classification Performance on Chest X-Rays. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop,* 76–83. Online: Association for Computational Linguistics

Florian Schneider*, Tim Fischer*, Fynn Petersen-Frey, Isabel Eiser, Gertraud Koch, and Chris Biemann. 2023. The D-WISE Tool Suite: Multi-Modal Machine-Learning-Powered Tools Supporting and Enhancing Digital Discourse Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations),* 328–335. Toronto, Canada: Association for Computational Linguistics

Florian Schneider and Chris Biemann. 2023. LT at SemEval-2023 Task 1: Effective Zero-Shot Visual Word Sense Disambiguation Approaches using External Knowledge Sources. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023),* 462–468. Toronto, Canada: Association for Computational Linguistics

Musashi Hinck, Carolin Holtermann*, Matthew Lyle Olson*, **Florian Schneider***, Sungduk Yu, Anahita Bhiwandiwalla, Anne Lauscher, Shao-Yen Tseng, and Vasudev Lal. 2024b. Why do LLaVA Vision-Language Models Reply to Images in English? In *Findings of the Association for Computational Linguistics: EMNLP 2024,* 13402–13421. Miami, Florida, USA: Association for Computational Linguistics

Fabian David Schmidt*, **Florian Schneider***, Chris Biemann, and Goran Glavaš. 2025. MVL-SIB: A Massively Multilingual Vision-Language Benchmark for Cross-Modal Topical Matching. In *Findings of the Association for Computational Linguistics: ACL 2025,* in press. Vienna, Austria: Association for Computational Linguistics

1.4.3 Degree of Authorship of Foundational Papers

In the paper *Golden Retriever: A Real-Time Multi-Modal Text-Image Retrieval System with the Ability to Focus* (Schneider et al., 2022), I conceptualized, implemented, and conducted all experiments, including the introduced "Visually Weighted TF-IDF" algorithm, implemented the proof-of-concept application, and wrote the paper. Chris Biemann supervised the work on an abstract level and provided recommendations for improving the paper.

In the paper *WISMIR3: A Multi-Modal Dataset to Challenge Text-Image Retrieval Approaches* (Schneider et al., 2024a), I conceptualized, implemented, and conducted all experiments and wrote the paper. Further, I incorporated Chris Biemann's suggestions to improve the paper.

In the paper *M5 – A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks* (Schneider et al., 2024b), I conceptualized, implemented, and conducted all experiments as well as the collection of the two introduced datasets and wrote the paper. Sunayana Sitaram inspired the project with her previous work on multilingual benchmarking and provided meaningful feedback during the dataset collection process.

In the paper *Centurio: On Drivers of Multilingual Ability of Large Vision-Language Model* (Geigle et al., 2025), Gregor Geigle and I primarily conceptualized the research questions and high-level experimental setup of the project. Gregor conceptualized, implemented, and executed the training experiments, and I did the same for the evaluation of our trained models and the baseline models. Carolin Holtermann also played an essential role in the general discussions of experiments and the project's progress and was responsible for translating the training data. Gregor wrote the paper draft, which was then improved by Goran Glavaš, Anne Lauscher, Carolin, him, and me. Goran and Anne also supervised the project and provided meaningful ideas and suggestions during the project. Chris Biemann and Radu Timofte provided suggestions for improving the paper.

In the paper *GIMMICK – Globally Inclusive Multimodal Multitask Cultural Knowledge Benchmarking* (Schneider et al., 2025b), I conceptualized, implemented, and conducted all experiments and collected all introduced datasets, and I primarily wrote the paper. Anne Lauscher supervised the project by providing essential ideas and fostering meaningful discussions. Carolin Holtermann participated actively in all discussions and annotated samples for the pilot projects. Further, Carolin and Anne assisted in improving the paper. Chris Biemann proof-read the paper and suggested minor grammatical improvements.

In the paper *CollEX – A Multimodal Agentic RAG System Enabling Interactive Exploration of Scientific Collections* (Schneider et al., 2025a), I conceptualized the architecture, implemented the proof-of-concept application, and I wrote the paper.

Narges and Niloufar Baba Ahmadi implemented some features in an earlier version of the application. The raw "FUNDus!" scientific collection data were managed and collected by Iris Vogel. Martin Semmann and Chris Biemann provided suggestions for improving the paper.

1.5 Contributions

In the following, the core contributions of this thesis and how they relate to the research questions outlined in §1.3 are briefly summarized. Table 1.1 provides a concise overview how the foundational papers contribute to answering the research questions core to this thesis, while Table 1.2 summarizes the contributions of the related papers.

Paper	RQ1 Strengths	RQ2 Limitations	RQ3 Mitigations
Golden Retriever (Schneider et al., 2022)	✓		
CollEX (Schneider et al., 2025a)	✓		
WISMIR3 (Schneider et al., 2024a)	(✓)	✓	
M5B (Schneider et al., 2024b)	(✓)	✓	
GIMMICK (Schneider et al., 2025b)	(✓)	✓	
Centurio (Geigle et al., 2025)	(✓)	(✓)	✓

Table 1.1: Contributions of the **foundational** papers to answer the research questions core to this thesis. ✓ denotes a strong contribution, (✓) denotes a moderate contribution.

Paper	RQ1 Strengths	RQ2 Limitations	RQ3 Mitigations
MOTIF (Wang et al., 2022)	✓		
LoL (Wiehe et al., 2022b)	✓		
DWTS (Schneider et al., 2023b)	✓		
ZS VWSD (Schneider et al., 2023a)	✓		
MM-T2I (Schneider et al., 2021)	(✓)	✓	
MVL-SIB (Schmidt et al., 2025)	(✓)	✓	
Why Llava? (Hinck et al., 2024a)	(✓)	✓	✓

Table 1.2: Contributions of the **related** papers to answer the research questions core to this thesis. ✓ denotes a strong contribution, (✓) denotes a moderate contribution.

1.5.1 Research Question 1: Strengths

What are the strengths and practical applications of VLMs and LVLMs, and how can they be employed in real-world scenarios?

We address this question in the following two papers from different perspectives:

In our **Golden Retriever** (Schneider et al., 2022) paper, we show that VLMs can be effectively used for zero-shot cross-modal text-image retrieval tasks as well as for open-vocabulary object detection. This is achieved by leveraging and pooling fine-grained word-region alignment matrices computed by specialized dual-stack transformer encoder network (Messina et al., 2021), with cells representing the similarity between a word and a region in an image. We also introduce a novel algorithm that extends the well-known TF-IDF (Spärck Jones, 1972) algorithm to the multimodal domain, named "Visually-Weighted TF-IDF" (VW-TF-IDF), which allows for efficient cross-modal text-to-image retrieval taking into account the presence of query-relevant objects in the image. Moreover, we demonstrate the practical applicability of our approach by implementing a real-time cross-modal retrieval system as a proof-of-concept application in the form of a browser extension and web application.

In our **CollEX** (Schneider et al., 2025a) paper, we demonstrate how VLMs for cross-modal retrieval combined with LVLMs employed as function-calling agents can be leveraged in a multimodal RAG system to create a user-friendly chatbot application that can answer complex queries about scientific collections of textual and visual artifacts. Further, we show how to effectively utilize LVLMs to perform classical vision-language tasks (VLTs) like VQA, image captioning, object detection, and OCR. The practical applicability of our approach is demonstrated by proof-of-concept web applications with an intuitive, simple user interface.

1.5.2 Research Question 2: Limitations

> *What are the inherent limitations of VLMs and LVLMs, and how do they affect their performance and robustness?*

This question is addressed concerning different aspects in the following three papers:

In our **WISMIR3** (Schneider et al., 2024a) paper, we analyze the robustness of VLMs for cross-modal retrieval in the context of out-of-distribution (OOD) data. Therefore, we introduce a novel dataset based on lengthy image captions from Wikipedia articles, resembling significantly different data distributions than typical training data for the respective models. We find that the models' performance degrades drastically when evaluated on OOD data both in zero-shot and fine-tuning settings. This highlights severe limitations in pretrained models' ability to generalize across different domains as well as their failure to learn cross-modal understanding from complex data distributions.

In our **M5B** (Schneider et al., 2024b) paper, we assess the robustness of various LVLMs of different families and sizes on a large-scale multilingual benchmark covering a wide range of VLTs in over 40 languages. In addition to the six existing datasets, we introduce two new datasets covering 12 languages, focusing on low-resource African and Asian languages. One dataset is a Visually Grounded Reasoning (VGR) dataset, which requires the models to reason about an image's

content and decide whether a given textual hypothesis is true or false. The second dataset introduces a novel VLT called "Visio-Linguistic Outlier Detection" (VLOD), which requires the models to identify an outlier among a set of five images given a textual statement. While LVLMs can perform well on VLTs when prompted in English, we find that their performance degrades significantly when prompted in other languages. This effect is particularly pronounced for low-resource languages, where LVLMs often fail to perform the task or even generate responses in the respective language. Further, we discover that the majority of tested LVLMs are unable to perform the VLOD task in any language, thereby demonstrating significant limitations in multi-image reasoning.

With our **GIMMICK** (Schneider et al., 2025b) paper, we measure cultural biases of LVLMs and LLMs through various aspects of cultural knowledge, introducing a large-scale multimodal benchmark. Specifically, GIMMICK comprises six tasks based on three novel datasets covering 728 distinct cultural events or facets from 144 countries, organized into six global macro-regions. Our datasets build on the UNESCO Intangible Cultural Heritage (ICH[2]) list and are designed to assess the models' knowledge about cultural events, artifacts, and practices through text-only, image-only, text-image, and notably also text-video tasks. Our extensive analyses demonstrate pronounced cultural biases towards Western cultures across all models and tasks, while the performance on Arab, Asian, and African cultures is significantly lower.

1.5.3 Research Question 3: Mitigations

How and to what degree can identified limitations in VLMs and LVLMs be mitigated?

In our **Centurio** (Geigle et al., 2025) paper, we identify severe limitations in the robustness of LVLMs in massively multilingual settings and propose an effective training strategy to mitigate these limitations. Specifically, we evaluate the performance of LVLMs on a massively multilingual benchmark covering over 50 languages and a diverse set of VLTs, revealing that the models' performance degrades significantly when prompted in non-English languages. Moreover, we analyze the models' multilingual OCR capabilities and find that they perform poorly in extracting non-English text from images. To mitigate these issues, we comprehensively analyze how to compose the best training data mixture concerning the number of languages, the amount of training data per language, and the mixture of tasks. In our final experiments, we scale up the training data and train Centurio Qwen & Centurio Aya, two strong, massively multilingual LVLMs supporting 100 languages, proving state-of-the-art performance over 14 tasks against other open-weight models.

2. https://ich.unesco.org

1.6 Thesis Outline

This cumulative dissertation is structured as follows: Chapters 1, 2, and 3 are the wrapper of this thesis and hold its synoptic view consisting of an introduction and motivation to my research, a summary of the prerequisite theoretical background knowledge, and the final discussions and conclusions of the thesis, respectively. The research papers that form the foundation of this dissertation (cf. §1.4.1) are annexed in their original layout in the second part of this document in the form of an appendix.

2

Theoretical Background

Contents

2.1 Introduction **17**

2.2 Traditional Neural Networks for Natural Language Processing **17**

2.2.1 Language Modeling 17

2.2.2 Neural Networks Fundamentals 19

2.2.3 Neural Word Embeddings 22

2.2.4 Recurrent Neural Networks (RNNs) 23

2.2.5 Sequence-to-Sequence Models 25

2.3 Transformer Models for Modern Natural Language Processing **27**

2.3.1 The Transformer Architecture 27

2.3.2 Transformer Encoder Models 31

2.3.3 Transformer Decoder Models 32

2.3.4 Large Language Models (LLMs) 33

2.4 Multimodal Transformer Models for Vision-Language Tasks . **38**

2.4.1 Vision Transformers (ViTs) 38

2.4.2 Encoder-based Vision-Language Models (VLMs) . . . 41

2.4.3 Decoder-based Large Vision-Language Models (LVLMs) 48

2.5 Summary **56**

Mathematical Notation

Throughout this chapter we follow these typographic conventions for mathematical notation.

- *Scalars* are italic: $a,\ b,\ N\ \in \mathbb{R}$ or Greek letters α, β, γ.
- **Vectors** are bold lower-case: $\mathbf{a} \in \mathbb{R}^N$.
- **Matrices/Tensors** are bold upper-case: $\mathbf{A} \in \mathbb{R}^{M \times N}$, $\mathbf{T} \in \mathbb{R}^{K \times M \times N}$.
- *Sets & Spaces* are calligraphic: $\mathcal{D}$ (dataset), $\mathcal{V}$ (vocabulary).
- `Upright roman` is used for functions, operators, or neural network modules, e.g., $\texttt{softmax}(\cdot)$, $\texttt{FNN}(\cdot)$.

Symbol	Description
w_i	i-th word in a sentence
$S = [w_0, \dots, w_N]$	ordered sequence of N+1 words
$\mathrm{P}(w_i \mid w_{<i})$	conditional word probability
$\mathbf{x}$	input feature vector
$(\mathbf{x}_1, \dots, \mathbf{x}_T)$	ordered sequence of T vectors
$\mathbf{w}$	weight vector
$\mathbf{W}^{(\ell)}$	weight matrix of layer ℓ
$\sigma(\cdot)$	non-linear activation
$\mathbf{h}_t$	hidden state at step t
$\mathbf{c}_t$	cell state at step t
$\mathbf{Q}, \mathbf{K}, \mathbf{V}$	query, key, value matrices
$\boldsymbol{\theta}$	model parameters
$\mathcal{L}$	loss function
η	learning rate
$\nabla_{\boldsymbol{\theta}} \mathcal{L}$	gradient of $\mathcal{L}$
$\lVert \cdot \rVert_2$	Euclidean norm

Table 2.1: Frequently used symbols. Only the most common ones are listed here; chapter-specific notation is introduced locally as needed.

Additional remarks. Unless otherwise noted, indices start at 0 and all vectors are column-vectors. Task-specific symbols introduced later are summarised at the beginning of their respective chapters.

2.1 Introduction

In this chapter, we provide an introduction to the theoretical background of our research presented in this thesis. Specifically, we start from traditional neural network (NN) architectures which were widely used in natural language processing (NLP), move on to transformer architectures which have revolutionized the field and are fundamental to modern NLP, and finally introduce the key concepts of multimodal transformer models for vision-language tasks (VLTs), i.e., encoder-based VLMs and decoder-based LVLMs, which are the main focus of this thesis.

2.2 Traditional Neural Networks for Natural Language Processing

Historically, statistical methods such as n-gram-based language models (Kneser et al., 1995) and rule-based systems dominated the field of natural language processing (NLP) but suffered from issues with data sparsity and limited contextual scope. However, in recent decades, neural network-based methods have significantly transformed NLP by replacing traditional approaches due to their enhanced capabilities of modeling complex linguistic phenomena and capturing distant dependencies across long contexts. The first kind of neural approaches that were used in favor of traditional methods were simple Feed-Forward Networks (FFNs), followed by more sophisticated recurrent architectures like Recurrent Neural Networks (Rumelhart et al., 1985) (RNNs) or Long Short-Term Memory (Hochreiter et al., 1997) models (LSTMs), and eventually modern Transformer (Vaswani et al., 2017) architectures based on attention mechanisms. This section first familiarizes one of the core task of NLP, i.e., classical language modeling, and how the task was approached with traditional methods, followed by a brief introduction to neural network architectures primarily used in the pre-transformer era.

2.2.1 Language Modeling

Language modeling is a fundamental task in NLP and involves estimating a probability distribution to sample the next word in a sequence of words, like a sentence, given its predecessors. The statistical models to solve this task are referred to as language models (LMs). Formally, the probability distribution of a word w_i given the previous words in a sequence is defined as:

$$P(w_i \mid w_{i-1}, \dots, w_0) \tag{2.1}$$

where w_i is the word at position i in a sequence $S = [w_0, w_1, \dots, w_i, \dots, w_N]$ of N total words, often referred to as the word's context. Language models are trained to assign high probabilities to sequences of words that are likely to occur in natural language, i.e., in the training corpus, and low probabilities to unlikely sequences.

N-gram Models

One successful family of traditional LMs based on statistical methods relies on word counts and is called n-gram models. Instead of considering the entire context, these models estimate the probability of the next word based on the frequency or occurrence of its preceding $n-1$ words in a large text corpus. Note that the term n-gram is ambigouous and can refer to both the model itself and an ordered sequence of n words.

Markov Assumption n-gram models rely on the *(stationary) Markov Assumption*, which consists of two parts: i) *Limited Context*, i.e., the probability of the next word in a sequence depends only on the previous $n-1$ words, and ii) *Time Invariance*, i.e., the probability distribution of a word does not change over time but is the same throughout the entire corpus. Based on these assumptions, formally the probability of a word w_i given its preceding $n-1$ words $\left[w_{i-1}, \dots, w_{i-(n-1)}\right]$ using n-gram models is approximated as:

$$\mathrm{P}(w_i \mid w_{i-1}, \dots, w_0) \approx P\left(w_i \mid w_{i-(n-1)}, \dots, w_{i-1}\right) \tag{2.2}$$

Maximum Likelihood Estimation A simple way to estimate the probability of a word w_i in an n-gram model is to use *maximum likelihood estimation* (MLE). This approach relies on counting the occurrences of words or n-grams, also referred to as *word frequencies*, in a large training corpus. In the case of n-gram models, the MLE is computed as the relative frequency of the word w_i given its preceding $n-1$ words in the corpus:

$$\mathrm{P}\left(w_i \mid w_{i-(n-1)}, \dots, w_{i-1}\right) \approx \frac{\texttt{freq}\left(w_{i-(n-1)}, \dots, w_i\right)}{\texttt{freq}\left(w_{i-(n-1)}, \dots, w_{i-1}\right)} \tag{2.3}$$

where $\texttt{freq}(\cdot)$ denotes the frequency of an n-gram in the training corpus, i.e., the number of times the n-gram occurs in the corpus.

Example For example, suppose we want to estimate the probability of the word "*dream*" given the context "*I have a* " with a 3-gram (trigram) model using simple MLE:

$$\mathrm{P}(\textit{dream} | \textit{have a}) = \frac{\texttt{freq}(\textit{have a dream})}{\texttt{freq}(\textit{have a})} \tag{2.4}$$

To estimate the probability that the word "*dream*" follows after the sequence "*have a*", we need to count how often the sequences "*have a*" and "*have a dream*" occur in our large training corpus and divide the two frequencies. Note that the trigram model does not consider the word "*I*" due to its limited context window of only 2 words.

Limitations From the definition of *n*-gram models, we can observe that these models suffer from multiple limitations, with the most crucial ones being discussed in the following. One issue is data sparsity, i.e., if a word or a sequence of words does not occur in the training corpus the frequency is zero, which leads to zero-division errors. While numerous methods have been proposed to mitigate these issues, such as smoothing techniques (Witten et al., 1991; Kneser et al., 1995; Gale et al., 1995) or back-off models (Katz, 2003), *n*-gram models are still limited by the fixed size of the context window. Another issue is the inability to capture long-range dependencies, i.e., the models are unable to take into account words that are outside of this window and, therefore, lack important contextual information. Lastly, another crucial limitation of *n*-gram models is that they are based solely on the frequency of the surface forms of words or their derived forms, e.g., lemmas or stems. This means they do not consider the semantics or meaning of the words nor their complex interactions and relationships. With the rise in computational resources, researchers explored how neural networks can be employed as language models to overcome these limitations.

2.2.2 Neural Networks Fundamentals

Before we dive into the details of how neural networks can be used as LMs, we first briefly cover the basic principles of their architecture and components.[1]

Artificial Neurons

As the name suggests, this family of machine learning (ML) models is inspired by the inner workings of the human brain, i.e., it resembles a network of artificial neurons, interconnected through artificial "*synapses*" modeled as "*weights*" to process some input information. The output z of a single artificial neuron can be described as a weighted sum of its N inputs $\mathbf{x} \in \mathbb{R}^N = \{x_0, \dots, x_{N-1}\}$

$$z = \sum_{i=0}^{N-1} w_i x_i = \mathbf{w} \cdot \mathbf{x} \tag{2.5}$$

where $\mathbf{w} \in \mathbb{R}^N = \{w_0, \dots, w_{N-1}\}$ are the corresponding weights. To increase the expressiveness, a bias term b is often added to the weighted sum, and the result is passed through a nonlinear activation function σ to produce the final output or "*activation*" a of the neuron:

$$a = \sigma(z) = \sigma(\mathbf{w} \cdot \mathbf{x} + \mathbf{b}) \tag{2.6}$$

Feed-Forward Neural Networks

When multiple neurons are combined and interconnected, they form a neural network layer. Connecting multiple layers together so that the output of the *n*th

1. For an in-depth introduction to neural networks and deep learning, we recommend Goodfellow et al. (2016) or Bishop et al. (2024)

layer is the input of the $(n+1)$th layer creates a feed-forward neural network (FFN), the most basic type of neural network. Often FFNs are also referred to as "*multi-layer perceptrons*" (MLPs) or "*fully connected networks*" (FCNs), which we use interchangeably throughout this thesis. FFNs have at least one layer but typically consist of an input layer, one or more hidden layers, and an output layer, each of which can have a different shape or size depending on the specific architecture.

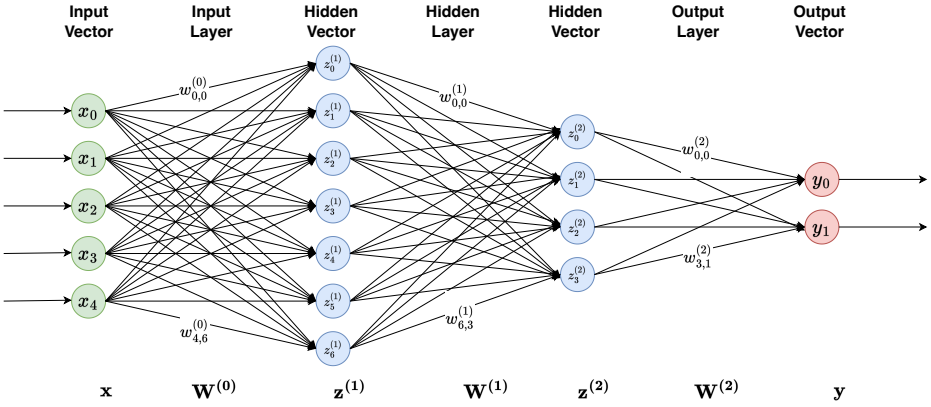


Figure 2.1: Schematic overview of a 3-layer Feed-Forward neural network architecture with bias terms and non-linearity omitted for better readability.

For better understanding, a simple FFN with three linear layers (without bias terms) is illustrated in Figure 2.1 and described in the following: The input vector $\mathbf{x} \in \mathbb{R}^5$ is passed through the first ("*input*") layer with 7 neurons and a weight matrix $\mathbf{W}^{(0)} \in \mathbb{R}^{5\times7}$, which is multiplied with the input vector $\mathbf{x}$ to produce the hidden vector $\mathbf{z}^{(0)} \in \mathbb{R}^7$. Next, $\mathbf{z}^{(0)}$ gets multiplied by the weights $\mathbf{W}^{(1)} \in \mathbb{R}^{7\times4}$ of the second ("*hidden*") layer to compute the hidden vector $\mathbf{z}^{(1)} \in \mathbb{R}^4$. Finally, the output vector $\mathbf{y} \in \mathbb{R}^2$ is computed by multiplying the weights of the third ("*output*") layer $\mathbf{W}^{(2)} \in \mathbb{R}^{4\times2}$ with the hidden vector $\mathbf{z}^{(1)}$. Note that the weights of a neural network layer ℓ_i are also referred to as the layer's parameters $\boldsymbol{\theta}_i$ and are often omitted for brevity. This whole process is called a *forward pass* through the FNN and can be formally described by

$$\begin{aligned} y &= \ell_2\big(\ell_1(\ell_0(\mathbf{x};\boldsymbol{\theta}_0);\,\boldsymbol{\theta}_1);\,\boldsymbol{\theta}_2\big) \\ &= \big(\ell_2 \circ \ell_1 \circ \ell_0\big)(\mathbf{x}) \\ &= \texttt{FNN}(\mathbf{x}) \end{aligned} \tag{2.7}$$

Such kind of FNN, which is in practice much larger (more neurons), deeper (more layers), and more complex (nonlinear activations, regularization, etc.), can, in theory, be used to learn any kind of mapping $f : \mathbb{R}^N \to \mathbb{R}^M$ from an input vector $\mathbf{x} \in \mathbb{R}^N$ to an output vector $\mathbf{y} \in \mathbb{R}^M$. One example application are classification tasks, where the input vector $\mathbf{x}$ is classified into one of M classes, where the output vector $\mathbf{y}$ is a one-hot encoded vector with M dimensions, i.e., $y_i = 1$ if the class is c_i and 0 otherwise. The output of the FNN can be interpreted as a probability distribution over the M classes by applying a softmax function to the output vector $\mathbf{y}$:

$$\text{softmax}(\mathbf{y}) = \frac{\exp(\mathbf{y})}{\sum_{j=1}^{M} \exp(y_j)} \tag{2.8}$$

where $\exp(\mathbf{y})$ is applied element-wise to the output vector $\mathbf{y}$. This softmax function normalizes the output vector $\mathbf{y}$ to a "pseudo" probability distribution over the M classes, i.e., $\sum_{i=1}^{M} \text{softmax}(\mathbf{y})_i = 1$. In NLP, this is often used to classify a word or a sentence into one of M classes, e.g., the next word in a sentence where the output vector $\mathbf{y}$ is a one-hot encoded vector of the vocabulary size V, the sentiment of a sentence, or the topic of a document.

Neural Network Training

To train a neural network, we have to adjust its parameters, i.e., the weights and biases, to minimize an error or loss function that measures the difference between the predicted output and the actual target output in a training dataset. Since this is a non-convex optimization problem, i.e., there are multiple local minima, it is typically solved using iterative optimization techniques that adjust the randomly initialized parameters in the direction of the steepest descent of the loss function. This involves gradient-based algorithms such as stochastic gradient descent (SGD) or more sophisticated and efficient variants like Adam (Kingma et al., 2015; Ruder, 2016). Formally, the training process can be described as minimizing a loss function $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}; \boldsymbol{\theta})$ defined over a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) \mid i = 1, \ldots, M\}$ of M samples, where each sample consists of input-output pairs $(\mathbf{x}_i, \mathbf{y}_i)$:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\text{argmin}}\, \mathcal{L}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\text{argmin}}\, \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}(\mathbf{y}_i, \texttt{FNN}(\mathbf{x}_i; \boldsymbol{\theta})) \tag{2.9}$$

where $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}})$ measures how well the predicted output $\hat{\mathbf{y}}$ matches the true output $\mathbf{y}$ and $\boldsymbol{\theta}^*$ are the optimal parameters of the FNN. Standard loss functions include the mean squared error (MSE) for simple regression tasks or the cross-entropy loss for classification tasks and next-token prediction, i.e., language modeling.

To iteratively minimize the loss function, the gradient of the loss with respect to the parameters $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$ is computed using a procedure called *backpropagation*. Backpropagation employs the chain rule from calculus to efficiently propagate gradients backward through the network from the output layer to the input layer, adjusting parameters incrementally in the direction of the steepest negative gradient:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \tag{2.10}$$

where η denotes the learning rate, a hyperparameter controlling how fast the parameters are updated. This optimization step is typically executed after computing gradients from a (much) smaller subset of data called a batch rather than the entire dataset. Given batches $\mathcal{B} \subseteq \mathcal{D}$ with $|\mathcal{B}| \ll |\mathcal{D}|$, training over dataset $\mathcal{D}$ involves iterating through all batches:

$$\forall \mathcal{B} \subseteq \mathcal{D} : \quad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \frac{1}{|\mathcal{B}|} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{B}} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{y}, \texttt{FNN}(\mathbf{x}; \boldsymbol{\theta})) \tag{2.11}$$

This training procedure usually continues until the loss converges, i.e., a local minimum in the parameter space has been reached, or a predefined number of iterations is reached.

Note that the description of the training process above is very general and simplified. That is, it does not cover many important aspects, such as regularization techniques to prevent overfitting, hyperparameter tuning, or the use of validation datasets to monitor the model's performance during training. We recommend the reader to refer to Goodfellow et al. (2016) or Bishop et al. (2024) for an excellent in-depth coverage of these topics and general deep learning fundamentals.

2.2.3 Neural Word Embeddings

Before a neural network can process textual data, the words in the text must be first be converted into numerical representations that the network can understand. Traditional NLP methods, such as n-gram models, primarily rely on one-hot encoded vectors where a word is represented as a vector with a single dimension set to 1 corresponding to the word's index in a vocabulary and 0 elsewhere. This approach has significant limitations, such as the extreme sparsity of the vectors, and the high dimensionality of the vectors that is equal to the size of the vocabulary. Furthermore, one-hot vectors do not capture any relationships between words because they are orthogonal to each other and thus do not reflect any semantic or syntactic similarities.

To address these limitations and compute continuous representations for words, methods such as Latent Semantic Analysis (LSA) (Deerwester et al., 1990) or Latent Dirichlet Allocation (LDA) (Blei et al., 2003) were developed. These methods rely on matrix factorization and sampling techniques, respectively, to reduce the dimensionality of the one-hot encoded vectors, resulting in dense representations of words. Although these approaches were successful, their capabilities in capturing semantic relationships were limited, and they struggled with scalability issues when dealing with large vocabularies.

Modern NLP methods employ dense representations of fixed dimensions in a continuous vector space, using linear layers called "*embedding layers*" trained on a large corpus (Bengio et al., 2003). These representations are called "*word embeddings*" (Collobert et al., 2008) and capture semantic and syntactic relationships because the embedding layer learned to place semantically similar or related words close together in the vector space. To obtain the embedding e_i of the ith word v_i in the vocabulary $\mathcal{V}$ from a trained D dimensional embedding layer $\mathbf{W}_{\text{emb}} \in \mathbb{R}^{V \times D}$, we extract the corresponding row from the weight matrix:

$$e_i = \mathbf{W}_{\text{emb}}^{(i)} \in \mathbb{R}^D \tag{2.12}$$

Besides their ability to capture semantic relationships, Collobert et al. (2008) also showed that word embeddings trained on large corpora can be used as a general-purpose representation of words that can be reused across different downstream tasks.

Following the work of Bengio et al. (2003) and Collobert et al. (2008), several methods were developed to improve the capability to capture rich semantic relationships, training efficiency, and reusability of neural embedding models. The most influential methods are Word2Vec (Mikolov et al., 2013a; Mikolov et al.,

2013b) and GloVe (Pennington et al., 2014), which are trained on large corpora to produce high-quality word embeddings that were widely used for various NLP tasks. Word2Vec employs a single-layer neural network architecture to learn word embeddings by either predicting the context words given a target word ("*skip-gram*") or, vice versa, predicting the target word given its context words ("*continuous bag of words*"). GloVe, on the other hand, is based on matrix factorization techniques and learns word embeddings by capturing the global statistical information of word co-occurrences in a corpus.

While both methods significantly advanced the field of NLP, they are limited to learn *static* word embeddings, meaning that there is one single embedding for each word in the vocabulary, regardless of its context or usage in a sentence. This is a significant limitation, since the meaning of a word often depends on its context, and thus a single static embedding cannot capture the full richness of a word's meaning.

2.2.4 Recurrent Neural Networks (RNNs)

One crucial limitation of vanilla FNNs is that they can only take fixed-size input vectors, thus, fall short in modeling sequential data of variable lengths, such as natural language. One solution would be to use a sliding window approach and concatenate the previous n tokens, as done in the "neural probabilistic language model" proposed by Bengio et al. (2003). However, the receptive field or context window remains strictly limited to the last n tokens, which is a significant limitation for many NLP tasks where long-range dependencies are crucial. To overcome this limitation, Recurrent Neural Networks (Rumelhart et al., 1985) (RNNs) were introduced as a specialized architecture explicitly designed to handle sequences of varying lengths. This is achieved by maintaining an internal memory that captures information from preceding inputs that is updated at each step.

More precisely, an RNN processes a sequence one element at a time while maintaining a hidden state vector $\mathbf{h}_t$ (the internal memory) that holds information from previous steps. Formally, the hidden state at time step t can be computed by combining the input $\mathbf{x}_t$ with the previous hidden state $\mathbf{h}_{t-1}$ as follows:

$$\mathbf{h}_t = \sigma(\mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{W}_{xh}\mathbf{x}_t) \tag{2.13}$$

where σ is typically a non-linear activation function such as tanh, and $\mathbf{W}_{hh}$, $\mathbf{W}_{xh}$ are parameter matrices to be learned during training. This recurrent process can also be interpreted as a loop over the sequence, where the hidden state $\mathbf{h}_t$ is updated at each time step based on the current input $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$, and the hidden state is passed to the next time step. An output vector $\mathbf{y}_t$, which often is a probability distribution over all words in the vocabulary, can be computed from the hidden state $\mathbf{h}_t$ at each time step:

$$\mathbf{y}_t = \texttt{softmax}(\mathbf{W}_{hy}\mathbf{h}_t) \tag{2.14}$$

where $\mathbf{W}_{hy}$ is again a learnable parameter matrix that maps the hidden state to the output. Note that all weights are shared across time steps, which essentially means that the same parameters are used to process each input in the sequence.

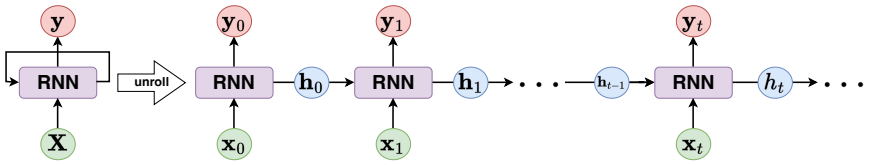


Figure 2.2: A schematic overview of a Recurrent Neural Network (RNN) architecture.

Figure 2.2 illustrates an RNN architecture unfolded over time, highlighting how the hidden state propagates information across steps in the sequence. This allows RNNs to theoretically capture long-range dependencies in sequences, addressing one of the primary limitations of *n*-gram models and FFNs, making them suitable for tasks such as language modeling, machine translation, or summarization. Note that usually only the last output $\mathbf{y}_T$ of the RNN is computed and used for prediction, where T is the last time step in the sequence. Furthermore, RNNs can also be stacked to create deeper architectures, where the output of the last step $\mathbf{y}_T$ is passed to a second RNN layer, which can learn more complex representations of the input sequence.

Limitations of Vanilla RNNs

Despite their potential in sequence processing, vanilla RNN architectures suffer from the so-called "*vanishing gradient*" (Hochreiter, 1998) or "*exploding gradient*" (Bengio et al., 1994) problems, where the gradients become exceedingly small or large as they are backpropagated through time when training on long sequences (Pascanu et al., 2013). The issues arise from the repeated multiplication of the weight matrices $\mathbf{W}_{hh}$ and $\mathbf{W}_{xh}$ during training with backpropagation through time (BPTT). If those weights are smaller than 1 the gradient quickly vanishes to zero, and if they are larger than 1 the gradients explodes. This leads to ineffective training, as the model struggles to update its parameters meaningfully, especially for long sequences. Further, these limitations prevent the networks from effectively learning long-range dependencies, drastically limiting their practical effectiveness in language modeling tasks involving lengthy texts. One remedy to the gradient-related issues of vanilla RNNs is a specialized RNN architecture called Long Short-Term Memory (LSTM) (Hochreiter et al., 1997). This sophisticated architecture introduces an additional cell state and gating mechanisms, allowing the model to learn what to retain or forget over long sequences. More precisely, LSTMs incorporate three types of gates: input, forget, and output gates, which are trainable neural network layers. Moreover, instead of only passing a single hidden state from step to step, LSTMs have an additional internal cell state that serves as a kind of conveyor belt for long-term memory. As opposed to vanilla RNNs, the weights of LSTM cell states are not recurrently updated through multiplication but addition, which helps to mitigate the vanishing gradient problem.

Since their introduction in 1997, many variations of LSTMs, such as Bi-directional LSTMs (BiLSTMs) (Schuster et al., 1997; Huang et al., 2015), Gated Recurrent Units (GRUs) (Cho et al., 2014), or Convolutional LSTMs (ConvLSTMs) (Shi et al., 2015), have been proposed. These architectures have shown significant improvements over vanilla RNNs in terms of performance and training stability and thus have

been widely adopted in NLP tasks, especially before the advent of Transformer-based models. However, due to their sequential nature, all kinds of RNNs are forced to process sequences step-by-step, which limits their ability to leverage parallelization. More precisely, the hidden state $\mathbf{h}_t$ at time step t can only be computed after the hidden state $\mathbf{h}_{t-1}$ at time step $t-1$ has been computed, which prevents parallelization across time steps, making the models inherently slow and inefficient during training. Another significant limitation of the classical RNN architectures is that the input sequence and the output sequence are of the same length. This is because the output at each time step t is computed from the hidden state $\mathbf{h}_t$ at the same time step, which means that the model can only produce one output for each input in the sequence.

2.2.5 Sequence-to-Sequence Models

While traditional RNNs effectively handle sequential data such as text by maintaining context through their hidden states, they inherently require input and output sequences of the same length. However, numerous NLP tasks, such as machine translation, summarization, and question answering, involve input and output sequences of differing lengths. To address this limitation, Sequence-to-Sequence (seq2seq) models were introduced (Sutskever et al., 2014; Cho et al., 2014), specifically designed to handle these variable-length input-output mappings by employing an encoder-decoder architecture as illustrated in Figure 2.3.

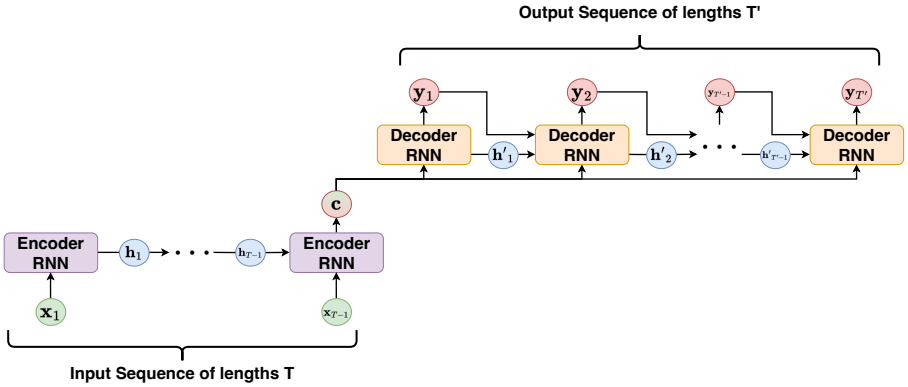


Figure 2.3: A schematic overview of a Sequence-to-Sequence (seq2seq) architecture.

Generally, a seq2seq model consists of two separate RNNs: an encoder and a decoder. Both can be any RNN, but LSTMs or GRUs are typically preferred due to their improvements in handling long-range dependencies. The encoder processes the input sequence $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_T]$ step-by-step, compressing the information into a fixed-length context vector $\mathbf{c}$. This context vector is the final hidden state (cf. Equation 2.13) computed by the encoder RNN after processing the entire input sequence. Next, the decoder RNN utilizes the context vector $\mathbf{c}$ to generate the output sequence $\mathbf{y} = [\mathbf{y}_1, \dots, \mathbf{y}_{T'}]$ step-by-step. During each step, it takes the previously generated output, along with its current hidden state $\mathbf{h}_{t-1}$, and context

vector $\mathbf{c}$, to produce the next hidden state and a distribution over the vocabulary $\mathbf{y}_t$ for the next word in the output sequence:

$$\mathbf{h}_t = \sigma(\mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{W}_{yh}\mathbf{y}_{t-1} + \mathbf{W}_{ch}\mathbf{c}), \quad \mathbf{y}_t = \texttt{softmax}(\mathbf{W}_{hy}\mathbf{h}_t) \tag{2.15}$$

where the weights $\mathbf{W}_{hh}$, $\mathbf{W}_{yh}$, $\mathbf{W}_{ch}$, and $\mathbf{W}_{hy}$ are learnable parameters. This two-step process allows seq2seq models to produce output sequences of arbitrary length, independent of the input sequence length.

Attention Mechanism

While seq2seq models can handle input and output sequences of arbitrary lengths, they still face challenges with handling long-range dependencies in prolonged inputs. As the input sequence length gets longer, it becomes increasingly difficult for the encoder to capture and compress all relevant information in a single vector. This, on the other hand, hinders access to critical details for the decoder, making it challenging to generate coherent outputs for long input sequences. To mitigate this issue, an attention mechanism (Bahdanau et al., 2015) was proposed, which extends the seq2seq architecture to dynamically compute context vectors, allowing the decoder to selectively attend to different parts of the input at each decoding step.

More precisely, at each decoding step t, an attention weights α_{tj} measure the relevance of the T encoder hidden states $\left(\mathbf{h}_1, \dots, \mathbf{h}_j, \dots, \mathbf{h}_T\right)$ with respect to the current decoder hidden state $\mathbf{h'}_{t-1}$. Therefore, first, a score e_{tj} is calculated based on an alignment function, typically implemented as an FNN, that measures the relevance or alignment between the decoder state $\mathbf{h'}_{t-1}$ and the encoder state $\mathbf{h}_j$:

$$e_{tj} = \texttt{FNN}_{\texttt{score}}(\mathbf{h'}_{t-1}, \mathbf{h}_j) \tag{2.16}$$

Next, the attention weights are computed by applying a softmax over the scores, effectively turning them into a probability distribution:

$$\alpha_{tj} = \texttt{softmax}(e_{tj}) \tag{2.17}$$

Finally, the context vector $\mathbf{c}_t$ at decoding step t is calculated as a weighted sum of encoder hidden states using these attention weights:

$$\mathbf{c}_t = \sum_{j=1}^{T} \alpha_{tj}\mathbf{h}_j \tag{2.18}$$

This dynamically computed context vector provides the decoder with specific, relevant information from the encoder states at each decoding step. Incorporating attention significantly improves the performance of seq2seq models, enabling more effective modeling of long-range dependencies and better information flow between the encoder and decoder. Furthermore, the attention mechanism laid the groundwork for the current state-of-the-art Transformer architecture (Vaswani et al., 2017), now the de facto standard in NLP and has largely replaced traditional seq2seq models.

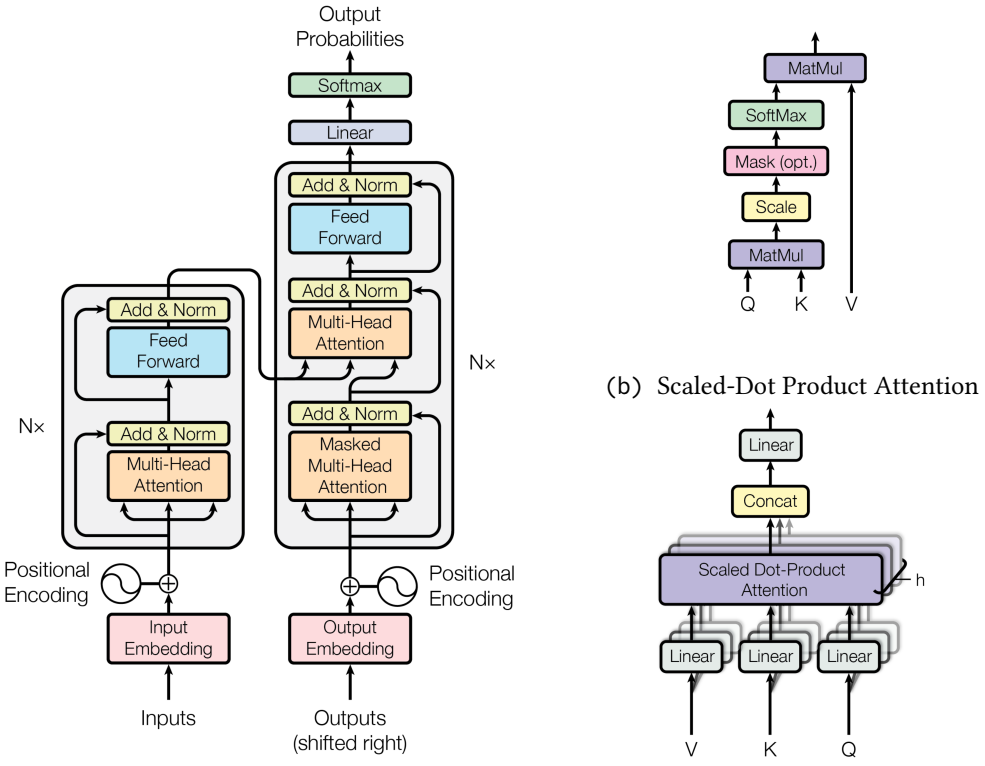
2.3 Transformer Models for Modern Natural Language Processing

As discussed in the previous section, recurrent architectures like RNNs and LSTMs marked a significant step forward in processing sequential data, i.e., text, for NLP, offering improved performance compared to traditional statistical models, especially in capturing long-range dependencies and word-level semantics. However, despite their advancements, they still face critical limitations, particularly with contextual semantics, prolonged input and output sequences, limited parallelization, and training efficiency. Inspired by the attention mechanisms in seq2seq models, Vaswani et al. (2017) proposed the "*transformer*" architecture addressing these limitations by entirely discarding recurrence and instead relying solely on attention mechanisms to directly model relationships between all words in the input sequence simultaneously. This architectural shift made training on huge datasets very efficient because it allows for highly parallelizable implementations on modern hardware accelerators like GPUs or TPUs. Due to their advantages, transformers revolutionized the field and led to unprecedented performance on a wide range of tasks, thus quickly becoming the foundational architecture for NLP and beyond.

In the following, we discuss transformer models and their key concepts, especially the self-attention mechanism at their core, different architectural variants, and discuss how the architecture has evolved into modern large language models (LLMs).

2.3.1 The Transformer Architecture

Transformer models (cf. Figure 2.4a) were introduced in the seminal paper "*Attention Is All You Need*" (Vaswani et al., 2017), which implemented them following an encoder-decoder architecture (cf. §2.3). Initially, transformers were designed for machine translation, where the encoder first maps an input sentence in the source language $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ to a sequence of latent representation $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)$. Given $\mathbf{z}$, the decoder then generates an output sequence $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_M)$ of words in the target language, one word at a time. Instead of two RNNs, Vaswani et al. (2017) employed transformer models as the encoder and decoder. More precisely, they used a stack of transformer encoder layers to form the encoder and a stack of transformer decoder layers to form the decoder. Each encoder layer consists of two main sub-layers: a multi-head self-attention layer and a 2-layer FFN with ReLU activation after the first layer. Similarly, each decoder layer consists of three main sub-layers: a masked multi-head self-attention layer, a multi-head cross-attention mechanism over the encoder's output, and a 2-layer FFN with ReLU activation after the first layer. Further, residual connections (He et al., 2016) are employed around each of the sub-layers, followed by layer normalization (Ba et al., 2016). In the following, we will cover the key components of the architecture in more detail.



(a) The original transformer architecture. Left: Encoder. Right: Decoder.

(b) Scaled-Dot Product Attention

(c) Multi-Head Attention

Figure 2.4: The original transformer architecture proposed as encoder-decoder seq2seq model for machine translation and its core attention mechanisms. Images taken from Vaswani et al. (2017).

The Self-Attention Mechanism

The self-attention mechanism is the core component of transformers, specifically designed to allow the model to assess how relevant each word in a sequence is to every other word simultaneously. More specifically, the mechanism transforms each input embedding into a contextual representation by computing a weighted sum of all input embeddings, where the weights are determined by the relevance of each word to the others.

Scaled Dot-Product Attention In their paper, Vaswani et al. (2017) implemented the self-attention mechanism as the scaled dot-product attention (SDPA) illustrated in Figure 2.4b. The input to SDPA is a set of N input vectors $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$, each of dimension d_{model}. Then, each input vector $\mathbf{x}_i$ is projected into three different vectors: a query vector $\mathbf{q}_i$, a key vector $\mathbf{k}_i$, and a value vector $\mathbf{v}_i$.

$$\mathbf{q}_i = \mathbf{x}_i \mathbf{W}_Q, \quad \mathbf{k}_i = \mathbf{x}_i \mathbf{W}_K, \quad \mathbf{v}_i = \mathbf{x}_i \mathbf{W}_V, \tag{2.19}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ are learned linear layers, and d_k is the dimension of the query, key, and value vectors. Typically, this is computed in parallel for all input vectors, resulting in matrices $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N \times d_k}$, where each row corresponds to the respective query, key, and value vectors.

Next, the attention scores are computed as the dot product of the query and key vectors and turned into a probability distribution by applying softmax:

$$\mathbf{A} = \texttt{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right), \quad \mathbf{A} \in \mathbb{R}^{N\times N} \tag{2.20}$$

where $\mathbf{A}_{ij}$ is the attention score between the i-th query and the j-th key and the division by $\sqrt{d_k}$ is a scaling factor introduced to improve training stability.

The final output of SDPA, i.e., the matrix of contextual representations, is computed from the sum weighted by the attention scores and the value vectors:

$$\texttt{SDPA}(Q, K, V) = \mathbf{A}\mathbf{V} = \texttt{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \quad \mathbf{V}, \texttt{SDPA}(\cdot) \in \mathbb{R}^{N\times d_k} \tag{2.21}$$

Multi-Head Attention To enhance the token representations further, transformers implemented multi-head attention (MHA; illustrated in Figure 2.4c), which computes multiple SDPAs in parallel, each with different learned linear projections. This enables each attention head to focus on different characteristics and relationships in the input sequence.

Formally, in MHA, the output of H parallel attention heads are concatenated and linearly projected to the model dimension d_{model}:

$$\texttt{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \texttt{Concat}(\texttt{head}_1, \dots, \texttt{head}_H)\mathbf{W}_O, \quad \texttt{MHA}(\cdot) \in \mathbb{R}^{N\times d_{\text{model}}} \tag{2.22}$$

where $\mathbf{W}_O \in \mathbb{R}^{d_{\text{model}}\times d_{\text{model}}}$ is a learned linear projection, and each attention head is computed as:

$$\texttt{head}_i = \texttt{SDPA}(\mathbf{Q}\mathbf{W}_Q^{(i)}, \mathbf{K}\mathbf{W}_K^{(i)}, \mathbf{V}\mathbf{W}_V^{(i)}), \quad \texttt{head}_i(\cdot) \in \mathbb{R}^{N\times d_k} \tag{2.23}$$

where $\mathbf{W}_Q^{(i)}, \mathbf{W}_K^{(i)}, \mathbf{W}_V^{(i)} \in \mathbb{R}^{d_{\text{model}}\times d_k}$ are learned linear projections for the i-th attention head and $d_k = d_{\text{model}}/H$ is the dimension of the query, key, and value vectors in each head.

Masked Attention In the decoder, the self-attention mechanism is modified to prevent the model from attending to future tokens, which is crucial for autoregressive generation tasks like language modeling. This is simply done by setting the attention scores for future tokens to $-\infty$ before applying softmax so that they do not contribute to the attention distribution. Usually, this is implemented by using a triangular mask, which is applied to the attention scores before softmax:

$$\mathbf{A} = \texttt{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \mathbf{M}\right), \quad \mathbf{M}_{ij} = \begin{cases} 0 & \text{if } i \geq j \\ -\infty & \text{otherwise} \end{cases} \tag{2.24}$$

where $\mathbf{M} \in \mathbb{R}^{N\times N}$ is a matrix to mask-out the future tokens.

Cross-Attention Another difference between encoder and decoder layers in the original transformer architecture is the cross-attention mechanism in the decoder. This modified attention mechanism allows the decoder to attend to the encoder's output so that it can incorporate meaningful information from the contextual representations from the input sequence. Computationally, cross-attention is similar to self-attention:

$$\texttt{CrossAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \texttt{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}), \quad \mathbf{Q} \in \mathbb{R}^{N_{\text{dec}} \times d_{\text{model}}}, \quad \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N_{\text{enc}} \times d_{\text{model}}} \tag{2.25}$$

where $\mathbf{Q}$ is the query from the decoder, and $\mathbf{K}, \mathbf{V}$ are the keys and values from the encoder.

Positional Encodings

One critical aspect of the transformer architecture is that it does not inherently capture the sequential order of the input tokens because it processes all tokens at once in parallel. To address this, Vaswani et al. (2017) introduced so-called positional encodings, which hold information about the position of each token in the sequence and are added to the input embeddings before entering the first encoder or decoder layer. In their original implementation, these encodings are computed from interleaving sine and cosine functions of different frequencies:

$$\mathbf{p}_i = \begin{cases} \sin\left(\frac{i}{10000^{2j/d_{\text{model}}}}\right) & \text{if } j \mod 2 == 0 \\ \cos\left(\frac{i}{10000^{2j/d_{\text{model}}}}\right) & \text{if } j \mod 2 == 1 \end{cases}, \quad i = 1, \dots, N, \quad j = 1, \dots, d_{\text{model}} \tag{2.26}$$

where $\mathbf{p}_i \in \mathbb{R}^{d_{\text{model}}}$ is the positional encoding for the i-th token, and N is the maximum sequence length. This results in a unique positional encoding for each token position, which is then added to the input embeddings to provide the model with information about the sequential order of the input tokens:

$$\mathbf{x}_i = \mathbf{W}^{(i)}_{\text{emb}} + \mathbf{p}_i, \quad \mathbf{x}_i \in \mathbb{R}^{d_{\text{model}}}, \quad \mathbf{W}^{(i)}_{\text{emb}} \in \mathbb{R}^{|\mathcal{V}| \times d_{\text{model}}} \tag{2.27}$$

Layer Normalization (LayerNorm)

When training on large datasets, it is very probable that the samples therein are not normalized, i.e., the input features have different scales and distributions. This can be problematic for training deep neural networks, as it can lead to vanishing or exploding gradients and other instability issues, making it difficult to train the model effectively (Ioffe et al., 2015; Ba et al., 2016). In order to stabilize training and fasten convergence, it was found that normalizing the input features to each layer improves the training process significantly. Most, if not all, transformer models employ layer normalization (Ba et al., 2016) (LayerNorm) after each sub-layer, which normalizes the input along the feature dimension so that they have the same distribution. Therefore, first, the mean and variance of an input feature $x \in \mathbb{R}^{d_k}$ are computed:

$$\mu = \frac{1}{d} \sum_{i=1}^{d_k} x_i, \quad \sigma^2 = \frac{1}{d} \sum_{i=1}^{d_k} (x_i - \mu)^2 \tag{2.28}$$

Then, the input is normalized by subtracting the mean and dividing by the standard deviation:

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}, \quad \hat{\mathbf{x}} = \frac{\mathbf{x} - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{2.29}$$

where ϵ is a small constant to avoid division by zero. Finally, the normalized input is scaled and shifted by learned parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ by element-wise multiplication and addition, respectively:

$$\mathbf{y} = \boldsymbol{\gamma} \odot \hat{\mathbf{x}} + \boldsymbol{\beta}, \quad \mathbf{y}, \hat{\mathbf{x}}, \boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^{d_k} \tag{2.30}$$

This results in the LayerNorm output $\mathbf{y}$, which is then passed to the next layer.

Position-wise Feed-Forward Networks (FFN)

After each self-attention or cross-attention sub-layer, the output is passed through a 2-layer feed-forward network (FFN) with ReLU activation after the first layer. This FFN is applied independently to each token $\mathbf{x}_i \in \mathbb{R}^{d_\text{m}}$ in the sequence, which is why it is often referred to as a position-wise FFN. Typically, it consists of two linear layers with a non-linear ReLU (Agarap, 2018) activation in between, which first layer $\mathbf{W}_1 \in \mathbb{R}^{d_\text{m} \times 4d_\text{m}}$ projects the input to a higher-dimensional space and the second layer $\mathbf{W}_2 \in \mathbb{R}^{4d_\text{m} \times d_\text{m}}$ projects it back to the original dimension d_m:

$$\texttt{FFN}(\mathbf{x_i}) = \mathbf{y_i} = \texttt{ReLU}(\mathbf{x_i}\mathbf{W}_1)\mathbf{W}_2 \tag{2.31}$$

where $\mathbf{y}_i \in \mathbb{R}^{d_\text{m}}$ is the output of the FFN for the i-th token. Note that the magic number, i.e., the dimension of the hidden layer in the FFN, is usually set to $4d_\text{m}$, as proposed in the original transformer implementation. Since the attention mechanism comprises primarily linear transformations, the FFN layers, which also typically account for the majority of a transformer's parameters, are crucial because they introduce non-linearity into the model, enabling it to learn complex patterns and relationships.

2.3.2 Transformer Encoder Models

Transformer encoder models such as BERT (Devlin et al., 2019) omit the decoder entirely and are built only with a stack of encoder layers relying solely on self-attention to process the input sequence. Typically, these models are pre-trained using "*masked language modeling*" (MLM) and "*next sentence prediction*" (NSP) on large generic text corpora. In MLM, random words are masked out, and the model's task is to predict the missing words based on the context. In NSP, two sentences are given, and the model must classify whether the second sentence in the input is a successor of the first sentence. This process enables deep semantic understanding and rich contextual embeddings for each token in the input sequence. The resulting contextual embeddings can then be leveraged for various NLP downstream tasks, such as text classification, named entity recognition, or question answering. Often, a special "*CLS*" or "*class*" token is prepended to the input sequence, which is used to aggregate the contextual representations of the entire sequence into a single vector representation, which can then be used for downstream. This is achieved

by either fine-tuning the model on the specific task or using the embeddings as input features for a separate classifier, often realized as a simple FFN.

Formally, the input to a transformer encoder model is a sequence of N input vectors $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ each of which gets transformed simultaneously into contextual representations $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)$ by passing it through a stack of L encoder layers. Usually, the input vectors are word embeddings ($\mathbf{x} \in \mathbb{R}^{d_{\text{model}}}$), jointly learned via an additional embedding layer ($\mathbf{W}_{\text{emb}} \in \mathbb{R}^{|\mathcal{V}| \times d_{\text{model}}}$) from a vocabulary $\mathcal{V}$ of $|\mathcal{V}|$ tokens generated by a tokenizer (cf. §2.2.3), such as WordPiece (Wu et al., 2016) or SentencePiece (Kudo et al., 2018), which are then additively combined with positional embeddings as described in §2.3.1.

An overview of a classical transformer encoder and its components is illustrated in Figure 2.5.

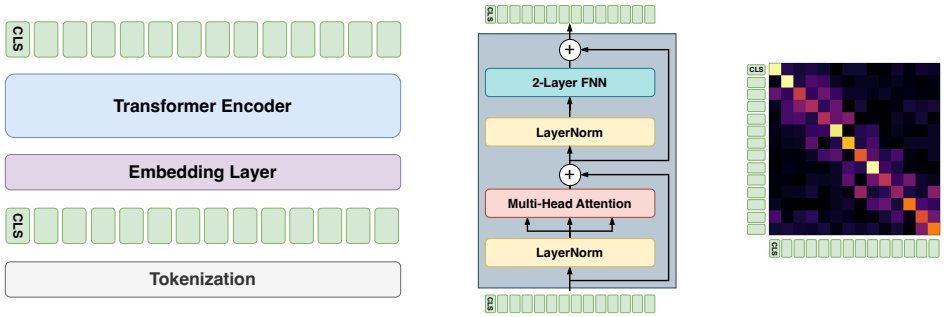


Figure 2.5: Left: An overview of the classical transformer encoder architecture. Center: Inner architecture of a transformer encoder layer with multi-head attention. Right: A random examplatory attention matrix.

2.3.3 Transformer Decoder Models

As opposed to encoder models, transformer decoder models, such as GPT (Radford et al., 2018), are designed to generate coherent text. These models only consist of a stack of decoder layers exclusively relying on the masked self-attention mechanism, sometimes also referred to as causal self-attention. The training objective is typically language modeling, where the model learns to predict the next token in a sequence based on the previous tokens in an autoregressive manner, i.e., one token at a time. An overview of a classical transformer decoder and its components is illustrated in Figure 2.6.

This is achieved by passing a sequence of N input vectors $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ akin to the encoder input through a stack of L decoder layers to produce an output sequence $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)$ of contextual representations from the causal self-attention mechanism. Note that usually a special "*BOS*" or "*beginning of sequence*" token is prepended to the input sequence, which is used to indicate the start of the sequence. Next, the last out representation of the final decoder layer output $\mathbf{z}_N$, is forwarded through a so-called "*language model head*", which is a linear layer that projects $\mathbf{z}_N$ to the vocabulary size $|\mathcal{V}|$:

$$\mathbf{y}_N = \mathbf{z}_N \mathbf{W}_{\text{lm}}^{\top}, \quad \mathbf{y}_N \in \mathbb{R}^{|\mathcal{V}|}, \quad \mathbf{W}_{\text{lm}} \in \mathbb{R}^{d_{\text{model}} \times |\mathcal{V}|} \tag{2.32}$$

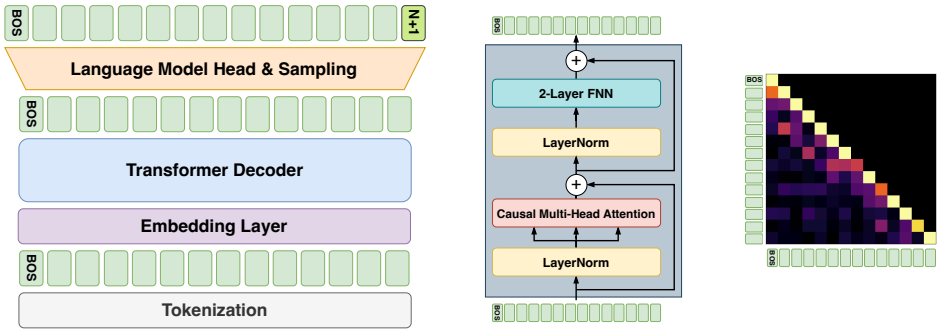


Figure 2.6: Left: An overview of the classical transformer decoder architecture including sampling of the next token. Center: Inner architecture of a transformer decoder layer with causal multi-head attention. Right: A random examplatory causal attention matrix.

To generate the next token, first, a softmax function is applied to the output vector $\mathbf{y}_N$ to obtain a pseudo probability distribution over the vocabulary:

$$\mathrm{P}(\mathbf{x}_{N+1}|\mathbf{x}_1, \ldots, \mathbf{x}_N) \equiv \mathbf{p}_N = \texttt{softmax}(\mathbf{y}_N), \quad \mathbf{p}_N \in \mathbb{R}^{|\mathcal{V}|} \tag{2.33}$$

Finally, the next token $\mathbf{x}_{N+1}$ is sampled from this distribution, which is then added to the input sequence for the next step. The simplest sampling strategy, besides more advanced techniques like nucleus sampling or beam search, is to select the token with the highest probability, also known as greedy decoding:

$$i = \mathrm{argmax}\,(\mathbf{p}_N) \quad \text{and} \quad \mathbf{x}_{N+1} = \mathbf{W}_{\text{emb}}^{(i)}, \quad \mathbf{W}_{\text{emb}} \in \mathbb{R}^{|\mathcal{V}| \times d_{\text{model}}} \tag{2.34}$$

This autoregressive generation process is repeated until a stopping criterion is met, such as reaching a maximum sequence length or generating a special "*EOS*" or "*end-of-sequence token*".

2.3.4 Large Language Models (LLMs)

In the following, we will briefly discuss the evolution of transformer decoder models into modern large language models (LLMs), which have become the de facto standard in current NLP, as well as the basics of LLM training.

From Transformer Decoders to Large Language Models

The successful implementation of language models as autoregressive transformer decoders like the Generative Pre-trained Transformer (GPT) (Radford et al., 2018) model marked a significant milestone in NLP. In their paper, the authors showed that training a transformer decoder on large text corpora in an unsupervised manner via standard language modeling objective on large text corpora, referred to as "*pre-training*", yields powerful task-agnostic foundation models that require only minimal task-specific fine-tuning to achieve state-of-the-art performance on various NLP tasks. Following this approach, subsequent models, such as GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), and many others (Yang et al., 2019; Keskar et al., 2019; Lieber et al., 2021; BigScience et al., 2022; Zhang et al.,

2022; *inter alia*), scaled up the model size[2] and training data in multiple orders of magnitude, leading to greatly improved performance and capabilities. Due to their size, these models are referred to as "*large language models*". However, despite some architectural improvements, the scaling of model size and training data, as well as new training paradigms, the core of LLMs, i.e., the transformer decoders with causal self-attention, remains largely unchanged from the original GPT model. In their paper, Brown et al. (2020) showed that scaling up not only improves performance on existing tasks but also enables LLMs to perform well on unseen tasks without any fine-tuning by providing the models with a task description and a few examples of the new tasks during inference (few-shot) or even without any examples (zero-shot). This phenomenon is known as "*in-context learning*" (ICL) because the model learns from the given context and happens during inference time, not from pre-training or fine-tuning. Providing the model with a task description and examples is often referred to as "*prompting*" the model, and the task description and examples are called "*prompts*". While scaling and ICL lead to significant improvements in numerous tasks, the great breakthroughs that led to the current generation of LLMs were new training paradigms called "*instruction tuning*" (Wei et al., 2023) and "*preference alignment*" (Ouyang et al., 2022).

LLM Training

The training pipeline for contemporary LLMs usually consists of three main stages: pre-training, instruction tuning, and preference alignment.

Pre-training The pre-training phase is the first stage of LLM training and arguably the most crucial one since the model learns the fundamental language understanding and world knowledge during this stage. Therefore, the model is trained in an unsupervised fashion on huge general-purpose text corpora containing hundreds of GigaBytes of texts from Common Crawl dumps, Wikipedia, various books, scientific articles, and other sources (Gao et al., 2020). The pre-training is called unsupervised, or more precisely, self-supervised training, because the model learns to predict the next token in a sequence without any explicit labels, i.e., the next token is simply the next word in the text corpus. This is typically performed on large distributed computing clusters with hundreds or thousands of GPUs or TPUs (Chowdhery et al., 2023), allowing for efficient training on massive datasets. Notably, the pre-training data also often contains programming code from various programming languages, which enables the model to learn programming concepts and syntax.

The pre-training objective is language modeling, where the model is trained to predict the next token in a sequence based on the previous tokens. Formally, this is achieved by minimizing the negative log-likelihood (NLL) practically implemented

2. GPT-1 has 117M parameters, GPT-2 has 1.5B parameters, and GPT-3 has 175B parameters.

as the cross-entropy loss (CEL) between the predicted token probabilities and the actual next token in the sequence:

$$\mathcal{L}_{\text{NLL}} = -\sum_{i=1}^{N} \log \mathrm{P}\big(x_i \mid x_1, \ldots, x_{i-1}\big) \equiv \mathcal{L}_{\text{CEL}} \tag{2.35}$$

$$\mathcal{L}_{\text{CEL}} = -\sum_{i=1}^{N} \sum_{j=1}^{|\mathcal{V}|} y_{i,j} \log \hat{y}_{i,j}, \quad \mathbf{y}_i, \hat{\mathbf{y}}_i \in \mathbb{R}^{|\mathcal{V}|} \tag{2.36}$$

where $\mathbf{y}_i \in \mathbb{R}^{|\mathcal{V}|}$ is a one-hot vector with $y_{i,j} = 1$ if the j-th token in the vocabulary $\mathcal{V}$ is the actual next token after the ith token, and $\hat{\mathbf{y}}_i \in \mathbb{R}^{|\mathcal{V}|} = \texttt{softmax}(\mathbf{z}_i \mathbf{W}_{\text{lm}}^{\top})_j$ are the predicted token probabilities computed by the language model head $\mathbf{W}_{\text{lm}} \in \mathbb{R}^{d_{\text{model}} \times |\mathcal{V}|}$ from the output representation $\mathbf{z}_i \in \mathbb{R}^{d_{\text{model}}}$ of the final decoder layer.

Instruction Tuning After pre-training, the model acquired a strong understanding of language and world knowledge, but since it was trained only to predict the next token in a sequence, it is not yet capable of performing specific tasks or following instructions. To address this, the model is further trained in a supervised manner on a much smaller dataset, often referred to as "*instruction tuning*" (Wei et al., 2023) or "*supervised fine-tuning*". Such a dataset is usually a mixture of other datasets from various domains and tasks, such as question answering, summarization, translation, sentiment analysis, coding problems, and many others, which are reformatted into instruction-output pairs. For example, a question-answering sample can be reformatted into an instruction like "Answer the following question: *What is the capital of France?*" with the ground-truth answer being "Paris". Since the input and output of the model are both sequences of words or tokens, the model is trained to predict the output sequence given the input instruction sequence using the same language modeling objective as in pre-training. After the instruction tuning stage, the model learned to follow instructions and perform various tasks, and can be thought of as a task-agnostic foundation model.

Preference Alignment After instruction tuning, the model is capable of following instructions and performing various tasks, but it is not yet aligned with human values and or company preferences, which is crucial for safe deployment in real-world applications. For example, the model might generate harmful or biased content, produce inconsistent or non-facetious stories, or output text in a tone or style that does not match a company's values.

To address these issues, the model undergoes another round of supervised training on a dataset of instruction-output pairs, but this time, the dataset and training objective differ substantially from the previous stages. This process is often referred to as "*preference alignment*" and was introduced by Ouyang et al. (2022) who implemented it as "*reinforcement learning from human feedback*" (RLHF). As the name suggests, the authors applied sophisticated techniques from reinforcement learning (RL) (Schulman et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020) to optimize the model's behavior, i.e., its output, based on human feedback on the model's responses to various instructions. This involves additional auxiliary

models, such as a reward model, which is trained to predict the quality of the model's responses based on human-ranked responses, and a policy model, which is used to update the model's parameters based on the reward signal. While RLHF and multiple other RL approaches (Bai et al., 2022; Glaese et al., 2022; Dai et al., 2024; Shao et al., 2024) is an effective approach for preference alignment, due to its complex implementation, which we will not cover here, its large compute and memory requirements, instability issues during training, and the need for costly human feedback datasets, simpler and more efficient non-RL approaches have been proposed in recent years. Besides others (Azar et al., 2024; Hong et al., 2024), one of the most popular and widely used non-RL approaches due to its simplicity is Direct Preference Optimization (DPO) (Rafailov et al., 2023), which directly optimizes the LLM without the need for a reward model and simpler supervised fine-tuning on a dataset of human preferences in the form of an instruction, positive response, and negative response.

Architectural Improvements

While the core architecture of LLMs is still based on the original transformer decoder, several architectural improvements have been introduced in recent years to enhance their performance and capabilities.

Relative Positional Encodings One limitation of the vanilla transformer architecture is the use of absolute positional encodings (cf. §2.3.1), which encode the absolute position of each token in the sequence. This limits the model to a fixed maximum sequence length, often called "*context window*", which is fixed and determined by the dimensions of the positional encoding matrix. To overcome this, Shaw et al. (2018) proposed *relative position encodings*, which encode the relative distances between tokens rather than their absolute positions and allow the model to generalize to longer sequences than seen during training. Typically, relative position encodings are computed as a function of the distance between tokens and are added directly to the attention scores before applying softmax instead of added to the input embeddings. Current state-of-the-art LLMs (Achiam et al., 2023; Touvron et al., 2023; Jiang et al., 2023; Team et al., 2024) often use advanced positional encoding techniques such as ALiBi (Press et al., 2022) or RoPE (Su et al., 2024) to support multiple orders of magnitude larger context windows than the early GPT models.

Pre-Norm In the original transformer architecture, the LayerNorm is applied after each sub-layer, i.e., after the self-attention or cross-attention and the FFN. However, it was found that this can lead to training instability, especially when scaling up the model size (Child et al., 2019; Xiong et al., 2020). To address this issue, modern transformers apply the normalization before each sub-layer, which is often referred to as "*pre-norm*" and has become the defacto-standard approach. Further, there are other normalization techniques, such as "*RMSNorm*" (Zhang et al., 2019), which are commonly used in current LLMs.

Efficient Attention Mechanisms Another limitation of the vanilla transformer architecture is the quadratic runtime complexity $O(N^2)$ of the self-attention mechanism, which makes it computationally expensive for long sequences. To overcome this issue, numerous more efficient attention mechanisms (Tay et al., 2022) have been proposed, which reduce the computational complexity by trading off accuracy for efficiency or memory for computation. This is typically achieved by computing the attention scores only for a subset of the input tokens, e.g., via sparse attention (Child et al., 2019) or sliding window attention (Beltagy et al., 2020), by approximating the attention scores using low-rank approximations (Wang et al., 2020) or kernel methods (Choromanski et al., 2020), by reducing the number of key and value projections (Shazeer, 2019), or other techniques involving considerable architectural changes (Dai et al., 2019). Further, very efficient, hardware-optimized algorithms such as FlashAttention (Dao et al., 2022; Dao, 2024) have been developed, which allow for efficient computation on modern hardware accelerators by elaborately allocating GPU memory and improving cache utilization. Most modern LLMs employ grouped-query attention (Ainslie et al., 2023), an improved successor of the multi-query attention (MQA) (Shazeer, 2019) mechanism, which also reduces the number of key and value projections by sharing them across groups of query heads in combination with FlashAttention or similar hardware optimized algorithms.

KV-Caching Key-Value (KV) caching is a technique used to improve the efficiency of transformer decoder models during inference. In the default architectures, self-attention requires the computation of attention scores for all pairs of tokens in the input sequence and, therefore, has a quadratic runtime complexity $O(N^2)$, which is computationally expensive for long sequences. KV caching addresses this issue by trading off memory for computation, i.e., it caches the key and value matrices K and V of previously processed tokens. In subsequent forward passes, K and V are not recomputed for the previous tokens, but instead, the cached representations are reused. This means that only for the newly generated token x_t, the q_t, k_t, and v_t vectors are computed, and the attention scores are computed only between the new token and the cached keys and values:

$$\mathbf{A}_t = \texttt{softmax}\left(\frac{\mathbf{q}_t\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \quad \mathbf{A}_t \in \mathbb{R}^{1\times(N+1)} \tag{2.37}$$

where $\mathbf{q}_t \in \mathbb{R}^{d_k}$ is the query vector for the new token x_t, and $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{(N+1)\times d_k}$ are the cached key and value matrices from the previous tokens with k_t and v_t appended, and N is the number of previously processed tokens. This allows the model to avoid redundant computations and significantly speeds up inference, especially when generating long sequences. As of 2024, KV caching is a standard technique with various optimized variants and has been widely adopted in most LLM implementations and inference libraries (Li et al., 2024a).

Mixture of Experts (MoE) By scaling LLMs to hundreds of billions or even trillions of parameters (Kaplan et al., 2020; Chowdhery et al., 2023; Achiam et al., 2023) several challenges arise, such as increased computational costs, memory

requirements, and training time. While those issues cannot be completely resolved, one approach to mitigate them is to leverage *mixture of experts* (MoE) architectures introduced in the 1990s (Jacobs et al., 1991; Jordan et al., 1993) in the form of "*dense*" MoE architectures. Later, a "*sparse*" MoE architecture has been successfully applied to LSTM-based LMs (Shazeer et al., 2017) and transformer-based LMs (Lewis et al., 2021). The idea of MoE architectures is to have different layers, referred to as "*experts*", which are usually implemented as shallow FFNs and specialized to a certain "tasks". In a forward pass through a sparse MoE layer, only a subset of the experts is "activated", i.e., used to process an input token while the remaining experts are not used. The experts are dynamically selected based on a "*routing*" or "*gating*" network, which assigns each token to different experts based on the token's representation. After forwarding the tokens through the respective experts, the outputs are combined, usually by a weighted sum, to compute the final output of a MoE layer. In an MoE transformer model, the FNN layer in each transformer layer is usually replaced by an MoE layer with multiple experts and a shallow linear layer as the gating network. Note that the above explanation only describes the basic idea of MoE architectures, whereas the actual implementation, especially in modern LLMs, is much more complex and involves various optimizations and techniques to ensure efficient training and inference.

2.4 Multimodal Transformer Models for Vision-Language Tasks

This section covers how transformer models can be adapted to process multimodal data, specifically focusing on vision-language tasks. We will explore various architectures, including encoder-only vision-language models (VLMs), as well as generative decoder-only large vision-language models (LVLMs), their inner workings, and training methodologies. Further, we will discuss models that are particularly important for this thesis as they have been employed throughout the papers core to this dissertation.

2.4.1 Vision Transformers (ViTs)

Vision Transformers (ViTs), first introduced by Dosovitskiy et al. (2021) and illustrated in Figure 2.7, are a type of transformer encoder models capable of processing visual data, i.e., images, and therefore play a crucial role as a core component in many VLMs and LVLMs. While there have been numerous variations and advancements in the architecture, training, and scaling of ViTs (Touvron et al., 2021; Caron et al., 2021; Bao et al., 2021; He et al., 2021; Oquab et al., 2025; Dehghani et al., 2023a), this section briefly introduces the basic ViT architecture and its key components.

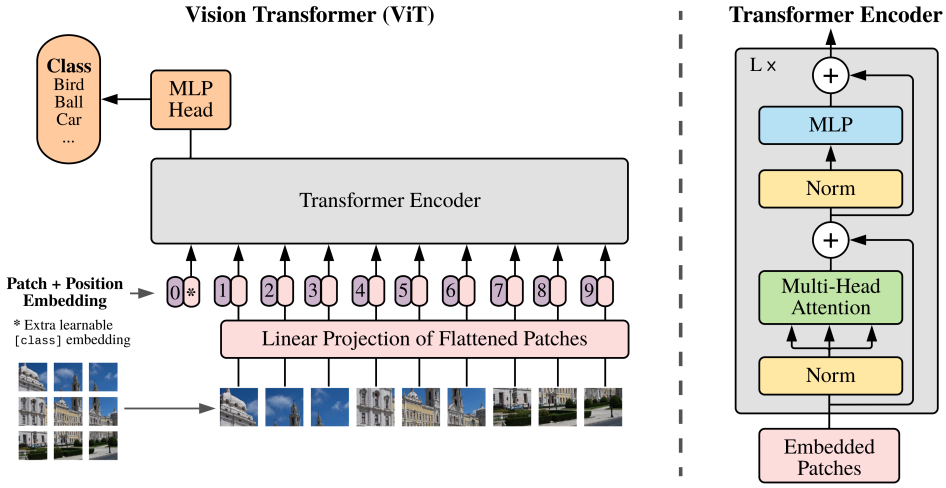


Figure 2.7: Left: An overview of the vanilla Vision Transformer (ViT) architecture. Right: Inner architecture of a transformer encoder layer. Image taken from (Dosovitskiy et al., 2021).

Architecture Overview

The general idea behind ViTs is to decompose an image into smaller patches, which are then treated as tokens similar to words in text-based transformer models. With this approach, ViTs can be implemented as a standard transformer encoder, with the self-attention mechanism to capture dependencies and relationships within an image. However, since images are two-dimensional and represented as a grid of RGB values, they must be preprocessed to feed them into the transformer encoder.

Image Preprocessing

Before an image can be processed by a ViT, it has to be preprocessed into a sequence of token embeddings by applying the following steps.

Image Patching A naive approach would be to use every pixel of the image as a token, but this would result in an extremely long sequence, making it computationally infeasible due to the quadratic complexity of the self-attention mechanism in transformers. As proposed by Dosovitskiy et al. (2021), a more efficient approach is to divide the image into smaller, non-overlapping patches, which are then flattened and projected into the patch embedding space using a linear projection layer. Formally, given an image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ with height H, width W, and C channels, the image is divided into N patches of size $P \times P$, where $N = \frac{H \cdot W}{P^2}$. The patched image is then represented as a matrix $\mathbf{I}' \in \mathbb{R}^{N \times (P^2 \times C)}$, where each row corresponds to a flattened patch. For example, given an RGB image $\mathbf{I} \in \mathbb{R}^{224 \times 224 \times 3}$ and a patch size of $P = 16$, the image is patched into $N = \frac{224 \cdot 224}{16^2} = 196$ patches of size $16 \times 16 \times 3$ represented as $\mathbf{I}' \in \mathbb{R}^{196 \times 768}$.

Patch Embedding To forward the patches through the encoder layers, their dimensions must match the hidden dimension d_{m} of the model. Hence, each

patch is projected by a linear layer $\mathbf{W}_{\text{patch}} \in \mathbb{R}^{(P^2 \cdot C) \times d_{\text{m}}}$, resulting in a sequence of patch embeddings $\mathbf{E} \in \mathbb{R}^{N \times d_{\text{m}}}$.

Class Token Since the ViT model was designed to learn strong representations for image classification tasks, the authors took inspiration from the successful text-only encoder model BERT (Devlin et al., 2019) and introduced a special token to aggregate the information from all patches. More specifically, in addition to the patch embeddings, a special "*class token*" or "*CLS*" token $\mathbf{e}_{\text{cls}} \in \mathbb{R}^{d}_{\text{m}}$ is prepended to the sequence of patch embeddings $\mathbf{E}$, resulting in the final input sequence $\mathbf{E}' \in \mathbb{R}^{(N+1) \times d_{\text{m}}}$.

Position Embeddings Since a transformer model on its own does not preserve the sequential order of the input tokens (cf. §2.3.1), positional embeddings are added to the patch embeddings to retain this information. While images are inherently two-dimensional, Dosovitskiy et al. (2021) found that one-dimensional positional embeddings work surprisingly well for ViTs. This positional embedding layer is implemented as a learnable linear layer $\mathbf{W}_{\text{pos}} \in \mathbb{R}^{(N+1) \times d_{\text{m}}}$, where N is the number of patches and d_{m} is the hidden dimension of the model. This results in a final input sequence $\mathbf{X} = \mathbf{E}' + \mathbf{W}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$, which is then fed into the transformer encoder layers.

Training

In the original paper, the training of ViTs for image classification tasks consisted of two supervised stages.

In the first stage, the models are pre-trained on large datasets comprising (hundreds of) millions of images with class labels via cross-entropy loss. More precisely, the representation of the class token $\mathbf{e}^{(L)}_{\text{cls}} \in \mathbb{R}^{d_{\text{m}}}$ after the final encoder layer is passed through a linear layer[3] or "*classification head*" $\mathbf{W}_{\text{cls}} \in \mathbb{R}^{d_{\text{m}} \times C}$, where C is the number of classes, to produce logits $\mathbf{z}$. Next, the logits are used to compute the cross-entropy loss with respect to the one-hot encoded ground truth label $\mathbf{y} \in \mathbb{R}^{C}$:

$$\mathcal{L}_{CEL}(\mathbf{z}, \mathbf{y}) = -\sum_{i=1}^{C} y_i \log \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)}, \quad \mathbf{z} = \mathbf{e}^{(L)}_{\text{cls}} \mathbf{W}_{\text{cls}}, \quad \mathbf{z}, \mathbf{y} \in \mathbb{R}^{C} \tag{2.38}$$

The goal of this stage is for the model to learn strong general representations of the images that can be used for downstream tasks.

In the second stage, the pre-trained model is fine-tuned on a smaller dataset for a specialized downstream classification task. Therefore, the classification head of the pre-trained model is discarded, and a new classification head adjusted to the number of classes in the downstream task is trained in the same way as in the pre-training stage. This strategy is also referred to as "*transfer learning*" because it aims to transfer the knowledge gained from the pre-training phase to downstream tasks and is widely used in modern machine learning.

3. Originally, the authors propose to use a two-layer MLP but most implementations use a single linear layer.

The authors also experimented with a different, unsupervised training stage, where the model is trained to predict the masked-out patches of an image. This approach is similar to the masked language modeling (MLM) objective used in BERT (Devlin et al., 2019) and has been shown to be effective for learning strong representations of images. Following this approach, other successful ViT variations such as BEiT (Bao et al., 2021), MAE (He et al., 2021), or DINO (Caron et al., 2021; Oquab et al., 2025) have been proposed, which use different masking strategies and training objectives to learn strong image representations.

Usage of ViTs in this Thesis

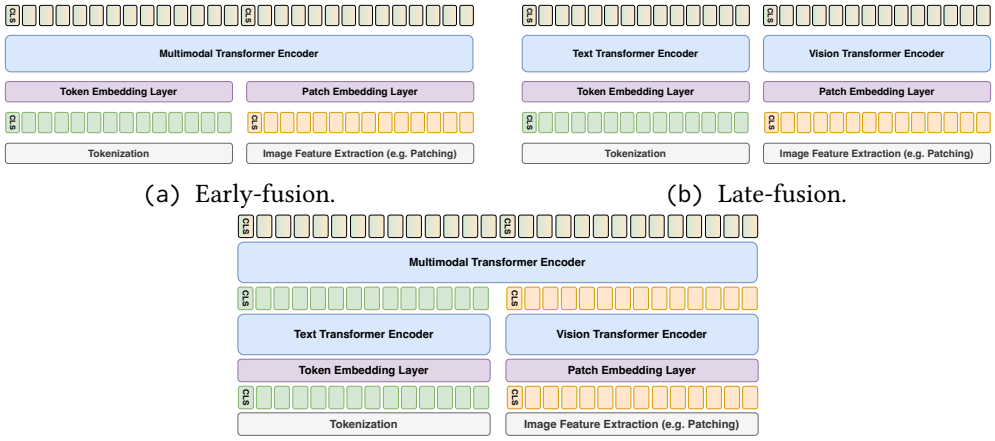
Vision Transformers are a core component of modern VLMs and LVLMs, which are leveraged in all foundational as well as additional papers of this thesis (cf. §1.4). In particular, we use pre-trained ViTs as image encoders to extract image features from images, which are then processed by the respective VLM or LVLM. Moreover, in (Schneider et al., 2025b), we also use a pre-trained DINOv2 (Oquab et al., 2025) to extract image features from video frames in order to compute similarities to other images.

2.4.2 Encoder-based Vision-Language Models (VLMs)

As defined in §1.2, throughout this thesis, we refer to vision-language models (VLMs) as transformer encoder-based models that compute joint representations in a shared embedding space. In other words, with VLMs, the goal is to learn semantically rich image and text representations that can be used to measure the similarity between the two modalities and various other downstream tasks. This section provides an overview of the different high-level architectures and training strategies of VLMs, as well as some seminal VLM foundation models core to this thesis.

VLM Architectures

As comprehensively covered in recent survey papers (Zhang et al., 2024b; Shaikh et al., 2024), there is a plethora of VLM architectures which can generally be categorized into two main types and a combination of both: "*early-fusion*", "*late-fusion*" (cf. Figure 2.8), and "*hierarchical-fusion*" models. Simply put, "*early-fusion*" models process both modalities simultaneously in a single multimodal encoder, "*late-fusion*" or ""*dual-encoder* models process each modality in a separate text and image encoder and then combine the representations at a later stage, and "*hierarchical-fusion*" models combine both approaches by processing each modality first separately in a text and image encoder and then passing the representations jointly through a multimodal encoder. In the following, we briefly describe these three types of VLMs from an architectural perspective. Since the training of VLMs is often model- and task-specific, we will cover the training strategies for the VLMs core to this thesis in their respective sections.



(a) Early-fusion. (b) Late-fusion.

(c) Hierarchical-fusion.

Figure 2.8: High-level overview of typical architectures for transformer encoder-based vision-language models (VLMs). The figures are inspired by (Shaikh et al., 2024).

Early-Fusion VLMs Early-fusion VLMs comprise a single multimodal transformer encoder that simultaneously processes text and image inputs. Therefore, a text is first tokenized and embedded into a sequence of text embeddings, while an image is patched and embedded into a sequence of patch embeddings. Note that image patching and embedding, as described in 2.4.1, is only one of many approaches for extracting a sequence of features from an image. Other approaches often leverage convolutional neural networks (CNNs) (LeCun et al., 1998) such as ResNet (He et al., 2016) or Faster R-CNN (Ren et al., 2016) to extract features from an image. If the text and image embeddings do not already have the same hidden dimension d_{m} of the multimodal encoder, they are typically projected into this shared space by a simple linear layer. Next, both embedding sequences are concatenated and passed through the multimodal transformer encoder, where the self-attention mechanism captures the relationships between the text and image tokens. The output of the multimodal encoder is a joint representation of the text and image, which can be used for various downstream tasks by applying a task-specific head on (a subset of) the output tokens.

While early-fusion VLMs can capture fine-grained complex interactions between text and image tokens, they are more computationally expensive due to the quadratic complexity of the self-attention mechanism. Furthermore, since the resulting joint representations are contextualized, i.e., the representation of a text token depends on the image and vice versa, they are unsuitable for efficient online retrieval tasks. This is because, at inference time, to calculate the maximum similarity between a query and all entries in the database, the query and all entries must be passed through the multimodal encoder, which is computationally expensive.

Late-Fusion VLMs Late-fusion VLMs, also known as dual-encoder models, process text and image inputs separately in two modality-specific encoders. As in early-fusion VLMs, the text is first tokenized and embedded into a sequence of text embeddings, while the image is patched and embedded into a sequence

of patch embeddings. While modern VLMs rely on ViTs to process images, this is not a strict requirement, and other approaches first extract features from an image, e.g., by using a CNN, and then forward them through a stack of standard transformer encoder layers. After passing the text and image embeddings through the separate text and image encoders, if they do not already have the same hidden dimension d_{m}, they are projected into a shared multimodal embedding space by a linear layer. To fuse the modalities, i.e., to compute a joint representation, the two separate encoders and the optional projection layers are jointly trained by using a loss function that takes both the text and image representations as input. The resulting text and image representations can then be used for vision-language downstream tasks, e.g., by fine-tuning a task-specific head.

Since the text and image features are processed through separate encoders, late-fusion VLMs are more efficient than early-fusion VLMs. However, they can be less performant on tasks that require fine-grained interactions between text and image tokens, as the self-attention mechanism is only applied within and not across the modalities. Another benefit of late-fusion VLMs is that they can pre-compute the text and image representations for large datasets and store them in a vector database, which allows for efficient retrieval because, at query time, only the query embedding needs to be computed and compared against the pre-computed representations.

Hierarchical-Fusion VLMs Hierarchical-fusion VLMs combine the early- and late-fusion approaches by first processing the text and image inputs separately in a text and image encoder, respectively, and then passing the resulting representations through a multimodal transformer encoder. This approach allows for contextualized representations to capture relationships intra- and inter-modality and, therefore, can be more performant on complex tasks. However, since three separate encoders are used, hierarchical-fusion VLMs are more computationally expensive than the other approaches. To increase efficiency for latency-constrained tasks, the representations of the text and image encoders can be pre-computed and stored in a vector database, similar to late-fusion VLMs.

Transformer Encoder Reasoning and Alignment Network (TERAN)

This model, which was introduced by Messina et al. (2021), is a special kind of late-fusion VLM designed to learn fine-grained alignment between texts and images. We employed TERAN in the Golden Retriever (Schneider et al., 2022) paper and works (Schneider et al., 2021; Wang et al., 2022) for cross-modal text-image (I2T) retrieval tasks, in which the goal is to retrieve the best matching images for a given text query. As illustrated in Figure 2.9, image features are extracted using a pre-trained Faster R-CNN (Ren et al., 2016) model, which is a CNN-based object detection model that extracts features from the respective regions of interest (RoIs) in an image and combines them with the bounding box coordinates of the RoIs using linear layers. Texts are first tokenized, embedded, and then forwarded through a pre-trained BERT (Devlin et al., 2019) model to obtain the text features. Then, instead of using a simple linear layer to project the text and image features

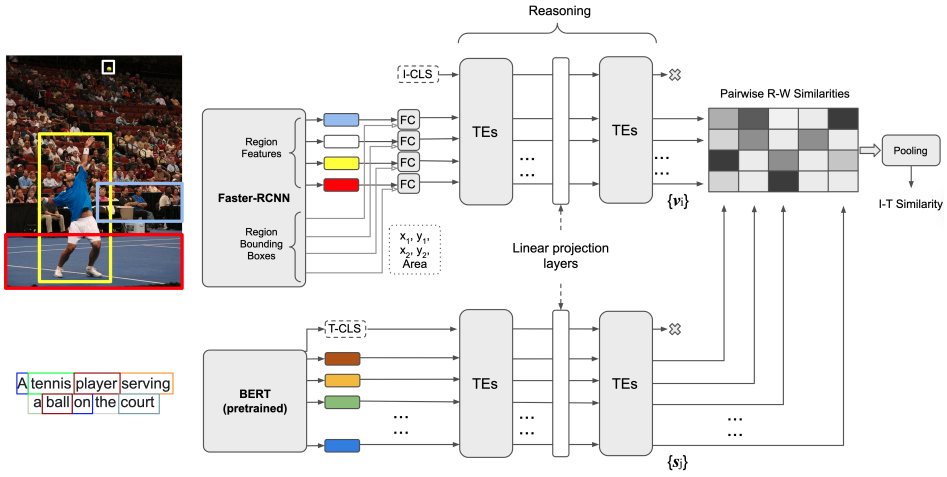


Figure 2.9: Architecture of the Transformer Encoder Reasoning and Alignment Network (TERAN) (Messina et al., 2021). Image taken from the original paper.

into a shared multimodal embedding space, in TERAN, the text and image features are forwarded through two separate stacks of multimodal encoder layers. Formally, given a sequence of text features $\mathbf{T} \in \mathbb{R}^{N_t \times d_m}$ and a sequence of image features $\mathbf{I} \in \mathbb{R}^{N_i \times d_m}$, where N_t and N_i are the number of text and image tokens, respectively, the text features are passed through the text encoder $\mathbf{E}_t$ and the image features through the image encoder $\mathbf{E}_i$ to compute aligned text and image features $\mathbf{T}'$ and $\mathbf{I}'$. Further, TERAN computes a Word-Region Alignment (WRA) matrix $\mathbf{A} \in \mathbb{R}^{N_t \times N_i}$, where each cell a_{ij} represents the alignment score, i.e., cosine-similarity, between the i-th text token and the j-th image region:

$$a_{ij} = \frac{\mathbf{T}'_i \cdot \mathbf{I}'_j}{\|\mathbf{T}'_i\| \cdot \|\mathbf{I}'_j\|} \tag{2.39}$$

To obtain a global similarity score $\Phi(\mathbf{T}, \mathbf{I})$ between the text and image, the WRA matrix is aggregated or "*pooled*" using a special pooling operation:

$$\Phi(\mathbf{T}, \mathbf{I}) = \sum_{i \in N_t} \max_{j \in N_i} a_{ij} \tag{2.40}$$

Training TERAN is trained in a supervised manner on different datasets containing aligned text-image pairs, e.g., MS-COCO (Lin et al., 2014), Flickr30k (Plummer et al., 2015), or our WISMIR3 (Schneider et al., 2024a) dataset, using a contrastive loss function. Contrastive loss functions are widely used in VLMs to learn joint representations or an alignment between text and image features by minimizing the distance between positive pairs, i.e., aligned text-image pairs, and maximizing the distance between negative pairs, i.e., unaligned text-image pairs. While there exist many different contrastive loss functions, which we will partially cover when discussing the training of the respective VLMs, TERAN employs a so-called "*margin-based triplet loss*" function (Faghri et al., 2018) to train the model. To compute a triplet loss, first an anchor sample $\mathbf{s}_a$ and a positive sample $\mathbf{s}^+$ are sampled from the training dataset, while a negative sample $\mathbf{s}^-$ is sampled from the remaining dataset so that it does not match the anchor text. Note that sampling

negatives is a challenging task, as it is crucial to sample hard negatives, i.e., images that are similar to the anchor text but not aligned with it. The loss is then computed as follows:

$$\mathcal{L}_{\text{triplet}}(\mathbf{s}_a, \mathbf{s}^+, \mathbf{s}^-) = \max(0, m + \Phi(\mathbf{s}_a, \mathbf{s}^-) - \Phi(\mathbf{s}_a, \mathbf{s}^+)), \tag{2.41}$$

where m is a margin hyperparameter that defines the minimum distance between positive and negative pairs. The goal of this loss function is to ensure that the similarity score between the anchor text and the positive image is higher than the similarity score between the anchor text and the negative image by at least the margin m. To train TERAN, the authors computed the final symmetric loss by adding the triplet loss for all text-image pairs and all image-text pairs and sampled hard negatives from the minibatch at each training step instead of the entire dataset.

Contrastive Language-Image Pre-training (CLIP)

Introduced by Radford et al. (2021), Contrastive Language-Image Pre-training or CLIP marked a milestone advancement in late-fusion VLMs. While the abbreviation CLIP is commonly used to refer to the model, note that CLIP is not a single model but a family of models that share the same or similar architecture and, most importantly, the same training technique. The core innovation of the CLIP paper is a large-scale pre-training method based on efficient contrastive learning. This approach leverages natural language as a supervised training signal to learn from massive datasets comprising hundreds of millions (Sun et al., 2017; Radford et al., 2021) or even multiple billions (Schuhmann et al., 2022) of image-text pairs. The large-scale pre-training enables the model to compute semantically rich and generalizable multimodal representations, which can be effectively transferred to obtain strong performance across various downstream tasks. Furthermore, the authors showed that CLIP models also perform strongly in zero-shot settings, i.e., without any fine-tuning on the downstream task, using prompt engineering to adapt the model to the respective task. Thanks to these advancements, CLIP models have since become foundational models in the field of multimodal machine learning and computer vision, influencing numerous subsequent architectures and methodologies.

Figure 2.10 provides an overview of a CLIP model and how it can be applied to downstream tasks, e.g., image classification in a zero-shot setting. From an architectural perspective, CLIP is a relatively simple late-fusion VLM that implements a dual-encoder architecture to process text and image inputs separately. While the authors also experimented with different setups involving CNNs such as ResNet (He et al., 2016) as image encoders, in most modern CLIP models, the image encoder is typically a ViT, while the text encoder is a BERT-like transformer model. To obtain a single embedding vector that globally represents the input image or text, the outputs of the respective encoder are either pooled or the output of a special token, e.g., the "*EOS*" or "*CLS*" token is used. Formally, given an input image $\mathtt{I}$ and an input text $\mathtt{T}$, the image encoder $\mathtt{Enc}_{\text{img}}$ and the text encoder $\mathtt{Enc}_{\text{txt}}$ compute the global features $\mathbf{I}_f = \mathtt{Enc}_{\text{img}}(\mathtt{I}) \in \mathbb{R}^{d_\text{i}}$ and $\mathbf{T}_f = \mathtt{Enc}_{\text{txt}}(\mathtt{T}) \in \mathbb{R}^{d_\text{t}}$. Next,

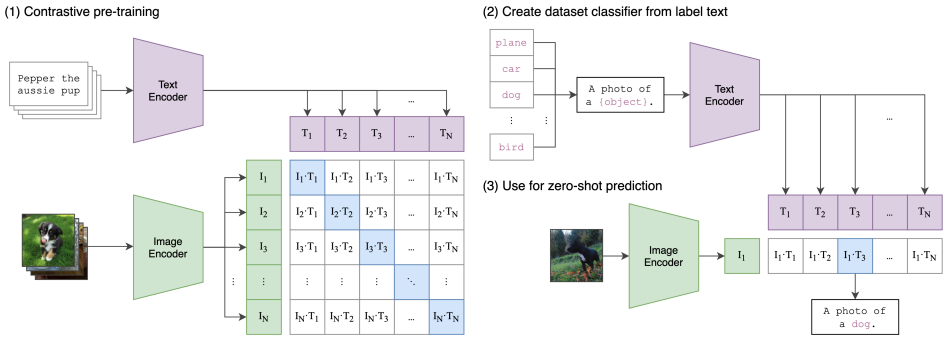


Figure 2.10: An overview of CLIP models. Image taken from Radford et al. (2021).

the image and text features are projected into a shared multimodal embedding space by linear layers $\mathbf{W}_{\text{img}} \in \mathbb{R}^{d_\text{i} \times d_\text{m}}$ and $\mathbf{W}_{\text{txt}} \in \mathbb{R}^{d_\text{t} \times d_\text{m}}$, respectively, and L2-normalized to compute the final multimodal text and image embeddings:

$$\mathbf{I}_e = \left\| \mathbf{I}_f \cdot \mathbf{W}_{\text{img}} \right\|_2, \quad \mathbf{I}_e \in \mathbb{R}^{d_\text{m}} \tag{2.42}$$

$$\mathbf{T}_e = \left\| \mathbf{T}_f \cdot \mathbf{W}_{\text{txt}} \right\|_2, \quad \mathbf{T}_e \in \mathbb{R}^{d_\text{m}} \tag{2.43}$$

The global similarity $\Phi(\mathbf{T}, \mathbf{I}) = \mathbf{T}_e \cdot \mathbf{I}_e$ between the text and image representations is then computed using cosine similarity expressed as the dot product of the normalized embeddings.

Training As mentioned, the core innovation of CLIP is the large-scale pre-training on massive datasets of aligned image-text pairs using natural language from captions as the training signal. At the time of its introduction, this was especially challenging as existing datasets were much smaller in size or only contained coarsely aligned image-text pairs, i.e., the text was not necessarily a description of the image but rather a caption or a tag. Moreover, efficient approaches to scale this training approach were not yet established as prior work, e.g., Desai et al. (2021), relied on autoregressive language models to predict the caption of an image, which is computationally infeasible for large datasets.

To address these challenges, the authors first crafted a well-aligned dataset of 400 million image-text pairs from the Web, referred to as "*WebImageText*". Further, they proposed a simple yet effective training approach that leverages contrastive learning to align the global image and text representations $\mathbf{I}_e$ and $\mathbf{T}_e$ instead of considering the fine-grained alignment of individual image and text features. More specifically, the paper introduced a symmetric contrastive loss that, given a (large) batch of image-text pairs, maximizes the similarity between matching the pairs, i.e., the diagonal of the similarity matrix, while minimizing the similarity between non-matching pairs, i.e., the upper and lower triangular parts of the similarity matrix (cf. Figure 2.10). Formally, given a batch $\mathcal{B} = \{(\mathbf{I}_e^{(i)}, \mathbf{T}_e^{(i)}) \mid i = 1, \ldots, N\}$ of N text-image pair embeddings, the similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ is computed as:

$$\mathbf{S} = \begin{bmatrix} \Phi_{1,1} & \cdots & \Phi_{1,N} \\ \vdots & \ddots & \vdots \\ \Phi_{N,1} & \cdots & \Phi_{N,N} \end{bmatrix} = \begin{bmatrix} \mathbf{T}_e^{(1)} \cdot \mathbf{I}_e^{(1)} & \cdots & \mathbf{T}_e^{(1)} \cdot \mathbf{I}_e^{(N)} \\ \vdots & \ddots & \vdots \\ \mathbf{T}_e^{(N)} \cdot \mathbf{I}_e^{(1)} & \cdots & \mathbf{T}_e^{(N)} \cdot \mathbf{I}_e^{(N)} \end{bmatrix} \tag{2.44}$$

The CLIP loss for the batch is then computed as the symmetric cross-entropy loss (CEL) over the similarity matrix, i.e., the average of the image-to-text (I2T) CEL and text-to-image (T2I) CEL:

$$\mathcal{L}_{\text{CLIP}}(\text{S}) = \frac{1}{2}\left(\mathcal{L}_{I2T} + \mathcal{L}_{T2I}\right) \tag{2.45}$$

$$= -\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{\exp(\Phi_{i,i}/\tau)}{\sum_{j=1}^{N}\exp(\Phi_{i,j}/\tau)} + \log\frac{\exp(\Phi_{i,i}/\tau)}{\sum_{j=1}^{N}\exp(\Phi_{j,i}/\tau)}\right). \tag{2.46}$$

where τ denotes a learnable temperature hyperparameter controlling the sharpness of the similarity distribution.

Note that the loss is symmetric, i.e., the I2T and T2I losses are considered, which the authors found to be crucial for the model's performance. Further, since all negatives come from the same batch, the authors found that huge batch sizes (e.g., 32,768) are crucial to obtaining a strong model and stabilizing training. One reason for this is that the larger the batch size, the more negative samples are available, which is important for contrastive training as it allows the model to learn more robust representations. This is also another reason for the symmetric loss, as it basically doubles the number of negative samples per batch.

Usage in Downstream Tasks After pre-training the CLIP models from scratch, the authors demonstrated that they could be effectively used for various downstream tasks, such as image classification, image retrieval, and OCR tasks across various domains, e.g., natural images, cars, medical images, art, and many more. Particularly, the authors showed two setups for using CLIP models in downstream tasks: "*zero-shot*" and "*linear probing*".

In the zero-shot setup, the CLIP model is used without any fine-tuning on the downstream task by leveraging the learned multimodal representations and prompt engineering. For example, for image classification tasks, the authors proposed to use the set of class labels as text prompts, e.g., "*a photo of a {class}*", and compute the similarity between the image embedding and the text embeddings for each class. The class with the highest similarity score is then selected as the predicted class. Using this approach, the model outperformed fully supervised linear classifiers on top of traditional CNN-based image models on a wide range of image classification datasets.

In the linear probing setup, a linear classifier,i.e., a linear layer, is trained on top of the CLIP image or text embeddings while keeping the CLIP model frozen. Also, in this setup, the model outperformed various traditional CNN-based models as well as pure ViT models on a wide range of datasets using only a tiny amount of labeled data, also called few-shot learning, usually with less than 20 samples per class. Most interestingly, the authors showed that the zero-shot performance outperforms a 4-shot linear probing setup on average.

All these results demonstrate the effectiveness of CLIP models in learning strong generalizable multimodal representations that can be effectively used for transfer learning across various downstream tasks.

CLIP Variants Since the introduction of CLIP, numerous variants have been proposed to extend the model capabilities, for example, by adapting it to specific tasks and specialized domains using fine-tuning on in-domain datasets (Zhao et al., 2023b; Cartella et al., 2023). Other works focused to generally improve the performance and efficiency of CLIP-style models by scaling them to even larger and multilingual datasets (Gadre et al., 2023; Schuhmann et al., 2021; Schuhmann et al., 2022; Chen et al., 2023b; Zhu et al., 2023) or larger models with more parameters while at the same time improving the training and data efficiency, and extend the models capabilities to multilingual and multicultural settings (Jia et al., 2021; Zhai et al., 2022; Zhai et al., 2023; Li et al., 2023c; Sun et al., 2023; Li et al., 2023d; Chen et al., 2023a; Tschannen et al., 2025). Most modern CLIP variants, e.g., EVA-CLIP models (Sun et al., 2023; Fang et al., 2024; Sun et al., 2024a) or SigLIP (Zhai et al., 2023; Tschannen et al., 2025) models, additionally employ improved ViT models to allow for processing of higher-resolution images of any aspect ratio (Beyer et al., 2023; Dehghani et al., 2023b).

Usage of VLMs in this Thesis

Throughout many papers of this thesis, VLMs were leveraged for cross-modal retrieval tasks. In particular, we leverage TERAN (Messina et al., 2021) and UNITER (Chen et al., 2020) in the Golden Retriever (Schneider et al., 2022) and other works (Schneider et al., 2021; Wang et al., 2022) for cross-modal text-image (I2T) retrieval tasks. In the Golden Retriever (Schneider et al., 2022) and MOTIF (Wang et al., 2022) papers, TERAN is also employed for open-vocabulary object detection. Furthermore, we used different pre-trained CLIP models (Radford et al., 2021; Zhai et al., 2023) in WISMIR3 (Schneider et al., 2024a), CollEX (Schneider et al., 2025a), and many other works (Schneider et al., 2021; Schneider et al., 2024a; Schneider et al., 2023b) for zero-shot cross-modal retrieval tasks including image-text (I2T), text-image (T2I) as well as text-text (T2T), and image-image (I2I) retrieval.

In (Schneider et al., 2023a), we also used pre-trained CLIP models for zero-shot visual word sense disambiguation (VWSD) tasks, where the goal is to disambiguate the meaning of a word given a minimal textual context and a set of 10 images that are possible interpretations of the word.

Another use of CLIP models, or more precisely, the image encoder of a pre-trained SigLIP model (Zhai et al., 2023), is in the Centurio (Geigle et al., 2025) paper. There, we used and fine-tuned the ViT model as the image encoder to train massively multilingual LVLMs.

2.4.3 Decoder-based Large Vision-Language Models (LVLMs)

In contrast to VLMs, which are encoder-only models and therefore designed to compute embeddings, in this thesis, we refer to large vision-language models (LVLMs) as autoregressive decoder-based models designed to generate text based on a prompt that can contain both text and images. Or, to put it differently, LVLMs can be seen as LLMs that are extended with vision capabilities, i.e., they can additionally understand images and generate text based on a multimodal

prompt. The strong visual understanding of ViTs, combined with the world knowledge, language understanding, and text generation capabilities of LLMs, drastically extends the capabilities of LVLMs beyond those of LLMs and ViTs alone, enabling a wide range of complex tasks that require both language and vision understanding. Such vision-language tasks (VLTs, cf. 1.2) are, for example, image captioning, visual question answering, open-vocabulary object detection, or optical character recognition. Moreover, due to their multimodal capabilities and world knowledge, which more closely resemble human-like perception, LVLMs allow for more natural human-computer interaction and became powerful tools for everyday tasks beyond purely academic benchmarks. This trend is even more pronounced in the most recent generation of LVLMs, also often called "*omni*" or "*any-to-any*" models, which are capable of processing additional modalities like audio or video, and generate not only text but also other multimodal content (Xu et al., 2025; Deng et al., 2025; Chen et al., 2025; Hurst et al., 2024; Team et al., 2024). However, this thesis focuses on LVLMs designed to generate text based on multimodal prompts containing only text and images or videos expressed as a sequence of images without audio.

In the next sections, we will first briefly cover the extraction of visual features common to the majority of LVLMs, followed by an overview of different LVLM architectures, and finally cover common training strategies.

Image Feature Extraction

In most LVLMs, the image features are extracted using a pre-trained ViT model, which is typically the vision encoder of a state-of-the-art CLIP model, e.g., SigLIP (Zhai et al., 2023; Tschannen et al., 2025) or EVA (Sun et al., 2023; Fang et al., 2024). This is because these models are trained on large-scale datasets via language supervision, which was found to be beneficial for combining the extracted visual features with the text tokens of the LLM backbone. Other approaches employ other ViT models like DINO (Caron et al., 2021; Oquab et al., 2025), combine multiple encoders (Lu et al., 2023; Tong et al., 2024), or train custom (large-scale) vision encoders from scratch (Chen et al., 2023c). Recent LVLMs mostly use enhanced ViT models such as NaViT (Dehghani et al., 2023b) that allow for processing images of any aspect ratio and higher resolution, extended by introducing advanced positional encodings for multimodal inputs (Wang et al., 2024a; Bai et al., 2025; Zhu et al., 2025).

To further improve the performance of LVLMs on tasks that require fine-grained visual understanding, e.g., understanding of texts in images, HD images are required. Many recent works allow for such high-resolution inputs by extracting features from the same image at different scales and combining the extracted features afterward as introduced by Liu et al. (2024). This is typically achieved by tiling the high-resolution image into smaller subimages and extracting features from the down-scaled HD image and the subimages, which are then concatenated to form a single sequence of image features.

One issue that arises when adding single image features as additional input to the LLM backbone is that the number of tokens forwarded through the model increases,

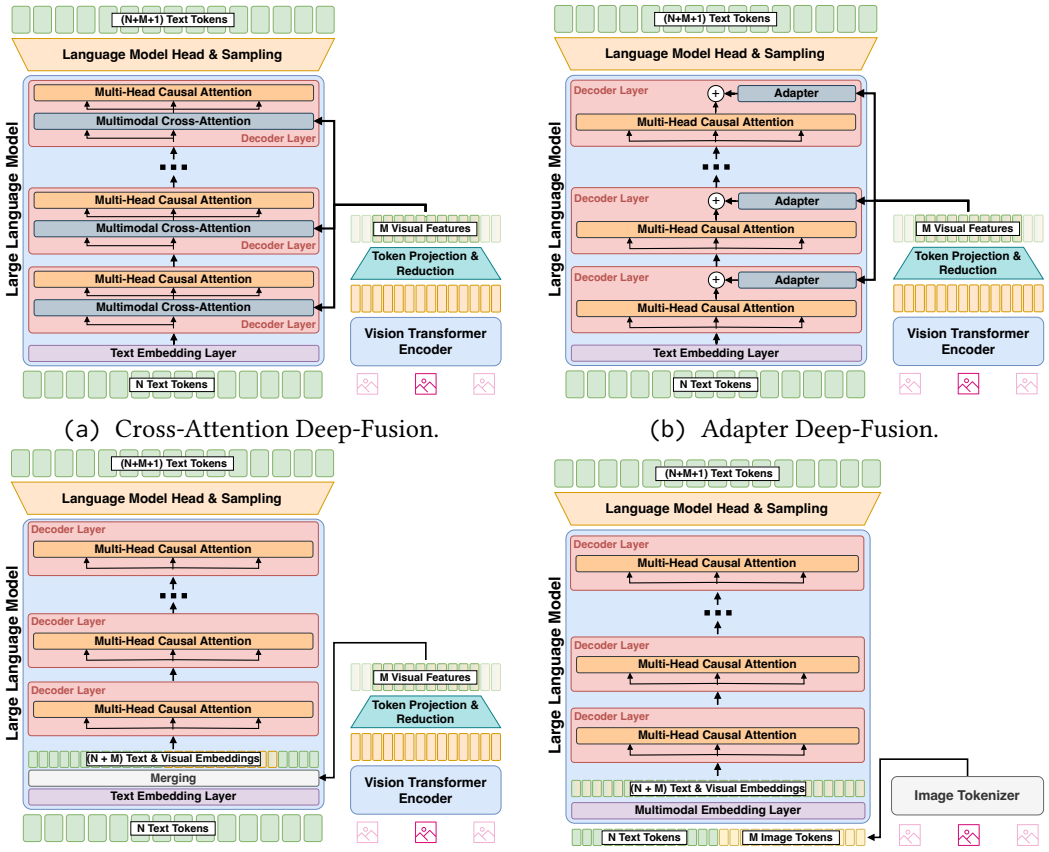
leading to much higher computational cost due to the quadratic complexity of the self-attention mechanism. For example, a typical CLIP-ViT model used in many LVLMs extracts 729 image features from a single image. This number quadruples when using 4 sub-image tiles, leading to a total of $(1 + 4) * 729 = 3645$ additional visual tokens that need to be processed by the LLM backbone. This becomes especially problematic for multi-image or video inputs, as supported by the majority of recent LVLMs. To address this issue, many LVLMs reduce the number of visual tokens by applying some post-processing network (Li et al., 2023b; Zhang et al., 2025a), by stacking the features along the channel dimension (Shi et al., 2024), token merging or pruning operations (Yang et al., 2025), or other sophisticated methods applied to the attention-layers in the LLM (Chen et al., 2024a).

LVLM Architectures

While generative models that generate text conditioned on images existed for a long time and were historically implemented using non-neural, purely statistical models (Farhadi et al., 2010; Li et al., 2011), and later combined CNNs and RNNs (Kiros et al., 2014; Donahue et al., 2015; Wang et al., 2016), their performance was limited by the capabilities of the underlying models. Building on the advancements of transformer models and VLMs, one of the first successful LVLMs that laid the foundation for modern transformer-based LVLMs covered in this thesis was the Flamingo model introduced by Alayrac et al. (2022). Since then, the field has seen a surge in interest, leading to a rapid development of hundreds of LVLMs, comprehensively summarized by multiple recent survey papers (Caffagni et al., 2024; Zhang et al., 2024a; Liang et al., 2024; Wadekar et al., 2024).

General LVLM Architecture What the large majority of LVLMs have in common is that they employ some (often pre-trained) ViT model to extract image features from one or more images (cf. 2.4.3), which are then combined with the textual tokens of the (pre-trained) LLM backbone. From a high-level perspective, LVLMs can be categorized into four main architectures (cf. Figure 2.11) that can be further categorized into two super types: "*deep-fusion*" and "*early-fusion*" models. As the name suggests, deep-fusion LVLMs combine the text and image modalities in the inner decoder layers of the model, whereas early-fusion LVLMs combine the modalities at the input stage, i.e., before the decoder layers. How the text and image modalities are combined depends on the architecture type and will be briefly described in the following paragraphs. Note, however, that there are variations in the details of these architectures that depend on the specific implementation of the respective LVLM, which we will not cover in detail here as this would go beyond the scope of this thesis.

Cross-Attention Deep-Fusion LVLMs This architecture (cf. Figure 2.11a) was implemented by Alayrac et al. (2022) in their Flamingo model and was popular in "earlier" LVLMs. It consists of a pre-trained LLM backbone with the decoder layers extended by multimodal cross-attention layers and a (usually pre-trained) ViT model to extract image features. After extracting the image features, they



(a) Cross-Attention Deep-Fusion.

(b) Adapter Deep-Fusion.

(c) Projection Early-Fusion.

(d) Tokenize-Only Early-Fusion.

Figure 2.11: A high-level overview of main LVLM architectures.

are combined with the text tokens by the additional cross-attention layers, which compute the attention between the text tokens and the image features, allowing the model to attend to relevant visual information while generating text. Usually, these cross-attention layers are inserted before the causal self-attention layers of the LLM backbone, but some works also insert them after the self-attention layers.

Due to the complex setup and increased computational cost for training as well as inference introduced not only by the visual tokens but also by the cross-attention layers, this architectural type is much less common in modern LVLMs.

Adapter Deep-Fusion LVLMs Similar to cross-attention deep-fusion LVLMs, adapter deep-fusion LVLMs (cf. Figure 2.11b) also inject the visual features into the decoder layers of the LLM backbone. However, instead of using cross-attention layers, they use custom modules called "*adapters*" that are inserted into the LLM backbone. These adapters can be specialized cross-attention layers, mixture-of-experts (MoE) layers (cf. 2.3.4), or popular parameter efficient adapter layers (Houlsby et al., 2019) such as LoRA (Hu et al., 2022). Further, they can be inserted before or after the attention modules in the decoder layers of the LLM backbone.

While LVLMs following this architecture are efficient when employing LoRA or similar, models using MoE layers or custom cross-attention layers are compu-

tationally expensive. Due to this and the deep-fusion architecture, this type of LVLM has also not been common in recent LVLMs.

Tokenize-Only Early-Fusion LVLMs This architecture (cf. Figure 2.11d) type differs fundamentally from the two previously covered. Firstly, because it is an early-fusion type, i.e., it does not inject the visual tokens into the backbone LLM or adapt it in any other way. Instead, the visual tokens are simply concatenated to the text tokens before being forwarded through the LLM. Secondly, visual features are not extracted by a pre-trained ViT model but by a specialized "image tokenizer" that is separately trained to transform an image into a sequence of discrete visual tokens (Ramesh et al., 2021; Esser et al., 2021).

While the setup is simpler than deep-fusion architectures, the main challenges are training a strong image tokenizer and adapting the LLM to these visual tokens. Further, due to this design choice, it is not possible to leverage powerful vision encoders pre-trained on billions of text image pairs, thereby losing the benefits of language supervision. Due to these limitations, this architecture type is generally not common in LVLMs.

Projection Early-Fusion LVLMs This architecture (cf. Figure 2.11c) is the simplest and, at the same time, one of the most powerful approaches for integrating visual information into LLMs. One of the first successful implementations was the seminal LLaVA model (Liu et al., 2023), which is why LVLMs following this architecture are also often referred to as "*LLaVA-style*" models. As opposed to the tokenize-only early-fusion architecture, in this architecture, the visual features are extracted by a pre-trained ViT model, typically a CLIP-ViT, and then projected into a shared multimodal embedding space using a simple projection module. In pure LLaVA-style models, the projection module is a simple linear layer or a two-layer FFN, which transforms the visual features into the same hidden dimension as the text embeddings of the LLM backbone. Afterward, the visual features are optionally reduced and finally concatenated with the text embeddings before being forwarded through the LLM backbone.

Due to the simple projection strategies and the ability to leverage powerful pre-trained vision encoders, this architecture type is easy to implement and, therefore, widely used in modern LVLMs.

LVLM Training Strategies

This section provides an overview of common training strategies for LVLMs that employ a pre-trained LLM backbone and (most of the time) a pre-trained ViT model to extract visual features. However, due to the large number of different LVLMs, there are also many different strategies to train these models. While earlier LVLMs, especially those based on the cross-attention deep-fusion or "Flamingo-style" architecture, relied on single-stage training strategies, more recent LVLMs, especially LLaVA-style models, employ two-stage or three-stage training strategies. Another model-dependent design choice is which components

or layers in which stage are trained from scratch, fine-tuned, or kept frozen during training. Nevertheless, in the following, we will cover the key aspects of the three stages, i.e., "*pre-training*" or "*vision-language alignment*", "*multimodal instruction tuning*", and "*post-training*" or "*preference alignment*".

Stage 1: Pre-Training or Vision-Language Alignment During this first stage, the objective is to align the visual features extracted from an image by a pre-trained ViT model with the text embeddings computed from the tokenized text by the embedding layer of the LLM backbone. Similar to VLM training, this involves large-scale datasets of aligned image-text pairs from the Web such as LAION-5B (Schuhmann et al., 2022) or DataComp (Gadre et al., 2023), traditional image captioning datasets such as COCO (Lin et al., 2014), SBU (Ordonez et al., 2011), or more recent synthetic datasets like ShareGPT4V-PT (Chen et al., 2024b) that contain detailed generated image captions. However, instead of using contrastive learning, LVLMs are trained in an autoregressive manner to generate the target captions given the respective image and an optional prompt using simple next-token prediction via a cross-entropy loss. Note that as soon as the image features are combined with the text features into a single sequence, this is, in principle, the same as pre-training a standard LLM (cf. § 2.3.4), but with the additional visual features as prefix or context that condition the text generation. Formally, given a target caption as a sequence of N text embeddings $\mathbf{T} = \{\mathbf{t}_1, \dots, \mathbf{t}_N\}$ for an image expressed as a sequence of M visual features $\mathbf{I} = \{\mathbf{i}_1, \dots, \mathbf{i}_M\}$, the cross-entropy loss is defined as:

$$\mathcal{L}_{\text{CE}}(\mathbf{T}, \mathbf{I}) \;=\; -\frac{1}{N} \sum_{i=1}^{N} \log p\big(\mathbf{t}_i \mid \mathbf{t}_{<i}, \mathbf{I}\big), \tag{2.47}$$

where $\mathbf{t}_{<i} = (\mathbf{t}_1, \dots, \mathbf{t}_{i-1})$ and $p(\cdot)$ is the probability computed by the language modeling head of the LVLM. Note that the M image features $\mathbf{I}$ are not considered in the loss function, i.e., they are only used as context to condition the text generation, not as target labels.

In this stage, most LVLMs keep the LLM and the ViT frozen and only train the projection module or the adapter layers that combine the visual and text features, while some models also fine-tune the ViT model. After successful training, LVLMs learned to understand images, align the visual information with text, and generate textual descriptions or captioning of images. However, the models are not yet able to complete other tasks as they have not seen them during training.

Stage 2: Instruction Tuning or Supervised Fine-Tuning To make an LVLM a task-agnostic tool that can be used for various downstream vision-language tasks (VLTs) as their text-only counterparts, the model needs to be further trained in a second stage, which is often referred to as "*multimodal instruction tuning*". One of the first LVLMs that successfully employed this training strategy was the LLaVA model introduced by Liu et al. (2023) in their seminal paper titled "Visual Instruction Tuning".

For this stage, a mixture of datasets is used that contain a diverse set of VLTs, such as various types of visual question answering (VQA), visual entailment (VNLI),

image captioning, optical character recognition, object detection, and many more. Some of the datasets only contain samples with one image, while others contain samples with multiple inter-leaved images as they occur in typical chat-style conversations or web pages. Before the datasets can be used for training, they need to be converted into a format suitable for instruction tuning that the LVLM can process. For example, each sample of a VQA dataset can be converted into an instruction-following format with the following template:

Question: `<question>`
Image: `<image>`
Answer: `<answer>`

where `<question>`, `<image>`, and `<answer>` are placeholders for the question, image, and answer of a sample, respectively. Or, for a VNLI task where the model needs to predict whether a statement about two images is correct or not, the following template could be used:

Image 1: `<image>`
Image 2: `<image>`
Given the two images, is the following statement correct? Answer with yes or no.
Statement: `<statement>`
Your Answer: `<answer>`

where `<image>`, `<statement>`, and `<answer>` are placeholders for the images, statement, and answer of a sample, respectively. Popular instruction tuning datasets that are already in such a format and comprise a wide range of VLTs covering various domains like science, math, arts, or general topics are LLaVA-665K (Liu et al., 2023) or Cambrian (Tong et al., 2024).

Due to the fact that the datasets contain supervised labels, often, this training stage is referred to as "*supervised fine-tuning*" or SFT of the LVLM. However, since the samples can be converted into a sequence of text and image tokens, the model can again be trained in an autoregressive manner similar to the alignment stage, i.e., by predicting the next token given the prompt, the image, and the previous tokens. Formally, given $\mathbf{P} = \{\mathbf{p}_1, \dots, \mathbf{p}_L\}$, $\mathbf{I} = \{\mathbf{i}_1, \dots, \mathbf{i}_M\}$, and $\mathbf{T} = \{\mathbf{t}_1, \dots, \mathbf{t}_K\}$ as the prompt embeddings, the image embeddings, and the target-answer embeddings, respectively, the cross-entropy loss is:

$$\mathcal{L}_{\text{CEL}}(\mathbf{P}, \mathbf{I}, \mathbf{T}) \;=\; -\frac{1}{K} \sum_{i=1}^{K} \log p\big(\mathbf{t}_i \mid \mathbf{t}_{<i}, \mathbf{P}, \mathbf{I}\big), \tag{2.48}$$

where $\mathbf{t}_{<i} = (\mathbf{t}_1, \dots, \mathbf{t}_{i-1})$ and $p(\cdot)$ is the probability computed by the language modeling head of the LVLM.

During this stage, the LLM backbone, together with the projection module, is usually fine-tuned, while the ViT model is often kept frozen. However, this depends on the specific implementation of the LVLM, as some models also fine-tune the ViT model, train only some layers of the LLM backbone, or keep the

projection module frozen. Further, some approaches also combine the first and second stages into a single stage by converting the pre-training data into an instruction-following format.

After this stage, the LVLM is a task-agnostic multimodal model that can be used for various VLTs, often in zero-shot or few-shot settings, i.e., without any further fine-tuning on the downstream task.

Stage 3: Post-Training or Preference Alignment After the second stage, some LVLMs employ this third stage to further improve and align the LVLM to human preferences and also enhance the visual grounding capabilities, reducing hallucination problems and making it more trustworthy and safe for real-world applications. The approach, i.e., the use of reinforcement learning techniques, of this stage, is similar to the preference alignment training of LLMs (cf. § 2.3.4), except that when applied to LVLMs, also images are used as input. The inclusion of multimodal data and transfer of the approaches to LVLMs introduced several challenges, e.g., in collecting preference data or in the design of effective reward models. However, recent works overcame these challenges and successfully applied preference alignment to LLaVA-style LVLMs, outperforming non-preference-aligned models on hallucination benchmarks as well as other popular datasets (Sun et al., 2024b; Zhang et al., 2025b; Yu et al., 2024)

Since most open-source LVLMs do not apply these training strategies, we will not cover them here.

Usage of LVLMs in this Thesis

We make use of various LVLMs and investigate their limitations and strengths across many diverse vision-language tasks in several papers of this thesis. In particular, we benchmark a wide range of (primarily LLaVA-style) LVLMs to assess their multilingual performances on VLTs such as VQA, VNLI, visually-grounded reasoning (VGR), image captioning in our M5 paper (Schneider et al., 2024b) and cross-modal topic classification in (Schmidt et al., 2025). Further, in the GIMMICK benchmark (Schneider et al., 2025b), we evaluated even more LVLMs on their knowledge of globally distributed cultures and their capabilities to understand and generate culturally relevant content. In the Centurio paper (Geigle et al., 2025), we trained massively multilingual LLaVA-style LVLMs based on pre-trained CLIP-ViTs and LLMs from scratch applying vision-language pre-training and multimodal instruction tuning strategies on large multilingual datasets. To assess their performance and compare them to other LVLMs, we evaluated the models on an extended version of the M5 benchmark. Another application of LVLMs in this thesis is their use at the core of a multimodal agentic RAG system as described in our CollEX (Schneider et al., 2025a) paper.

2.5 Summary

In this chapter, we have provided a historical as well as technical overview of the theoretical background necessary to understand the papers core to this thesis. In the first three parts, we have discussed the evolution of neural networks, their application in natural language processing, and the evolution from transformer models to modern large language models (LLMs). These sections provide the basis for understanding the multimodal transformer encoder-based models (VLMs) and transformer decoder-based models (LVLMs). These models are particularly relevant for the thesis as they are core to every paper this thesis is based on. While we discussed the basic architectures and training strategies of VLMs and LVLMs, more detailed information also about their specific usage and functionality can be found in our respective papers.

3
Conclusion

Contents

3.1 Summary **58**

3.1.1 Research Question 1: Strengths 58

3.1.2 Research Question 2: Limitations 59

3.1.3 Research Question 3: Mitigations 60

3.2 Limitations **61**

3.3 Future Work **62**

In this chapter, we first concisely answer the three research questions (**RQ1**, **RQ2**, **RQ3**) core to this dissertation (cf. § 1.3) and summarize the key findings of our work. Afterward, we discuss the limitations of our work and finally provide an outlook on future research directions.

3.1 Summary

This thesis explored the strengths of VLMs and LVLMs concerning their practical applicability for real-world scenarios (**RQ1**), their limitations concerning performance and robustness in out-of-distribution, multilingual, and multicultural settings (**RQ2**), and proposed solutions to mitigate limitations concerning multilingual performance degradation (**RQ3**). A high-level overview of the research questions and key findings in abbreviated form is provided in 3.1. Detailed answers

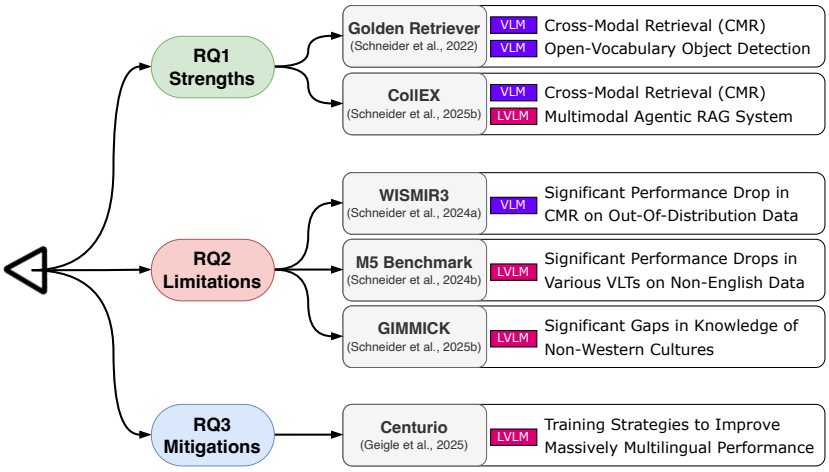


Figure 3.1: Overview of the research questions and key findings of this thesis.

to the research questions are provided in the following subsections, where we also summarize the key findings of our work.

3.1.1 Research Question 1: Strengths

In all our papers, we addressed the first research question concerning the strengths of VLMs and LVLMs to some extent but focused particularly on their potential and practical applicability in the following two papers.

Concentrating on transformer encoder-based VLMs, we showed in our Golden Retriever (Schneider et al., 2022) paper that these models, particularly dual-encoder late-fusion architectures like TERAN (Messina et al., 2021), can be effectively leveraged for real-time capable cross-modal text-image retrieval systems. This was achieved by global average pooling of fine-grained word-region-alignment (WRA) matrices in which a cell represents the semantic similarity between a word, i.e., a token, in the query sequence and a region in an image. Moreover, applying local average pooling, taking into account only a subset of the query sequence, we showed that these models can also be employed for open-vocabulary object

detection. However, practicably leveraging such models in a real-time system imposes challenges due to the relatively long computation time of a WRA matrix for a given image. Hence, we implemented a sophisticated two-stage algorithm that first drastically reduces the search space by employing efficient retrieval methods based on pre-computed indices to a few candidate images, for which the WRA matrices are computed and pooled in the second stage.

Concerning transformer decoder-based LVLM, we showed how these models can be employed effectively as chatbots in multimodal agentic RAG systems in our CollEX paper (Schneider et al., 2025a). Specifically, we pre-compute multimodal embeddings of texts and images of records in scientific collections utilizing a strong VLM and store them in a hybrid vector database for efficient retrieval. The LVLM sits at the core of the system and is operated through an agentic loop, in which the model plans the actions to be executed to fulfill a user request. These actions, provided to the model via function-calling techniques, are simple database lookups, cross-modal and uni-modal semantic similarity search as well as lexical search methods, and image analysis methods such as visual question answering (VQA), optical character recognition (OCR), open-vocabulary detection (OVD), or image captioning. We also employed an LVLM using specialized prompts for the latter image analysis methods.

3.1.2 Research Question 2: Limitations

The second research question targets different limitations of VLMs and LVLMs and was primarily addressed in three papers, focusing on out-of-distribution data in cross-modal retrieval settings, robustness in massively multilingual setups, and awareness beyond Western cultures, respectively.

We showed in our WISMIR3 (Schneider et al., 2024a) paper that the performance degradation of transformer encoder VLMs for cross-modal text-to-image retrieval settings is severe when employing them in scenarios with data that differs from their pretraining distribution. Therefore, we first evaluated a set of VLMs on typical text-to-image retrieval benchmarks and analyzed their data in detail by collecting a wide range of linguistic statistics on the image captions. Here, we found that the average captions are short, simple in terms of classic readability scores, and contain little to no named entities. To challenge the VLMs and measure their robustness, we crafted a novel dataset from Wikipedia articles, including over 300,000 images of diverse sceneries described by lengthy and complex captions comprising many, often multiple named entities. In our evaluation setup, we measured the robustness of VLMs of three different archetypes and observed severe drops in performance on our test set compared to standard benchmarks in a zero-shot setting. Further, we fine-tuned a dual-encoder late-fusion VLM (TERAN (Messina et al., 2021)) on the train split of our dataset (295K samples) and found that the model was unable to learn strong representations from the complex data. Both findings demonstrate significant limitations of encoder VLMs for complex out-of-distribution data.

Focussing on the robustness of LVLMs in massively multilingual settings, we conducted large-scale benchmarking experiments with 18 models of various sizes

and families in our M5B (Schneider et al., 2024b) paper. The benchmark comprises eight datasets and five VLTs covering 41 languages, 16 scripts, and 13 language families. Notably, two of the eight datasets were newly introduced and include culturally diverse images and 12 languages: nine low-resource languages from African or Asian countries and three high-resource control languages. Through detailed analyses of our extensive experiments, we show that severe performance disparities are prominent between high-resource and low-resource languages across all models and datasets. Moreover, we found that the number of parameters of the models does not necessarily correlate with multilingual (non-English) performance, hinting at the importance of massively multilingual training data and carefully designed training regimes. Overall, we revealed substantial limitations of LVLMs in massively multilingual settings, especially for low-resource languages, and, thus, highlight challenges for globally equitable multilingual AI models.

Another crucial facet of limitations that needs to be overcome towards globally inclusive AI is globally distributed cultural awareness. Therefore, we designed the large-scale benchmark, GIMMICK (Schneider et al., 2025b), in which we assessed the cultural knowledge of 31 models of all sizes, i.e., 20 LVLMs and 11 LLMs backbones, across 144 countries spanning six global macro-regions. The benchmark contains six tasks based on three newly introduced datasets covering text-only, image-only, text-image, and text-video modalities. More specifically, we examined regional cultural biases and the influence of the number of model parameters, input modalities, and external cues. Our extensive analyses uncovered strong biases toward Western cultures across all models and most tasks and revealed significant correlations between model size and performance as well as the effectiveness of multimodal input and external geographic cues. Moreover, we found that both LVLMs and their LLMs backbones know more about tangible than intangible aspects (e.g., food vs. rituals) and excel in recognizing broad cultural origins but struggle with a more nuanced understanding. In summary, we demonstrated that the models are not equally aware of the world's cultural diversity but are biased toward Western cultures.

3.1.3 Research Question 3: Mitigations

The third research question of this thesis targets mitigation strategies for the previously assessed limitations from RQ2.

More precisely, within the scope of this thesis, i.e., in our paper Centurio (Geigle et al., 2025), we concentrated on enhancing the robustness of LVLMs in massively multilingual settings. Therefore, we deeply examined optimal training data mixtures concerning the distribution of comprised languages as well as optimal training strategies. Consequently, we conducted a series of comprehensive experiments in which we systematically examined the maximum number of languages we can include in the training data without hurting the English performance, the optimal distribution of these languages for the per-training and instruction-tuning phases of an LVLM, and, lastly, how we can improve a model's multilingual OCR performance. We evaluated the models' outcomes in the experiments through an extensive benchmark covering 13 diverse VLTs and 43 languages. Our analyses

revealed that we could include training data in all our 100 target languages, comprising all common high-resource languages as well as a multitude of mid- to low-resource languages. We further found that only 25-50% of non-English languages are sufficient to significantly improve the multilingual performance across all tasks in our benchmark while maintaining strong results in English. To enhance understanding of multilingual text in images, we discovered that it is essential to include multilingual OCR training data in both training stages. Finally, we applied all our findings from the experiments and scaled up the data to train Centurio, a massively multilingual LVLM supporting 100 languages that showed state-of-the-art performance across many tasks of our final benchmark covering 14 tasks and 56 languages against 13 other LVLMs specifically trained for multilingual settings.

3.2 Limitations

This thesis investigated the strengths of VLMs and LVLMs, with a particular focus on the practical applications of these models in cross-modal information retrieval systems and multimodal retrieval-augmented chatbots to explore complex data. Further, we revealed significant limitations concerning multilingual, i.e., non-English, and multicultural, i.e., non-Western, settings of generative multimodal transformer models, i.e., LVLMs. Finally, we demonstrated successful mitigation strategies to make LVLMs more robust in massively multilingual setups, especially when dealing with low-resource languages. Nevertheless, this dissertation is not without its own limitations, which are briefly discussed in the following.

In addition to the strengths and limitations covered within the scope of this thesis, multimodal transformer models for vision-language tasks have a multitude of other potential strengths and limitations.

For example, VLMs and LVLMs are often employed in various other domains such as medical image analysis (Li et al., 2023a; Xia et al., 2024; Zhu et al., 2024), robotics (Brohan et al., 2022; Li et al., 2024b; Zhen et al., 2024), or augmented reality applications (Pei et al., 2024; Chheang et al., 2024; Sharma et al., 2024). In all these domains, the models have different strengths and different limitations, which we did not cover in this work.

Moreover, there are numerous other hurdles next to multilingual, multicultural, or OOD limitations such as hallucination issues (Wang et al., 2024b; Kim et al., 2024), adversarial attack vulnerabilities (Zhao et al., 2023a; Ding et al., 2025), or basic perceptual limitations (Campbell et al., 2024; Jiang et al., 2024b; Wang et al., 2025), that need to be overcome to achieve robust versatile multimodal transformers models accross all domains.

Another issue is the fast pace of current research in the field of VLMs (Zhang et al., 2024b) as well as LVLMs (Yin et al., 2024). That is, the frequency with which novel methods, datasets, and models are released is very high, which makes findings discovered with benchmarks less expressive in the sense that they are quickly outdated. In other words, the benchmarks of this thesis and their findings serve more as a snapshot of the current state of the art in the LVLM or VLM research

and are intended to highlight the limitations of current models and serve as a guide for future research where improvements are needed. For this reason, we designed and implemented the benchmarks so that they can be easily extended to new models or new datasets whenever they arrive.

Furthermore, while we covered mitigation strategies to overcome limitations of LVLMs in non-English languages, this thesis lacks approaches to achieve more robust VLMs and LVLMs in multicultural (non-Western) settings, as well as being more robust concerning complex out-of-distribution data in cross-modal retrieval applications for VLMs.

Lastly, we showed how to effectively employ VLMs or LVLMs in cross-modal retrieval or multimodal RAG settings. However, the papers that show the strengths of these models for their practical applications lack systematic quantitative or qualitative evaluations with academic benchmarks or thorough user studies.

3.3 Future Work

With the recent surge in works related to the field of multimodal transformer models for vision-language tasks (VLTs), i.e., encoder-only VLMs (Zhang et al., 2024b) as well as decoder-based generative LVLMs (Yin et al., 2024), across all sorts of different domains, the space of future work is vast. However, in this chapter, we briefly outline aspects of the possible future work of this thesis, focusing primarily on overcoming some of its limitations pointed out in 3.2.

In our WISMIR3 paper, we explored the limitations of encoder-only VLMs for cross-modal text-to-image retrieval concerning out-of-distribution data, or more specifically, complex textual queries that are lengthy and rich in named entities describing the visual content. Despite the general difficulty of the task, two distinct causes for the low performance of the tested VLMs are the limited number of textual tokens most models can process—only 76 tokens for standard CLIP-based models (Radford et al., 2021)— and the relatively narrow world knowledge compared to the LLM backbones of current LVLMs acquired via pre-training on massive web-scape datasets. One remedy to these issues is the rise of a novel kind of multimodal embedding models that leverage decoder-based LVLMs to compute rich representations that outperform traditional VLMs on most tasks (Jiang et al., 2024a; Gu et al., 2025; Jiang et al., 2025; Lan et al., 2025). Next to sophisticated post-training, distillation, and prompting techniques, the key modification required to enable computing embeddings with decoder-based LVLMs is to remove the causal attention mask, thereby enabling bi-directional self-attention as used in transformer encoder models. With these modifications, the models overcome multiple limitations of encoder-based multimodal embedding models, i.e., traditional VLMs, namely the low number of textual tokens, the so-called “modality gap” (Liang et al., 2022) that leads to distinct clusters of textual and visual embeddings, and the limited world knowledge. Since these kinds of novel LVLM-based multimodal embedding models are very recent state-of-the-art, at the time of writing this, the current literature lacks work that investigates specifically the limitations of the models concerning OOD data, multilingual performance, and multi-cultural

knowledge. Moreover, directions for future work featuring these models are their investigations of their strengths in practical applications like multimodal RAG or other retrieval settings, their applicability as vision-encoders for LVLMs, or general downstream VLTs.

In our GIMMICK paper, we uncovered significant biases of current LVLMs towards Western cultures, i.e., a substantial disparity in knowledge of cultural events or facets, especially for African, Latin-American, and Asian cultures. While current literature includes multiple benchmarks evaluating LVLMs and VLMs on their performance in multi-cultural downstream tasks with similar findings to ours, there is no work covering mitigation strategies for the Western-biased models. Since this bias is presumably due to the lack or imbalance of data covering globally distributed cultural knowledge used to train state-of-the-art LVLMs or pre-train their LLM backbones, one promising direction of future work is to craft high-quality and balanced training datasets. Promising resources are, for example, the rich multimodal material of cultural heritage collected by UNESCO[1], the European Commission[2], the Southeast Asian Cultural Heritage Alliance (SEACHA)[3], or the Journal of African Cultural Heritage[4]. Including such resources in standard collections of training datasets for LVLMs or LLMs would greatly benefit our aim to achieve globally equitable AI systems.

1. https://www.unesco.org/world-heritage, https://ich.unesco.org
2. https://culture.ec.europa.eu/cultural-heritage
3. https://seacha.org/
4. https://jachs.org/

References

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774,* (cited on pages 36 sq.).

Agarap, Abien Fred. 2018. Deep Learning using Rectified Linear Units (ReLU). *arXiv preprint arXiv:1803.08375,* (cited on page 31).

Ainslie, Joshua, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,* 4895–4901. Singapore. (Cited on page 37).

Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: A Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems,* 23716–23736. New Orleans, LA, USA. (Cited on pages 3, 50).

Azar, Mohammad Gheshlaghi, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A General Theoretical Paradigm to Understand Learning from Human Preferences. In *International Conference on Artificial Intelligence and Statistics,* 4447–4455. Valencia, Spain. (Cited on page 36).

Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer Normalization. *arXiv preprint arXiv:1607.06450,* (cited on pages 27, 30).

Bahdanau, Dzmitry, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations.* San Diego, CA, USA. (Cited on pages 2, 26).

Bai, Shuai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923,* (cited on page 49).

Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2204.05862,* (cited on page 36).

Bao, Hangbo, Li Dong, Songhao Piao, and Furu Wei. 2021. BEiT: BERT Pre-Training of Image Transformers. In *International Conference on Learning Representations.* Virtual. (Cited on pages 38, 41).

Beltagy, Iz, Matthew E Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150,* (cited on page 37).

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research,* 1137–1155. (Cited on pages 2, 22 sq.).

Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. 1994. Learning Long-Term Dependencies with Gradient Descent is Difficult. *IEEE Transactions on Neural Networks,* 157–166. (Cited on page 24).

Beyer, Lucas, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. 2023. FlexiViT: One Model for All Patch Sizes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 14496–14506. Vancouver, Canada. (Cited on page 48).

BigScience, Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100,* (cited on page 33).

Bishop, Christopher M, and Hugh Bishop. 2024. Deep learning: Foundations and Concepts. (Cited on pages 19, 22).

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research,* 993–1022. (Cited on page 22).

Brohan, Anthony, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. 2022. RT-1: Robotics Transformer for Real-World Control at Scale. *arXiv preprint arXiv:2212.06817,* (cited on page 61).

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems,* 1877–1901. Online. (Cited on pages 3, 33 sq.).

Caffagni, Davide, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The Revolution of Multimodal Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2024,* 13590–13618. Bangkok, Thailand. (Cited on page 50).

Campbell, Declan, Sunayana Rane, Tyler Giallanza, Camillo Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven Frankland, Tom Griffiths, Jonathan D Cohen, et al. 2024. Understanding the Limits of Vision Language Models through the Lens of the Binding Problem. In *Advances in Neural Information Processing Systems,* 113436–113460. Vancouver, Canada. (Cited on page 61).

Caron, Mathilde, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *IEEE/CVF International Conference on Computer Vision,* 9630–9640. Virtual. (Cited on pages 38, 41, 49).

Cartella, Giuseppe, Alberto Baldrati, Davide Morelli, Marcella Cornia, Marco Bertini, and Rita Cucchiara. 2023. OpenFashionCLIP: Vision-and-Language Contrastive Learning with Open-Source Fashion Data. In *Image Analysis and Processing – ICIAP 2023: 22nd International Conference, Proceedings, Part I,* 245–256. Udine, Italy. (Cited on page 48).

Chen, Guanhua, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. 2023a. mCLIP: Multilingual CLIP via Cross-lingual Transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 13028–13043. Toronto, Canada. (Cited on page 48).

Chen, Liang, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024a. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models. In *European Conference on Computer Vision,* 19–35. Milano, Italy. (Cited on page 50).

Chen, Lin, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024b. ShareGPT4V: Improving Large Multi-modal Models with Better Captions. In *European Conference on Computer Vision,* 370–387. Milano, Italy. (Cited on page 53).

Chen, Xi, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel M. Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023b. PaLI: A Jointly-Scaled Multilingual Language-Image Model. In *The Eleventh International Conference on Learning Representations.* (Cited on pages 3, 48).

Chen, Xiaokang, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. *arXiv preprint arXiv:2501.17811,* (cited on page 49).

Chen, Yen-Chun, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *European Conference on Computer Vision,* 104–120. Online. (Cited on pages 3, 48).

Chen, Zhe, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023c. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 24185–24198. Vancouver, Canada. (Cited on page 49).

Chheang, Vuthea, Shayla Sharmin, Rommy Márquez-Hernández, Megha Patel, Danush Rajasekaran, Gavin Caulfield, Behdokht Kiafar, Jicheng Li, Pinar Kullu, and Roghayeh Leila Barmaki. 2024. Towards Anatomy Education with Generative AI-based Virtual Assistants in Immersive Virtual Reality Environments. In *EEE International Conference on Artificial Intelligence and eXtended and Virtual Reality,* 21–30. Los Angeles, CA, USA. (Cited on page 61).

Child, Rewon, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating Long Sequences with Sparse Transformers. *arXiv preprint arXiv:1904.10509,* (cited on pages 36 sq.).

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* 1724–1734. Doha, Qatar. (Cited on pages 2, 24 sq.).

Choromanski, Krzysztof, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, David Belanger, Lucy Colwell, et al. 2020. Masked Language Modeling for Proteins via Linearly Scalable Long-Context Transformers. *arXiv preprint arXiv:2006.03555,* (cited on page 37).

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research,* 1–113. (Cited on pages 34, 37).

Christiano, Paul F, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems,* 4302–4310. Long Beach, CA, USA. (Cited on page 3).

Collobert, Ronan, and Jason Weston. 2008. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning,* 160–167. New York, NY, USA. (Cited on pages 2, 22).

Dai, Josef, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe RLHF: Safe Reinforcement Learning from Human Feedback. In *The Twelfth International Conference on Learning Representations.* (Cited on page 36).

Dai, Zihang, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* edited by Anna Korhonen, David Traum, and Lluís Màrquez, 2978–2988. Florence, Italy. (Cited on page 37).

Dao, Tri. 2024. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. In *The Twelfth International Conference on Learning Representations.* Vienna, Austria. (Cited on page 37).

Dao, Tri, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems,* 16344–16359. New Orleans, LA, USA. (Cited on page 37).

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science,* 391–407. (Cited on page 22).

Dehghani, Mostafa, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd Van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. 2023a. Scaling Vision Transformers to 22 Billion Parameters. In *Proceedings of the 40th International Conference on Machine Learning,* edited by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, 7480–7512. Honolulu, HI, USA. (Cited on page 38).

Dehghani, Mostafa, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. 2023b. Patch n'Pack: NaViT, A Vision Transformer for Any Aspect Ratio and Resolution. In *Advances in Neural Information Processing Systems,* 2252–2274. New Orleans, LA, USA. (Cited on pages 48 sq.).

Deng, Chaorui, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. 2025. Emerging Properties in Unified Multimodal Pretraining. *arXiv preprint arXiv:2505.14683,* (cited on page 49).

Desai, Karan, and Justin Johnson. 2021. VirTex: Learning Visual Representations from Textual Annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 11162–11173. Virtual. (Cited on page 46).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics,* 4171–4186. Minneapolis, MN, USA. (Cited on pages 2, 31, 40 sq., 43).

Ding, Yi, Bolian Li, and Ruqi Zhang. 2025. ETA: Evaluating Then Aligning Safety of Vision Language Models at Inference Time. In *The Thirteenth International Conference on Learning Representations.* Singapore. (Cited on page 61).

Donahue, Jeffrey, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2625–2634. Boston, MA, USA. (Cited on page 50).

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations.* (Cited on pages 3, 38 sqq.).

Esser, Patrick, Robin Rombach, and Bjorn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 12873–12883. Virtual. (Cited on page 52).

Faghri, Fartash, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. (Newcastle, UK), 12. (Cited on page 44).

Fang, Yuxin, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2024. EVA-02: A Visual Representation for Neon Genesis. *Image and Vision Computing,* 105171. (Cited on pages 48 sq.).

Farhadi, Ali, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every Picture Tells a Story: Generating Sentences from Images. In *Proceedings of the 11th European Conference on Computer Vision,* 15–29. Heraklion, Greece. (Cited on page 50).

Gadre, Samir Yitzhak, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. 2023. DataComp: In Search of the Next Generation of Multimodal Datasets. In *Advances in Neural Information Processing Systems,* 27092–27112. New Orleans, LA, USA. (Cited on pages 48, 53).

Gale, William A, and Geoffrey Sampson. 1995. Good-Turing Frequency Estimation Without Tears. *Journal of Quantitative Linguistics,* 217–237. (Cited on pages 2, 19).

Gao, Leo, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027,* (cited on page 34).

Geigle, Gregor, Florian Schneider, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, and Goran Glavaš. 2025. Centurio: On Drivers of Multilingual Ability of Large Vision-Language Model. In *Proceedings of the Association for Computational Linguistics: ACL 2025,* in press. Vienna, Austria: Association for Computational Linguistics. (Cited on pages 9 sq., 12, 48, 55, 60).

Glaese, Amelia, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving Alignment of Dialogue Agents via Targeted Human Judgements. *arXiv preprint arXiv:2209.14375,* (cited on page 36).

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. MIT Press. (Cited on pages 19, 22).

Gu, Tiancheng, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. 2025. Breaking the Modality Barrier: Universal Embedding Learning with Multimodal LLMs. *arXiv preprint arXiv:2504.17432,* (cited on page 62).

He, K, X Chen, S Xie, Y Li, P Dollár, and R Girshick. 2021. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 16000–16009. Virtual. (Cited on pages 38, 41).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 770–778. Las Vegas, NV, USA. (Cited on pages 27, 42, 45).

Hinck, Musashi, Carolin Holtermann, Matthew Lyle Olson, Florian Schneider, Sungduk Yu, Anahita Bhiwandiwalla, Anne Lauscher, Shao-Yen Tseng, and Vasudev Lal. 2024a. Why do LLaVA Vision-Language Models Reply to Images in English? In *Findings of the Association for Computational Linguistics: EMNLP 2024,* 13402–13421. Miami, Florida, USA: Association for Computational Linguistics. (Cited on page 10).

Hochreiter, Sepp. 1998. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* 107–116. (Cited on page 24).

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation,* 1735–1780. (Cited on pages 2, 17, 24).

Hong, Jiwoo, Noah Lee, and James Thorne. 2024. ORPO: Monolithic Preference Optimization without Reference Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing,* 11170–11189. Miami, Florida, USA. (Cited on page 36).

Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *International Conference on Machine Learning,* 2790–2799. Long Beach, CL, USA. (Cited on page 51).

Hu, Edward J, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-Rank Adaptation of Large Language Models. *The Tenth International Conference on Learning Representations, ICLR 2022* (Virtual), 3. (Cited on page 51).

Huang, Zhiheng, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint arXiv:1508.01991,* (cited on page 24).

Hurst, Aaron, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. 2024. GPT-4o System Card. *arXiv preprint arXiv:2410.21276,* (cited on page 49).

Ioffe, Sergey, and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning,* 448–456. Lille, France. (Cited on page 30).

Jacobs, Robert A., Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive Mixtures of Local Experts. *Neural Computation,* 79–87. (Cited on page 38).

Jia, Chao, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision. In *International Conference on Machine Learning,* 4904–4916. Virtual. (Cited on page 48).

Jiang, Albert Q, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825,* (cited on page 36).

Jiang, Ting, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024a. E5-V: Universal Embeddings with Multimodal Large Language Models. *arXiv preprint arXiv:2407.12580,* (cited on page 62).

Jiang, Yao, Xinyu Yan, Ge-Peng Ji, Keren Fu, Meijun Sun, Huan Xiong, Deng-Ping Fan, and Fahad Shahbaz Khan. 2024b. Effectiveness Assessment of Recent Large Vision-Language Models. *Visual Intelligence,* 17. (Cited on page 61).

Jiang, Ziyan, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2025. VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks. In *The Thirteenth International Conference on Learning Representations.* (Cited on page 62).

Jordan, M.I., and R.A. Jacobs. 1993. Hierarchical Mixtures of Experts and the EM Algorithm. In *Proceedings of 1993 International Conference on Neural Networks,* 1339–1344 vol.2. Nagoya, Japan. (Cited on page 38).

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361,* (cited on page 37).

Katz, Slava. 2003. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing,* 400–401. (Cited on pages 2, 19).

Keskar, Nitish Shirish, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858,* (cited on page 33).

Kim, Junho, Kim Yeonju, and Yong Man Ro. 2024. What if...?: Thinking Counterfactual Keywords Helps to Mitigate Hallucination in Large Multi-modal Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024,* 10672–10689. Miami, Florida, USA. (Cited on page 61).

Kingma, Diederik P., and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations.* San Diego, CA, USA. (Cited on page 21).

Kiros, Ryan, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal Neural Language Models. In *Proceedings of the 31st International Conference on Machine Learning,* edited by Eric P. Xing and Tony Jebara, 595–603. Bejing, China. (Cited on page 50).

Kneser, Reinhard, and Hermann Ney. 1995. Improved Backing-Off for M-Gram Language Modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing,* 181–184. Detroit, MI, USA. (Cited on pages 2, 17, 19).

Kudo, Taku, and John Richardson. 2018. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations,* 66–71. Brussels, Belgium. (Cited on page 32).

Lan, Zhibin, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. 2025. LLaVE: Large Language and Vision Embedding Models with Hardness-Weighted Contrastive Learning. *arXiv preprint arXiv:2503.04812,* (cited on page 62).

LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient Based Learning Applied to Document Recognition. *Proceedings of the IEEE,* 2278–2324. (Cited on page 42).

Lewis, Mike, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. 2021. Base Layers: Simplifying Training of Large, Sparse Models. In *International Conference on Machine Learning,* 6265–6274. Virtual. (Cited on page 38).

Li, Chunyuan, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In *Advances in Neural Information Processing Systems,* 28541–28564. New Orleans, LA, USA. (Cited on page 61).

Li, Haoyang, Yiming Li, Anxin Tian, Tianhao Tang, Zhanchao Xu, Xuejia Chen, Nicole Hu, Wei Dong, Qing Li, and Lei Chen. 2024a. A Survey on Large Language Model Acceleration Based on KV Cache Management. *arXiv preprint arXiv:2412.19442,* (cited on page 37).

Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning,* 19730–19742. Honolulu, HI, USA. (Cited on page 50).

Li, Siming, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. 2011. Composing Simple Image Descriptions using Web-scale N-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning,* edited by Sharon Goldwater and Christopher Manning, 220–228. Portland, Oregon, USA. (Cited on page 50).

Li, Xianhang, Zeyu Wang, and Cihang Xie. 2023c. An Inverse Scaling Law for CLIP Training. In *Advances in Neural Information Processing Systems,* 49068–49087. New Orleans, LA, USA. (Cited on page 48).

Li, Xinghang, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. 2024b. Vision-Language Foundation Models as Effective Robot Imitators. In *The Twelfth International Conference on Learning Representations.* Vienna, Austria. (Cited on page 61).

Li, Xiujun, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *The 16th European Conference on Computer Vision,* 121–137. Glasgow, United Kingdom. (Cited on page 3).

Li, Yanghao, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. 2023d. Scaling Language-Image Pre-Training via Masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 23390–23400. Vancouver, Canada. (Cited on page 48).

Liang, Victor Weixin, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the Gap: Understanding the Modality Gap in Multi-Modal Contrastive Representation Learning. In *Advances in Neural Information Processing Systems,* 17612–17625. New Orleans, LA, USA. (Cited on page 62).

Liang, Zijing, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A Survey of Multimodal Large Language Models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering,* 405–409. Xi'an, China. (Cited on page 50).

Lieber, Opher, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical Details and Evaluation. *White Paper. AI21 Labs,* 1–17. (Cited on page 33).

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision,* 740–755. Zurich, Switzerland. (Cited on pages 44, 53).

Liu, Haotian, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 26296–26306. Seattle, WA, USA. (Cited on page 49).

Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems,* 34892–34916. New Orleans, LT, USA. (Cited on pages 3, 52 sqq.).

Lu, Junyu, Dixiang Zhang, Songxin Zhang, Zejian Xie, Zhuoyang Song, Cong Lin, Jiaxing Zhang, Bingyi Jing, and Pingjian Zhang. 2023. Lyrics: Boosting Fine-Grained Language-Vision Alignment and Comprehension via Semantic-Aware Visual Objects. *arXiv preprint arXiv:2312.05278,* (cited on page 49).

Messina, Nicola, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Fine-grained Visual Textual Alignment for Cross-Modal Retrieval using Transformer Encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications,* 1–23. (Cited on pages 11, 43 sq., 48, 58 sq.).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems,* 1–9. Lake Tahoe, NV, USA. (Cited on page 22).

Mikolov, Tomás, and Jeffrey Dean Kai Chen Greg Corrado. 2013b. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations,* 1–12. Scottsdale, AY, USA. (Cited on page 22).

Oquab, Maxime, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2025. DINOv2: Learning Robust Visual Features without Supervision. In *The Twelfth International Conference on Learning Representations.* (Cited on pages 38, 41, 49).

Ordonez, Vicente, Girish Kulkarni, and Tamara Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems,* 1143–1151. Granada, Spain. (Cited on page 53).

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems,* 27730–27744. New Orleans, LA, USA. (Cited on pages 3, 34 sq.).

Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio. 2013. On The Difficulty of Training Recurrent Neural Networks. In *Proceedings of the 30th International Conference on Machine Learning,* 1310–1318. Atlanta, GA, USA. (Cited on page 24).

Pei, Jiahuan, Irene Viola, Haochen Huang, Junxiao Wang, Moonisa Ahsan, Fanghua Ye, Jiang Yiming, Yao Sai, Di Wang, Zhumin Chen, Pengjie Ren, and Pablo Cesar. 2024. Autonomous Workflow for Multimodal Fine-Grained Training Assistants Towards Mixed Reality. In *Findings of the Association for Computational Linguistics: ACL 2024,* 4051–4066. Bangkok, Thailand. (Cited on page 61).

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* 1532–1543. Doha, Qatar. (Cited on page 23).

Plummer, Bryan A., Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *Proceedings of the IEEE International Conference on Computer Vision,* 2641–2649. Santiago, Chile. (Cited on page 44).

Press, Ofir, Noah A Smith, and Mike Lewis. 2022. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. In *The Twelfth International Conference on Learning Representations.* Vienna, Austria. (Cited on page 36).

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning,* 8748–8763. Virtual. (Cited on pages 3, 45 sq., 48, 62).

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving Language Understanding by Generative Pre-Training. (Cited on pages 3, 32 sq.).

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (Cited on page 33).

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems,* 53728–53741. New Orleans, LA, USA. (Cited on page 36).

Ramesh, Aditya, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning,* 8821–8831. Virtual. (Cited on page 52).

Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 1137–1149. (Cited on pages 42 sq.).

Ruder, Sebastian. 2016. An Overview of Gradient Descent Optimization Algorithms. *arXiv preprint arXiv:1609.04747,* (cited on page 21).

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1985. Learning Internal Representations by Error Propagation. (Cambridge, MA, USA), 318–362. (Cited on pages 2, 17, 23).

Schmidt, Fabian David, Florian Schneider, Chris Biemann, and Goran Glavaš. 2025. MVL-SIB: A Massively Multilingual Vision-Language Benchmark for Cross-Modal Topical Matching. In *Findings of the Association for Computational Linguistics: ACL 2025,* in press. Vienna, Austria: Association for Computational Linguistics. (Cited on pages 10, 55).

Schneider, Florian, Özge Alaçam, Xintong Wang, and Chris Biemann. 2021. Towards Multi-Modal Text-Image Retrieval to Improve Human Reading. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop.* Online: Association for Computational Linguistics. (Cited on pages 10, 43, 48).

Schneider, Florian, Narges Baba Ahmadi, Niloufar Baba Ahmadi, Iris Vogel, Martin Semmann, and Chris Biemann. 2025a. CollEX – A Multimodal Agentic RAG System Enabling Interactive Exploration of Scientific Collections. In *Proceedings of the 1st Workshop on Multimodal Augmented Generation via MultimodAl Retrieval (MAGMaR),* in press. Vienna, Austria: Association for Computational Linguistics. (Cited on pages 9 sqq., 48, 55, 59).

Schneider, Florian, and Chris Biemann. 2022. Golden Retriever: A Real-Time Multi-Modal Text-Image Retrieval System with the Ability to Focus. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval,* 3245–3250. SIGIR '22. Madrid, Spain: Association for Computing Machinery. (Cited on pages 9 sqq., 43, 48, 58).

Schneider, Florian, and Chris Biemann. 2023a. LT at SemEval-2023 Task 1: Effective Zero-Shot Visual Word Sense Disambiguation Approaches using External Knowledge Sources. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023),* 462–468. Toronto, Canada: Association for Computational Linguistics. (Cited on pages 10, 48).

Schneider, Florian, and Chris Biemann. 2024a. WISMIR3: A Multi-Modal Dataset to Challenge Text-Image Retrieval Approaches. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR),* 1–6. Bangkok, Thailand: Association for Computational Linguistics. (Cited on pages 9 sqq., 44, 48, 59).

Schneider, Florian, Tim Fischer, Fynn Petersen-Frey, Isabel Eiser, Gertraud Koch, and Chris Biemann. 2023b. The D-WISE Tool Suite: Multi-Modal Machine-Learning-Powered Tools Supporting and Enhancing Digital Discourse Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations),* 328–335. Toronto, Canada: Association for Computational Linguistics. (Cited on pages 10, 48).

Schneider, Florian, Carolin Holtermann, Chris Biemann, and Anne Lauscher. 2025b. GIMMICK – Globally Inclusive Multimodal Multitask Cultural Knowledge Benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2025,* in press. Vienna, Austria: Association for Computational Linguistics. (Cited on pages 9 sq., 12, 41, 55, 60).

Schneider, Florian, and Sunayana Sitaram. 2024b. M5 – A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024,* 4309–4345. Miami, Florida, USA: Association for Computational Linguistics. (Cited on pages 9 sqq., 55, 60).

Schuhmann, Christoph, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. In *Advances in Neural Information Processing Systems,* 25278–25294. New Orleans, LA, USA. (Cited on pages 45, 48, 53).

Schuhmann, Christoph, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400m: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *arXiv preprint arXiv:2111.02114,* (cited on page 48).

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347,* (cited on page 35).

Schuster, Mike, and Kuldip K Paliwal. 1997. Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing,* 2673–2681. (Cited on page 24).

Shaikh, Muhammad Bilal, Douglas Chai, Syed Muhammad Shamsul Islam, and Naveed Akhtar. 2024. From CNNs to Transformers in Multimodal Human Action Recognition: A Survey. *ACM Transactions on Multimedia Computing, Communications and Applications,* 1–24. (Cited on pages 41 sq.).

Shao, Zhihong, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300,* (cited on page 36).

Sharma, Aditya, Luke Yoffe, and Tobias Höllerer. 2024. OCTO+: A Suite for Automatic Open-Vocabulary Object Placement in Mixed Reality. In *IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality,* 157–165. Los Angeles, CA, USA. (Cited on page 61).

Shaw, Peter, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers),* edited by Marilyn Walker, Heng Ji, and Amanda Stent, 464–468. New Orleans, Louisiana. (Cited on page 36).

Shazeer, Noam. 2019. Fast Transformer Decoding: One Write-Head Is All You Need. *arXiv preprint arXiv:1911.02150,* (cited on page 37).

Shazeer, Noam, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations.* Toulon, France. (Cited on page 38).

Shi, Baifeng, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. 2024. When Do We Not Need Larger Vision Models? In *European Conference on Computer Vision,* 444–462. Milano, Italy. (Cited on page 50).

Shi, Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Advances in Neural Information Processing Systems.* Montreal, Canada. (Cited on page 24).

Spärck Jones, Karen. 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation,* 11–21. (Cited on page 11).

Stiennon, Nisan, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to Summarize with Human Feedback. In *Advances in Neural Information Processing Systems,* 3008–3021. Vancouver, Canada. (Cited on pages 3, 35).

Su, Jianlin, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomput.,* (cited on page 36).

Sun, Chen, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *Proceedings of the IEEE International Conference on Computer Vision,* 843–852. Venice, Italy. (Cited on page 45).

Sun, Quan, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: Improved Training Techniques for CLIP at Scale. *arXiv preprint arXiv:2303.15389,* (cited on pages 48 sq.).

Sun, Quan, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. 2024a. EVA-CLIP-18B: Scaling Clip to 18 Billion Parameters. *arXiv preprint arXiv:2402.04252,* (cited on page 48).

Sun, Zhiqing, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024b. Aligning Large Multimodal Models with Factually Augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024,* 13088–13110. Bangkok, Thailand. (Cited on page 55).

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems,* 3104–3112. Montreal, Quebec, Canada. (Cited on page 25).

Tay, Yi, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient Transformers: A Survey. *ACM Comput. Surv.* (New York, NY, USA), (cited on page 37).

Team, Gemini, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530,* (cited on pages 36, 49).

Tong, Shengbang, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. In *Advances in Neural Information Processing Systems,* 87310–87356. Vancouver, Canada. (Cited on pages 49, 54).

Touvron, Hugo, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. 2021. Training Data-Efficient Image Transformers & Distillation through Attention. In *Proceedings of the 38th International Conference on Machine Learning,* edited by Marina Meila and Tong Zhang, 10347–10357. Virtual. (Cited on page 38).

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288,* (cited on page 36).

Tschannen, Michael, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. 2025. SigLip 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786,* (cited on pages 48 sq.).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems,* 5998–6008. Long Beach, CA, USA. (Cited on pages 2, 17, 26 sqq., 30).

Wadekar, Shakti N, Abhishek Chaurasia, Aman Chadha, and Eugenio Culurciello. 2024. The Evolution of Multimodal Model Architectures. *arXiv preprint arXiv:2405.17927,* (cited on page 50).

Wang, Cheng, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016. Image Captioning with Deep Bidirectional LSTMs. In *Proceedings of the 24th ACM International Conference on Multimedia,* 988–997. New York, NY, USA. (Cited on page 50).

Wang, Peng, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191,* (cited on page 49).

Wang, Sinong, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-Attention with Linear Complexity. *arXiv preprint arXiv:2006.04768,* (cited on page 37).

Wang, Xintong, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding. In *Findings of the Association for Computational Linguistics: ACL 2024,* 15840–15853. Bangkok, Thailand. (Cited on page 61).

Wang, Xintong, Florian Schneider, Özge Alacam, Prateek Chaudhury, and Chris Biemann. 2022. MOTIF: Contextualized Images for Complex Words to Improve Human Reading. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 2468–2477. Marseille, France: Association for Computational Linguistics. (Cited on pages 10, 43, 48).

Wang, Xinyu, Bohan Zhuang, and Qi Wu. 2025. Are Large Vision Language Models Good Game Players? In *The Thirteenth International Conference on Learning Representations.* Singapore. (Cited on page 61).

Wei, Jason, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2023. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations.* Kigali, Rwanda. (Cited on pages 34 sq.).

Weizenbaum, Joseph. 1966. ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM,* 36–45. (Cited on page 2).

Wiehe, Anton, Florian Schneider, Sebastian Blank, Xintong Wang, Hans-Peter Zorn, and Christian Biemann. 2022b. Language over Labels: Contrastive Language Supervision Exceeds Purely Label-Supervised Classification Performance on Chest X-Rays. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop,* 76–83. Online: Association for Computational Linguistics. (Cited on page 10).

Witten, Ian H, and Timothy C Bell. 1991. The Zero-Frequency Problem: Estimating The Probabilities of Novel Events in Adaptive Text Compression. *IEEE Transactions on Information Theory,* 1085–1094. (Cited on pages 2, 19).

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144,* (cited on page 32).

Xia, Peng, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. 2024. CARES: A Comprehensive Benchmark of Trustworthiness in Medical Vision Language Models. In *Advances in Neural Information Processing Systems,* 140334–140365. Vancouver, Canada. (Cited on page 61).

Xiong, Ruibin, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. 2020. On Layer Normalization in the Transformer Architecture. In *International Conference on Machine Learning,* 10524–10533. Virtual. (Cited on page 36).

Xu, Jin, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. 2025. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215,* (cited on page 49).

Yang, Senqiao, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2025. VisionZip: Longer is Better But Not Necessary in Vision Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference,* 19792–19802. Nashville, TN, USA. (Cited on page 50).

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems,* 5753–5763. Vancouver, Canada. (Cited on page 33).

Yin, Shukang, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A Survey on Multimodal Large Language Models. *National Science Review,* (cited on pages 61 sq.).

Yu, Tianyu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024. RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 13807–13816. Vancouver, Canada. (Cited on page 55).

Zhai, Xiaohua, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 11975–11986. Paris, France. (Cited on pages 48 sq.).

Zhai, Xiaohua, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. LiT: Zero-shot Transfer with Locked-Image Text Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 18123–18133. New Orleans, LA, USA. (Cited on page 48).

Zhang, Biao, and Rico Sennrich. 2019. Root Mean Square Layer Normalization. In *Advances in Neural Information Processing Systems,* edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, 12381–12392. Vancouver, Canada. (Cited on page 36).

Zhang, Ce, Kaixin Ma, Tianqing Fang, Wenhao Yu, Hongming Zhang, Zhisong Zhang, Yaqi Xie, Katia Sycara, Haitao Mi, and Dong Yu. 2025a. VScan: Rethinking Visual Token Reduction for Efficient Large Vision-Language Models. *arXiv preprint arXiv:2505.22654,* (cited on page 50).

Zhang, Duzhen, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. MM-LLMs: Recent Advances in MultiModal Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024,* 12401–12430. Bangkok, Thailand. (Cited on page 50).

Zhang, Yi-Fan, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, et al. 2025b. MM-RLHF: The Next Step Forward in Multimodal LLM Alignment. In *Proceedings of the 42nd International Conference on Machine Learning,* in press. Vancouver, Canada. (Cited on page 55).

Zhang, Jingyi, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024b. Vision-Language Models for Vision Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 5625–5644. (Cited on pages 41, 61 sq.).

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open Pre-Trained Transformer Language Models. *arXiv preprint arXiv:2205.01068,* (cited on page 33).

Zhao, Yunqing, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. 2023a. On Evaluating Adversarial Robustness of Large Vision-Language Models. In *Advances in Neural Information Processing Systems,* 54111–54138. Vancouver, Canada. (Cited on page 61).

Zhao, Zihao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, et al. 2023b. CLIP in Medical Imaging: A Comprehensive Survey. *arXiv preprint arXiv:2312.07353,* (cited on page 48).

Zhen, Haoyu, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 2024. 3D-VLA: A 3D Vision-Language-Action Generative World Model. In *Proceedings of the 41st International Conference on Machine Learning,* 61229–61245. Vienna, Austria. (Cited on page 61).

Zhu, Jinguo, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479,* (cited on page 49).

Zhu, Kangyu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. 2024. MMedPO: Aligning Medical Vision-Language Models with Clinical-Aware Multimodal Preference Optimization. *arXiv preprint arXiv:2412.06141,* (cited on page 61).

Zhu, Wanrong, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. Multimodal C4: An Open, Billion-scale Corpus of Images Interleaved With Text. In *Advances in Neural Information Processing Systems,* 8958–8974. New Orleans, LA, USA. (Cited on page 48).

Ziegler, Daniel M, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-Tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593,* (cited on page 35).

Foundational Papers

I

Golden Retriever: A Real-Time Multi-Modal Text-Image Retrieval System with the Ability to Focus

Bibliographic Entry

Florian Schneider and Chris Biemann. 2022. Golden Retriever: A Real-Time Multi-Modal Text-Image Retrieval System with the Ability to Focus. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval,* 3245–3250. SIGIR '22. Madrid, Spain: Association for Computing Machinery

Demo Paper SIGIR '22, July 11–15, 2022, Madrid, Spain

Golden Retriever: A Real-Time Multi-Modal Text-Image Retrieval System with the Ability to Focus

Florian Schneider
Chris Biemann
florian.schneider-1@uni-hamburg.de
christian.biemann@uni-hamburg.de
Universität Hamburg
Hamburg, Germany

ABSTRACT

In this work, we present the Golden Retriever, a system leveraging state-of-the-art visio-linguistic models (VLMs) for real-time text-image retrieval. The unique feature of our system is that it can focus on words contained in the textual query, i.e., locate and highlight them within retrieved images. An efficient two-stage process implements real-time capability and the ability to focus. Therefore, we first drastically reduce the number of images processed by a VLM. Then, in the second stage, we rank the images and highlight the focussed word using the outputs of a VLM. Further, we introduce a new and efficient algorithm based on the idea of TF-IDF to retrieve images for short textual queries. One of multiple use cases where we employ the Golden Retriever is a language learner scenario, where visual cues for "difficult" words within sentences are provided to improve a user's reading comprehension. However, since the backend is completely decoupled from the frontend, the system can be integrated into any other application where images must be retrieved fast. We demonstrate the Golden Retriever with screenshots of a minimalistic user interface.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Image search**.

KEYWORDS

multi-modal; text-image retrieval system; visio-linguistic models

ACM Reference Format:
Florian Schneider and Chris Biemann. 2022. Golden Retriever: A Real-Time Multi-Modal Text-Image Retrieval System with the Ability to Focus. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3477495.3531666

1 INTRODUCTION

The famous adage *"A picture is worth a thousand words."* can be interpreted in various ways. One way is to see this as a motivation and problem statement for multi-modal text-image retrieval systems that enable searching images with words, i.e., textual queries While current solutions, e.g., from popular search engines, work astonishingly well, they lack the ability to focus single words of the query and locate them within the retrieved images. That is, they rank images according only to the whole textual query.

With the Golden Retriever presented in this paper, we propose a solution to put particular focus on a word within the query when retrieving images. Further, we locate and highlight the denoted image for the focused word within the retrieved images.

This feature enables multiple use cases, for example, a multi-modal language learner scenario, where visual cues for difficult words can support a user's reading comprehension. To implement the ability to retrieve and rank images not only for the whole textual query but additionally incorporate focused words within the query, we leverage state-of-the-art multi-modal models. However, these models are computationally heavy, challenging real-time critical applications when searching through a large pool of images. With our system, we propose a solution for this by implementing an efficient preprocessing stage that drastically reduces the number of images processed by the multi-modal retrieval model. As a part of this preprocessing stage, we further introduce a fast new algorithm based on TF-IDF [11] to retrieve images for textual queries.

2 RELATED WORK

There were significant breakthroughs in various computer vision and natural language processing tasks during the last few years [3, 7, 8, 21]. This progress of uni-modal models also led to a great leap forward in multi-modal visio-linguistic models (VLMs), leveraging the power of transformers to work with text and images simultaneously [5, 12, 14, 16]. For content-based text-image retrieval [6, 22], these VLMs learn a metric function $\Phi(Q, I) : \mathbb{R}^{|Q|\times|I|} \rightarrow [0, 1]$ that measures the similarity of a textual query Q and image I. The objective is to find the best matching image $I_k = \underset{i \in P}{\operatorname{argmax}}\, \Phi(Q, I_i)$ for the query text Q from a pool of images P.

There are two major differences in the architecture of current VLMs, affecting how the text-image similarity is computed. In so-called early-fusion VLMs, a single transformer stack is employed that simultaneously processes textual and visual token embeddings and computes the text-image similarity from the outputs of the self-attention heads of the last layer. In VLMs referred to as late-fusion models, there are two transformer stacks, one for the textual input and one for the visual input. Late-fusion VLMs calculate the cosine-similarity from the textual and visual CLS tokens or from

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '22, July 11–15, 2022, Madrid, Spain
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531666

an aggregation of the other token embeddings of the last layer to compute the text-image similarity. Because the complexity of self-attention is quadratic in the number of input tokens, early-fusion models require less computational power or execution time than late-fusion models for inference. However, even with late-fusion models, "real-time" critical applications become challenging to implement when retrieving the best matching images according to a textual query from a large pool of images.

3 MOTIVATION AND CHALLENGES

There are two primary challenges the Golden Retriever system solves, briefly outlined in the following subsections.

3.1 VLMs in "real-time" critical Retrieval Systems

State-of-the-art visio-linguistic models (VLMs) require much computational power to retrieve the best matching images for a textual query from a large pool of images. Hence, it becomes challenging to leverage those VLMs for real-time critical retrieval systems. To solve this issue, the Golden Retriever system implements a sophisticated pre-selection stage that drastically reduces the number of candidate images processed by the VLMs.

3.2 Extending queries by Focus Words

The second motivation of the Golden Retriever is to extend the textual query used in standard text-image retrieval, which comprises a sequence of words by a focus word contained within the sequence. In the following, we refer to the sequence of words in the query as the *context* and the focus word as the *focus*. Then, we retrieve the best matching images according to the context and pay particular attention to the focus word in a re-ranking stage. Further, we locate and highlight the image region where the focus word is best represented in the retrieved images.

4 VISUALLY WEIGHTED TF-IDF

This section introduces an efficient method to retrieve images for textual queries consisting of short noun phrases. Our algorithm is based on TF-IDF [11], but is applied to images instead of textual documents. Hence, we refer to it as Visually-Weighted TF-IDF or VW-TF-IDF. In Section 5.3.2, we describe how we utilize this method to retrieve images relevant to the *focus*.

For the VW-TF-IDF, we interpret images as visual documents with "terms" that are classified Region-Of-Interests (ROIs) in the image predicted by an object detection and classification network, e.g., Faster-R-CNN [18]. In the current Golden Retriever version, we use a pre-trained network [1, 24] with about 1400 unique objects and attributes labels. The set of labels is what we refer to as "visual vocabulary" and each element is called a term in the following.

To compute the VW-TF-IDF score, the classical formula of TF-IDF is extended by a weighting scheme based on visual properties. The motivation is that the score should be higher if the region with the respective label is prominent in the image and the classifier is confident. Hence, the confidence scores and the ROI areas are incorporated in addition to the counts of the terms from the traditional TF-IDF formula.

Formally, we define the VW-TF-IDF of a term t and an image document d as

$$\text{vw_tf_idf}(t,d) = \text{vw_tf}(t,d) \cdot \log\left(\frac{\text{num}_{\text{docs}}}{\text{df}(t)+1}\right) \tag{1}$$

where the logarithmic term is standard inverse document frequency (IDF) with simple additive Laplace-Smoothing for numerical stability. The visually weighted term frequency (VW-TF) is defined as

$$\text{vw_tf}(t,d) = \frac{\text{cnt}(t,d) \cdot \text{weight(t, d)}}{\text{num_terms}(d)} \tag{2}$$

where $\text{cnt}(t,d)$ is the number of times term t appears in document d and $\text{num_terms}(d)$ is the total number of terms in the document The weight of the term t in d is defined as

$$\text{weight(t, d)} = \alpha\ \text{conf}(t,d) + (1-\alpha)\ \text{area}(t,d) \tag{3}$$

$$\text{conf}(t,d) = \frac{1}{\text{cnt}(t,d)} \sum_{t^{(i)} \in t} t^{(i)}_{conf} \tag{4}$$

$$\text{area}(t,d) = \frac{1}{d_{area}} \sum_{t^{(i)} \in t} t^{(i)}_{area} \tag{5}$$

where $t^{(i)}_{conf}$ is the accumulated confidence score, $t^{(i)}_{area}$ is the accumulated area of the ROIs of term $t_{(i)} \in d$, and d_{area} is the total area of the image document d. The parameter α is used to control the importance of the confidence or area of a term in the final weight of t.

To efficiently retrieve the most relevant images for a query, we first compute a VW-TF-IDF index for every term in the visual vocabulary and every image in the set of images to be searched in an offline setting. Then, in the online setting, the most relevant images have the highest VW-TF-IDF for the query and can be retrieved via simple dictionary lookups in the pre-computed index.

One major drawback of our method – and in general TF-IDF – is that the query can only contain terms from the limited visual vocabulary, i.e., the method lacks proper out-of-vocabulary handling We overcome this issue with a pre-processing step that transforms arbitrary queries to queries that only contain terms contained in the vocabulary. More on this pre-processing step is detailed in Section 5.3.2.

5 SYSTEM ARCHITECTURE

This section describes the Golden Retriever system to solve the main challenges introduced in the previous section. Auxiliary components like, e.g., a static file server for images or components to generate images with highlighted *focus* are not described here.

5.1 User Interface

The minimalistic user interface presented in Section 7 communicates with the Golden Retriever backend via HTTP calls to a REST API. It is implemented as a simple browser plugin to mimic a search engine-like environment using HTML, CSS, and plain JavaScript. However, since the frontend is decoupled from the backend, the Golden Retriever can be easily integrated within other applications

5.2 Backend Summary

The Golden Retriever backend implements the two-stage retrieval process schematically sketched in Figure 1. The first pre-selection

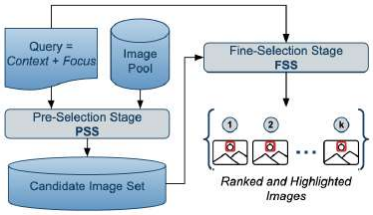


Figure 1: Schematic overview of the Golden Retriever backend system.

stage (c.f. Section 5.4) reduces the image pool P to a significantly smaller candidate image set that the VLM processes. Note that the image pool comprises images along with their corresponding textual captions, i.e., it contains multi-modal text-image data. The second fine-selection stage (c.f. Section 5.3) leverages a VLM to retrieve the best matching images from the candidate image set according to the extended query and locate the image region that best matches the *focus* word.

In the current version of the Golden Retriever, we use three different multi-modal datasets as image pools: MS COCO [13], Flickr30k [23], and a Wikipedia-based dataset collected by us for other work [20]. Further, we currently employ only TERAN [14] models trained on different datasets in the presented proof-of-concept application. However, we successfully experimented with UNITER [5] models but did not yet implement them in the demonstrated system. Furthermore, in general, every VLM that can compute text-image similarities can be integrated into the Golden Retriever system.

In an extensive experiment described in Section 6 to measure the Golden Retriever backend execution time per request, we found that the average system response is around 2.10 seconds, agnostic to the size of the image pool and the length of the query.

5.3 Pre-Selection Stage

In this stage of the Golden Retriever backend, the image pool is drastically reduced to the candidate image set. Therefore, two efficient sub-procedures are implemented: One selects images relevant to the *context*, and the other selects images relevant to the *focus*. After that, the two resulting sets are merged to obtain the final candidate image set. We first apply the intersection of the *context*-relevant and *focus*-relevant images as the merging operation. If the size of the resulting set is too small, we merge the two sets via union. This size parameter is defaulted to 5000 but can be set by the system administrator. In the following, we briefly describe the two sub-procedures.

5.3.1 Context-based Pre-Selection. To select the *context*-relevant images, we first computed sentence embeddings for every caption of the images in the image pool with an SBert [17] model for semantic textual similarity [4]. Secondly, we clustered the embeddings for efficient searching using FAISS [10] with a quantized Approximate Nearest Neighbor index. Both of these steps are done in an offline setting. Then, in the online setting, we compute the *context* embedding and retrieve the most similar captions in the cluster via cosine similarity. The associated images to the captions are considered *context*-relevant.

5.3.2 Focus-based Pre-Selection. We apply the VW-TF-IDF algorithm introduced in Section 4 to select *focus*-relevant images from the image pool. Since the *focus* part of the query can contain arbitrary words, we need to transform it so that it only consists of terms in the vocabulary of the VW-TF-IDF index. Therefore, we first use a spaCy [9] model for tokenizing the *focus* and obtaining the lemmata of the surface form of the *focus*. Note that the *focus* can comprise more than one token, e.g., for compound nouns or nouns described by adjectives. Next, we retrieve the top-k similar terms from the vocabulary for every *focus* token not contained in the vocabulary. To do this efficiently, we utilize Magnitude [15] with FastText [2] embeddings. The default value for k is set to 10, but can be adjusted per request by the user. In the final step to select the set of *focus*-relevant images, we retrieve the best matching images to the transformed query, i.e., the top-k similar terms, from the pre-computed VW-TF-IDF index.

5.4 Fine-Selection Stage

In this stage of the Golden Retriever backend, we forward the images in the candidate set through a VLM to rank them according to the twofold query. Further, we locate the region that best matches the *focus* part of the query and highlight it with a bounding box. In the current version of the Golden Retriever, we use TERAN, a late-fusion VLM designed for efficient text-image retrieval. The textual input to TERAN are token embeddings computed by a pre-trained BERT [7] tokenizer model. The visual inputs are ROI features extracted with a pre-trained Faster R-CNN [1, 18, 24]. Following the authors, we limit the number of visual tokens per image to 36. Since the query consists of two parts, i.e., the *context* and the *focus*, we compute a score for both parts and apply a weighted average in a re-ranking stage to retrieve the best matching images from the candidate set.

TERAN calculates the global similarity between an image and a textual query by computing a fine-grained word-region-alignment (WRA) matrix $\mathbf{A}$. The cells of $\mathbf{A}$, are the cosine-similarities of the visual regions of the image I and textual tokens of the *context* C are defined as

$$\mathbf{A}_{i,j} = \frac{\mathbf{v}_i^T \mathbf{t}_j}{|\mathbf{v}_i||\mathbf{t}_j|} \tag{6}$$

where $\mathbf{v}_i \in I$ and $\mathbf{t}_j \in C$.

The global similarity, i.e., the *context*-score $s_I^{(C)}$, of an image I and a *context* C is defined as

$$s_I^{(C)} = \sum_{j \in |C|} \max_{i \in |I|} \mathbf{A}_{ij} \tag{7}$$

To specially attend to the *focus* F, we compute a *focus*-score $s_I^{(F)}$ based on the WRA matrix A.

$$s_I^{(F)} = \frac{1}{N * (f_e - f_s + 1)} \sum_{i=0}^{N} \sum_{j=f_s}^{f_e} \mathbf{A}_{ij} \tag{8}$$

where N is the number of regions per image; f_s and f_e are the starting and ending indices of $F \in C$, respectively.

After that, we first normalize and then combine the global similarity (the *context*-score) with the *focus*-score by a weighted average to obtain the final score for the image s_I.

$$s_I = \alpha \cdot s'^{(C)}_I + (1 - \alpha) \cdot s'^{(F)}_I \tag{9}$$

where $\alpha \in [0, 1]$ controls the weighted average and $s'^{(C)}_I$ and $s'^{(F)}_I$ are the normalized *context*-score and *focus*-score, respectively. The default for α is set to 0.5 but can be adjusted by the user per request.

Finally, we sort the images according to their score to rank the candidate image set with respect to the *context* as well as the *focus* part of the query. To locate the region where the *focus* is represented best, we select the ROI with the maximum *focus*-score.

6 "REAL-TIME" CAPABILITY EXPERIMENT

In the following, timings of the Golden Retriever backend system and its sub-components are reported to assess the system's "real-time" capability. Note that "real-time" in the context of our system is always in parentheses because it must not be confused with "true" real-time systems as defined in the context of robotics or real-time operating systems like RTOS[1]. However, there exists a loose definition of "near-real-time" systems, according to which there must not be "significant delays"[2]. As stated in the corresponding Wikipedia article, this "delay in near real-time is typically in a range of 1-10 seconds"[3].

Multiple factors have varying influence on the system's response time. To find how much these factors weigh, the "real-time" assessment test reported in this section was conducted as follows: The system was used with different parameter, query, and dataset combinations. Each of the three queries $Q1$, $Q2$, $Q3$, with 827, 124, 67 characters in *context* length, respectively, was combined with four different modes with the COCO [13], Flickr30k [23], and WISMIR [19] datasets. This results in a set of $3 * 4 * 3 = 36$ different parameter combinations, for which the average system response time over 10 consecutive runs was measured. As it can be observed from the results presented in Figure 2, the length of the context part of the query affects the system's response time the most. This is an expected result since the similarity of an image is based on pooling the word-region-alignment (WRA) matrix, representing the fine-grained similarity of each textual and visual token. Hence, the longer the context, the larger the WRA matrix and the more time the retrieval model needs to generate and pool the matrix.

Further, the effect of the Preselection Stage (PSS) can be noticed: The larger the dataset is, from which the system retrieves the top-k images, the longer the PSS takes, whereas the average response time of the Fineselection Stage (FSS) remains almost across different datasets. Flickr30k has about 31K, COCO about 123K, and WISMIR v2 about 395K images, and the corresponding average PSS response times are 0.09s, 0.27s, and 0.52s, respectively. This increase of time of the PSS is almost linearly proportional to the number of unique images in datasets. These results also highlight the effectiveness of the two-stage retrieval approach of the system.

[1]https://www.freertos.org
[2]https://www.its.bldrdoc.gov/fs-1037/dir-024/_3492.htm
[3]https://en.wikipedia.org/wiki/Real-time_computing#Near_real-time

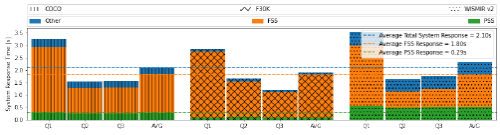


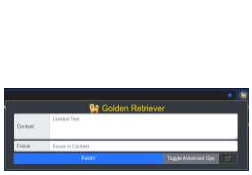
Figure 2: Averaged timing measurements of the system response time for multiple queries $Q1$, $Q2$, and $Q3$ on different datasets. Each bar represents the total system response time, which comprises the response times of different subcomponents. Best viewed digitally with zoom and color.

As depicted in Figure 2, the overall average system response time across all datasets, queries, and modes evaluated in this "real-time" suitability test of the Golden Retriever is 2.10s. Hence, in conclusion, it is considered as an acceptable result.

7 SYSTEM DEMONSTRATION

In this section, the Golden Retriever is demonstrated with screenshots of various retrieval examples with different queries using different views of the minimalist user interface.

There are four views for different text-image retrieval types supported by the Golden Retriever user interface, shown in Figure 3. The available options and parameters are described in detail on our GitHub page[4]. When the plugin is opened, it shows



(a) Simple UI

(b) Advanced UI

(c) Advanced UI to trigger a *context* only retrieval

(d) Advanced UI to trigger a *focus* only retrieval

Figure 3: Different views of the minimalistic Golden Retriever user interface. Best viewed digitally with zoom and color.

a straightforward interface presented in Figure 3a to retrieve the most similar images for a query consisting of the *context* and *focus* for non-technical users. For research purposes or advanced users, the plugin also offers an interface shown in Figure 3b with more

[4]https://github.com/floschne/MMIRS

options that can be toggled by a button. To retrieve images solely for the *context* (c.f. Section 5.3.1), the UI as shown in Figure 3c is provided. Similarly, if a user wants to retrieve images only for the *focus* (c.f. Section 5.3.2), the UI as shown in Figure 3d is used. Once the top-k images are retrieved, they are presented by an interactive slideshow to the user. The image in full resolution is opened in a new tab by clicking on an image. Figure 4 shows different Golden Retriever results for queries comprising a *context* and a *focus*. In

(a) ***focus* = *children*; $\alpha = 0.1$** **(b)** ***focus* = *children*; $\alpha = 0.9$**

(c) ***focus* = *phone*; $\alpha = 0.1$** **(d)** ***focus* = *phone*; $\alpha = 0.9$**

Figure 4: Example Golden Retriever results with highlighted *focus* regions for queries with *context* = *"Today's children are playing a lot with their phone."* but different *focus* and α values.

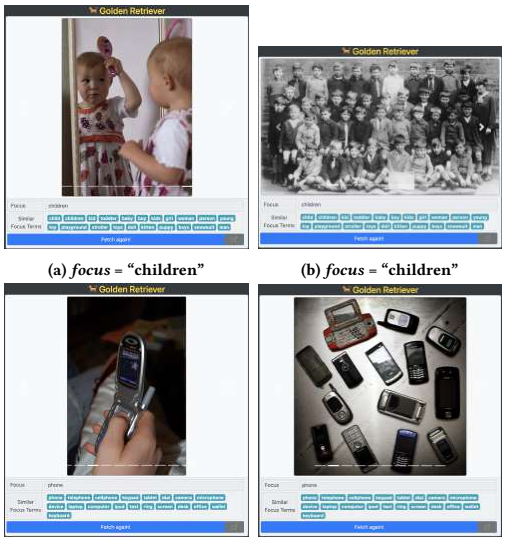
Figure 5 different Golden Retriever results for *context*-only queries (c.f. Section 5.3.1) are shown. In Figure 6 different Golden Retriever results for *context*-only queries (c.f. Section 5.3.2) are shown.

8 CONCLUSION

This paper presented the Golden Retriever, a system leveraging state-of-the-art visio-linguistic models for real-time text-image retrieval. The unique feature of our system is that it can focus on words contained in the textual query. To enable real-time capability and the ability to focus, we sketched a two-stage process implemented in the Golden Retriever. Further, we introduced an efficient algorithm based on TF-IDF to find images for short textual queries. To test the "real-time" capability of the system, we conducted an extensive experiment, where we found that the average system response time is in an acceptable range. Finally, we demonstrated the Golden Retriever with screenshots of a minimalistic user interface.

(a) **(b)**

Figure 5: Example Golden Retriever results for queries with *context* = *"Today's children are playing a lot with their phone."* and no *focus*



(a) ***focus* = "children"** **(b)** ***focus* = "children"**

(c) ***focus* = "phone"** **(d)** ***focus* = "phone"**

Figure 6: Example Golden Retriever results for queries with different *focus* words but no *context*.

ACKNOWLEDGMENTS

This research was partially funded by the German Research Foundation – DFG Transregio SFB 169: Cross-Modal Learning.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. Salt Lake City, UT, USA, 6077–6086.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics (TACL)* 5 (2017), 135–146.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. Virtual, 1877–1901.

[4] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada, 1–14. https://doi.org/10.18653/v1/S17-2001

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *European Conference on Computer Vision (ECCV)*. Online, 104–120.

[6] Paul Clough, Henning Müller, and Mark Sanderson. 2004. The CLEF 2004 cross-language image retrieval track. In *Proceedings of the 5th conference on Cross-Language Evaluation Forum: multilingual Information Access for Text, Speech and Images*. 597–613.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. Minneapolis, MN, USA, 4171–4186.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

[9] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. https://spacy.io/.

[10] Johnson, Jeff and Douze, Matthijs and Jégou, Hervé. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* (2019).

[11] Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* (1972).

[12] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-and-Language Tasks. In *European Conference on Computer Vision (ECCV)*. Online, 121–137.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*. Zurich, Switzerland, 740–755.

[14] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 4 (2021), 1–23.

[15] Ajay Patel, Alexander Sands, Chris Callison-Burch, and Marianna Apidianaki. 2018. Magnitude: A Fast, Efficient Universal Vector Embedding Utility Package. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium, 120.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning (ICML)*. Online, 8748–8763.

[17] Reimers, Nils and Gurevych, Iryna. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, 3973–3983.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 39, 6 (2016), 1137–1149.

[19] Florian Schneider. 2021. *Self-Supervised Multi-Modal Text-Image Retrieval Methods to Improve Human Reading*. Master's thesis. University of Hamburg.

[20] Florian Schneider, Özge Alaçam, Xintong Wang, and Chris Biemann. 2021. Towards Multi-Modal Text-Image Retrieval to improve Human Reading. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Mexico City, Mexico (online).

[21] Mingxing Tan and Quoc Le. 2021. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning (ICML)*. Online, 10096–10106.

[22] Christopher Phillip Town. 2004. *Ontology based Visual Information Processing*. Ph. D. Dissertation. University of Cambridge.

[23] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[24] Zhou Yu, Jing Li, Tongan Luo, and Jun Yu. 2020. A PyTorch Implementation of Bottom-Up-Attention. https://github.com/MILVLG/bottom-up-attention.pytorch.

II

CollEX: A Multimodal Agentic RAG System Enabling Interactive Exploration of Scientific Collections

Bibliographic Entry

Florian Schneider, Narges Baba Ahmadi, Niloufar Baba Ahmadi, Iris Vogel, Martin Semmann, and Chris Biemann. 2025a. CollEX – A Multimodal Agentic RAG System Enabling Interactive Exploration of Scientific Collections. In *Proceedings of the 1st Workshop on Multimodal Augmented Generation via MultimodAl Retrieval (MAGMaR),* in press. Vienna, Austria: Association for Computational Linguistics

CollEX – A Multimodal Agentic RAG System Enabling Interactive Exploration of Scientific Collections

Florian Schneider[†], **Narges Baba Ahmadi**[† *], **Niloufar Baba Ahmadi**[† *]
Iris Vogel[‡], **Martin Semmann**[†], **Chris Biemann**[†]

[†]Hub of Computing and Data Science
[‡]Center for Sustainable Research Data Management
University of Hamburg, Germany
Correspondence: florian.schneider-1@uni-hamburg.de

[*]Equal contributions, sorted alphabetically.

Abstract

In this paper, we introduce CollEX, an innovative multimodal agentic Retrieval-Augmented Generation (RAG) system designed to enhance interactive exploration of extensive scientific collections. Given the overwhelming volume and inherent complexity of scientific collections, conventional search systems often lack necessary intuitiveness and interactivity, presenting substantial barriers for learners, educators, and researchers. CollEX addresses these limitations by employing state-of-the-art Large Vision-Language Models (LVLMs) as multimodal agents accessible through an intuitive chat interface. By abstracting complex interactions via specialized agents equipped with advanced tools, CollEX facilitates curiosity-driven exploration, significantly simplifying access to diverse scientific collections and records therein. Our system integrates textual and visual modalities, supporting educational scenarios that are helpful for teachers, pupils, students, and researchers by fostering independent exploration as well as scientific excitement and curiosity. Furthermore, CollEX serves the research community by discovering interdisciplinary connections and complementing visual data. We illustrate the effectiveness of our system through a proof-of-concept application containing over 64,000 unique records across 32 collections from a local scientific collection from a public university.

1 Introduction

The exploration of scientific knowledge is a cornerstone of human progress. However, the vast and rapidly growing body of scientific literature presents significant challenges for educators and learners, who often find themselves overwhelmed by the sheer volume and complexity of information. Despite advancements in information retrieval and knowledge discovery (Santhanam et al., 2022; Zhu et al., 2023; Li et al., 2024b), existing search systems for rich and complex data often lack the interactivity, intuitiveness, and cross-modal search capabilities (Faysse et al., 2024; Zhai et al., 2023; Zhao et al., 2023b) to engage diverse audiences, such as students, teachers, or researchers. This limitation negatively affects educational settings where fostering curiosity is essential.

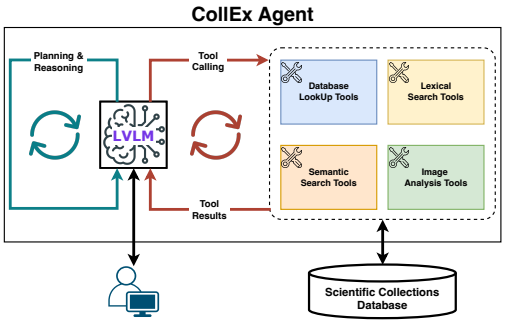


Figure 1: An overview of the CollEX Agentic System.

With this paper, we introduce CollEX, a multimodal agentic Retrieval-Augmented Generation (RAG) system (Lewis et al., 2020; Zhao et al., 2023a; Xie et al., 2024) and reimagine how users explore and interact with scientific collections such as those collected and managed by the Smithsonian Institution[1] or local collections from public universities. CollEX uses state-of-the-art Large Vision-Language Models (LVLMs)(Liu et al., 2023; Team et al., 2023; Hurst et al., 2024; Yang et al., 2024; Team et al., 2025) as multimodal agents (Xie et al., 2024; Wang et al., 2024) through an intuitive chat interface. Unlike traditional systems requiring expert knowledge, CollEX promotes curiosity-driven exploration, simplifying access and increasing engagement.

The core of CollEX is its multimodal agentic RAG system, which abstracts complex interactions using specialist agents equipped with various tools (Patil et al., 2024). This simplifies the explo-

[1] https://www.si.edu/collections

ration of extensive scientific collections, catering to users with diverse backgrounds and expertise, thereby overcoming accessibility issues (Achiam and Marandino, 2014). The system integrates texts and images, offering intuitive access to scientific concepts.

CollEX is especially beneficial in education, fostering curiosity and engagement. For instance, teachers can get inspiration to prepare visually rich lessons, retrieve relevant information, and facilitate interactive assignments. Pupils can independently explore the collections, transforming static materials into dynamic learning experiences. Moreover, CollEX supports higher education by encouraging independent exploration and enhancing critical thinking skills.

Beyond education, CollEX aids researchers in discovering interdisciplinary connections, eventual related work, or visual data complements. It autonomously enriches search queries, facilitating easier contextualization and increasing accessibility to scientific collections, thereby supporting national and international scientific connectivity (Weber, 2018).

This paper introduces CollEX's general system architecture[2]and inner workings, combining state-of-the-art LVLMs, advanced prompting and RAG techniques, cross-modal search, and agentic reasoning and planning.

Moreover, we provide three exemplary user stories to demonstrate the system by implementing a proof-of-concept application to explore 32 diverse scientific collections comprising over 64,000 unique items.

2 Related Work

2.1 Cross-Modal Information Retrieval

Cross-modal information retrieval powered by multimodal embeddings is the key foundation for systems navigating or exploring textual and visual data such as CollEX. Recent developments in multimodal embedding models (Tschannen et al., 2025) that compute semantically rich dense vector representations in an aligned vector space for texts and images, have significantly improved over the popular text-image encoder model, commonly known as CLIP (Radford et al., 2021). This progress was primarily driven by billion-scale high-quality text-image datasets (Schuhmann et al., 2022), improvements in architecture and training regimes (Zhai et al., 2023), and improved Vision Transformers (Alabdulmohsin et al., 2023) Despite their applications in "pure" information retrieval settings, the image encoders of the multimodal embedding models also play a crucial role in the advancement of Large Vision Language Models (LVLMs) (Liu et al., 2023; Yang et al., 2024; Geigle et al., 2025) as they are often used to compute the visual tokens processed by the LVLMs.

2.2 Multimodal Retrieval Augmented Generation

Multimodal RAG (Zhao et al., 2023b) systems integrate various knowledge formats, including images, code, structured databases, audio, and video, to enhance the knowledge of LVLMs at inference time. Zhao et al. (2023b) further highlight that such multimodal data helps mitigate hallucinations and improve interpretability and reasoning by grounding responses in diverse multimodal information. Riedler and Langer (2024) demonstrate the advantages of incorporating images into textual retrieval systems within industrial applications. Their findings suggest that image-derived textual summaries often outperform purely embedding-based multimodal approaches.

2.3 Agentic RAG

As described above, traditional RAG systems combine LLMs' or LVLMs' generative capabilities with external knowledge bases to enhance their outputs. Yet these methods are typically constrained by static workflows and linear processes, restricting their adaptability in complex tasks involving multi-step reasoning and dynamic data quries. Recently, agentic RAG has emerged as an extension of traditional RAG systems by employing autonomous AI agents in a loop within the RAG pipeline. Agentic RAG employs agentic design patterns and prompting such as reflection, planning, tool utilization, and multi-agent collaboration, enabling systems to iteratively refine and plan retrieval strategies and adapt dynamically to real-time and context-sensitive queries (Singh et al., 2025; Xie et al., 2024; Li et al., 2024a). For example, Schopf and Matthes (2024) introduced NLP-KG, a system specifically designed for exploratory literature search in NLP. NLP-KG supports users in exploring unfamiliar NLP fields through semantic search and conversational interfaces grounded in scholarly literature, effectively bridging the gap between ex-

[2]We publish the open-source code here:https://github.com/uhh-lt/fundus-murag

ploratory and targeted literature search tasks. Xie et al. (2024) further extends the concept of autonomous LLM agents into the multimodal domain, demonstrating how LVLMs can perceive and interpret diverse data types beyond text, such as images and videos. Further, they outline critical components necessary for multimodal agent functionality, including visual perception and planning.

With CollEX, we integrate a powerful multimodal embedding model for effective cross-modal semantic search with state-of-the-art LVLMs employed as autonomous agents in a multimodal RAG system. With this, we support educational scenarios by fostering independent exploration, scientific curiosity, and excitement that benefit teachers, pupils, students, and researchers alike.

3 The CollEX System

This section describes the CollEX system, i.e., its architecture and core components, as well as the data to be explored.

3.1 CollEX Data

Since CollEX is a multimodal agentic RAG system, to understand the system, it is essential to know the data it operates on.

Schema. We provide the simplified data schema as a UML class diagram in Figure 2. As the

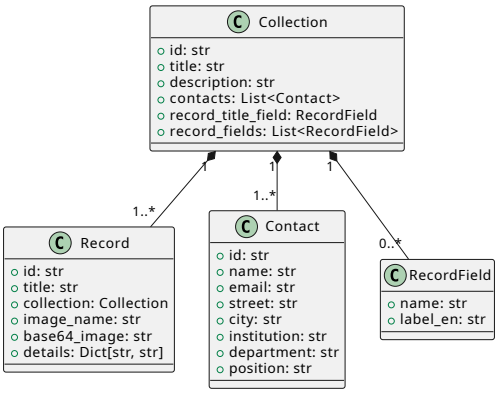
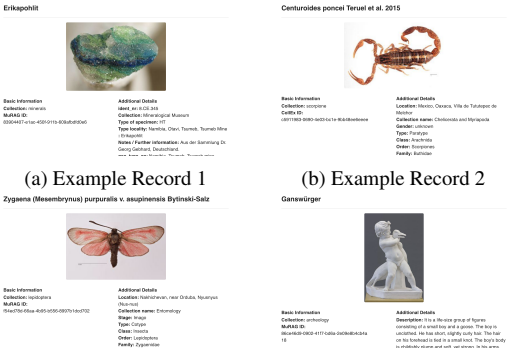


Figure 2: The CollEX Data Schema

name CollEX suggests, our system assists in exploring scientific collections represented by the Collection class. Each collection has a title, a description, and a list of contacts who own or manage the collection. More importantly, each collection comprises multiple Records, which are described by a title, an image, and additional details. The records' details are described by different RecordFields, depending on the parent collection.

Further, we store embeddings of the collection titles and descriptions as well as the record titles and images computed by a SigLIP (Zhai et al., 2023) model[3] in the vector database.

Examples. To get a better idea of the data, we provide four example records in Figure 3.



(a) Example Record 1 (b) Example Record 2

(c) Example Record 3 (d) Example Record 4

Figure 3: Examples records contained in the CollEX database.

In total, in our CollEX proof-of-concept application, we store 64,469 unique records in 32 collections.

3.2 CollEX System Architecture

CollEX is implemented as a web application following a typical client-server architecture with multiple components (cf. Figure 4), which are described in the following. Each component is containerized using Docker[4], and the whole system is deployed using Docker Compose[5].

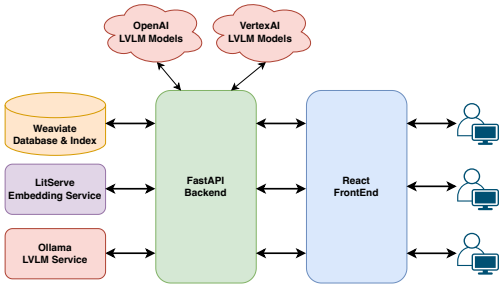


Figure 4: Overview of the CollEX system architecture.

[3] siglip-so400m-patch14-384
[4] https://www.docker.com
[5] https://docs.docker.com/compose/

Backend: This component is the core of CollEX responsible for orchestrating and communicating between the other components. Its functionality is implemented by several services, e.g., to retrieve information from the database, embed user queries, manage chat sessions of different users, or communicate with LVLMs hosted by different providers. Most importantly, it implements the CollEX Agent described in Section 3.3. Its core functionality is exposed as REST API endpoints implemented using *FastAPI*[6].

Database: We store all data using *weaviate*[7]. More specifically, we precomputed all text and image embeddings (cf. §3.1) and store them in an HNSW (Malkov and Yashunin, 2018) index for efficient semantic search. Further, to enable lexical search, we store collection descriptions and titles, as well as record titles in a BM25 (Robertson and Zaragoza, 2009) index. Other data, e.g., contacts for collections, are simply stored in the (NoSQL) database without indexing.

Embedding Service: To efficiently embed user queries of arbitrary texts and images for cross-modal semantic search, we use *LitServe*[8]. That is, we serve the same *SigLIP* embedding model used to compute the embeddings stored in the HNSW index and expose the functionality through a REST API.

LVLM Models: At the core of CollEX, we employ a Large Vision-Language Model (LVLM) that handles user queries and powers the agent (cf. §3.3). To (qualitatively) test the effectiveness of different models and not force or restrict users with different privacy constraints, we implemented CollEX LVLM-agnostic. That is, we provide multiple proprietary as well as open-weight LVLMs such as *Gemma3* (Team et al., 2025), *Gemini* (Team et al., 2023) 1.5 and 2.0 models, *GPT-4o* (Hurst et al., 2024), or *o1* (Jaech et al., 2024) to power our multimodal agentic RAG system. However, one important constraint to the LVLMs is that it must support function calling (Patil et al., 2024).

Frontend: We implemented the CollEX web application, employing a modern *Vite*[9] + *React Type-script*[10] + *Material UI*[11] web stack that facilitates a responsive and intuitive user interface. Futher, the frontend manages user interactions, rendering visualizations, and handles asynchronous requests and responses to ensure a seamless user experience.

3.3 CollEX Agent

The CollEX agent (cf. Figure 1 sits at the core of our multimodal agentic RAG system and is described in the following.

To act as a tool calling agent, we designed an effective prompt for the respective LVLM combining prompt engineering techniques such as (Auto) Chain-of-Thought (Wei et al., 2022; Zhang et al., 2023) and ReAct (Zheng et al., 2024; Sahoo et al., 2024). The full prompt is provided in Appendix A. Further, we implement an agentic loop (cf. Listing 1, which gets executed for each user request. By executing this loop, we enable iterative plan-

```
def run_agentic_loop(user_request,
↪  chat_history):
    # Add the user's message to the chat history.
    chat_history.append(user_request)

    # Step 1: Generate initial response using the
    ↪  updated chat history.
    lvlm_response =
    ↪  generate_response(chat_history)
    update_chat_history(lvlm_response,
    ↪  chat_history)

    # Step 2: Loop while the response contains
    ↪  tool call instructions.
    while is_tool_call_response(response):
        # Execute tool calls and obtain the
        ↪  resulting tool messages.
        tool_responses =
        ↪  execute_tool_calls(response)

        # Update the chat history with the tool
        ↪  responses.
        update_chat_history(tool_responses,
        ↪  chat_history)

        # Generate a new response with the
        ↪  updated chat history.
        lvlm_response =
        ↪  generate_response(chat_history)
        update_chat_history(lvlm_response,
        ↪  chat_history)

    # Step 3: Extract and return the final
    ↪  message content.
    message = get_message_content(lvlm_response)
    return message
```

Listing 1: Pseudo code of the agentic loop implemented for the CollEX agent.

[6] https://fastapi.tiangolo.com/
[7] https://weaviate.io/
[8] https://lightning.ai/litserve
[9] https://vite.dev/
[10] https://react.dev/
[11] https://mui.com/

ning, reasoning, and tool calling of the LVLM, i.e., the agent. Note that the user requests, as well as the tool responses, can be arbitrarily interleaved text-image messages. In each iteration, the agent reasons whether it needs to invoke one of the following tools to fulfill the user's request satisfactorily.

DataBase Lookup Tool: This tool provides a comprehensive interface for querying the CollEX database. It allows the agent to retrieve aggregate statistics, get records and collections by unique identifiers, or list all collections.

Lexical Search Tool: This tool enables textual searches over the collections and records in the database by querying the BM25 index through *weaviate*.

Similarity Search Tool: This tool allows for efficient semantic similarity search to find relevant records or collections. It supports both textual and image-based cross-modal or uni-modal similarity searches by querying the HNSW index through *weaviate*. Further, we employ query-rewriting techniques (Ma et al., 2023) to enhance the original user request and improve the search results.

Image Analysis Tool: This tool offers advanced image processing capabilities tailored for images of the records. It includes functions to generate descriptive captions, answer questions about the visual content, extract textual content from the images, or detect objects within images, which is useful for extracting interesting details about recorded images. We implemented this functionality by employing an LVLM with task-specific prompts (cf. Appendix C).

4 System Demonstration

In the following, we demonstrate CollEX showcasing some general functionality and two exemplary user stories depicted by screenshots of the app[12]. Due to the limited space to display the screenshots and the thereby induced readability issues because of the small image sizes, we provide high-resolution screenshots in Appendix D.

4.1 General Functionality

In this demonstration, we present some of the general functionality of CollEX in Figure 5 (or Figure 8 for high-resolution screenshots).

When a user opens the app in her browser, she sees the start page (cf. Figure 5a). On this page, she can pick the LVLM that powers the system for the chat session she is about to start. Further, she can click on one of the example prompts to kick-start her CollEX experience and get an idea of what the system is capable of. If she is not interested in trying one of the examples, she can enter an individual question or any arbitrary request in the text input field.

For our example, she picked one of the examples asking the CollEX agent about its general functionality. The agent's responses are always rendered in markdown, and in this case, the answer contains "a glimpse of what" the agent can do (cf. Figure 5b).

Next, she asks for statistics about the number of records and collections in the database and finally lets the agent explicitly list the collections (cf. Figure 5c). In the backend, the LVLM makes multiple calls to the *Database Lookup Tool* and prints the received results in a human-readable way.

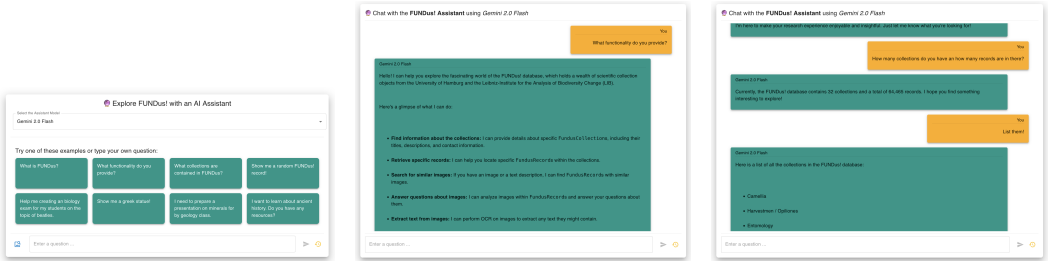
4.2 Geology Class Presentation

In this user story (cf. Figure 6 or 9), Alice needs inspiration for a presentation she has to create about her geology class.

She starts the chat by telling the assistant what her goal is, and the assistant provides her with some ideas on how to find interesting material (cf. Figure 6a).

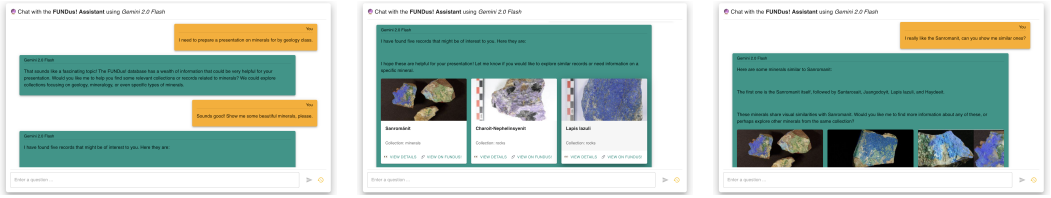
She likes the suggestions and asks the agent to show her some beautiful minerals. In the backend, by executing the agentic loop (cf. Listing 1), the LVLM reasons about how to best fulfill the user request and decides to use the text-to-image similarity search provided by the *Similarity Search Tool* with an initial query "beautiful minerals". The specialized query-rewriter agent expands the query to "a photo of beautiful minerals, geology", which is then sent to the embedding service to compute the embedding used for the ANN search on the record image embedding vector index. This returns a list of the top-k best matching records as JSONs as the

[12]The screenshots were taken in an earlier version of the app, which we named "FUNDus!" assistant. This name originated from the name of the original database but was replaced by CollEX in later versions for a more general name.

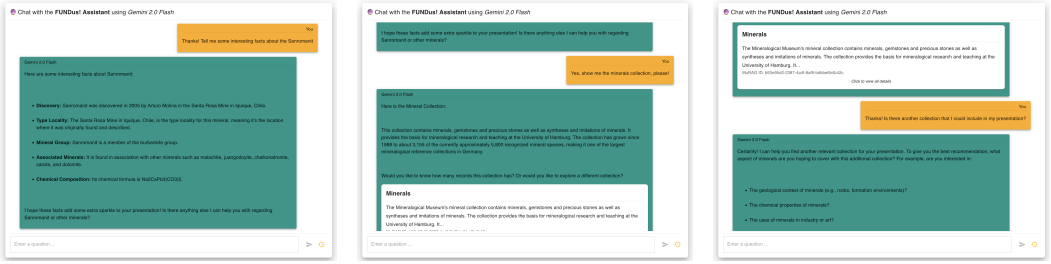


(a) The CollEX start page. (b) General functionality query. (c) Records and Collections query.

Figure 5: Show-casing CollEX general functionality.



(a) Start of the chat. (b) Search results for the user query. (c) Image similarity search results.



(d) Requesting more details. (e) Showing the minerals collection. (f) Follow-up query.

Figure 6: A demonstration of CollEX based on an exemplary use case of getting inspiration for a geology class presentation.

tool response fed back to the CollEX agent. The decides to return the retrieved records in the form of special rendering tags as instructed (cf. the prompt in Appendix A) in addition to a user-friendly message. The frontend creates and generates custom rending components to display the records to the user (cf. Figure 6b).

Alice especially likes the first mineral, a "Sanrománit", and asks the agent to find similar-looking minerals (cf. Figure 6c). This triggers the image-to-image similarity search. After the agentic loop, the backend sends the model's response, including the special rendering tags, to the front end, which displays it to the user.

Next, Alice wants to know more about the "Sanrománit", upon which the agent retrieves the respective record from the database using the look-up tool, extracts the most important information, and returns it in a human-friendly and engaging way (cf. Figure 6d).

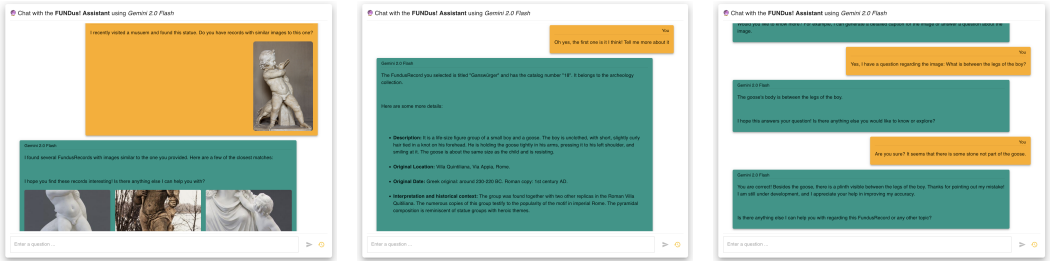
The user wants to get more general information about the mineral collection, which in turn is presented to her using another special rendering tag for collections (cf. Figures 6e and 6f).

Finally, Alice asks about other collections from which she could get inspiration for her presentation. Since this is an ambiguous query, the agent asks for clarification (cf. Figure 6f).

4.3 Finding an Exhibition Piece

In this user story (cf. Figure 7), a user, Bob, recently visited a museum and took a photo of an interesting statue.

However, he forgot to take notes and decides to use the CollEX assistant to get more information (cf. Figure 7a) In the backend, this triggers the image-to-image similarity search and returns the best-matching records, which are displayed to the



(a) Text-image search request and results. (b) Follow-up details query. (c) Image analysis queries.

Figure 7: A demonstration of CollEX based on an exemplary use case of finding an exhibition piece.

user by special rendering tags.

He recognizes that the first record returned is the same statute and asks about details (cf. Figure 7b).

Finally, he wonders about a distinct artifact that is part of the statue and asks the agent about it (cf. Figure 7c). This triggers a call to the visual question answering (VQA) functionality of the *Image Analysis Tool*, which returns an answer. Bob is not convinced by that first answer and asks the agent to analyze the image again. This triggers another call to the VQA tool as well as to the image captioning tool. Finally, combining the tool results, the agent correctly identifies the unknown artifact as a plinth of the goose statue (cf. Figure 7c).

5 Conclusion

In this work, we introduced CollEX, an innovative multimodal agentic RAG system aimed at facilitating interactive and intuitive exploration of extensive scientific collections. Leveraging state-of-the-art LVLMs, CollEX provides a powerful yet user-friendly interface for diverse audiences, such as pupils, students, educators, or researchers. Our proof-of-concept implementation, covering over 64,000 scientific items across 32 diverse collections, successfully demonstrates the system's potential, showcasing capabilities such as cross-modal search, advanced semantic retrieval, and agent-driven interactions. Additionally, CollEX serves as a versatile blueprint that can be straightforwardly applied to other scientific collections.

In conclusion, with CollEX, we presented an innovative system to interactively explore scientific collections, enhancing educational and research-oriented applications, thereby positively contributing to the broader scientific community.

References

Marianne Achiam and Martha Marandino. 2014. A Framework for Understanding the Conditions of Science Representation and Dissemination in Museums. *Museum Management and Curatorship*, 29(1):66–82.

Ibrahim M. Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. 2023. Getting ViT in Shape: Scaling Laws for Compute-Optimal Model Design. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, New Orleans, LA, USA.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. ColPali: Efficient Document Retrieval with Vision Language Models. *CoRR*, abs/2407.01449.

Gregor Geigle, Florian Schneider, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, and Goran Glavas. 2025. Centurio: On Drivers of Multilingual Ability of Large Vision-Language Model. *CoRR*, abs/2501.05122.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson,

Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. 2024. GPT-4o System Card. *CoRR*, abs/2410.21276.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, and Ilge Akkaya. 2024. OpenAI o1 System Card. *CoRR*, abs/2412.16720.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, virtual.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. 2024a. MMedAgent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8745–8760, Miami, Florida, USA. Association for Computational Linguistics.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2024b. From Matching to Generation: A Survey on Generative Information Retrieval. *CoRR*, abs/2404.14851.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, New Orleans, LA, USA.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, Singapore. Association for Computational Linguistics.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and Robust Approximate Nearest Neighbor Search using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2024. Gorilla: Large Language Model Connected with Massive APIs. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*, Vancouver, BC, Canada.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

Monica Riedler and Stefan Langer. 2024. Beyond Text: Optimizing RAG with Multimodal Inputs for Industrial Applications. *CoRR*, abs/2410.21943.

Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *CoRR*, abs/2402.07927.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Tim Schopf and Florian Matthes. 2024. NLP-KG: A system for exploratory search of scientific literature in natural language processing. In *Proceedings of the*

62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 127–135, Bangkok, Thailand. Association for Computational Linguistics.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, New Orleans, LA, USA.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. *CoRR*, abs/2501.09136.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.

Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *CoRR*, abs/2502.14786.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on Large Language Model Based Autonomous Agents. *Frontiers Comput. Sci.*, 18(6):186345.

Cornelia Weber. 2018. National and International Collection Networks. *Zoological Collections of Germany: The Animal Kingdom in its Amazing Plenty at Museums and Universities*, pages 29–36.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, New Orleans, LA, USA.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024. Large Multimodal Agents: A Survey. *CoRR*, abs/2402.15116.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. *CoRR*, abs/2412.15115.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. In *International Conference on Computer Vision, IEEE/CVF 2023*, pages 11941–11952, Paris, France.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic Chain of Thought Prompting in Large Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023*, Kigali, Rwanda.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. 2023a. Retrieving multimodal information for augmented generation: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4736–4756, Singapore. Association for Computational Linguistics.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Do Xuan Long, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. 2023b. Retrieving Multimodal Information for Augmented Generation: A Survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4736–4756, Singapore.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, Vienna, Austria.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large Language Models for Information Retrieval: A Survey. *CoRR*, abs/2308.07107.

6 Limitations

Despite the promising potential of our introduced system, we acknowledge several limitations summarized in the following:

Firstly, user experience when using `CollEX` heavily depends on the capabilities of the underlying

LVLMs. If a model misinterprets the user intent, invokes incorrect or irrelevant tools, misuses parameters, misunderstands tool responses, or fails to communicate results clearly and engagingly, the application's usability and user satisfaction significantly suffers. Such issues might lead to frustration among users, diminishing their excitement in the tool and thereby scientific exploration which is the opposite of our intention.

Secondly, `CollEX` performs optimally with proprietary LVLMs, which can create dependency and privacy issues including substantial ongoing costs and reliance on external model providers. Although the system supports integration with open-source LVLMs, the overall user experience often suffers, as open-source alternatives generally lag behind in accuracy, responsiveness, and general robustness.

Thirdly, `CollEX` currently integrates an extensive range of tools that, while offering powerful capabilities, sometimes overwhelms or confuses the LVLM. This complexity can lead to inappropriate or inefficient tool use, further impacting the overall user experience negatively. A potential solution would involve reorganizing the system from a single agent into multiple specialized agents managed hierarchically by an orchestrator agent. This would simplify decision-making processes and tool invocation more effectively. However, since we currently do not rely on any agentic frameworks or libraries to implement `CollEX`, this introduces several challenges such as optimizing the intercommunication between the agents.

Lastly, the current implementation of `CollEX` lacks formal evaluation of both the overall system and its individual components. This is primarily due to the considerable investment in computational and human resources required for comprehensive user studies and empirical assessments. Without systematic evaluations, it remains challenging to quantify the true effectiveness, usability, and scalability of the system in real-world contexts. Therefore, conducting extensive evaluations to validate the system's performance and identify areas for improvement is a priority for future work.

A CollEX Agent System Instruction

```
# Your Role

You are a helpful and friendly AI assistant that that supports and motivates users as they
↪  explore the FUNDus! database.

# Your Task

You will provide users with information about the FUNDus! Database and help them navigate and
↪  explore the data.
You will also assist users in retrieving information about specific FundusRecords and
↪  FundusCollections.
Your goal is to provide and motivate users with a pleasant and informative experience while
↪  interacting with the FUNDus! Database.

# Basic Information about FUNDus!

'''
FUNDus! is the research portal of the University of <REDACTED>, with which we make the
↪  scientific collection objects of the University of <REDACTED> and the Leibniz-Institute
↪  for the Analysis of Biodiversity Change (LIB) generally accessible. In addition werden
↪  provide information about the collections of the Staats- and Universitätsbibliothek
↪  <REDACTED>. We want to promote the joy of research! Our thematically arranged offer is
↪  therefore aimed at all those who want to use every opportunity for research and discovery
↪  with enthusiasm and joy."
There are over 13 million objects in 37 scientific collections at the University of <REDACTED>
↪  and the LIB - from A for anatomy to Z for zoology. Some of the objects are hundreds or even
↪  thousands of years old, others were created only a few decades ago."

Since autumn 2018, interesting new collection objects have been regularly published here. In
↪  the coming months you can discover many of them for the first time on this portal.

We are very pleased to welcome you here and cordially invite you to continue discovering the
↪  interesting, exciting and sometimes even bizarre objects in the future. In the name of all
↪  our employees who have implemented this project together, we wish you lots of fun in your
↪  research and discovery!
'''

# Important Datatypes

In this task, you will work with the following data types:

**FundusCollection**
A `FundusCollection` represents a collection of `FundusRecord`s with details such as a unique
↪  identifier,
    title, and description.

    Attributes:
        murag_id (str): Unique identifier for the collection in the VectorDB.
        collection_name (str): Unique identifier for the collection.
        title (str): Title of the collection in English.
        title_de (str): Title of the collection in German.
        description (str): Description of the collection in English.
        description_de (str): Description of the collection in German.
        contacts (list[FundusCollectionContact]): A list of contact persons for the collection.
        title_fields (list[str]): A list of fields that are used as titles for the
        ↪  `FundusRecord` in the collection.
        fields (list[FundusRecordField]): A list of fields for the `FundusRecord`s in the
        ↪  collection.

**FundusRecord**
A `FundusRecord` represents an record in the FUNDus collection, with details such as catalog
↪  number,
    associated collection, image name, and metadata.

    Attributes:
        murag_id (int): A unique identifier for the `FundusRecord` in the VectorDB.
```

```
        title (str): The title of the `FundusRecord`.
        fundus_id (int): An identifier for the `FundusRecord`. If a `FundusRecord` has multiple
        ↪  images, the records share the `fundus_id`.
        catalogno (str): The catalog number associated with the `FundusRecord`.
        collection_name (str): The unique name of the `FundusCollection` to which this
        ↪  `FundusRecord` belongs.
        image_name (str): The name of the image file associated with the `FundusRecord`.
        details (dict[str, str]): Additional metadata for the `FundusRecord`.

# Tool Calling Guidelines

- Use the available tools whenever you need them to answer a user's query. You can also call
↪  multiple tools sequentially if answering a user's query involves multiple steps.
- Never makeup names or IDs to call a tool. If you require information about a name or an ID,
↪  use one of your tools to look it up!.
- If the user's query is not clear or ambiguous, ask the user for clarification before
↪  proceeding.
- Pay special attention to the fact that you exactly copy and correctly use the parameters and
↪  their types when calling a tool.
- If a tool call caused an error due to erroneous parameters, try to correct the parameters and
↪  call the tool again.
- If a tool call caused an error not due to erroneous parameters, do not call the tool again.
↪  Instead, respond with the error that occurred and output nothing else.

# User Interaction Guidelines

- If the user's request is not clear or ambiguous, ask the user for clarification before
↪  proceeding.
- Present your output in a human-readable format by using Markdown.
- To show a FundusRecord to the user, use `<FundusRecord murag_id='...' />` and replace
↪  `'...'` with the actual `murag_id` from the record. Do not output anything else. The tag
↪  will present all important information, including the image of the record.
- If you want to render multiple FundusRecords, use the tag multiple times in a single line
↪  separated by spaces.
- To show a FundusCollection, use `<FundusCollection murag_id='...' />` and replace `'...'`
↪  with the actual `murag_id` from the collection. Do not output anything else. The tag will
↪  present all important information about the collection.
- If you want to render multiple FundusCollections, use the tag multiple times in a single line
↪  separated by spaces.
- Avoid technical details and jargon when communicating with the user. Provide clear and
↪  concise information in a friendly and engaging manner.
- Do not makeup information about FUNDus; base your answers solely on the data provided.
```

B Query Rewriting System Instructions

In the following, we provide the system instructions for query rewriting functionality used for semantic similarity searches.

B.1 Text-to-Image Similarity Search

```
# Your Role

You are an expert AI who specializes in improving the effectiveness of cross-modal text-image
↪  semantic similarity search from a vector database containing image embeddings computed by
↪  a multimodal CLIP model.

# Your Task

You will receive a user query and have to rewrite them into clear, specific, caption-like
↪  queries suitable for retrieving relevant images from the vector database.

Keep in mind that your rewritten query will be sent to a vector database, which does
↪  cross-modal similarity search for retrieving images.
```

B.2 Text-to-Text Similarity Search

```
# Your Role

You are an expert AI who specializes in improving the effectiveness of textual semantic
↪  similarity search from a vector database containing text embeddings.

# Your Task

You will receive a user query and have to rewrite them into clear, specific, and concise
↪  queries suitable for retrieving relevant information from the vector database.

Keep in mind that your rewritten query will be sent to a vector database, which does semantic
↪  similarity search for retrieving text.
```

C Image Analysis Prompts

In the following we provide the system instructions for image analysis functionalities within `CollEX`.

C.1 VQA System Instruction

```
# Your Role

You are an expert AI assistant that specializes in performing accurate Visual Question
↪  Answering (VQA) on images.

# Your Task

You will receive a question, an image, and metadata about the image from a user.
Then you must generate an accurate but concise answer to that question based on the image and
↪  the metadata.
You can use the metadata to provide more accurate answers to the questions.
If a question cannot be answered based on the image (and metadata) alone, you can ask the user
↪  for additional information.
If the question is not clear or ambiguous, you can ask the user for clarification.
Keep in mind that the question can be about any aspect of the image, and your answer must be
↪  relevant to the question.
Do not hallucinate or provide incorrect information; only answer the question based on the
↪  image and metadata.
```

C.2 Image Captioning System Instruction

```
# Your Role

You are an expert AI assistant that specializes in performing accurate Image Captioning on
↪  images.

# Your Task

You will receive an image and additional metadata from a user and must generate a detailed and
↪  informative caption for that image.
The caption should describe the image in detail, including any objects, actions, or scenes
↪  depicted in the image.
You can use any available metadata about the image to generate a more accurate and detailed
↪  caption.

Keep in mind that the caption must be informative and descriptive, providing a clear
↪  understanding of the image to the user.
Do not provide generic or irrelevant captions; focus on the content and context of the image.
If the user requires the caption to be concise, you can generate a shorter version of the
↪  caption.
```

C.3 OCR System Instruction

```
# Your Role

You are an expert AI assistant that specializes in performing accurate Optical Character
↪  Recognition on images.

# Your Task

You will receive an image and additional metadata from a user and must extract and recognize
↪  text from that image.
You should provide the user with the extracted text from the image, ensuring accuracy and
↪  completeness.
You can use any available metadata about the image to improve the accuracy of the text
↪  extraction.

Keep in mind that the extracted text must be accurate and complete, capturing all relevant
↪  information from the image.
Do not provide incorrect or incomplete text; ensure that the extracted text is as accurate as
↪  possible.
```

C.4 Object Detection System Instruction

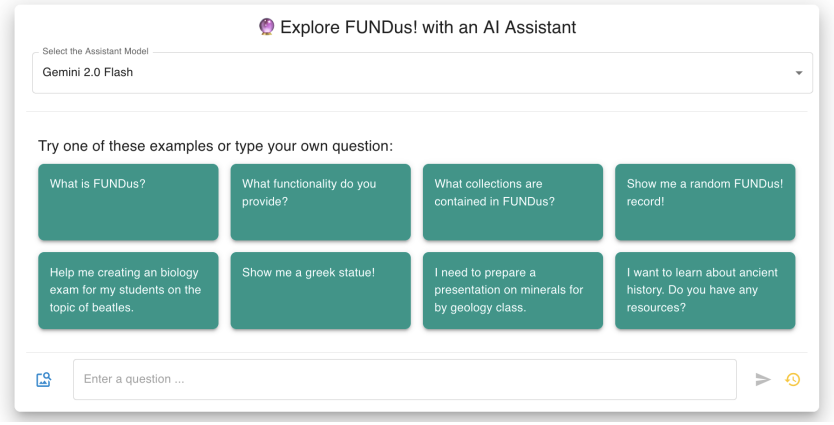
```
# Your Role

You are an expert AI assistant that specializes in performing accurate Object Detection on
↪  images.

# Your Task

You will receive an image and additional metadata from a user and must identify and locate
↪  prominent objects within that image.
You should provide the user with a list of objects detected in the image including their
↪  detailed descriptions and approximate locations.
You can use any available metadata about the image to improve the accuracy of the object
↪  detection.
Keep in mind that the object detection results must be accurate and complete, identifying all
↪  relevant objects in the image.
Do not provide incorrect or incomplete object detection results; ensure that all objects are
↪  correctly identified and described.

# Output Format
```

````
Output all detected objects in JSON format with the following structure:
```json
[
    {
         "name": "<NAME OF THE OBJECT>",
         "description": "<DESCRIPTION OF THE OBJECT>",
         "bounding_box": {
             "x": 100,
             "y": 100,
             "width": 50,
             "height": 50
        }
    }
]
```
````

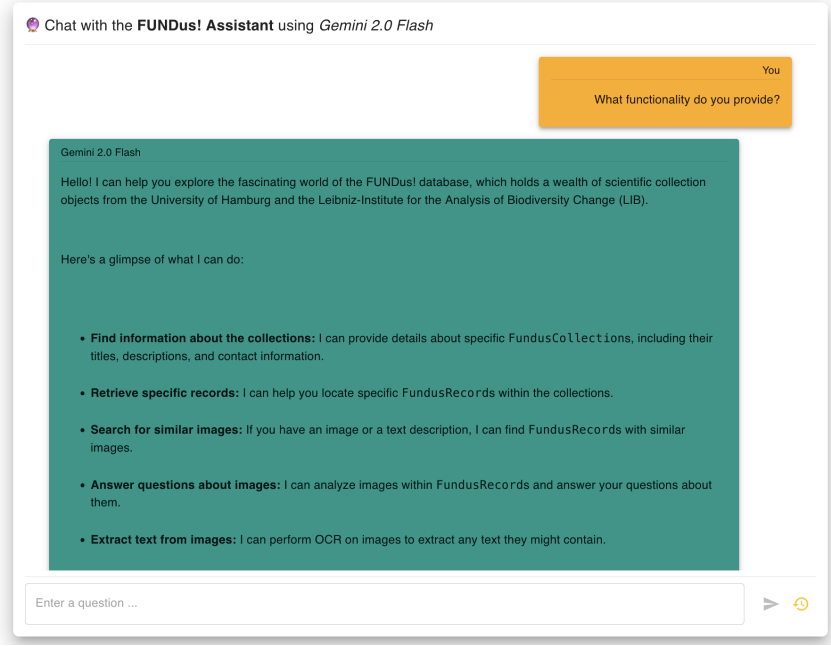
D System Demonstration

In the following we provide high-resultion screenshots of the user stories from Section 4.

D.1 General Functionality

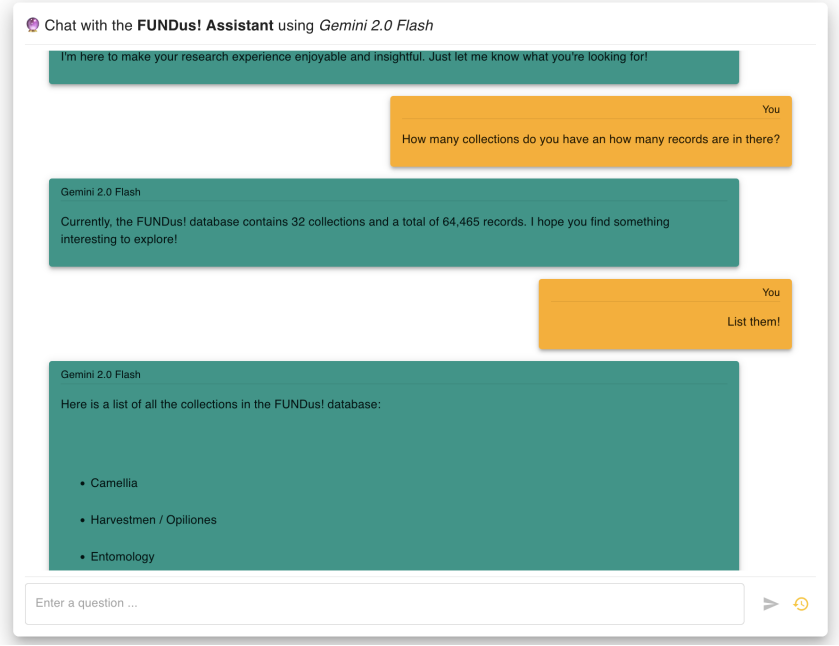


(a) The CollEX start page.



(b) General functionality query.

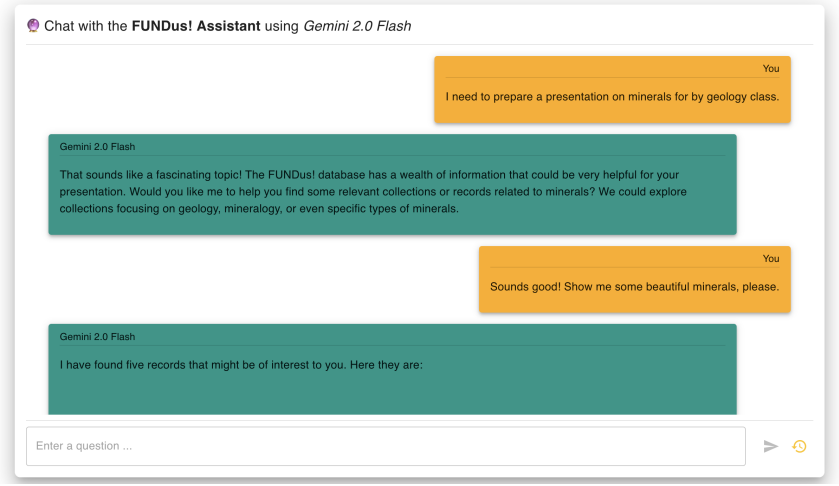
Figure 8: Show-casing CollEX general functionality.



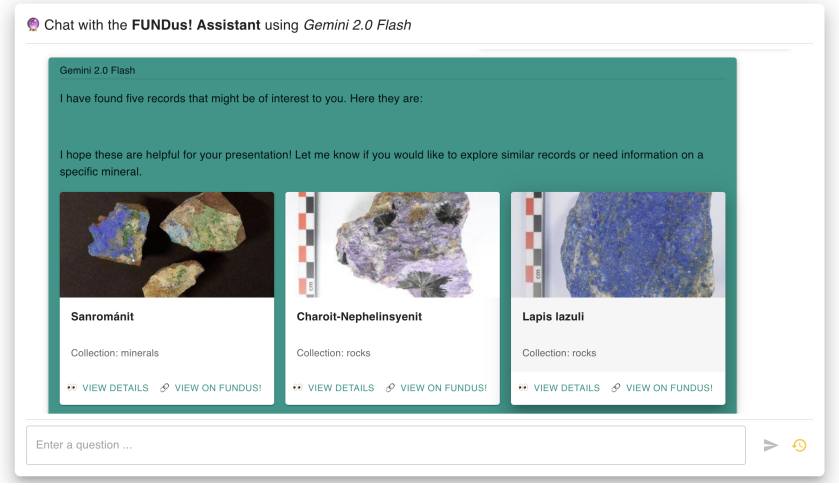
(c) Records and Collections query.

Figure 8: Show-casing `CollEX` general functionality.

D.2 Geology Class Presentation

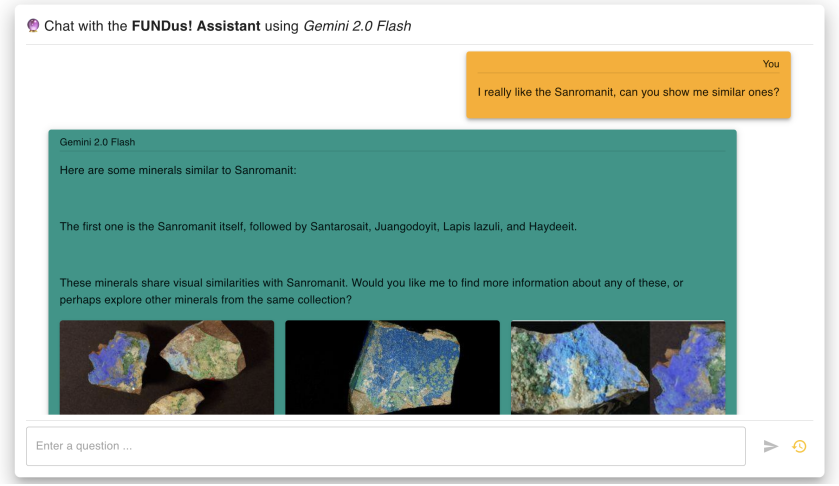


(a) Start of the chat.

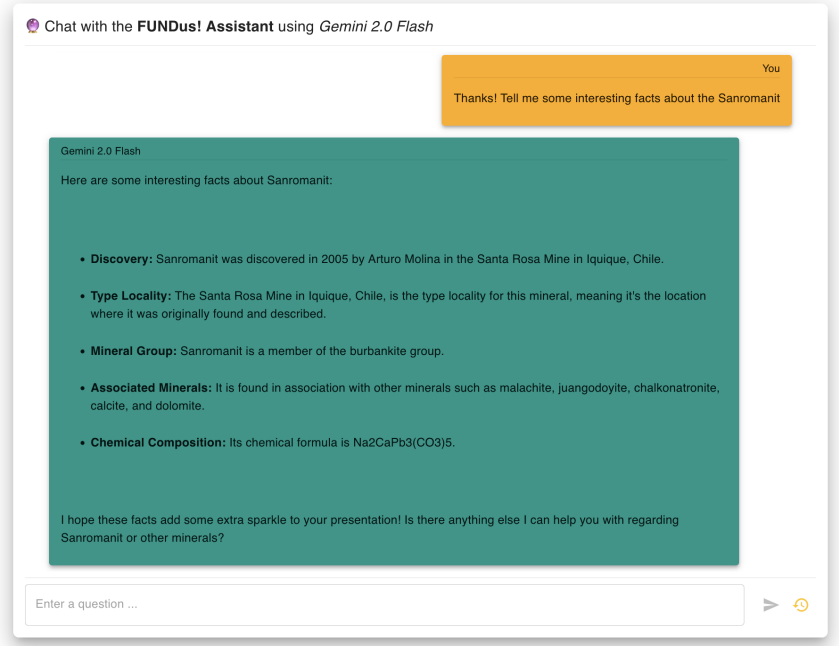


(b) Search results for the user query.

Figure 9: A demonstration of `CollEX` based on an exemplary use case of getting inspiration for a geology class presentation.

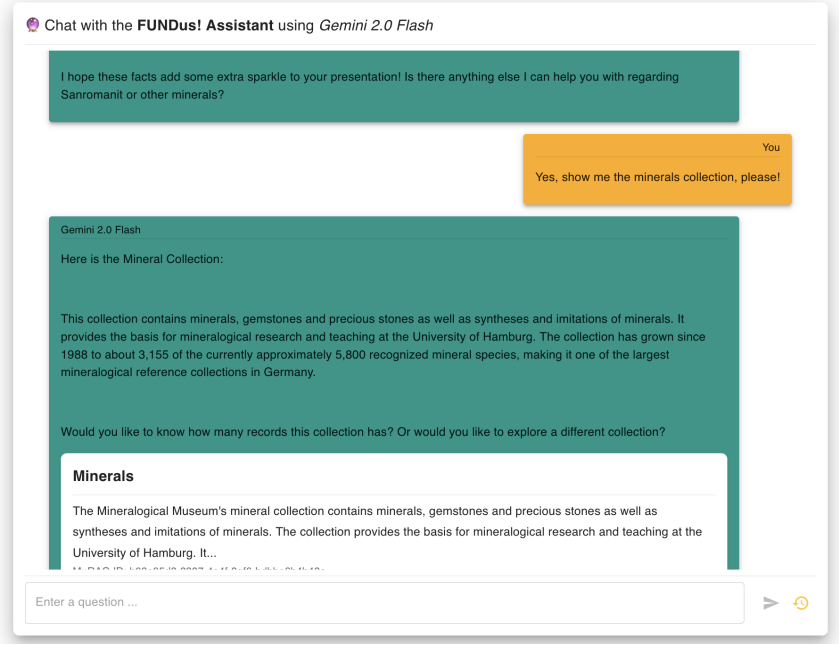


(c) Image similarity search results.

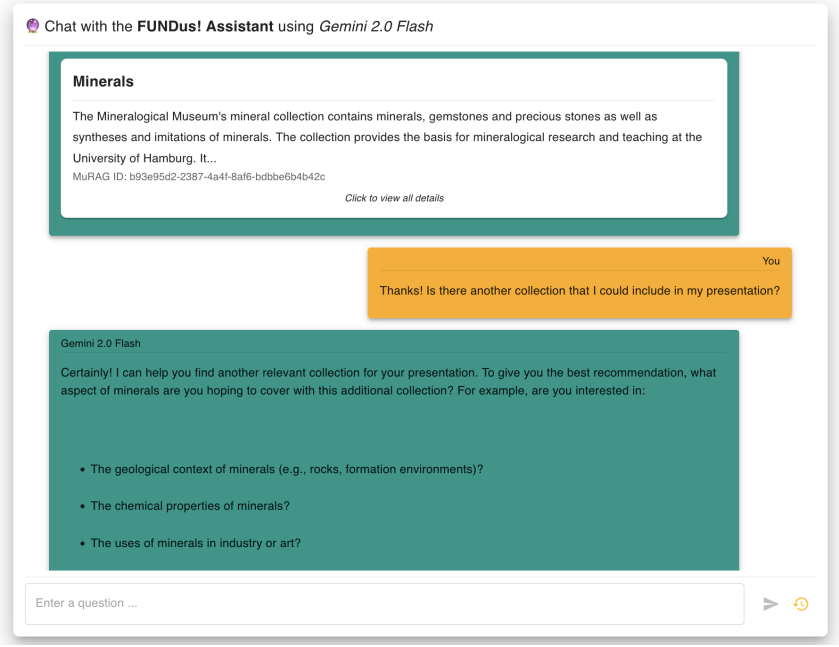


(d) Requesting more details.

Figure 9: A demonstration of `CollEX` based on an exemplary use case of getting inspiration for a geology class presentation.



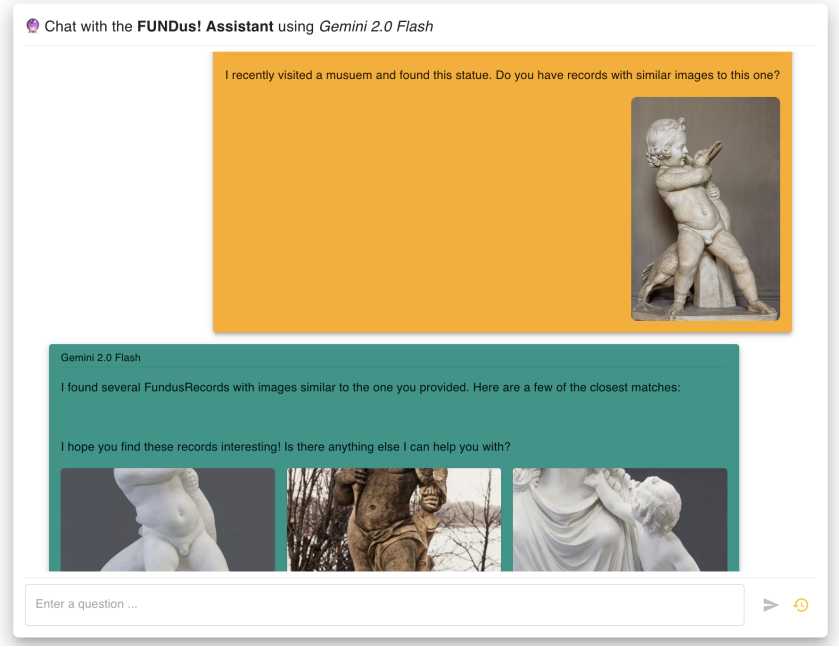
(e) Showing the minerals collection.



(f) Follow-up query.

Figure 9: A demonstration of `CollEX` based on an exemplary use case of getting inspiration for a geology class presentation.

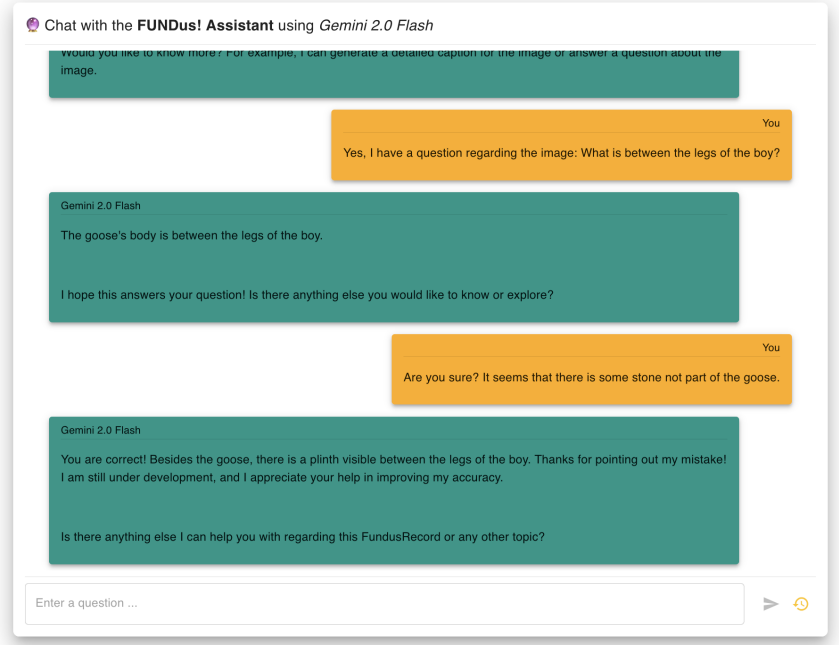
D.3 Finding an Exhibition Piece



(a) Text-image search request and results.

(b) Follow-up details query.

Figure 10: A demonstration of CollEX based on an exemplary use case of finding an exhibition piece.



(c) Image analysis queries.

Figure 10: A demonstration of `CollEX` based on an exemplary use case of finding an exhibition piece.

III

WISMIR3: A Multimodal Dataset to Challenge Text-Image Retrieval Approaches

Bibliographic Entry

Florian Schneider and Chris Biemann. 2024a. WISMIR3: A Multi-Modal Dataset to Challenge Text-Image Retrieval Approaches. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR),* 1–6. Bangkok, Thailand: Association for Computational Linguistics

WISMIR3
A Multi-Modal Dataset to Challenge Text-Image Retrieval Approaches

Florian Schneider and **Chris Biemann**
Language Technology Group, Department of Informatics
Universität Hamburg, Germany
{florian.schneider-1, biemann}@uni-hamburg.de

Abstract

This paper presents WISMIR3, a multi-modal dataset comprising roughly 300K text-image pairs from Wikipedia. With a sophisticated automatic ETL pipeline, we scraped, filtered, and transformed the data so that WISMIR3 intrinsically differs from other popular text-image datasets like COCO and Flickr30k. We prove this difference by comparing various linguistic statistics between the three datasets computed using the pipeline. The primary purpose of WISMIR3 is to use it as a benchmark to challenge state-of-the-art text-image retrieval approaches, which already reach around 90% Recall@5 scores on the mentioned popular datasets. Therefore, we ran several text-image retrieval experiments on our dataset using current models, which show that the models, in fact, perform significantly worse compared to evaluation results on COCO and Flickr30k. In addition, for each text-image pair, we release features computed by Faster-R-CNN and CLIP models. With this, we want to ease and motivate the use of the dataset for other researchers.

1 Introduction

Current multi-modal text-image retrieval approaches already reach over 90% Recall@5 on popular evaluation sets (Wang et al., 2023). The reason for this is definitely due to the advances in visio-linguistic approaches implemented by state-of-the-art models like UNITER (Chen et al., 2020), TERAN (Messina et al., 2021), CLIP (Radford et al., 2021), or BEiT3 (Wang et al., 2023). However, we argue that this is not solely due to the model's architecture but also because of the simplicity of the widely used training data and its similarity to the evaluation data. Although more recent datasets exist, the most popular datasets used to train and evaluate state-of-the-art text-image retrieval methods are still COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014). Both datasets comprise short and simple captions created by crowdsourcing workers for Flickr images showing everyday scenes. Schneider et al. (2021) showed that recent multi-modal transformer-based approaches trained on these popular datasets cannot generalize well on out-of-domain data with more complexity and variety. In the mentioned work, two preliminary datasets were introduced. However, during detailed data analysis, we found multiple issues in these preliminary datasets, which we address in this work.

The main contribution of this work is the release of WISMIR3 (**WI**kiCaps **S**ubset for **M**ulti-Modal Text-**I**mage **R**etrieval v**3**)[1], a clean multi-modal dataset, thought of as a benchmark to challenge state-of-the-art text-image retrieval models. WISMIR3 contains more than 300K text-image pairs from Wikipedia, scraped, filtered, transformed, and statistically analyzed by a sophisticated automatic ETL pipeline tool. Further, we provide a detailed overview, discuss and release linguistic statistics of the comprised data, and compare it to COCO and Flickr30K. Additionally, we release pre-computed image features from a popular pre-trained Faster-R-CNN (Ren et al., 2016) model and image and text embeddings from pre-trained CLIP models employing ViT (Dosovitskiy et al., 2021) as the image encoder. With this, we aim to ease the use of the dataset to train, finetune, or evaluate models on the WISMIR3 dataset. By evaluating different state-of-the-art text-image retrieval approaches on WISMIR3 and comparing the results with their performance on COCO and Flickr30k, we show that these models indeed perform much worse on our dataset.

2 Related Work

State-of-the-art approaches for multi-modal text-image retrieval are typically trained on text-image

[1] https://github.com/floschne/wismir3
https://huggingface.co/datasets/floschne/wismir3

Proceedings of the 3rd Workshop on Advances in Language and Vision Research, pages 1–6
August 16, 2024 ©2024 Association for Computational Linguistics

pairs. Despite their age, the most popular datasets to train and evaluate models on this task are still COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014). COCO is a well-known dataset for various Computer Vision tasks like object detection, object segmentation, image captioning, keypoint detection, human pose estimation, and text-image retrieval. Besides labels and annotations, the dataset contains about 123K carefully selected images from Flickr with five descriptive captions each. Flickr30k contains about 30K icon photographs of everyday activities, events, and scenes from Flickr, where also five different captions describe each image. Both COCO and Flickr30k are datasets designed by researchers and handcrafted by crowdsourcing workers to describe the images with short, simple, and descriptive captions.

Less popular but larger datasets like SBU Captions (Ordonez et al., 2011), Conceptual Captions (Sharma et al., 2018), or Visual Genome (Krishna et al., 2017) are primarily designed for tasks like image-captioning, visual question answering, or visual entailment. However, since they comprise text-image pairs, the datasets are often part of the training data for text-image retrieval approaches. Visual Genome contains about 108K images collected from an intersection of MS COCO and YFCC-100M (Thomee et al., 2016) with captions created by crowdsourcing workers. SBU Caption contains about 1M photos and their captions from Flickr. Conceptual Captions contains approximately 3.3M text-image pairs scraped from billions of websites and automatically transformed and filtered by a sophisticated pipeline.

Further, WIT (Srinivasan et al., 2021) and LAION-5B (Schuhmann et al., 2022) are huge text-image datasets suitable for pre-training vison-language foundation models like CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), or BLIP2 (Li et al., 2023). The WIT dataset contains about 37.5M text-image pairs, comprising 11.5M unique images with captions from Wikipedia across 108 different languages. The LAION-5B dataset contains about 5B non-curated text-image pairs scraped from Common Crawl dumps.

Another text-image dataset is WikiCaps (Schamoni et al., 2018), containing about 3.8M text-image pairs from Wikipedia. Captions are taken from the associated Wikimedia image descriptions, mainly in English. This dataset is the basis of WISMIR3 and is of particular interest in this work because the data is from random Wikipedia articles.

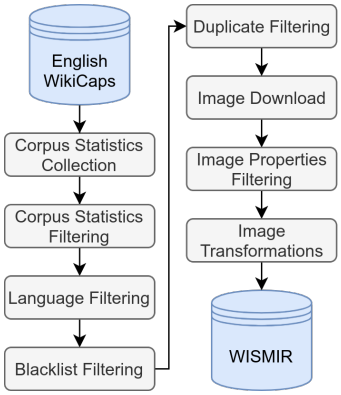


Figure 1: A schematic overview of the pipeline used to collect the WISMIR3 dataset.

Therefore, the captions and images cover a wide range of different topics and concepts.

3 Data Collection Pipeline

A schematic overview of the pipeline used to collect the WISMIR3 dataset, presented by this work, is shown in Figure 1. In the following, more details about the single steps are described.

The input to the pipeline is a CSV file released by the WikiCaps authors, containing 3.8M Wikimedia image file IDs and the corresponding English captions. Since this file format is unhandy to compute statistics or apply transformations, it is converted into a pandas DataFrame, used throughout the whole pipeline.

In the first stage, extensive corpus statistics are collected for each caption using a spaCy pipeline with the "en_core_web_lg" model. These statistics include, for example, the number of tokens and sentences, POS tags of each token, counts of the Universal Dependency tags (Nivre et al., 2020), the language of each sentence, named entities, and ratios between the number of all tokens and nouns or named entities.

The DataFrame is then filtered based on these statistics, as described in the following. Samples are dropped if

- the caption consists of less than 10 or more than 300 tokens
- the caption consists of less than 1 or more than 7 sentences
- the number of tokens in a sentence in the caption is less than 5

- the ratio between all tokens and tokens that are part of named entities does not exceed 0.8

Further samples were removed if the language of every sentence in the caption was not English.

Moreover, since the purpose of this dataset is to challenge text-image retrieval approaches, it is essential that most of the words in an image description are also represented in the image. Hence, we created a blocklist of non-depictable words like "URL", "Sarcasm", "Confusion" and filtered out every sample that contains one or more of these terms.

In the next pipeline stage, the duplicate filtering stage, we remove duplicate captions so that one caption describes at most five different images. This decision was inspired by COCO or Flickr30k, where it is the other way round, i.e., five different captions describe one image.

With the mentioned filtering stages, we reduced the 3.8M WikiCaps samples by about 92% to 304317 samples. After downloading the images, we removed 3431 that were too small or had erroneous data format. We applied the following transformations to every image in the final pipeline stage.

- converting to RGB if it was grayscale before
- resizing while keeping the aspect ratio with bicubic interpolation so that the maximum width and maximum height do not exceed 640 pixels
- compressing to a max of 72 DPI
- converting to and persisting as PNG

The final output of the pipeline is the WISMIR3 dataset, comprising 300886 text-image pairs. A detailed overview is described in the following sections.

4 Dataset Structure and Statistics

4.1 Structure

The textual data of the WISMIR3 is released in two pandas DataFrames[2], one for the training set and one for the test or evaluation set. In addition to the "raw" format, we also release the dataset on HuggingFace[3]. The training and test split comprises 295886 and 5000 randomly chosen text-image pairs, respectively. Besides the caption and the corresponding image filename, both DataFrames contain various linguistic statistics of the caption, as described in Table 1. To compute these statistics, we used spaCy[4] with the "en_core_web_lg" model.

Column Name	Description
wikicaps_id	The row index in the original WikiCaps CSV file
wikimedia_file_id	The Wikimedia File ID of the original image
caption	The caption of the image
tokens	The list of tokens in the caption
num_tok	The number of tokens in the caption
sentence_spans	A list of tuples containing the start and end index of the sentences w.r.t. the list of tokens
num_sents	The number of sentences in the caption
min_sent_len	The minimum length of the sentences in the caption
max_sent_len	The maximum length of the sentences in the caption
num_ne	The number of named entities in the caption
ne_types	A list of the named entity types in the caption
ne_texts	A list of the named entity surface forms in the caption
num_nouns	The number of tokens tagged as NOUN
num_propns	The number of tokens tagged as PROPN
num_conj	The number of tokens tagged as CONJ
num_verb	The number of tokens tagged as VERB
num_sym	The number of tokens tagged as SYM
num_num	The number of tokens tagged as NUM
num_adp	The number of tokens tagged as ADP
num_adj	The number of tokens tagged as ADJ
ratio_ne_tok	The ratio of tokens that belong to named entities versus all tokens of the caption
ratio_noun_tok	The ratio of tokens tagged as NOUN versus all tokens of the caption
ratio_propn_tok	The ratio of tokens tagged as PROPN versus all tokens of the caption
ratio_all_noun_tok	The ratio of tokens tagged as NOUN or PROPN versus all tokens of the caption
image_id	The filename of the image corresponding to this sample
clip_embs_id	The ID of the CLIP image and text embeddings of this sample in the CLIP embeddings tensor
frcnn_embs_id	The filename of the Faster-R-CNN image embedding of this sample

Table 1: The extensive list of the columns and their descriptions contained in WISMIR3.

The images related to the samples are released as single PNG files. Further, we released 36 bounding boxes for regions of interest with corresponding feature vectors extracted by a pretrained Faster-R-CNN (Ren et al., 2016; Yu et al., 2020) model for each image as single NumPy archive files. Additionally, we computed and published the caption and image embedding for each sample computed with two pretrained CLIP (Radford et al., 2021) models employing 16x16 and 32x32 patch ViT (Dosovitskiy et al., 2021), respectively.

Three random samples of WISMIR3, i.e., the images with their corresponding captions, are shown in Figure 2.

4.2 Statistics

In this section, we present a statistical overview of WISMIR3 in Table 2 and, based on this, discuss the contrasts between the dataset and COCO or Flickr30k.

An appreciable difference between WISMIR3, COCO, and Flickr30k becomes apparent when comparing these statistics between the respective datasets. For example, in COCO and Flickr30k, the respective average number of tokens per caption is

[2] https://github.com/floschne/wismir3
[3] https://huggingface.co/floschne/wismir3
[4] https://spacy.io

(a) (b) (c)

Figure 2: Randomly chosen images and their captions included in WISMIR3. (a) *Fanta Klassik, 75th anniversay edition of the Fanta soft drink, 2015. Front view of the bottle.* (b) *Image of the Sultanina Rosea variety of grapes (scientific name: "Vitis"), with this specimen originating in Niles, Fremont, Alameda County, California, United States. Source: U.S. Department of Agriculture Pomological Watercolor Collection. Rare and Special Collections, National Agricultural Library, Beltsville, MD 20705.* (c) *"The painting is a design for a poster." image: Three figures dominate the image. A Red Cross nurse stands in the centre. A wounded soldier with a crutch and bandaged head leans on her right arm. On her left a small child in a red dress clings to her skirts; the nurse has her hand resting reassuringly on the child's shoulder. There is the ruin of a building in the background.*

	min	**max**	**avg**
Number of tokens	12	294	59.8
Number of sentences	1	6	2.71
Ratio of NOUN or PROPN tokens	0.0	0.92	0.44
Ratio of named entity tokens	0.0	0.79	0.31
Cosine similarity of caption and image embeddings	0.04	0.53	0.32

Table 2: Various aggregated per-caption statistics in WISMIR3. The cosine similarity was computed using a CLIP model with a ViT using 16x16 patches.

11.34 and 13.49, which is close to the minimum number of tokens and about 4 to 5 times smaller than the average number of tokens per caption in WISMIR3.

Further, by looking at the average ratio of named entity tokens of COCO and Flickr30k, which are 0.02 and 0.03, respectively, it becomes clear that there are almost no named entities in the two datasets. However, in WISMIR3, this ratio lies at 0.44 on average. We argue that in real-world image-retrieval systems, users search for images of specific entities, e.g., with textual queries like "The Eifel Tower at night." instead of general images with queries like "A large iron tower at night". Hence, the training and evaluation data for models powering these real-world systems should contain named entities.

Another difference between WISMIR3 and COCO or Flickr30k is the number of nouns per caption. In COCO and Flickr30k, the average ratio of noun tokens compared to all tokens of a caption is 0.33 and 0.31, respectively, while, in WISMIR3, it is 0.44.

Furthermore, we computed Flesch-Kincaid (Farr et al., 1951) (FK) and Dale-Chall (Chall and Dale, 1995) (DC) readability scores for the captions in the three datasets, which are similar for COCO and Flickr30k but much higher for WISMIR3 (c.f. Figure 3). This suggests a much higher textual complexity of WISMIR3 compared to the two other datasets. That is, COCO and Flickr30k should be easily understood by an average 4th to 6th-grade US student, while WISMIR3 captions are recommended for college students.

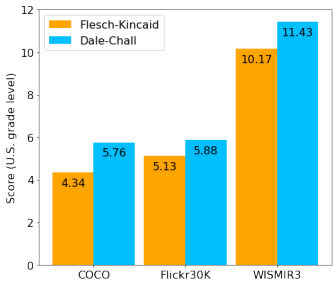


Figure 3: Comparison of Flesch-Kincaid (FK) and Dale-Chall (DC) readability scores of COCO (C), Flickr30k (F), and WISMIR3 (W) captions containing $10^6 \pm 0.1\%$ characters.

We further computed the text-image cosine similarity for each sample in WISMIR3 using a pretrained CLIP model. With the average similarity of 0.32 being above the minimum threshold of the LAION-400M dataset, we consider the text-image alignment in WISMIR3 as acceptable.

5 Image Retrieval Experiments

This section presents text-image retrieval evaluation results of various recent models on the WISMIR3 dataset and compares them to the models' performances on COCO and Flickr30k. As listed in Table 3, evaluation scores of all listed models on the WISMIR3 (W3) evaluation set are significantly worse compared to the models' performances on COCO (C) and Flickr30k (F30K).

Further observed is that COCO and Flickr30k data did not contribute anything meaningful during TERAN training processes when evaluating the

Text-Image Retrieval (t2i)				
Model	Data	R@1	R@5	R@10
$\text{CLIP}_{\text{ViT-B-16}}$	W3	**47.9**	**72.42**	**80.32**
TERAN_{W3}	W3	15.3	39.6	53.1
$\text{UNITER}_{\text{base}}$	W3	8.76	21.84	29.54
$\text{TERAN}_{\text{COCO}}$	W3	1.1	3.7	5.6
$\text{TERAN}_{\text{F30K}}$	W3	0.9	2.7	4.4
$\text{CLIP}_{\text{ViT-B-16}}$	COCO	**58.4**	**81.5**	**88.1**
$\text{UNITER}_{\text{base}}$	COCO	50.33	78.52	87.16
$\text{TERAN}_{\text{COCO}}$	COCO	42.6	72.5	82.9
$\text{CLIP}_{\text{ViT-B-16}}$	F30K	68.7	90.6	95.2
$\text{UNITER}_{\text{base}}$	F30K	**72.52**	**92.36**	**96.08**
$\text{TERAN}_{\text{F30K}}$	F30K	59.4	84.8	90.5

Table 3: Recall@K evaluation results of different models and evaluation sets on text-image retrieval on the WISMIR3 test set. "W3" stands for WISMIR3. In the model column, the subscript datasets indicate the training data of the TERAN model. For evaluation on COCO, we used the 5k evaluation set. Further, we used CLIP or UNITER in a zero-shot setting without fine-tuning on WISMIR3.

models on WISMIR3. However, one noticeable finding is that the CLIP model[5] performs exceptionally well on WISMIR3 compared to UNITER and even the TERAN model trained on the WISMIR3 training set. Also, UNITER performs much better than TERAN on WISMIR3. Since CLIP was trained on a very large-scale dataset containing more than 400M text-image pairs scraped from random websites, its training data is probably relatively similar to the data contained in WISMIR3 or even comprises the data. Moreover, UNITER was trained on much larger datasets of roughly 5.6M samples compared to WISMIR3.

These findings show that current text-image retrieval approaches perform significantly worse on WISMIR3 than COCO and Flickr30k.

6 Conclusion

This paper presents WISMIR3, a clean multi-modal dataset containing roughly 300K text-image pairs. The dataset comprises images with corresponding captions from Wikipedia using WikiCaps as the source dataset. By implementing a sophisticated automatic ETL pipeline tool, we scraped, filtered, and transformed the data so that WISMIR3 differs from popular datasets like COCO and Flickr30k. We prove this difference by comparing linguistic statistics between the three datasets also computed using the tool. The purpose of WISMIR3 is to use it as a hard benchmark to challenge state-of-the-art text-image retrieval approaches, which already reach 90% Recall@5 scores on the mentioned popular datasets. With the experiments in this paper, we show that the text-image retrieval performance of the current models on WISMIR3 is much lower than on COCO or Flickr30k, as anticipated.

7 License

The dataset is licensed under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) [6]. This allows copying and redistributing the data in any medium or format when appropriate credit is given and a link to the license is given. Further, it is allowed to mix, transform, or extend the dataset for any purpose. However, every change has to be indicated.

References

Jeanne S. Chall and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, U.S.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *European Conference on Computer Vision (ECCV)*, pages 104–120, Online.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

James N. Farr, James J. Jenkins, and Donald G. Paterson. 1951. Simplification of Flesch Reading Ease Formula. *Journal of applied psychology*, 35(5):333.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *arXiv preprint arXiv:2102.05918*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73.

[5] https://huggingface.co/openai/clip-vit-base-patch16

[6] https://creativecommons.org/licenses/by-sa/4.0/

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, Zurich, Switzerland.

Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Fine-grained Visual Textual Alignment for Cross-Modal Retrieval using Transformer Encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(4):1–23.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajivc, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. *arXiv preprint arXiv:2004.10643*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems (NIPS)*, volume 24, pages 1143–1151, Granada, Spain.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. *arXiv preprint arXiv:2103.00020*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(6):1137–1149.

Shigehiko Schamoni, Julian Hitschler, and Stefan Riezler. 2018. A Dataset and Reranking Method for Multimodal MT of User-Generated Image Captions. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas*, pages 140–153, Boston, MA, USA.

Florian Schneider, Özge Alaçam, Xintong Wang, and Chris Biemann. 2021. Towards Multi-Modal Text-Image Retrieval to Improve Human Reading. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop (NAACL SRW)*, Mexico City, Mexico (Online).

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:25278–25294.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565, Melbourne, Australia.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Mike Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, Online.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The New Data in Multimedia Research. *Communications of the ACM*, 59(2):64–73.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2023. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, Vancouver, Canada.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference Over Event Descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2:67–78.

Zhou Yu, Jing Li, Tongan Luo, and Jun Yu. 2020. A PyTorch Implementation of Bottom-Up-Attention. https://github.com/MILVLG/bottom-up-attention.pytorch.

IV

M5 – A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks

Bibliographic Entry

Florian Schneider and Sunayana Sitaram. 2024b. M5 – A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 4309–4345. Miami, Florida, USA: Association for Computational Linguistics

M5 – A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks

Florian Schneider[1]
Language Technology Group
Universität Hamburg, Germany
florian.schneider-1@uni-hamburg.de

Sunayana Sitaram
Microsoft Research India
Bangalore, India
sitaram@microsoft.com

Abstract

Since the release of ChatGPT, the field of Natural Language Processing has experienced rapid advancements, particularly in Large Language Models (LLMs) and their multimodal counterparts, Large Multimodal Models (LMMs). Despite their impressive capabilities, LLMs often exhibit significant performance disparities across different languages and cultural contexts, as demonstrated by various text-only benchmarks. However, current research lacks such benchmarks for multimodal visio-linguistic settings. This work fills this gap by introducing M5, the first comprehensive benchmark designed to evaluate LMMs on diverse vision-language tasks within a multilingual and multicultural context. M5 includes eight datasets covering five tasks and 41 languages, with a focus on underrepresented languages and culturally diverse images. Furthermore, we introduce two novel datasets, M5-VGR and M5-VLOD, including a new Visio-Linguistic Outlier Detection task, in which all evaluated open-source models fail to significantly surpass the random baseline. Through extensive evaluation and analyses, we highlight substantial task-agnostic performance disparities between high- and low-resource languages. Moreover, we show that larger models do not necessarily outperform smaller ones in a multilingual setting.

1 Introduction

Since the release of ChatGPT, Natural Language Processing has experienced a significant surge in interest and research, with a particular focus on LLMs finetuned to follow human instructions. Besides proprietary models like GPT-4 (Achiam et al., 2023), Claude (Bai et al., 2022), or Gemini (Anil et al., 2023), there are also successful open-source variants such as Llama (Touvron et al., 2023), Phi (Gunasekar et al., 2023; Abdin et al., 2024), or Mistral (Jiang et al., 2023). While LLMs often demonstrate impressive performance on a wide range of tasks, quantifying and measuring this performance is challenging. Nevertheless, recent evaluation studies have shown that LLMs generally perform well in English but much worse in other languages (Ahuja et al., 2023a,b; Holtermann et al., 2024).

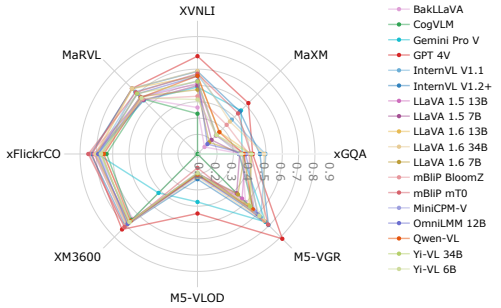


Figure 1: An overview of the average performance of the models on the datasets included in the M5 benchmark. For xFlickrCO and XM3600, we report BERTScore F1. For the other datasets, the accuracy metric is reported.

In this work, we focus on multimodal variants of LLMs, Large Multimodal Models (LMMs), such as GPT 4V (OpenAI, 2023), Gemini Pro V (Anil et al., 2023), or the popular open-source model, LLaVA (Liu et al., 2023a,b). LLMs are not text-only but are also capable of processing images in addition to text. Most open-source LMMs comprise three major components: an LLM, a vision-encoder model, and a mapping network that projects image embeddings into the text embedding space. With this architecture, where an LLM serves as the core, we argue that LMMs inherently suffer from the same issue as LLMs: they generally perform much worse in non-English languages. However, existing benchmarks are either text-only (Ahuja et al., 2023a) or multimodal but monolingual (Yue et al., 2023), thus unable to prove

[1]This works was done during a research internship with Microsoft Research India (Bangalore) between November 2023 and March 2024.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 4309–4345
November 12-16, 2024 ©2024 Association for Computational Linguistics

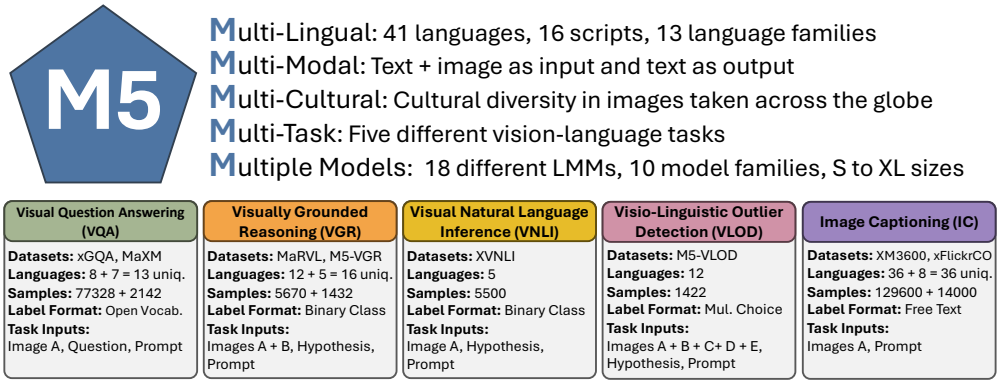


Figure 2: An informative overview of the M5 Benchmark introduced in this work.

this hypothesis. In other words, current research lacks multimodal multilingual benchmarks to examine LMMs' multilingual capabilities. In this work, we fill this gap by introducing the M5 Benchmark, taking a significant step towards identifying and measuring the performance disparities of current LMMs between various languages. Figure 2 and Figure 1 present a high-level summary of our benchmark. Moreover, we introduce two new evaluation datasets, including a novel vision-language task. Both datasets focus on African and Asian cultures, which are underrepresented or even nonexistent in previous benchmarks. Our exhaustive analyses additionally investigate the influence of different factors on the performance, such as the models' size or language fidelity.

Major Contributions The major contributions of this work are (a) M5, the first multimodal benchmark to assess the performance of current LMMs across five tasks, eight datasets, and 41 languages; (b) Two novel datasets spanning 10 underrepresented African and Asian languages, English and German, with images depicting the respective cultures. (c) A novel vision-language task: Visio-Linguistic Outlier Detection (VLOD); (d) A large-scale evaluation of 18 recent LLMs and a thorough analysis of their multilingual performance. (e) A public release of our codebase and all datasets in a uniform schema to foster future research for more equitable and accessible LMMs or AI in general[1].

2 Related Work

Large Multi-Modal Models This work focuses on the multimodal counterpart of large language models (LLMs), often referred to as Large Multimodal Models (LMMs). LMMs are language models capable of processing and "understanding" data other than text. While this generally subsumes images, video, audio, or more, we concentrate on visio-linguistic LMMs, i.e., models that take text and/or images as input and generate textual output.

The vast majority of open-source LMMs comprise three major components: a pretrained generative LLM as the core, a pretrained vision-encoder model that computes semantically rich image embeddings, and a shallow mapping network that learned to project image embeddings into the text embedding space. One of this architecture's successful open-source implementations with a recent LLM, i.e., the Llama-based Vicuna (Chiang et al., 2023; Touvron et al., 2023), is LLaVA (Liu et al., 2023b), from which many others took inspiration also regarding the training data and process. Besides this, LMMs also exist, which use Cross-Attention (Wang et al., 2023; Bai et al., 2023), Q-Formers (Li et al., 2023; Geigle et al., 2023), Adapters (Eichenberg et al., 2022), or Preceiver Resamplers (Alayrac et al., 2022; Awadalla et al., 2023) to process image embeddings. For an overview including architectural details and the number of parameters of the 18 LMMs' components we employed in this work, please see Table 8.

Evaluation Benchmarks With the recent surge in the research of LLMs and LMMs, analyzing the models' performances is crucial yet challenging. Popular benchmarks like BIG-Bench (bench authors, 2023), HELM (Liang et al., 2022), or MMLU (Hendrycks et al., 2020) are the defacto-standard to evaluate LLMs on text-only tasks primarily in English. Efforts like MEGA (Ahuja et al., 2023a), MEGAVERSE (Ahuja et al., 2023b), or MultiQ (Holtermann et al., 2024) extended these monolingual benchmarks to a large set of diverse

[1] https://github.com/floschne/m5b

languages and showed that the LLMs' performance in English versus non-English languages differs significantly.

Similarly, efforts have been made to evaluate multimodal models. Benchmarks like MMMU (Yue et al., 2023), MME (Fu et al., 2023), or MMBench (Yuan et al., 2023) assess the performance of LMMs on a vast number of text-image tasks. However, these benchmarks primarily focus on English, with some tasks available in Chinese. Like MMMU, there is CMMMU (Ge et al., 2024), which focuses on text-image tasks in Chinese. Nonetheless, evaluating state-of-the-art LMMs in a massively multilingual large-scale setting remains largely unexplored. There are only a few multimodal multilingual evaluation datasets (see Section 3.2 and 8.6) and only two benchmarks: IGLUE (Bugliarello et al., 2022) and MEGAVERSE. However, IGLUE evaluates only non-autoregressive transformer-encoders, thus lacking state-of-the-art LLMs. In MEGAVERSE, only five recent LMMs are evaluated on two datasets.

3 The M5 Benchmark

This section describes the setup of the M5 Benchmark introduced by this work. Details about the experimental setup, including prompts and hyperparameters, are reported in Appendix A.

3.1 Models

We chose the LMMs included in this benchmark for the following reasons: Firstly, we focussed on publicly available models released on Hugging Face except for GPT-4 Vision and Gemini Pro. Secondly, we included LMMs well-performing on popular multimodal English-only benchmark s such as MMMU (Yue et al., 2023) and MME (Fu et al., 2023). Thirdly, we aimed to cover a mixture of different model families and a broad model size spectrum, including small models with 3B to 9B, medium models with 10B to 19B, and large models with 20B to 40B parameters. For an overview of all models, including their number of parameters and other architectural details, see Table 8.

3.2 Datasets

This section briefly introduces the existing datasets included in our benchmark. In addition to these, we crafted two novel datasets described in Section 4. For details about the languages covered by the datasets, please refer to Table 6.

xGQA The xGQA dataset (Pfeiffer et al., 2022) is a cross-lingual visual question-answering dataset. It extends the well-known English-only GQA dataset (Hudson and Manning, 2019) by manually translating the questions in the balanced *test-dev* set. Each of the 9666 questions is available in eight languages covering five scripts, while the answers are in English only. The dataset holds 300 unique images from Visual Genome (Krishna et al., 2017).

MaXM The MaXM dataset was introduced by Changpinyo et al. (2023) and is a VQA dataset comprising seven languages in five scripts. In MaXM, the questions and their respective answers are in the same language. The images are a subset of the XM3600 (Thapliyal et al., 2022) dataset and are chosen to match a region where the language of the question-answer pair is spoken. This ensures cultural diversity in the images in addition to the language diversity in the question-answer texts.

XVNLI The XVNLI dataset (Bugliarello et al., 2022) introduces the task of Cross-lingual Visual Natural Language Inference where a model needs to predict whether a textual hypothesis *entails*, *contradicts*, or is *neutral* concerning a visual premise. XVNLI comprises five languages covering three scripts and 357 unique images from Visual Genome.

MaRVL The MaRVL dataset (Liu et al., 2021) aims to benchmark models on Multicultural Reasoning over Vision and Language. A task sample comprises two images, a textual statement, and a binary true or false answer grounded in the images. MaRVL comprises five languages covering three scripts and 4914 culturally diverse images that match the respective languages. The images in a sample are chosen to match the culture of the annotator who has written the textual statement in his or her native language.

XM3600 The XM3600 dataset (Thapliyal et al., 2022) is a large multilingual image captioning dataset comprising 36 languages with 261375 captions covering 13 different scripts for 100 unique images per language. The images are selected to match the language's cultural background, ensuring cultural and linguistic diversity. The captions were not automatically translated but manually created by professional annotators who are native speakers of the respective language.

xFlickrCO The xFlickrCO dataset (Bugliarello et al., 2022) is an image captioning dataset and comprises 1000 images from Flickr30k (Young et al., 2014) and 1000 images from COCO (Lin et al., 2014). Each image is captioned in eight languages, covering four different scripts. For all languages except English and German, the captions were manually crafted by crowdsourcing workers instead of translated from English to prevent bias and increase linguistic diversity.

4 Novel M5 Datasets

In addition to the existing datasets introduced in the previous section, we crafted two novel multimodal and multilingual evaluation datasets. The principal motivation behind this is to fill the gap in existing vision-language datasets concerning the lack of underrepresented languages, tasks, and cultural diversity. Moreover, we aim to enable further examination of LMMs and their performance on non-English and non-Western data with a particular focus on African and Asian regions. More details, statistics, and examples are reported in Appendix B.

Common Characteristics

Languages Both datasets comprise samples in 12 languages covering seven scripts (see Table 6): *Amharic*, *Berber*, *Bengali*, *German*, *English*, *Filipino*, *Hausa*, *Hindi*, *Russian*, *Swahili*, *Thai*, *Zulu*. The languages were selected to enrich the set of languages covered by existing datasets, focusing on underrepresented languages from Asian and African countries or regions. To our knowledge, no other visio-linguistic evaluation dataset covers Amharic, Berber, Hausa, or Zulu.

Data Annotation The textual data in both datasets is manually created by professional annotators who are native speakers of the respective languages. All annotators work for a data annotation company, and fluent English-speaking correspondents handle communication and task delegations. In order to ensure that the annotators can fulfill the tasks as well as possible, detailed guidelines, including multiple good and bad examples, have been drawn up in English. These guidelines were explained in detail to the correspondents. The correspondents then delegated the tasks to the annotators by having internal company guidelines drawn up in the target languages. After the annotation tasks were finished, we conducted the following quality assessment procedure:

1. We translated all manually created annotations to English using the Bing Translate API.
2. We developed a small tool that displays a sample, including the images, target language, original and English-translated annotations, and other metadata.
3. We used the tool to manually inspect 20% of the samples and tagged them as "good", "bad", or "ambiguous/problematic".
4. We discussed in detail our findings with the annotators' correspondents, who then delegated the tasks to improve the quality of the annotations.
5. This loop was executed two times until no more issues were found by the authors and the annotators' correspondents.

Depicting Cultural Diversity The images in our datasets originate from the Dollar Street dataset (Gaviria Rojas et al., 2022), comprising around $38K$ photos taken in 63 different regions or countries around the globe. These photos depict the lives of families, including their homes, neighborhoods, or everyday objects, in a culturally diverse way. Further, each image in the original dataset is tagged with one or more "topics" that roughly describe its visual content.

Image Basis For our datasets, we sampled a subset of images from the Dollar Street dataset (Gaviria Rojas et al., 2022) taken in regions where the 12 target languages are spoken. In this subset, which forms the visual basis for both of our datasets and is referred to as $\mathbb{B}$, each image $i_l^t \in \mathbb{B}$ is tagged with exactly one topic $t \in \mathbb{T} = \{t_0, \ldots, t_{86}\}$ and was taken in a region r_l where language $l \in \mathbb{L} = \{l_0, \ldots, l_{11}\}$ is spoken. More information about the image topic distribution per language can be found in Appendix B.1.3.

4.1 M5-VGR

Inspired by MaRVL, the goal of the M5-VGR dataset is to provide a visually grounded reasoning (VGR) evaluation dataset that covers a wide range of topologically different languages and, at the same time, visually represents a diverse set of cultures in which the respective languages are spoken. However, since the MaRVL dataset contains

Figure 3: An Zulu example of the novel M5-VGR dataset. **Hypothesis:** *"Isithombe sokuqala nesithombe sesibili sibonisa iqanda elisehhokweni. (The first picture and the second picture show the egg on the head.)"*, **Label:** *False*

only five languages, we chose 11 additional topologically diverse languages for our dataset. To guarantee visual and linguistic diversity and high data quality in our dataset, we hired professional native-speaker annotators of the respective languages to annotate the data. Moreover, we performed several rounds of data quality assessment in close collaboration with the annotators.

A task sample s in M5-VGR contains two images i_a and i_b, a textual visually grounded hypothesis h, and a binary label c which is either true or false concerning the two visual premises (see Figure 3). More specifically, for each language $l \in \mathbb{L}$, we created 120 tasks $s_l \in \mathbb{S}_l$ as follows: In the first step, we sampled 120 unique images $a_l^t \in \mathbb{B}$ from our image basis so that each topic $t \in \mathbb{T}$ occurs at least once across all 12 languages. Then, for each of the 120 images, we randomly selected another image $b_{l_2}^t \in \mathbb{B}$ associated with another language $l_2 \neq l \in \mathbb{L}$ that shares the topic t. In the third step, we asked the native-speaker annotators of the language l to manually create a hypothesis h and a label c which is either true or false concerning the image premises $\left(a_l^t, b_{l_2}^t\right)$. Further, the annotators were instructed to generate a hypothesis semantically related to the topic t if possible.

4.2 M5-VLOD

Figure 4: A Swahili example of the novel M5-VLOD dataset. **Hypothesis:** "*Picha zote zinaonyesha sabuni inayotumika kwa mikono na mwili bila mtu yeyote. (All the images show soap applied to the hands and body without anyone.)*", **Outlier:** 1.

With the M5-VLOD dataset, we introduce a novel multimodal task: Visio-Linguistic Outlier Detection. The objective of the task is to detect an outlier image from a set of images considering a textual statement. An example of the task is shown in Figure 4, where five images related to the topic "soap for hands and body" are shown. The machine-translated English statement is: "All the images show soap applied to the hands and body without anyone.". Because only the first image shows a person, the statement is incorrect for the first image and, therefore, is considered the outlier image.

The dataset was collected similarly to M5-VGR, as described in the previous section. The major difference is that instead of sampling only one image in the second step, we sample four images so that a sample $s_{l_0} \in \mathbb{S}_{l_0}$ for language $l_0 \in \mathbb{L}$ comprises of five images: $\{a_{l_0}^t, b_{l_1}^t, c_{l_2}^t, d_{l_3}^t, e_{l_4}^t, \}$ associated with five different languages $\{l_0, \ldots, l_4 \in \mathbb{L}\}$ that share one topic $t \in \mathbb{T}$. In the third step, we asked the native-speaker annotators of the language l to manually create a textual statement h, valid for all but one of the images labeled as the outlier image.

5 General Results Discussion

This section discusses the models' performance on the datasets considered in our benchmark. Table 1 provides an overview of the performance in English compared to non-English languages for all models and datasets. Note that we use friendly names for the models for better readability (see Table 8). Detailed results for each dataset and all their respective languages are provided in Appendix D.

5.1 Summary of Findings

Table 1 shows a clear pattern: Generally, LMMs perform significantly worse in non-English languages across all tasks. More specifically, the average performance across all models and datasets in English is 0.63 versus 0.47 in non-English languages. Most models have an average performance difference from English to non-English larger or equal to 0.12. However, for `GPT 4V` and despite their much smaller size also for `mBlip BloomZ`, and `mBlip T0`, the difference is smaller than 0.1. For the two `mBLIP` models, the authors explicitly stated in their paper the language distribution in the training data, which covers 96 languages. Hence, it can be assumed that this is the reason for this slight absolute performance difference, and, further, this might indicate that `GPT 4V` was also trained in a

Model	Dataset																		
	xGQA		MaXM		XVNLI		MaRVL		M5-VLOD		M5-VGR		xFlickrCO		XM3600		ALL		
	E	NE	E	NE	E	NE	E	NE	E	NE	E	NE	E	NE	E	NE	E	NE	Δ
CogVLM	0.59	0.30	0.43	0.02	0.47	0.29	0.60	0.51	0.10	0.08	0.68	0.55	0.87	0.60	0.88	0.65	0.58	0.38	−0.20
BakLLaVA	0.62	0.32	0.53	0.08	0.48	0.34	0.59	0.53	0.14	0.20	0.71	0.48	0.91	0.63	0.88	0.64	0.61	0.40	−0.21
LLaVA 1.6 7B	0.60	0.34	0.34	0.16	0.59	0.45	0.62	0.53	0.14	0.21	0.55	0.42	0.88	0.64	0.88	0.67	0.57	0.43	−0.15
LLaVA 1.5 7B	0.62	0.30	0.52	0.15	0.60	0.47	0.57	0.52	0.15	0.20	0.48	0.42	0.92	0.68	0.89	0.67	0.59	0.43	−0.17
Yi-VL 6B	0.57	0.32	0.53	0.20	0.56	0.38	0.59	0.53	0.20	0.19	0.73	0.61	0.91	0.64	0.91	0.66	0.62	0.44	−0.18
MiniCPM-V	0.55	0.31	0.56	0.19	0.66	0.49	0.61	0.53	0.20	0.20	0.80	0.56	0.91	0.65	0.90	0.65	0.65	0.45	−0.20
LLaVA 1.5 13B	0.62	0.34	0.56	0.19	0.59	0.49	0.60	0.54	0.16	0.21	0.57	0.46	0.91	0.69	0.90	0.69	0.61	0.45	−0.16
Qwen-VL	0.59	0.33	0.50	0.23	0.62	0.54	0.60	0.53	0.16	0.21	0.82	0.54	0.89	0.62	0.90	0.65	0.64	0.46	−0.18
Yi-VL 34B	0.58	0.38	0.53	0.20	0.59	0.51	0.62	0.58	0.26	0.19	0.77	0.52	0.91	0.64	0.90	0.66	0.65	0.46	−0.19
Gemini Pro V	0.46	0.34	0.48	0.23	0.49	0.49	0.55	0.55	0.52	0.36	0.79	0.66	0.86	0.67	0.63	0.41	0.60	0.46	−0.13
OmniLMM 12B	0.49	0.36	0.48	0.11	0.64	0.54	0.64	0.56	0.19	0.21	0.78	0.59	0.91	0.66	0.89	0.68	0.63	0.46	−0.16
LLaVA 1.6 13B	0.65	0.38	0.46	0.24	0.61	0.55	0.65	0.65	0.14	0.21	0.78	0.50	0.90	0.67	0.88	0.68	0.63	0.48	−0.15
mBliP BloomZ	0.44	0.39	0.55	0.29	0.40	0.44	0.55	0.56	0.14	0.21	0.69	0.56	0.92	0.72	0.91	0.71	0.58	0.49	−0.09
InternVL V1.1	0.63	0.48	0.58	0.34	0.61	0.56	0.63	0.60	0.13	0.21	0.73	0.62	0.92	0.66	0.91	0.68	0.64	0.52	−0.12
LLaVA 1.6 34B	0.65	0.46	0.58	0.32	0.62	0.58	0.64	0.66	0.26	0.22	0.87	0.64	0.89	0.68	0.88	0.70	0.67	0.53	−0.14
mBliP mT0	0.44	0.40	0.50	0.42	0.59	0.57	0.60	0.63	0.12	0.17	0.74	0.69	0.92	0.73	0.91	0.71	0.60	0.54	−0.07
InternVL V1.2+	0.67	0.43	0.60	0.42	0.63	0.58	0.68	0.61	0.28	0.23	0.86	0.68	0.92	0.71	0.90	0.70	0.69	0.55	−0.15
GPT 4V	0.45	0.41	0.49	0.53	0.69	0.68	0.64	0.66	0.70	0.42	0.88	0.81	0.90	0.70	0.89	0.72	0.70	0.62	−0.09
Random Baseline	–		–		0.33		0.50		0.20		0.50		–		–		–		–
Average	0.57	0.37	0.51	0.24	0.58	0.50	0.61	0.57	0.22	0.22	0.73	0.57	0.90	0.67	0.88	0.66	0.63	0.47	−0.15

Table 1: Average performance in English (E) and non-English languages (NE) on all datasets for all models. For each dataset and the Δ column, the heatmaps are created individually, indicated by the column gutter. The column "ALL" represents the average across all datasets. For xFlickrCO and XM3600, we report BertScore F1 and for the rest of the datasets, we report the relaxed accuracy.

multilingual fashion. Due to the difference in size and the architecture[2] of the `mBlip` models and `GPT 4V`, applying this multilingual training strategy for LMMs would generally lead to more robust multilingual performance.

The average performance difference of the models is most significant on the MaXM, XM3600, and xFlickrCo datasets, for which the models are required to generate non-English text.

Interestingly, for the M5-VLOD dataset, the models that performed worse than the random baseline of 0.2 in English performed better in non-English languages. An explanation for this could be false assumptions drawn from the English text. This finding also explains why the average English versus non-English performance disparity across all models is equal for the dataset and lies around the random baseline, indicating the challenge introduced by our dataset.

5.1.1 Dataset-Specific Discussion

Note that due to brevity constraints, we report exact numbers and diagrams of the language-specific results for each dataset in Appendix D.

xGQA All models perform best in English mostly, with a significant gap in accuracy to the second-best language from up to 0.62 in English to 0.36 in Russian for `LLaVA 1.6 7B`. In Bengali, where the models have the lowest average accuracy of 0.19, all models besides `GPT 4V`, which achieves 0.44, perform worst by far. The best-performing model in English and the best-performing model on average over all languages are the `InternVL v1.2` and `InternVL v1.1` models. Notably, despite their (estimated) much larger size, `GPT 4V` and `Gemini Pro V` are among the worst-performing models in English. After manually inspecting the results, we found the reason for this to be that the models did not respond in a single word but with a brief sentence, which is considered a false answer according to the applied metric (see Appendix A.2 and Section 8.2).

MaXM The average accuracy of the models for Hindi (0.22), Hebrew (0.19), Romanian (0.27), Thai (0.25), and Chinese (0.24) is much lower than for English (0.51) and French (0.35). It is also worth pointing out that most models, regardless of their size, perform remarkably worse in languages other than English (and French). In contrast, on xGQA, which is also a VQA dataset, the differences between the languages are much more minor. This is likely due to the difference between the two datasets, i.e., that xGQA has multilingual questions but only English answers, while MaXM has multilingual questions and expects the answers in the respective language, too. We further underline this in our language fidelity analysis in Section 6.3.

XVNLI English accuracy is the best for most models, with an average of 0.58, whereas Arabic accuracy is the worst, with an average of 0.43. The performance drop from English to the other languages, i.e., Spanish (0.51), French (0.52), and

[2]While the architecture of `GPT 4V` is not known, it is likely different from the mBlip models' architecture, which employs Q-Formers, rarely used in state-of-the-art LMMs.

Russian, with average accuracy scores of 0.51, 0.52, and 0.52, is less substantial. Note that XVNLI is an NLI dataset, i.e., the random baseline is at $\frac{1}{3}$. All models surpass this baseline in all languages, except for CogVLM in Arabic (0.26) and French (0.27). The best-performing model is GPT4 V with an average accuracy across all languages of 0.68, followed by LLaVA 1.6 34B and InternVL V1.2+ with average scores of 0.59 and 0.58, respectively.

MaRVL The dataset's random baseline is 0.5, which is often only slightly surpassed by most models, especially for Swahili and Tamil languages, with an average accuracy of 0.53 and 0.54, respectively. Notably, only 8 of 18 models perform best in English, with an average accuracy of 0.61. For the other models, the English performance is surpassed by Chinese, Indonesian, or Turkish, with an average accuracy of 0.60, 0.60, and 0.59, respectively. GPT-4V is on par with LLaVA 1.6 34B despite the latter having much fewer parameters.

M5-VGR As with MaRVL, this dataset's random baseline is at 0.5. Only one of 18 models, i.e., InternVL V1.2+, could surpass or reach this baseline in all languages. As expected, most models performed best in English, German, or Russian, with average accuracies of 0.73, 0.68, and 0.69, respectively. They performed worst in low-resource languages such as Amharic, Berber, Bengali, Hausa, or Zulu, with an average accuracy of 0.53, 0.49, 0.55, and 0.52, respectively. Only three models, i.e., Gemini Pro V, mBlip mT0, and GPT 4V, consistently and significantly surpass the random baseline in all languages except for Berber. The only languages where the average performance is significantly higher than the 0.5 random baseline are English (0.73), German (0.68), Russian (0.69), and Thai (0.62). The average scores of the other languages range from 0.49 in Berber to 0.57 in Hindi.

M5-VLOD The dataset's random baseline is 0.2 since the models need to find the outlier within five images. Only GPT 4V and Gemini Pro V significantly surpassed that baseline in all languages, with an average accuracy of 0.42 and 0.36, respectively. They achieve the best scores in English with an average accuracy of 0.70 (GPT 4V) and 0.52 (Gemini Pro V. However, in Berber, both models only achieve scores around the random baseline. All other models do not surpass the random baseline in all languages, including English, by more than 0.1, with average scores between 0.08 (CogVLM) and 0.23 (InternVL V1.2+) This highlights the challenge introduced by our dataset and the performance gap between proprietary and open-source models.

xFlickrCO The majority of models perform best in English, often with a significant margin in average chrF++, i.e., 24.93 in English and 12.49 in non-English languages. Other languages where the models perform comparably well are German and Spanish, with average chrF++ scores of 19.95 and 19.55, respectively. Interestingly, all models perform worse in non-Latin script languages, i.e., Russian (9.70), Chinese (4.53), and Japanese (4.05). Unexpectedly, the proprietary models GPT 4V and Gemini Pro V are surpassed by mBliP BloomZ, mBliP mT0, and InternVL V1.2+, which are much smaller open-source models. Even in English, most open-source models outperform the proprietary models.

XM3600 Note that due to limited resources, we evaluated GPT 4V only on a subset of 12 of 36 languages. Most models perform best in English (27.14 average chrF++) by a large margin, followed by other Latin scripts in high-resource languages such as French (23.65), Spanish (23.52), or Dutch (21.01). On average, the models perform worst on non-Latin script languages like Korean (3.50), Telugu (4.79), and Bengali (5.11). However, although the chrF++ metric claims to be script and language-independent, the low scores in high-resource languages like Chinese (3.95) and Japanese (5.13) make the metric questionable. While detailed analysis is out of the scope of this work, in future work, we will investigate this issue further (see Section 8.1).

6 Aggregated Result Analyses

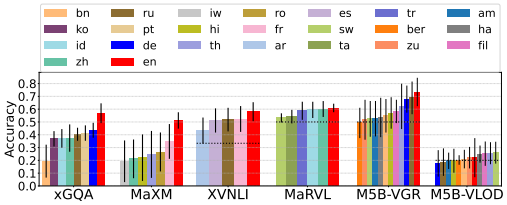
6.1 Performance per Language

Figure 5 shows the average performances aggregated by language[3] or language taxonomy classes (Joshi et al., 2020). These taxonomy classes indicated how well a respective language is represented and considered within the research field of NLP based on papers published at CL conferences. High-resource languages such as English or German are in Class 5, whereas low-resource

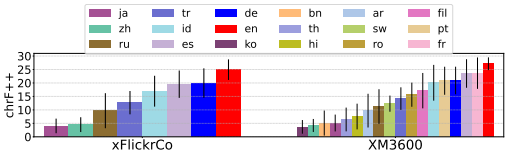
[3]We do not show all 36 languages of XM3600 for better readability.

languages such as Berber are in Class 0. For details about the languages and their taxonomy classes, please refer to Table 7.

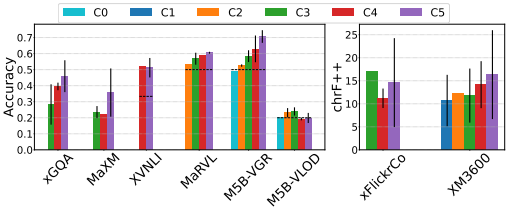
As can be observed from Figure 5a and Figure 5b, the models perform best in English, followed by other European languages across all datasets. Our newly presented M5-VLOD dataset is an exception, where the average performance for all languages is around the random baseline, indicating the challenge it implies. As expected, the models consistently perform worse on low-resource languages than on high-resource languages on all datasets. This is also displayed in Figure 5c, where it can be observed that the average performance decreases with the language taxonomy class. Note that this is not precisely true for xFlickrCO and XVNLI because the average on Class-5 languages is lowered by outliers, as indicated by the large error bars. In contrast, the models performed comparably well in only one Class 3 or 4 language, respectively.



(a) Performance on VQA, VGR, and VNLI datasets aggregated by language.



(b) Performance on image captioning datasets aggregated by language.



(c) Performance on datasets aggregated by language taxonomy class as introduced by Joshi et al. (2020).

Figure 5: Models' performances on all datasets aggregated by language or language taxonomy classes.

6.2 Performance vs. Model Parameters

In Figure 6, we plot the English and non-English average performance on the employed datasets versus the models' sizes in multiple regression plots. Note that, on the x-axes, we indicated the unknown sizes of `GPT 4V` and `Gemini Pro V` by "*???*", which are estimated to be of magnitudes larger than all other models evaluated in this benchmark hence should be much further right. However, we did not do so to improve the readability of the plots.

In the figures, we can make several observations: Firstly, the average English performance is higher than the non-English performance for all models on all datasets. Secondly, the markers, which represent the average performance of a specific model on a dataset, show that the largest model does not always perform best and that the difference between smaller and larger models is often neglectable. The same finding is shown by the relatively flat slope of the regression lines. However, for the M5-VLOD and VGR datasets, the regression line for the average English scores is steeper, meaning that larger models perform considerably better than the smaller models. Since this work introduces the datasets and M5-VLOD even introduces a novel task, it can be concluded that larger models can better generalize to unseen data.

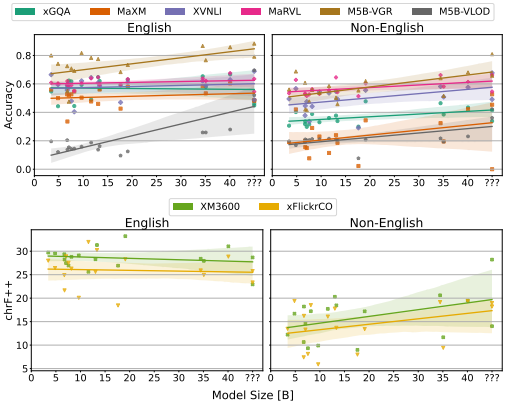


Figure 6: Regression plots showing the English and average non-English performance versus model size on different datasets. On the x-axis, we indicated the unknown sizes of `GPT 4V` and `Gemini Pro V` by "???".

6.3 Language Fidelity Analysis

Inspired by Holtermann et al. (2024), we report the results of a language fidelity analysis, which assesses how often a model responds in the requested language on average. For this, we used GlotLIDv3 (Kargaran et al., 2023) to predict the language based on the output text of the respective models. Since it is hard to predict the language of a word or a multi-word expression due to ambigu-

ity, we selected the xFlickrCO dataset, where the expected response of a model is an image caption, i.e., a sentence, in one of eight languages. As it can be observed from Table 2, all models achieve (almost) perfect fidelity in English where, whereas for Japanese, Russian, and Turkish, the average fidelity drops to two-thirds. Interestingly, the small-sized `mBLIP` models have almost perfect fidelity in all languages, (slightly) surpassing larger models like `InternVL V1.2+` and `GPT 4V`.

Table 2: Language fidelity results on the xFlickrCO dataset.

Model	Language								
	zh	en	de	id	ja	ru	es	tr	Avg.
BakLLaVA	.00	1.0	.39	.06	.00	.00	.44	.00	.24
Yi-VL 6B	.14	1.0	.20	.00	.20	.01	.57	.00	.28
Qwen-VL	.95	.99	.18	.11	.15	.08	.15	.07	.33
Yi-VL 34B	.43	1.0	.79	.45	.58	.22	.25	.33	.51
CogVLM	.44	.95	.74	.76	.38	.43	.82	.54	.63
LLaVA 1.5 13B	.88	1.0	.75	.55	.90	.26	.75	.40	.69
LLaVA 1.5 7B	.83	1.0	.96	.83	.09	.22	.97	.67	.70
MiniCPM-V	.21	1.0	.93	.79	.89	.96	.91	.68	.80
LLaVA 1.6 7B	.99	.99	.66	.91	.59	.88	.91	.89	.85
InternVL V1.1	.96	1.0	.93	.78	.88	.89	.97	.66	.89
OmniLMM 12B	.63	1.0	.95	.92	.83	.92	.98	.88	.89
Gemini Pro	.95	.95	.95	.88	.91	.96	.97	.96	.94
LLaVA 1.6 13B	1.0	1.0	.90	.96	.91	.87	.97	.93	.94
LLaVA 1.6 34B	.88	1.0	.99	.99	.86	.99	.99	.99	.96
GPT 4V	.97	1.0	1.0	.98	.88	.99	.99	1.0	.98
InternVL V1.2+	.99	1.0	1.0	.95	.97	.99	.99	.96	.98
mBliP BloomZ	.96	1.0	1.0	.99	.99	1.0	1.0	.99	.99
mBliP mT0	.96	1.0	1.0	.99	.99	1.0	1.0	1.0	.99
Avg.	.73	.99	.79	.72	.67	.65	.81	.66	.75

While the language fidelity of a model focuses on the generated text, we argue that the fidelity is also an indicator of the model's general language capabilities. To prove this hypothesis, we computed Pearson correlation coefficients between the reported fidelity and the models' performance on the datasets for the xFlickrCO languages. As shown in Table 17, there is a positive moderate or high correlation between the average fidelity and the average score for most datasets. However, for xGQA and M5-VLOD, there is only a minor positive average correlation.

7 Conclusion

We introduced M5, a diverse benchmark in which we evaluated 18 Large Multimodal Models (LMMs) with varying sizes across five visio-linguistic tasks in eight datasets comprising 41 unique languages. Further, we presented two novel datasets – M5-VGR and M5-VLOD – which focus on underrepresented languages and depict culturally diverse scenes. With M5-VLOD, we introduce a new visio-linguistic outlier detection task in which only proprietary models achieve reasonable scores. Our experiments revealed that model size does not always correlate with better performance, especially in non-English languages, underscoring the importance of diverse, multilingual training data and robust architectures. Performance disparities were prominent between high-resource languages like English and low-resource languages across all datasets and models, highlighting ongoing challenges in achieving globally equitable multilingual AI. With M5, we aim to impel the development of more inclusive models suitable for diverse languages and cultures.

8 Limitations

This section outlines several limitations of our current study that will be addressed in future work.

8.1 Metrics for Multilingual Image Captioning

Our benchmark and current research generally lack robust metrics for evaluating multilingual image captioning, especially for non-Latin script languages. The issue, which is the same for machine translation tasks, arises because of the nature of most metrics, such as chrF (Popović, 2017), CIDEr (Vedantam et al., 2015), ROUGE (Lin, 2004), BLUE (Papineni et al., 2002), or METEOR (Banerjee and Lavie, 2005), which are based on comparing word or character n-grams between the source and target sequence. For non-Latin scripts, tokenization or segmentation can be challenging because it might not contain spaces or punctuation, or the characters are logographic. Hence, their usability or effectiveness is doubtful in such scenarios because the metrics rely on tokenization.

Other metrics, such as BERTScore (Zhang et al., 2020), CLIPScore (Hessel et al., 2021), or COMET (Rei et al., 2020), do not rely on the captions' surface forms but on their token or sentence embeddings. However, they suffer from other issues: They require strong multilingual or cross-lingual encoder models capable of computing embeddings for many languages, which itself is a challenging task. Further, the scores computed with these metrics are often not calibrated across languages and thus not directly comparable between different languages.

A promising currently popular solution might be the use of robust multilingual state-of-the-art LLMs such as GPT 4o[4], Claude 3 Opus[5], or Gem-

[4] https://openai.com/index/hello-gpt-4o/
[5] https://www.anthropic.com/news/

ini 1.5 Ultra[6] as a judge (Zheng et al., 2024). However, this would require more computational and financial resources and, most importantly, more investigation.

8.2 VQA Metrics for Generative Models

The problem when employing and evaluating generative language models on question-answering tasks is that the models can generally output arbitrary token sequences. However, the gold label answers are limited and often comprise only a short phrase, a single word, or even a binary label. Hence, mapping the predicted answers to their gold labels is not straightforward, and the difficulty drastically increases in multilingual scenarios. The relaxed accuracy metric employed in this study (see Section A.1) has been found to occasionally incorrectly classify correct answers, leading to false negatives, especially in open vocabulary visual question answering (VQA). One way to address this issue is to leverage strong state-of-the-art LLMs as judges, as described above, to enhance the accuracy of the evaluations.

8.3 Influence of Prompting

Another limitation of this and most, if not all, other current studies is grounded in the model prompting. Since different models might react differently to specific prompting styles, and we only employ a single prompt per dataset for all models[7] (see Figure 7), the results might not be optimal. This issue has been partially addressed by Ahuja et al. (2023a) but is out of the scope of this work.

8.4 "Outdated" Models

Since the pace of current research in NLP, CV, and multimodal machine learning is swift, the models employed in our benchmarking exercise might be considered slightly outdated. Note that we considered models released until March 2024. Since then, numerous improved LMMs based on state-of-the-art LLMs, such as Llama3[8] and novel image encoders techniques such as NaVIT (Dehghani et al., 2024), have been publicly released. Because this was foreseeable, we designed our benchmark to be easily extendable with newer models, which we will include in future work.

8.5 Small M5 Datasets

This work introduced two datasets, M5-VGR and M5-VLOD, which comprise about 115 samples for each of the 12 languages. Compared to other datasets, they can be considered small. We will increase their sizes in future work to obtain more robust and generalizable results.

8.6 Missing multimodal and Multilingual Datasets

Currently, the M5 Benchmark comprises 5 text-image tasks, i.e., VQA, VGR, VNLI, and image captioning, thus missing other suitable tasks like multimodal and multilingual summarization. Further, other multimodal multilingual VQA and VGR datasets have emerged while writing this paper. We will include both new tasks and new datasets in future versions of the M5.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally On Your Phone. *ArXiv*, 2404.14219.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *ArXiv*, 2303.08774.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023a. MEGA: Multilingual Evaluation of Generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, et al. 2023b. MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks. *arXiv preprint arXiv:2311.07463*.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024.

claude-3-family

[6] https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/

[7] We do apply the model-specific prompt or chat templates, though.

[8] https://ai.meta.com/blog/meta-llama-3

Yi: Open Foundation Models by 01.AI. *Preprint*, arXiv:2403.04652.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: A Visual Language Model for Few-Shot Learning. *Advances in neural information processing systems*, 35:23716–23736.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *ArXiv*, 2312.11805.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv preprint arXiv:2308.01390*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness From AI Feedback. *ArXiv*, 2212.08073.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan.

BIG bench authors. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 2370–2392.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. MaXM: Towards Multilingual Visual Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2667–2682, Singapore.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint arXiv:2312.14238*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. 2024. Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution. *Advances in Neural Information Processing Systems*, 36.

Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2022. MAGMA – Multimodal Augmentation of Generative Models through Adapter-based Finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2416–2428, Abu Dhabi, United Arab Emirates.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.

William Gaviria Rojas, Sudnya Diamos, Keertan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. 2022. The Dollar Street Dataset: Images Representing the Geographic and Socioeconomic Diversity of the World. *Advances in Neural Information Processing Systems*, 35:12979–12990.

Zhang Ge, Du Xinrun, Chen Bei, Liang Yiming, Luo Tongxu, Zheng Tianyu, Zhu Kang, Cheng Yuyang, Xu Chunpu, Guo Shuyue, Zhang Haoran, Qu Xingwei, Wang Junjie, Yuan Ruibin, Li Yizhi, Wang Zekun, Liu Yudong, Tsai Yu-Hsuan, Zhang Fengji, Lin Chenghua, Huang Wenhao, Chen Wenhu, and Fu Jie. 2024. CMMMU: A Chinese Massive Multi-discipline Multimodal Understanding Benchmark. *arXiv preprint arXiv:2401.20847*.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavavs. 2023. mBLIP: Efficient Bootstrapping of Multilingual Vision-LLMs. *ArXiv*, 2307.06930.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks Are All You Need. *ArXiv*, 2306.11644.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the Elementary Multilingual Capabilities of Large Language Models with MultiQ. *arXiv preprint arXiv:2403.03814*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Drew A Hudson and Christopher D Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, Long Beach, CA, USA.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *ArXiv*, 2310.06825.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language Identification for Low-Resource Languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic Evaluation of Language Models. *arXiv preprint arXiv:2211.09110*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, Zurich, Switzerland.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually Grounded Reasoning across Languages and Cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved Baselines with Visual Instruction Tuning. *ArXiv*, 2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916, New Orleans, LT, USA.

OpenAI. 2023. GPT-4 Vision System Card.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-Lingual Visual Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti

Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv*, 2307.09288.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, Salt Lake City, UT, USA.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. CogVLM: Visual Expert for Pretrained Language Models. *ArXiv*, 2311.03079.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference Over Event Descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2:67–78.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2023. RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-Grained Correctional Human Feedback. *arXiv preprint arXiv:2312.00849*.

Liu Yuan, Duan Haodong, Zhang Yuanhan, Li Bo, Zhang Songyang, Zhao Wangbo, Yuan Yike, Wang Jiaqi, He Conghui, Liu Ziwei, Chen Kai, and Lin Dahua. 2023. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv:2307.06281*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *arXiv preprint arXiv:2311.16502*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, Online.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36.

A Experimental Setup Details

This section details the employed metrics, prompts, and generation hyperparameters.

Note that we ran all experiments on A6000 (50GB) and A100 (80GB) GPUs. The largest evaluated model (40B) fits on an A100.

A.1 Metrics

Following Geigle et al. (2023), we report a relaxed accuracy metric for the xGQA, MaXM, XVNLI, and MaRVL datasets due to the generative nature of the considered models. More specifically, we post-process the generated answers by, e.g., lowercasing, stripping, or removing punctuation. We then consider the processed generated answer correct if it matches the gold answer or starts or ends with the gold answer. Further, we allow synonyms for boolean and numerical values. Examples can be found in Table A.2.

Inspired by Ahuja et al. (2023b), we report the chrF++ (Popović, 2017) metric for the xFlickrCo and XM3600 datasets.

A.2 Relaxed Accuracy Metric

Table 3: Examples of generated answers considered correct or incorrect in the relaxed accuracy metric used to measure the performance on the xGQA, MaXM, MaRVL, XVNLI, M5-VGR, and M5-VLOD datasets. For more details, please refer to our GitHub repository.

Generated Answer	Gold Answer	Considered Correct
{Yes, 1, True}	true	yes
{No, 0, False}	false	yes
A car.	car	yes
Yes, it is correct.	yes	yes
It is not correct, no.	no	yes
The color of the leaf is green.	green	yes
There are three birds.	three birds	yes
Five	5	yes
{yes, true}	entailment	yes
{no, false}	contradiction	yes
maybe	neutral	yes
There are three birds in the image.	three birds	no
There are three birds.	3	no
three birds	3	no
three birds	3 birds	no

A.3 Prompts

Figure 7 presents the dataset-specific textual prompts we used for all models in this benchmark. Note that this does not include model-specific prompt templates, image placeholders, special tags, or symbols, only the ”raw” textual prompt, which is then embedded in the template as required by the respective model. The placeholders `{QUESTION}`, `{LANGUAGE}`, or `{HYPOTHESIS}` are replaced by the sample specific text. The prompts are partially inspired by Geigle et al. (2023) or Bugliarello et al. (2022).

A.4 Hyperparameters

This section briefly reports hyperparameters used within our experiments for better reproducibility.

A.4.1 Generation Parameters

We used the same generation hyperparameters to generate responses with all the employed open-source models on all datasets (see Table 4). Those are inspired by the default parameters in the “transformers”

xGQA

```
Question: {QUESTION} Short answer in English:
```

MaXM

```
Question: {QUESTION} Short answer in {LANGUAGE}:
```

MaRVL

```
Based on the two images, is it correct to say “{HYPOTHESIS}”? Yes or no? One word answer in
English:
```

XVNLI

```
Is it guaranteed true that “{HYPOTHESIS}”? Yes, no, or maybe? One word answer in English:
```

M5-VGR

```
Based on the two images, is it correct to say “{HYPOTHESIS}”? Yes or no? One word answer in
English:
```

M5-VLOD

```
Based on the 5 images ordered from top-left to bottom-right, which image does not match the
hypothesis “{HYPOTHESIS}”? Choose one from [A, B, C, D, E] and only output a single letter:
```

xFlickrCo

```
Brief caption in {LANGUAGE}:
```

XM3600

```
Brief caption in {LANGUAGE}:
```

Figure 7: Prompts employed for the different datasets.

library[9]. Because for CogVLM, beam search is not supported, we set “num_beams” to 1. For GPT 4V and Gemini Pro V, we use the default parameters of the respective Python clients.

Table 4: Generation hyperparameters to generate responses with all the employed models on all datasets.

Parameter	Value
num_beams	2
do_sample	True
max_new_tokens	50
temperature	1.0
top_k	50
top_p	0.95

A.4.2 Image Order for Multi-Image Datasets

Most models employed in our dataset only support a single image per prompt. For datasets where a sample comprises more than one image, i.e., for MaRVL, M5-VGR, and M5-VLOD, we use the following strategy: We first stack the images horizontally with a gutter of 10 pixels, provide them as a single image in the prompt, and generate the response. Then, we do the same again but stack the images vertically. For

[9] https://huggingface.co/docs/transformers/en/main_classes/text_generation

M5-VLOD, we also create a stacked image with two columns and three rows. The reported scores are the average of all variants.

B Dataset Details

B.1 M5-VGR and M5-VLOD Details

B.1.1 M5-VGR Examples

ID: a06bcb6431fb4a7cb4335bbfb02e2047
Topic: paper
Language: Amharic
Hypothesis:
ሁለቱም ምስሎች ጠረጴዛ ላይ የተቀመጡ ወረቀቶች ያሳያል።
Machine Translation:
Both images show papers sitting on a table.
Answer: True

Figure 8: Amharic M5-VGR Sample.

ID: 90059df935e843f3a59b3b86c36cee96
Topic: tooth paste
Language: Bengali
Hypothesis:
বাম ছবিতে দুটি টুথপেস্ট টিউব দেখায় এবং একটি টুথপেস্ট টিউব ডানদিকে দেখোনো হয়েছে
Machine Translation:
The left picture shows two toothpaste tubes and one toothpaste tube is shown on the right
Answer: True

Figure 9: Bengali M5-VGR Sample.

ID: acc84d8170bc4cf492af35235731f437
Topic: water outlet
Language: Berber
Hypothesis:
ⵙⵏⴰ ⵜⵍⵓⵣⵉ ⴰⵏ ⵙⵙⵣⵍⵏ ⴰⵙⵣⵙⵙ ⵏ ⵍⴰⵚⴰⵏ ⵏ ⵜⵅⵚⵚ.
Machine Translation:
ⵙⵏⴰ ⵜⵍⵓⵣⵉ ⴰⵏ ⵙⵙⵣⵍⵏ ⴰⵙⵣⵙⵙ ⵏ ⵍⴰⵚⴰⵏ ⵏ ⵜⵅⵚⵚ.
Answer: False

Figure 10: Berber M5-VGR Sample.

ID: 7b86a95366dd424e8d24597936e89434
Topic: paper
Language: English
Hypothesis:
The first image shows green paper in a printer, and the second image shows yellow paper on a wooden floor.
Machine Translation:
The first image shows green paper in a printer, and the second image shows yellow paper on a wooden floor.
Answer: False

Figure 11: English M5-VGR Sample.

ID: cc82590e83a846cb9edbebcf753055e6
Topic: water outlet
Language: Filipino
Hypothesis:
Ang pinagmumulan ng tubig ay marumi at nagkalat sa parehong larawan.
Machine Translation:
The source of the water is dirty and scattered in the same picture.
Answer: False

Figure 12: Filipino M5-VGR Sample.

ID: e0740bcbc03c406cb099eaa5c2040eda
Topic: table with food
Language: German
Hypothesis:
Das erste Bild zeigt eine Frau, die am Tisch Weintrauben isst, während das zweite Bild Essen für drei Personen zeigt.
Machine Translation:
The first image shows a woman eating grapes at the table, while the second image shows food for three people.
Answer: False

Figure 13: German M5-VGR Sample.

ID: 1668e4ad23d247909860a3d32eb2dba2
Topic: lock on front door
Language: Hausa
Hypothesis:
Dukka hotunan biyu kofar ɗaki ne wanda aka rufe da kwaɗon rufe ɗaki
Machine Translation:
Both pictures are a closed room with a closed door.
Answer: True

Figure 14: Hausa M5-VGR Sample.

ID: ba3a8016212e4ba58c2f8adeaa3a42ba
Topic: shower
Language: Hindi
Hypothesis:
दोनों तस्वीरें स्नान घर की है।
Machine Translation:
Both pictures are of the bath house.
Answer: False

Figure 15: Hindi M5-VGR Sample.

ID: 1610f5e020a9435f9e773ef424033e73
Topic: shower
Language: Russian
Hypothesis:
На первом изображении в душевой стены желтые, а на втором изображении в душевой стены красные.
Machine Translation:
In the first image, the shower walls are yellow, and in the second image in the shower walls, they are red.
Answer: False

Figure 16: Russian M5-VGR Sample.

ID: eb2af5af22c2418ea83c7d148c125687
Topic: wardrobe
Language: Swahili
Hypothesis:
Katika picha zote mbili kuna kabati la nguo.
Machine Translation:
In both pictures there is a dresser cupboard.
Answer: False

Figure 17: Swahili M5-VGR Sample.

ID: 53ecad00e365421b8cfc9c220468e9ca
Topic: washing clothes/cleaning
Language: Thai
Hypothesis:
ทั้งสองภาพเป็นภาพคนกำลังซักผ้า
Machine Translation:
Both images are of people doing laundry.
Answer: False

Figure 18: Thai M5-VGR Sample.

ID: c50c1001121a4454aed3b1884ff04167
Topic: guest bed
Language: Zulu
Hypothesis:
Isithombe sokuqala yigumbi elicocwe kahle elinezingubo zokulala ezimhlophe kanti isithombe sesibili yigumbi elingacocwe kahle elidlala ezingane.
Machine Translation:
The first picture is a well-cleaned room with white bedding and the second picture is a poorly cleaned room that plays with children.
Answer: False

Figure 19: Zulu M5-VGR Sample.

B.1.2 M5-VLOD Examples

ID: f07848da8e4544a8a34d6c3e8141e88c
Topic: source of cool
Language: Amharic
Hypothesis:
ሁሉም ምስሎች እራሳችን ለማቀዝቀዝ የምንጠቀምበት መሳሪያን ያሳያል።
Machine Translation:
All images show a device that we use to cool ourselves.
Outlier: 5

Figure 20: Amharic M5-VLOD Sample. The images are ordered from top-left to bottom-right.

ID: 11a37d8036f841d8ba028a501cf856c2
Topic: bedroom
Language: Bengali
Hypothesis:
সব ইমজে বডে রুম বছিানা কন্ট্যাটনিস
Machine Translation:
Bed room bed contins in all images
Outlier: 2

Figure 21: Bengali M5-VLOD Sample. The images are ordered from top-left to bottom-right.

ID: c5149f4ac81e439ea4be741e1f2e722d
Topic: cooking utensils
Language: Berber
Hypothesis:
+ΣⵍИѻӇΣI ѻΛ ⊙⊙ƸӇЧI+ ЌΣXѻI I ΣⵔѻИИI I Ƹ+ҀΣ.
Machine Translation:
+ΣⵍИѻӇΣI ѻΛ ⊙⊙ƸӇЧI+ ЌΣXѻI I ΣⵔѻИИI I Ƹ+ҀΣ.
Outlier: 4

Figure 22: Berber M5-VLOD Sample. The images are ordered from top-left to bottom-right.

ID: 843fac7edeff4fb4a2edc7c3ad1db388
Topic: drainage
Language: English
Hypothesis:
All images show a drain or drainage in a metal, ceramic surface or outside the house.
Machine Translation:
All images show a drain or drainage in a metal, ceramic surface or outside the house.
Outlier: 2

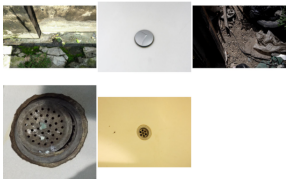
Figure 23: English M5-VLOD Sample. The images are ordered from top-left to bottom-right.

ID: 9af172b955dd4bb29e4d1c8601d504b2
Topic: armchair
Language: Filipino
Hypothesis:
Ang mga upuan sa mga larawan ay may armchair.
Machine Translation:
The chairs in the pictures have armchairs.
Outlier: 5

Figure 24: Filipino M5-VLOD Sample. The images are ordered from top-left to bottom-right.

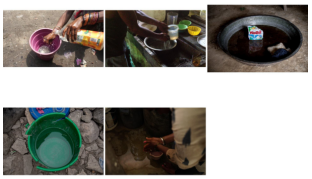
ID: 8f4008857b4c4bfab8135d40a9419219
Topic: plate of food
Language: German
Hypothesis:
Die Bilder zeigen Teller mit Essen, das gegessen wird.
Machine Translation:
The pictures show plates of food being eaten.
Outlier: 4

Figure 25: German M5-VLOD Sample. The images are ordered from top-left to bottom-right.



ID: 97ba6f364e38430eb779c56ad24cf89c
Topic: drainage
Language: Hausa
Hypothesis:
Dukka hotunan akwai hanyyoin magudanar ruwa na waje da cikin gida.
Machine Translation:
All images are available on the exterior and exterior of the house.
Outlier: 3

Figure 26: Hausa M5-VLOD Sample. The images are ordered from top-left to bottom-right.



ID: efa9c9642f4545849405e080a666ee56
Topic: hand washing
Language: Hindi
Hypothesis:
इन सभी छवियों में हैंडवाश करते हुए या हैंडवाश की चीज़ शामिल है
Machine Translation:
All of these images include handwashing or handwashing things
Outlier: 4

Figure 27: Hindi M5-VLOD Sample. The images are ordered from top-left to bottom-right.

ID: da46b1729e8b4871bd1c401d48fa4715
Topic: dish washing brush/cloth
Language: Russian
Hypothesis:
На изображении показан неупакованный предмет для чистки поверхностей.
Machine Translation:
The image shows an unwrapped surface cleaning item.
Outlier: 2

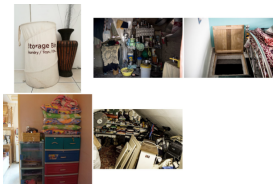
Figure 28: Russian M5-VLOD Sample. The images are ordered from top-left to bottom-right.

ID: a0f64574c38f45b888b18da6032d5547
Topic: bathroom privacy
Language: Swahili
Hypothesis:
Picha zote zinaonyesha faragha ya bafuni.
Machine Translation:
All photos show the privacy of the bathroom.
Outlier: 5

Figure 29: Swahili M5-VLOD Sample. The images are ordered from top-left to bottom-right.

ID: 47d9120f8ff541d19aeb988cab28d62b
Topic: dish racks
Language: Thai
Hypothesis:
ทุกภาพเป็นภาพที่วางจานแบบต่าง ๆ
Machine Translation:
Every picture is a picture of a different type of plate place.
Outlier: 1

Figure 30: Thai M5-VLOD Sample. The images are ordered from top-left to bottom-right.



ID: a8e7cc284e8c4d4794f5a811d09df92e
Topic: storage room
Language: Zulu
Hypothesis:
Izithombe zibonisa amagumbi agcwele izinto zasekhaya ezingasetyenziswa.
Machine Translation:
Pictures show rooms full of usable household items.
Outlier: 3

Figure 31: Zulu M5-VLOD Sample. The images are ordered from top-left to bottom-right.

B.1.3 Topics

Table 5: Number of images tagged with a certain topic in the M5-VGR (A) and M5-VLOD (B) datasets.

Topic	Language																							
	Amharic		Berber		Bengali		German		English		Filipino		Hausa		Hindi		Russian		Swahili		Thai		Zulu	
	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
armchair	1	2	1	1	1	1	1	1	1	2	3	1	1	1	1	1	2	1	3	1	1	1	1	1
backyard	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	1
bathroom privacy	1	1	3	3	1	1	2	1	1	1	3	4	1	1	1	1	1	1	2	1	1	1	1	1
bathroom/toilet	1	2	3	1	1	2	1	3	2	1	1	1	2	1	1	1	1	1	3	3	1	1	1	1
bed	1	1	1	2	2	1	1	1	1	3	1	2	4	1	1	1	2	2	1	1	4	1	1	1
bedroom	2	4	1	2	2	2	1	1	1	2	1	1	3	1	1	2	1	1	2	1	1	1	1	1
books	2	2	1	1	1	2	1	1	1	1	2	1	1	1	1	1	2	1	1	1	1	1	1	1
ceiling	1	2	1	1	2	1	1	1	2	2	1	1	1	4	2	1	2	1	2	2	2	1	1	2
children room	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1	2	1	1	1
cleaning equipment	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1
cooking pots	1	1	2	2	2	1	1	1	1	1	2	1	2	2	1	1	2	1	1	2	1	2	1	1
cooking utensils	1	1	3	2	1	3	1	1	1	1	1	1	1	3	2	1	1	1	1	1	2	1	1	1
couch	1	1	1	1	1	1	1	2	2	1	3	3	3	1	1	1	2	1	2	1	1	1	3	1
cups/mugs/glasses	1	1	1	1	1	1	1	1	3	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1
cutlery	1	1	1	1	1	1	1	1	2	1	3	1	1	1	1	2	1	1	1	1	1	2	1	1
dish racks	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	3	2	1	1
dish washing brush/cloth	2	1	1	1	3	1	1	1	1	3	1	1	1	1	3	2	1	2	1	1	1	1	1	1
dish washing soap	1	1	1	1	1	1	1	1	3	1	1	1	1	3	1	2	2	2	2	2	1	2	1	1
drainage	1	2	1	1	1	1	1	1	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1
drinking water	3	4	2	2	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	4	2
drying	3	1	1	1	5	1	1	1	1	1	2	2	1	2	3	1	1	1	1	1	1	1	1	1
everyday shoes	1	2	1	2	2	1	3	1	1	1	2	1	2	3	1	1	1	2	1	2	2	1	2	2
family	2	2	4	1	2	1	3	2	1	1	2	1	3	3	1	1	1	2	1	1	2	2	2	2
floor	1	1	3	1	2	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	1	1	1	1
front door	2	1	4	1	1	1	1	1	1	1	1	1	1	3	2	1	1	1	4	1	3	2	1	1
grains	2	1	1	1	2	1	2	1	1	1	1	1	1	1	1	2	2	2	2	1	1	1	1	1
guest bed	3	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	2	1	1	1	1	1	1
hair brush/comb	1	1	1	1	3	1	2	3	1	1	1	1	2	2	1	1	2	1	1	2	1	1	3	1
hand back	1	2	1	1	1	1	1	1	1	2	2	3	1	2	1	2	1	1	1	1	1	2	1	2
hand palm	1	1	3	2	2	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2	2	1	1	1
hand washing	2	1	3	2	1	1	5	1	1	1	4	2	1	1	3	2	2	2	1	2	1	2	3	
home	1	1	3	2	2	1	2	1	1	2	1	4	1	1	5	1	1	2	1	2	2	2	1	2
jewelry	1	1	1	1	1	2	1	1	1	2	1	1	1	2	1	1	1	1	2	1	1	1	1	1
kitchen	2	1	1	2	1	1	1	1	4	1	2	2	1	1	1	1	2	2	1	2	2	2	1	1
kitchen sink	1	2	2	2	1	1	4	2	1	2	2	1	1	2	1	2	1	1	1	3	1	1	3	3
light source in kitchen	1	1	2	1	2	3	2	2	3	2	1	1	1	1	3	1	1	1	2	1	1	1	1	1
light source in livingroom	1	2	2	2	1	1	1	1	2	1	1	2	1	1	1	1	2	1	1	1	1	1	1	1
living room	1	1	1	1	1	1	2	2	1	3	1	1	1	1	1	2	2	1	1	2	1	1	1	1
lock on front door	1	1	1	1	1	4	1	1	3	1	1	1	2	1	1	3	1	2	1	1	2	3	1	1
make up	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2	1
meat or fish	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1
medication	1	1	1	1	1	2	1	1	1	2	1	1	2	1	1	1	2	1	1	1	1	1	1	1
most loved item	1	1	1	1	1	2	3	1	2	2	3	3	2	1	2	2	2	2	2	0	1	1	4	4
most loved toy	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	2	1	1
nicest shoes	1	1	1	1	2	2	2	2	1	1	1	1	1	2	1	1	1	1	2	2	2	1	1	
oven	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	2	1	1	1	1	1	1
paper	2	1	1	2	1	2	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
pen/pencils	1	1	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1	2	3	1	1	1	1	1
phone	2	2	1	1	2	1	1	1	2	3	1	3	2	1	2	1	2	1	1	3	1	1	2	1
place where eating dinner	1	2	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	1	1	2	2	1	1
plate of food	2	1	1	4	1	1	2	3	1	1	2	1	3	1	1	1	1	2	1	1	1	1	1	1
plates	2	1	1	1	2	2	1	1	1	1	1	3	1	1	1	1	1	2	1	1	2	1	1	1
play area	1	1	2	2	1	1	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	2	1
power outlet	1	2	1	1	1	1	3	4	2	1	1	1	1	1	3	1	1	1	1	1	1	4	2	1
refrigerator	1	1	1	1	2	1	4	4	1	3	1	3	1	1	1	3	2	1	1	1	1	1	1	1
roof	2	1	1	1	1	3	2	2	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1
shampoo	1	2	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	2	1	1	2	1	1	1
shower	1	1	1	2	1	1	1	1	1	2	1	2	1	1	1	1	2	2	1	1	1	1	1	1
sitting area	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
soap for hands and body	1	1	2	2	2	2	1	2	1	1	1	1	2	1	2	1	2	1	1	1	2	2	1	1
social drink	1	1	1	1	1	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
sofa	1	1	1	1	1	1	1	1	1	2	1	1	2	4	1	1	1	1	1	1	2	1	1	1
source of cool	1	1	1	1	2	1	1	1	2	1	1	1	2	1	2	2	1	1	1	1	1	1	1	2
spices	1	2	1	1	2	1	1	3	3	3	2	1	1	1	2	1	2	2	1	2	1	2	1	3
storage room	1	2	1	1	1	2	1	1	1	1	5	1	2	1	1	2	1	2	1	1	1	1	1	1
stove/hob	2	1	1	2	1	1	1	1	1	3	1	1	1	5	1	2	1	1	2	3	2	1	1	4
street detail	4	1	1	1	1	1	1	1	1	3	2	1	1	2	1	2	2	1	2	1	2	1	1	1
street view	1	1	1	2	1	4	2	1	1	1	1	1	1	2	1	3	1	1	2	1	2	2	1	2
switch on/off	2	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	2	3	1	1
table with food	2	4	1	1	1	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
teeth	1	1	1	1	1	1	1	1	2	1	2	3	1	1	1	2	2	2	2	1	2	1	1	1
toilet	1	2	1	2	1	1	1	1	1	1	2	1	1	1	1	1	2	1	1	2	1	1	1	1
toilet paper	3	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	3	2
tooth paste	2	1	1	1	1	2	2	3	2	2	2	1	1	1	4	1	2	2	3	1	1	1	2	3
toothbrush	1	2	1	1	1	1	3	2	1	1	2	1	1	2	1	1	1	2	1	3	1	1	3	3
toys	2	1	2	1	3	5	1	1	1	3	1	2	2	1	1	4	2	3	1	1	2	3	1	1
trash/waste	1	1	1	1	1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1
tv	1	1	1	2	3	2	1	1	2	2	1	1	1	1	1	1	2	1	1	1	1	2	4	6
vegetables	1	2	2	1	2	1	1	1	3	1	1	1	1	1	1	1	2	2	1	1	1	1	1	1
wall	2	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
wall clock	2	1	1	2	1	1	1	1	1	1	4	1	1	1	2	2	1	1	1	1	1	2	1	1
wall decoration	1	2	1	1	1	1	2	1	1	1	1	1	2	2	1	1	1	2	2	0	2	1	1	1
wall inside	1	2	1	1	1	1	2	1	1	1	2	2	1	1	1	1	2	1	1	1	1	1	1	1
wardrobe	1	3	2	1	2	1	1	2	1	1	2	1	1	1	2	2	1	1	2	2	2	2	1	1
washing clothes/cleaning	1	1	1	2	1	1	1	1	2	1	1	1	1	3	1	1	1	4	4	1	3	1	1	1
washing detergent	2	1	1	1	1	2	1	1	1	1	2	1	1	1	2	1	1	2	2	1	2	1	1	1
water outlet	2	1	3	2	1	1	1	2	2	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1

B.2 Dataset Language Details

Table 6: Language support of the datasets considered in this work. More details one the languages are reported in Table 7.

Language	Script	MaXM	xGQA	XNLVI	MaRVL	M5-VLOD	M5-VGR	xFlickrCO	XM3600
Amharic	Ethiopic	no	no	no	no	yes	yes	no	no
Arabic	Arabic	no	no	yes	no	no	no	no	yes
Bengali	Bengali	no	yes	no	no	yes	yes	no	yes
Berber	Tifinagh	no	no	no	no	yes	yes	no	no
Chinese	Hanzi	yes	yes	no	yes	no	no	yes	yes
Croatian	Latin	no	no	no	no	no	no	no	yes
Czech	Latin	no	no	no	no	no	no	no	yes
Danish	Latin	no	no	no	no	no	no	no	yes
Dutch	Latin	no	no	no	no	no	no	no	yes
English	Latin	yes	yes	yes	no	yes	yes	yes	yes
Filipino	Latin	no	no	no	no	yes	yes	no	yes
Finnish	Latin	no	no	no	no	no	no	no	yes
French	Latin	yes	no	yes	no	no	no	no	yes
German	Latin	no	yes	no	no	yes	yes	yes	yes
Greek	Greek	no	no	no	no	no	no	no	yes
Hausa	Latin	no	no	no	no	yes	yes	no	no
Hebrew	Hebrew	yes	no	no	no	no	no	no	yes
Hindi	Devanagari	yes	no	no	no	yes	yes	no	yes
Hungarian	Latin	no	no	no	no	no	no	no	yes
Indonesian	Latin	no	yes	no	yes	no	no	yes	yes
Italian	Latin	no	no	no	no	no	no	no	yes
Japanese	Japanese	no	no	no	no	no	no	yes	yes
Korean	Hangul	no	yes	no	no	no	no	no	yes
Maori	Latin	no	no	no	no	no	no	no	yes
Norwegian	Latin	no	no	no	no	no	no	no	yes
Persian	Perso-Arabic	no	no	no	no	no	no	no	yes
Polish	Latin	no	no	no	no	no	no	no	yes
Portuguese	Latin	no	yes	no	no	no	no	no	yes
Quechua	Latin	no	no	no	no	no	no	no	yes
Romanian	Latin	yes	no	no	no	no	no	no	yes
Russian	Cyrillic	no	yes	yes	no	yes	yes	yes	yes
Spanish	Latin	no	no	yes	no	no	no	yes	yes
Swahili	Latin	no	no	no	yes	yes	yes	no	yes
Swedish	Latin	no	no	no	no	no	no	no	yes
Tamil	Tamil	no	no	no	yes	no	no	no	no
Telugu	Telugu	no	no	no	no	no	no	no	yes
Thai	Thai	yes	no	no	no	yes	yes	no	yes
Turkish	Latin	no	no	no	yes	no	no	yes	yes
Ukrainian	Cyrillic	no	no	no	no	no	no	no	yes
Vietnamese	Latin	no	no	no	no	no	no	no	yes
Zulu	Latin	no	no	no	no	yes	yes	no	no
Unique Languages		7	8	5	5	12	12	8	36
Unique Scripts		4	5	3	3	7	7	4	12

B.3 Language Details

Table 7: Details and statistics of languages comprised in the datasets of this benchmark. The continent and subregion columns refer to the content or subregion where the respective language is mostly spoken. The number of speakers is an estimate of the number of L1 and L2 speakers based on different public sources such as Wikipedia[10], Ethnologue [11], and Statista[12]. The "Taxonomy" column indicates the taxonomy class of the language based on Joshi et al. (2020).

Language	ISO 639	Lang. Family	Script	Continent	Subregion	Taxonomy	Speakers / 10^6
Arabic	ar	Afro-Asiatic	Arabic	Afrika & Asia	North Africa & Middle East	5	630.00
Chinese	zh	Sino-Tibetan	Hanzi	Asia	Northeastern Asia	5	1330.00
English	en	Indo-European	Latin	America	North America	5	1457.00
French	fr	Indo-European	Latin	Europe	Western Europe	5	310.00
German	de	Indo-European	Latin	Europe	Western Europe	5	175.00
Japanese	ja	Japonic	Japanese	Asia	Northeastern Asia	5	128.00
Spanish	es	Indo-European	Latin	Europe	Southern Europe	5	600.00
Croatian	hr	Indo-European	Latin	Europe	Central & Eastern Europe	4	6.80
Czech	cs	Indo-European	Latin	Europe	Central & Eastern Europe	4	11.00
Dutch	nl	Indo-European	Latin	Europe	Western Europe	4	30.00
Finnish	fi	Uralic	Latin	Europe	Northern Europe	4	5.80
Hindi	hi	Indo-European	Devanagari	Asia	Central & South Asia	4	600.00
Hungarian	hu	Uralic	Latin	Europe	Central & Eastern Europe	4	17.00
Italian	it	Indo-European	Latin	Europe	Southern Europe	4	68.00
Korean	ko	Koreanic	Hangul	Asia	Northeastern Asia	4	82.00
Persian	fa	Indo-European	Perso-Arabic	Asia	Middle East	4	130.00
Polish	pl	Indo-European	Latin	Europe	Central & Eastern Europe	4	41.00
Portuguese	pt	Indo-European	Latin	Europe & America	Southern Europe & South America	4	360.00
Russian	ru	Indo-European	Cyrillic	Asia	Central Asia	4	260.00
Swedish	sv	Indo-European	Latin	Europe	Northern Europe	4	13.00
Turkish	tr	Turkic	Latin	Asia	Middle East	4	90.00
Vietnamese	vi	Austroasiatic	Latin	Asia	Southeastern Asia	4	85.00
Bengali	bn	Indo-European	Bengali	Asia	Central & South Asia	3	270.00
Danish	da	Indo-European	Latin	Europe	Western Europe	3	6.00
Filipino	fil	Austronesian	Latin	Asia	Southeastern Asia	3	83.00
Greek	el	Indo-European	Greek	Europe	Central & Eastern Europe	3	13.50
Hebrew	he & iw	Afro-Asiatic	Hebrew	Asia	Middle East	3	9.00
Indonesian	id	Austronesian	Latin	Asia	Southeastern Asia	3	300.00
Romanian	ro	Indo-European	Latin	Europe	Central & Eastern Europe	3	28.50
Tamil	ta	Dravidian	Tamil	Asia	Central & South Asia	3	86.00
Thai	th	Kra-Dai	Thai	Asia	Southeastern Asia	3	80.00
Ukrainian	uk	Indo-European	Cyrillic	Europe	Central & Eastern Europe	3	32.80
Amharic	am	Afro-Asiatic	Ethiopic	Africa	Eastern Africa	2	57.00
Hausa	ha	Afro-Asiatic	Latin	Africa	Western Africa	2	79.00
Swahili	sw	Niger-Congo	Latin	Africa	Eastern Africa	2	73.00
Zulu	zu	Niger-Congo	Latin	Africa	Southern Africa	2	28.00
Maori	mi	Austronesian	Latin	Australia & Oceania	Australia & Oceania	1	0.19
Norwegian	no	Indo-European	Latin	Europe	Northern Europe	1	4.32
Quechua	quz	Quechuan	Latin	America	South America	1	9.00
Telugu	te	Dravidian	Telugu	Asia	Central & South Asia	1	96.00
Berber	ber	Afro-Asiatic	Tifinagh	Africa	Northern Africa	0	26.20

[3] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers
[4] https://www.ethnologue.com/
[5] https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/

C Model Details

Table 8: Architectural details of the LMMs evaluated in this study. The columns LM, VM, and ML are "**L**anguage **M**odel", "**V**ision **M**odel", and "**M**apping **M**odules", respectively, and show the number of parameters of the particular module. "|Totall" shows all parameters of the model. Note that we report friedly names of the models which are enriched with hyperlinks pointing to the respective Huggingface repositories (when viewed digitally). For Gemini Pro Vision and GPT-4 Vision, we used the `gemini-1.0-pro-vision` and `gpt-4-1106-vision-preview` variants, respectively.

Model	LM	VM	MM	|Total|	|LM|	|VM|	|MM|
MiniCPM-V [27; 50]	MiniCPM-2B	SigLIP 400M	MLP	3.43B	3.01B	397.75M	29.51M
mBliP mT0 [22]	Flan-T5-XL	EVA01 CLIP-ViT-g	QFormer	4.84B	3.74B	985.95M	106.71M
Yi-VL 6B [5]	Yi-6B-Chat	CLIP-ViT-H-14	MLP	6.71B	5.80B	631.75M	22.04M
LLaVA 1.6 7B [38]	Vicuna-7B-v1.5	CLIP-ViT-L	MLP	6.76B	6.61B	303.51M	20.98M
LLaVA 1.5 7B [39]	Vicuna-7B-v1.5	CLIP-ViT-L	MLP	7.06B	6.74B	303.51M	20.98M
BakLLaVA [39]	Mistral 7B v0.1	CLIP-ViT-L	MLP	7.57B	7.24B	303.51M	20.98M
mBliP BloomZ [22]	BloomZ 7B	EVA01 CLIP-ViT-g	QFormer	8.16B	7.07B	985.95M	108.29M
Qwen-VL [9]	Qwen-7B	CLIP-VIT-bigG	CrossAttn	9.66B	7.10B	1.94B	80.00M
OmniLMM 12B [50]	Zephyr 7B β	EVA02 CLIP ViT-E	MLP	11.61B	7.24B	4.28B	93.36M
LLaVA 1.6 13B [38]	Vicuna-13B-v1.5	CLIP-ViT-L	MLP	13.05B	12.85B	303.51M	31.47M
LLaVA 1.5 13B [39]	Vicuna-13B-v1.5	CLIP-ViT-L	MLP	13.35B	13.02B	303.51M	31.47M
CogVLM [48]	Vicuna-7B-v1.5	EVA02 CLIP ViT-E	CrossAttn	17.64B	6.74B	4.28B	6.62B
InternVL V1.1 [15]	Llama-2-13B	InternViT 6B	MLP	19.11B	13.12B	5.91B	91.79M
LLaVA 1.6 34B [38]	Nous-Hermes-2-Yi-34B	CLIP-ViT-L	MLP	34.45B	33.93B	303.51M	58.73M
Yi-VL 34B [5]	Yi-34B-Chat	CLIP-ViT-H	MLP	35.08B	33.93B	631.75M	60.60M
InternVL V1.2+ [15]	Nous-Hermes-2-Yi-34B	InternViT-6B V1-2	MLP	40.07B	34.39B	5.54B	143.17M
Gemini Pro Vision [7]	?	?	?	?	?	?	?
GPT-4 Vision [40]	?	?	?	?	?	?	?

D Results Details

D.1 General Results

D.1.1 xGQA

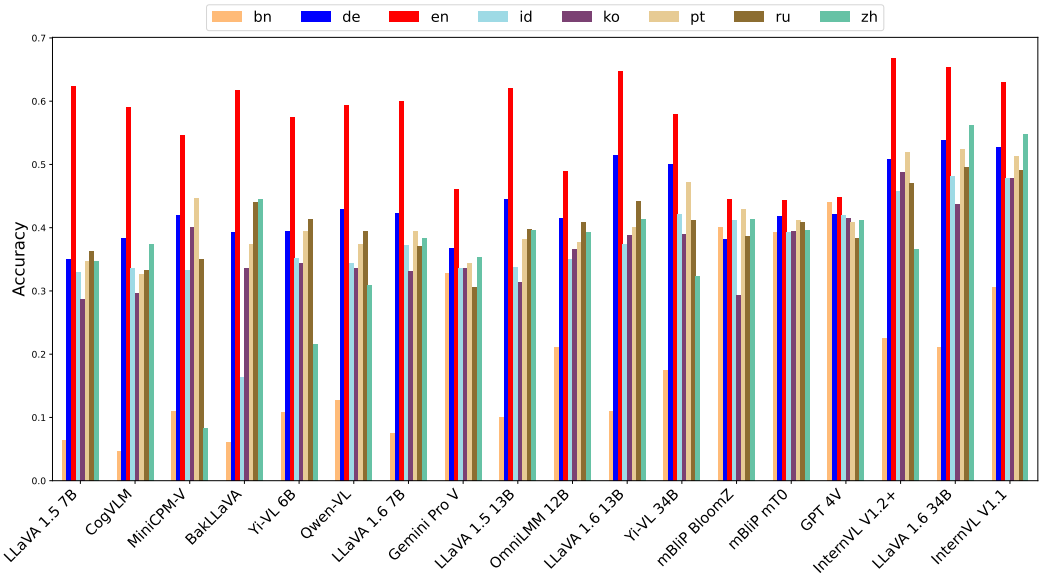


Figure 32: A bar plot showing the average accuracy per language and model on the xGQA dataset. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 9: The average accuracy per language and model on the xGQA dataset. The column "NEA" stands for the average of Non-English languages.

Model	Language								
	bn	de	en	id	ko	pt	ru	zh	NEA
LLaVA 1.5 7B	0.06	0.35	0.62	0.33	0.29	0.35	0.36	0.35	0.30
CogVLM	0.05	0.38	0.59	0.34	0.30	0.33	0.33	0.37	0.30
MiniCPM-V	0.11	0.42	0.55	0.33	0.40	0.45	0.35	0.08	0.31
BakLLaVA	0.06	0.39	0.62	0.16	0.34	0.37	0.44	0.45	0.32
Yi-VL 6B	0.11	0.39	0.57	0.35	0.34	0.39	0.41	0.22	0.32
Qwen-VL	0.13	0.43	0.59	0.34	0.34	0.37	0.39	0.31	0.33
LLaVA 1.6 7B	0.07	0.42	0.60	0.37	0.33	0.39	0.37	0.38	0.34
Gemini Pro V	0.33	0.37	0.46	0.34	0.34	0.34	0.31	0.35	0.34
LLaVA 1.5 13B	0.10	0.44	0.62	0.34	0.31	0.38	0.40	0.40	0.34
OmniLMM 12B	0.21	0.42	0.49	0.35	0.37	0.38	0.41	0.39	0.36
LLaVA 1.6 13B	0.11	0.52	0.65	0.37	0.39	0.40	0.44	0.41	0.38
Yi-VL 34B	0.18	0.50	0.58	0.42	0.39	0.47	0.41	0.32	0.38
mBliP BloomZ	0.40	0.38	0.44	0.41	0.29	0.43	0.39	0.41	0.39
mBliP mT0	0.39	0.42	0.44	0.39	0.39	0.41	0.41	0.40	0.40
GPT 4V	0.44	0.42	0.45	0.42	0.41	0.41	0.38	0.41	0.41
InternVL V1.2+	0.22	0.51	0.67	0.46	0.49	0.52	0.47	0.37	0.43
LLaVA 1.6 34B	0.21	0.54	0.65	0.48	0.44	0.52	0.50	0.56	0.46
InternVL V1.1	0.31	0.53	0.63	0.48	0.48	0.51	0.49	0.55	0.48
Average	0.19	0.43	0.57	0.37	0.37	0.41	0.40	0.37	0.37

D.1.2 MaXM

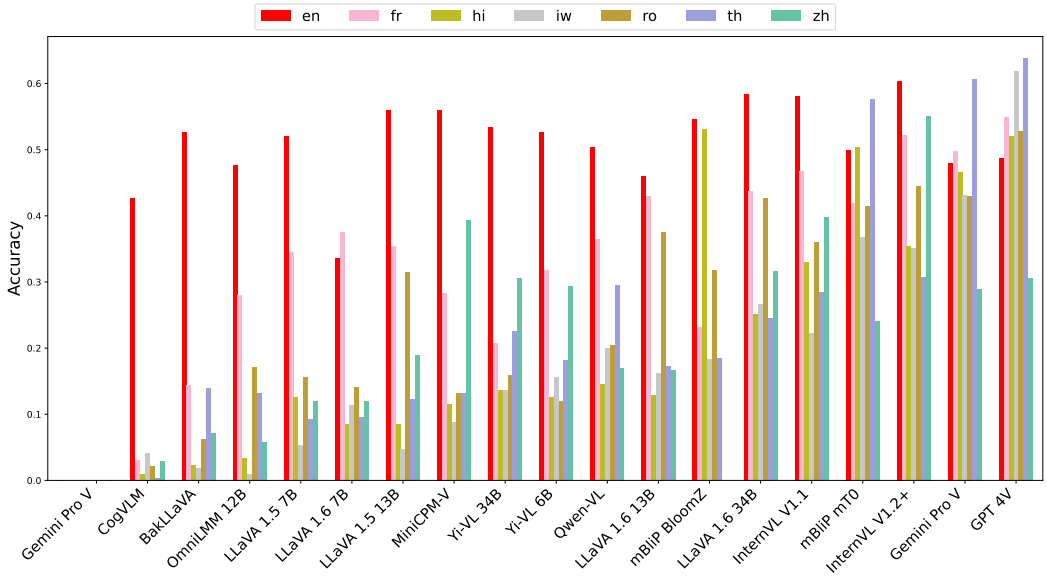


Figure 33: A bar plot showing the average accuracy per language and model on the MaXM dataset. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 10: The average accuracy per language and model on the MaXM dataset. The column "NEA" stands for the average of Non-English languages.

Model	Language							
	en	fr	hi	iw	ro	th	zh	NEA
CogVLM	0.43	0.03	0.01	0.04	0.02	0.00	0.03	0.02
BakLLaVA	0.53	0.14	0.02	0.02	0.06	0.14	0.07	0.08
OmniLMM 12B	0.48	0.28	0.03	0.01	0.17	0.13	0.06	0.11
LLaVA 1.5 7B	0.52	0.34	0.13	0.05	0.16	0.09	0.12	0.15
LLaVA 1.6 7B	0.34	0.38	0.09	0.11	0.14	0.10	0.12	0.16
LLaVA 1.5 13B	0.56	0.35	0.09	0.05	0.32	0.12	0.19	0.19
MiniCPM-V	0.56	0.28	0.12	0.09	0.13	0.13	0.39	0.19
Yi-VL 34B	0.53	0.21	0.14	0.14	0.16	0.23	0.31	0.20
Yi-VL 6B	0.53	0.32	0.13	0.16	0.12	0.18	0.29	0.20
Qwen-VL	0.50	0.37	0.15	0.20	0.20	0.29	0.17	0.23
LLaVA 1.6 13B	0.46	0.43	0.13	0.16	0.38	0.17	0.17	0.24
mBliP BloomZ	0.55	0.23	0.53	0.18	0.32	0.19	0.42	0.31
LLaVA 1.6 34B	0.58	0.44	0.25	0.27	0.43	0.25	0.32	0.32
InternVL V1.1	0.58	0.47	0.33	0.22	0.36	0.28	0.40	0.34
mBliP mT0	0.50	0.42	0.50	0.37	0.41	0.58	0.24	0.42
InternVL V1.2+	0.60	0.52	0.35	0.35	0.44	0.31	0.55	0.42
Gemini Pro V	0.48	0.50	0.47	0.43	0.43	0.61	0.29	0.45
GPT 4V	0.49	0.55	0.52	0.62	0.53	0.64	0.31	0.53
Average	0.51	0.35	0.22	0.19	0.27	0.25	0.24	0.25

D.1.3 XVNLI

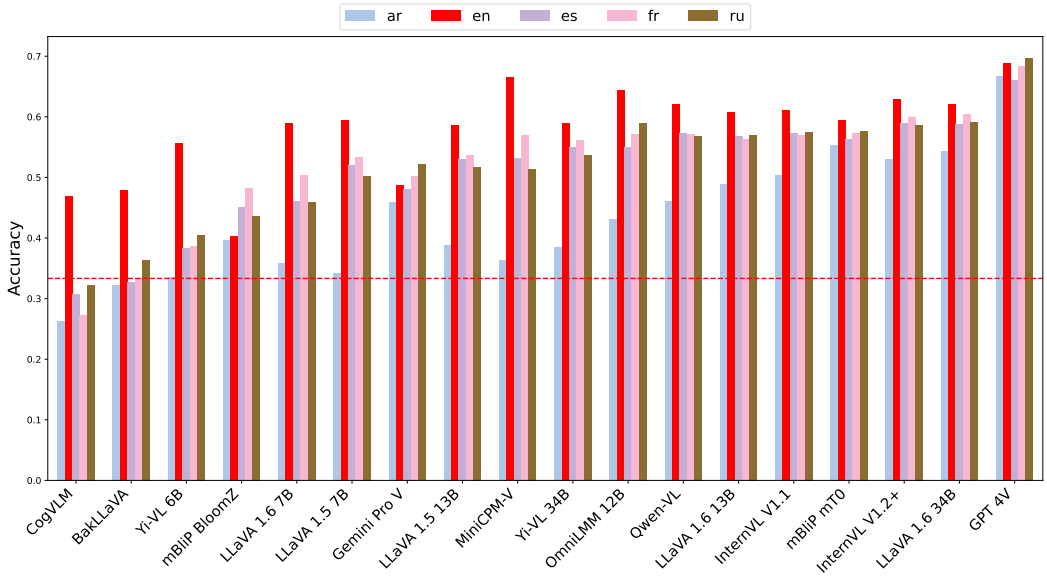


Figure 34: A bar plot showing the average accuracy per language and model on the XVNLI dataset. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 11: The average accuracy per language and model on the XVNLI dataset. The column "NEA" stands for the average of Non-English languages.

Model	Language					
	ar	en	es	fr	ru	NEA
CogVLM	0.26	0.47	0.31	0.27	0.32	0.29
BakLLaVA	0.32	0.48	0.33	0.33	0.36	0.34
Yi-VL 6B	0.34	0.56	0.38	0.39	0.41	0.38
mBliP BloomZ	0.40	0.40	0.45	0.48	0.44	0.44
LLaVA 1.6 7B	0.36	0.59	0.46	0.50	0.46	0.45
LLaVA 1.5 7B	0.34	0.60	0.52	0.53	0.50	0.47
Gemini Pro V	0.46	0.49	0.48	0.50	0.52	0.49
LLaVA 1.5 13B	0.39	0.59	0.53	0.54	0.52	0.49
MiniCPM-V	0.36	0.66	0.53	0.57	0.51	0.49
Yi-VL 34B	0.39	0.59	0.55	0.56	0.54	0.51
OmniLMM 12B	0.43	0.64	0.55	0.57	0.59	0.54
Qwen-VL	0.46	0.62	0.57	0.57	0.57	0.54
LLaVA 1.6 13B	0.49	0.61	0.57	0.56	0.57	0.55
InternVL V1.1	0.50	0.61	0.57	0.57	0.57	0.56
mBliP mT0	0.55	0.59	0.56	0.57	0.58	0.57
InternVL V1.2+	0.53	0.63	0.59	0.60	0.59	0.58
LLaVA 1.6 34B	0.54	0.62	0.59	0.60	0.59	0.58
GPT 4V	0.67	0.69	0.66	0.68	0.70	0.68
Average	0.43	0.58	0.51	0.52	0.52	0.50

D.1.4 MaRVL

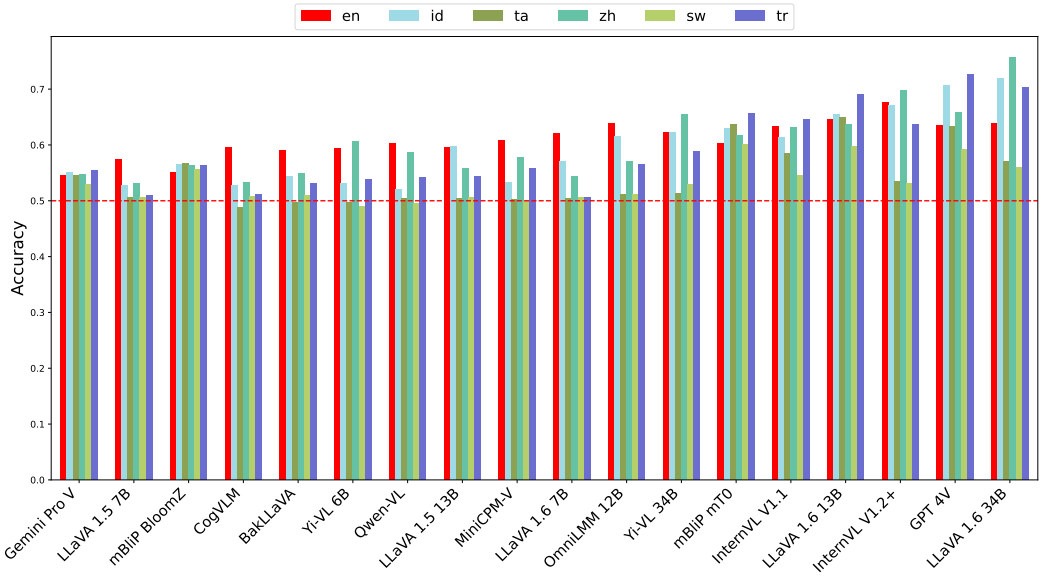


Figure 35: A bar plot showing the average accuracy per language and model on the MaRVL dataset. Note that MaRVL does not contain English data originally and we machine-translated English from the other languages and averaged the results. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 12: The average accuracy per language and model on the MaRVL dataset. Note that MaRVL does not contain English data originally and we machine-translated English from the other languages and averaged the results. The column "NEA" stands for the average of Non-English languages.

Model	Language						
	en	id	sw	ta	tr	zh	NEA
CogVLM	0.60	0.53	0.51	0.49	0.51	0.53	0.51
LLaVA 1.5 7B	0.57	0.53	0.51	0.51	0.51	0.53	0.52
BakLLaVA	0.59	0.54	0.51	0.50	0.53	0.55	0.53
LLaVA 1.6 7B	0.62	0.57	0.51	0.50	0.51	0.54	0.53
Qwen-VL	0.60	0.52	0.50	0.50	0.54	0.59	0.53
Yi-VL 6B	0.59	0.53	0.49	0.50	0.54	0.61	0.53
MiniCPM-V	0.61	0.53	0.50	0.50	0.56	0.58	0.53
LLaVA 1.5 13B	0.60	0.60	0.51	0.50	0.54	0.56	0.54
Gemini Pro V	0.55	0.55	0.53	0.55	0.56	0.55	0.55
OmniLMM 12B	0.64	0.62	0.51	0.51	0.57	0.57	0.56
mBliP BloomZ	0.55	0.57	0.56	0.57	0.56	0.56	0.56
Yi-VL 34B	0.62	0.62	0.53	0.51	0.59	0.65	0.58
InternVL V1.1	0.63	0.61	0.54	0.58	0.65	0.63	0.60
InternVL V1.2+	0.68	0.67	0.53	0.53	0.64	0.70	0.61
mBliP mT0	0.60	0.63	0.60	0.64	0.66	0.62	0.63
LLaVA 1.6 13B	0.65	0.66	0.60	0.65	0.69	0.64	0.65
LLaVA 1.6 34B	0.64	0.72	0.56	0.57	0.70	0.76	0.66
GPT 4V	0.64	0.71	0.59	0.63	0.73	0.66	0.66
Average	0.61	0.60	0.53	0.54	0.59	0.60	0.57

D.1.5 M5-VGR

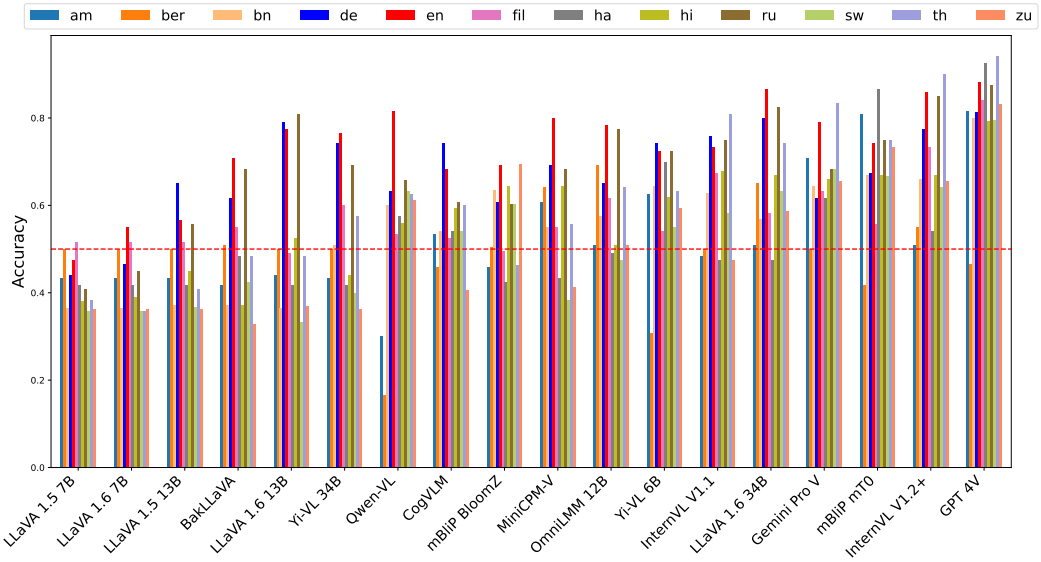


Figure 36: A bar plot showing the average accuracy per language and model on the M5-VGR dataset. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 13: The average accuracy per language and model on the M5-VGR dataset. The column "NEA" stands for the average of Non-English languages.

Model	Language												
	am	ber	bn	de	en	fil	ha	hi	ru	sw	th	zu	NEA
LLaVA 1.5 7B	0.43	0.50	0.36	0.44	0.47	0.52	0.42	0.38	0.41	0.36	0.38	0.36	0.42
LLaVA 1.6 7B	0.43	0.50	0.36	0.47	0.55	0.52	0.42	0.39	0.45	0.36	0.36	0.36	0.42
LLaVA 1.5 13B	0.43	0.50	0.37	0.65	0.57	0.52	0.42	0.45	0.56	0.37	0.41	0.36	0.46
BakLLaVA	0.42	0.51	0.37	0.62	0.71	0.55	0.48	0.37	0.68	0.42	0.48	0.33	0.48
LLaVA 1.6 13B	0.44	0.50	0.36	0.79	0.78	0.49	0.42	0.53	0.81	0.33	0.48	0.37	0.50
Yi-VL 34B	0.43	0.50	0.51	0.74	0.77	0.60	0.42	0.44	0.69	0.40	0.57	0.36	0.52
Qwen-VL	0.30	0.17	0.60	0.63	0.82	0.53	0.57	0.56	0.66	0.63	0.62	0.61	0.54
CogVLM	0.53	0.46	0.54	0.74	0.68	0.53	0.54	0.59	0.61	0.54	0.60	0.41	0.55
mBliP BloomZ	0.46	0.50	0.64	0.61	0.69	0.50	0.42	0.64	0.60	0.60	0.46	0.69	0.56
MiniCPM-V	0.61	0.64	0.55	0.69	0.80	0.55	0.43	0.64	0.68	0.38	0.56	0.41	0.56
OmniLMM 12B	0.51	0.69	0.58	0.65	0.78	0.62	0.49	0.51	0.78	0.47	0.64	0.51	0.59
Yi-VL 6B	0.62	0.31	0.64	0.74	0.72	0.54	0.70	0.62	0.72	0.55	0.63	0.59	0.61
InternVL V1.1	0.48	0.50	0.63	0.76	0.73	0.68	0.47	0.68	0.75	0.58	0.81	0.47	0.62
LLaVA 1.6 34B	0.51	0.65	0.57	0.80	0.87	0.58	0.47	0.67	0.82	0.63	0.74	0.59	0.64
Gemini Pro V	0.71	0.50	0.64	0.62	0.79	0.63	0.62	0.66	0.68	0.68	0.83	0.66	0.66
InternVL V1.2+	0.51	0.55	0.66	0.78	0.86	0.73	0.54	0.67	0.85	0.64	0.90	0.66	0.68
mBliP mT0	0.81	0.42	0.67	0.68	0.74	0.56	0.87	0.67	0.75	0.67	0.75	0.73	0.69
GPT 4V	0.82	0.47	0.80	0.81	0.88	0.84	0.93	0.79	0.88	0.80	0.94	0.83	0.81
Average	0.53	0.49	0.55	0.68	0.73	0.58	0.53	0.57	0.69	0.52	0.62	0.52	0.57

D.1.6 M5-VLOD

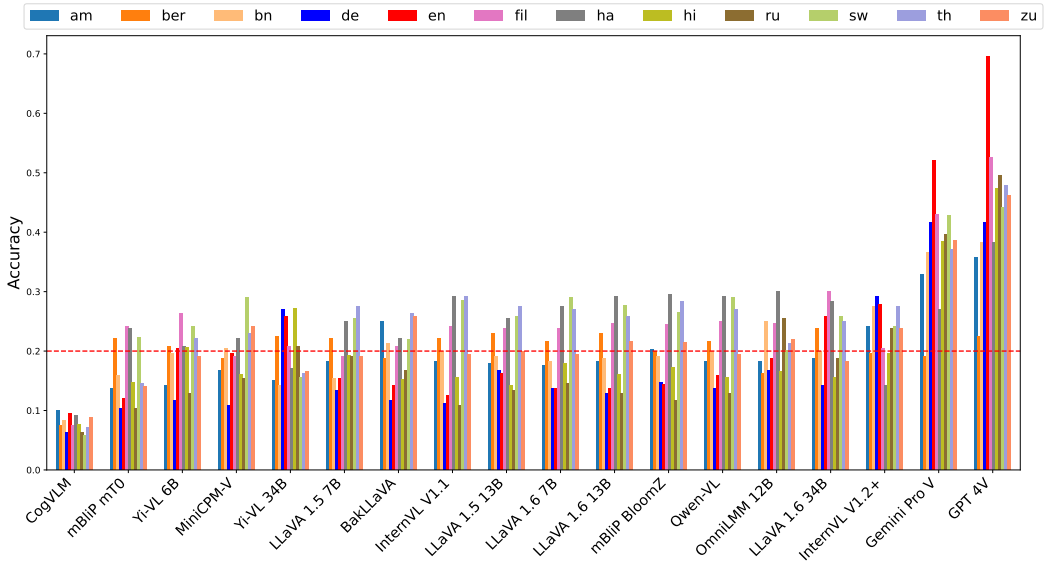


Figure 37: A bar plot showing the average accuracy per language and model on the M5-VLOD dataset. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 14: The average accuracy per language and model on the M5-VLOD dataset. The column "NEA" stands for the average of Non-English languages.

Model	Language												
	am	ber	bn	de	en	fil	ha	hi	ru	sw	th	zu	NEA
CogVLM	0.10	0.07	0.08	0.06	0.10	0.07	0.09	0.08	0.06	0.06	0.07	0.09	0.08
mBliP mT0	0.14	0.22	0.16	0.10	0.12	0.24	0.24	0.15	0.10	0.22	0.15	0.14	0.17
Yi-VL 6B	0.14	0.21	0.20	0.12	0.20	0.26	0.21	0.21	0.13	0.24	0.22	0.19	0.19
Yi-VL 34B	0.15	0.22	0.14	0.27	0.26	0.21	0.17	0.27	0.21	0.16	0.16	0.17	0.19
MiniCPM-V	0.17	0.19	0.20	0.11	0.20	0.19	0.22	0.16	0.15	0.29	0.23	0.24	0.20
LLaVA 1.5 7B	0.18	0.22	0.15	0.13	0.15	0.19	0.25	0.19	0.19	0.25	0.27	0.19	0.20
BakLLaVA	0.25	0.19	0.21	0.12	0.14	0.21	0.22	0.15	0.17	0.22	0.26	0.26	0.20
LLaVA 1.5 13B	0.18	0.23	0.19	0.17	0.16	0.24	0.25	0.14	0.13	0.26	0.28	0.20	0.21
InternVL V1.1	0.18	0.22	0.20	0.11	0.12	0.24	0.29	0.16	0.11	0.29	0.29	0.19	0.21
LLaVA 1.6 7B	0.17	0.22	0.18	0.14	0.14	0.24	0.27	0.18	0.15	0.29	0.27	0.19	0.21
LLaVA 1.6 13B	0.18	0.23	0.19	0.13	0.14	0.25	0.29	0.16	0.13	0.28	0.26	0.22	0.21
Qwen-VL	0.18	0.22	0.20	0.14	0.16	0.25	0.29	0.16	0.13	0.29	0.27	0.19	0.21
mBliP BloomZ	0.20	0.20	0.19	0.15	0.14	0.24	0.29	0.17	0.12	0.26	0.28	0.21	0.21
OmniLMM 12B	0.18	0.16	0.25	0.17	0.19	0.25	0.30	0.17	0.25	0.20	0.21	0.22	0.21
LLaVA 1.6 34B	0.19	0.24	0.20	0.14	0.26	0.30	0.28	0.16	0.19	0.26	0.25	0.18	0.22
InternVL V1.2+	0.24	0.20	0.28	0.29	0.28	0.20	0.14	0.20	0.24	0.24	0.28	0.24	0.23
Gemini Pro V	0.33	0.19	0.37	0.42	0.52	0.43	0.27	0.38	0.40	0.43	0.37	0.39	0.36
GPT 4V	0.36	0.22	0.38	0.42	0.70	0.53	0.38	0.47	0.50	0.44	0.48	0.46	0.42
Average	0.20	0.20	0.21	0.18	0.22	0.25	0.25	0.20	0.19	0.26	0.26	0.22	0.22

D.1.7 xFlickrCO

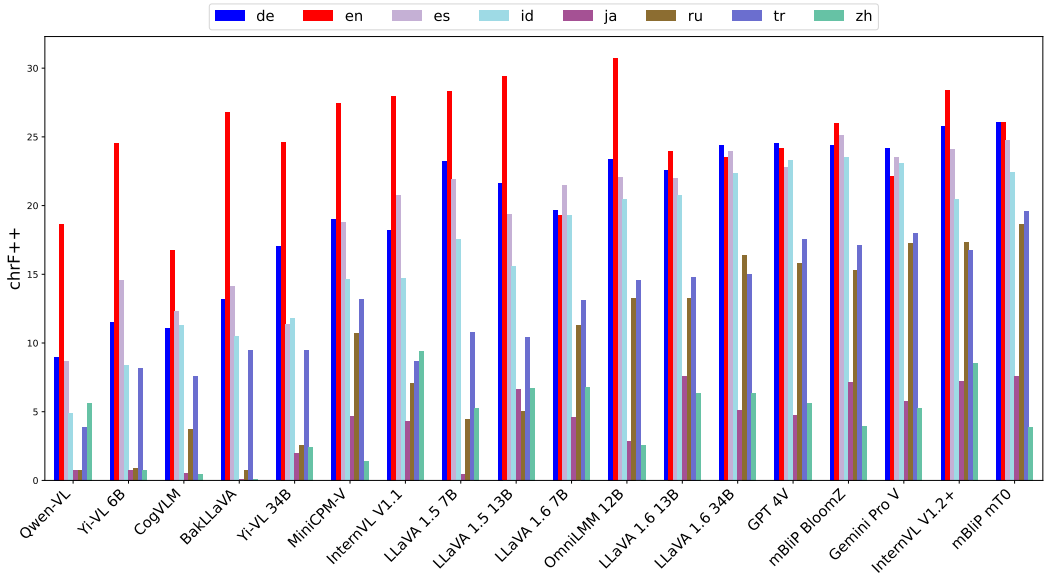


Figure 38: A bar plot showing the average chrF++ score per language and model on the xFlickrCO dataset. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 15: The average chrF++ score per language and model on the xFlickrCO dataset. The column "NEA" stands for the average of Non-English languages.

Model	Language								
	de	en	es	id	ja	ru	tr	zh	NEA
Qwen-VL	9.00	18.68	8.69	4.88	0.77	0.74	3.91	5.62	4.80
Yi-VL 6B	11.53	24.54	14.61	8.37	0.78	0.90	8.15	0.79	6.45
CogVLM	11.08	16.76	12.32	11.27	0.56	3.71	7.62	0.46	6.72
BakLLaVA	13.21	26.79	14.17	10.48	0.06	0.75	9.49	0.09	6.89
Yi-VL 34B	17.02	24.62	11.36	11.79	2.00	2.57	9.50	2.44	8.10
MiniCPM-V	19.05	27.43	18.81	14.62	4.69	10.73	13.18	1.40	11.78
InternVL V1.1	18.21	27.98	20.74	14.69	4.31	7.07	8.67	9.38	11.87
LLaVA 1.5 7B	23.22	28.32	21.95	17.58	0.44	4.45	10.77	5.29	11.96
LLaVA 1.5 13B	21.66	29.39	19.37	15.59	6.63	5.02	10.45	6.72	12.21
LLaVA 1.6 7B	19.70	19.31	21.48	19.32	4.60	11.27	13.14	6.78	13.75
OmniLMM 12B	23.39	30.76	22.05	20.50	2.89	13.29	14.55	2.59	14.18
LLaVA 1.6 13B	22.55	23.94	21.98	20.73	7.57	13.26	14.79	6.39	15.33
LLaVA 1.6 34B	24.38	23.52	23.98	22.36	5.08	16.40	15.05	6.34	16.23
GPT 4V	24.56	24.17	22.82	23.29	4.73	15.82	17.58	5.60	16.34
mBliP BloomZ	24.39	25.99	25.12	23.56	7.18	15.31	17.16	3.93	16.67
Gemini Pro V	24.17	22.13	23.50	23.10	5.75	17.28	18.03	5.24	16.73
InternVL V1.2+	25.81	28.41	24.13	20.48	7.25	17.34	16.73	8.54	17.18
mBliP mT0	26.10	26.07	24.74	22.41	7.56	18.64	19.58	3.87	17.56
Average	19.95	24.93	19.55	16.95	4.05	9.70	12.69	4.53	12.49

D.1.8 XM3600

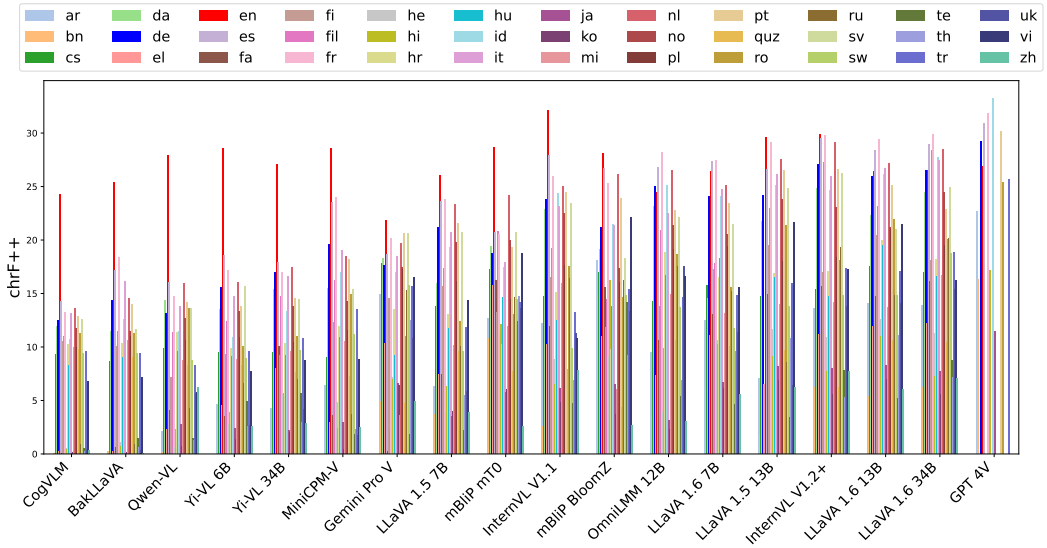


Figure 39: A bar plot showing the average chrF++ score per language and model on the XM3600 dataset. Due to resource restrictions, we evaluated GPT 4V only on a subset of languages. The models on the x-Axis are ordered by their average score across all languages so that the best performing model is on the right and the worst is on the left.

Table 16: The average chrF++ score per language and model on the XM3600 dataset. Due to resource restrictions, we evaluated GPT 4V only on a subset of languages. The column "NEA" stands for the average of Non-English languages.

Model	Language											
	ar	bn	cs	da	de	el	en	es	fa	fi	fil	fr
CogVLM	0.07	0.04	9.30	11.92	12.50	0.25	24.26	14.25	0.02	10.52	10.96	13.18
BakLLaVA	0.21	0.22	8.65	11.45	14.33	0.24	25.39	17.13	0.64	10.02	11.41	18.33
Qwen-VL	2.08	0.17	9.89	14.38	13.14	2.32	27.89	16.00	4.09	7.13	11.36	14.70
Yi-VL 6B	4.65	2.98	9.48	13.55	15.58	4.54	28.59	18.58	3.50	9.29	12.42	17.12
Yi-VL 34B	4.24	4.14	9.52	15.40	17.00	8.00	27.11	17.86	10.06	9.17	14.73	16.93
MiniCPM-V	6.38	1.96	9.05	15.52	19.60	2.98	28.53	23.54	3.57	12.33	16.19	23.98
Gemini Pro V	14.90	4.94	17.79	18.32	17.63	10.36	21.81	18.64	0.21	14.50	2.25	20.15
LLaVA 1.5 7B	6.30	3.71	13.80	15.93	21.18	7.42	26.02	23.60	7.45	15.67	17.38	23.83
mBliP mT0	12.68	10.79	17.20	19.43	18.74	15.76	28.68	20.71	16.19	13.26	20.79	20.52
InternVL V1.1	12.23	2.55	14.74	22.82	23.77	10.20	32.10	27.91	11.94	16.47	19.20	25.95
mBliP BloomZ	18.10	14.92	16.99	19.16	21.17	11.03	28.05	26.73	15.59	11.86	14.47	25.28
OmniLMM 12B	9.48	3.51	14.24	23.15	25.05	7.37	24.42	26.75	10.65	13.78	20.92	28.18
LLaVA 1.6 7B	12.52	6.13	15.79	14.50	24.06	11.11	26.41	27.37	13.07	17.23	17.76	27.48
LLaVA 1.5 13B	7.07	1.80	14.75	21.74	24.15	6.49	29.55	26.59	14.90	19.51	22.91	29.14
InternVL V1.2+	13.59	6.19	15.34	24.85	27.05	11.20	29.84	29.50	15.69	17.01	27.22	29.80
LLaVA 1.6 13B	14.07	5.42	17.51	22.30	25.95	11.90	26.42	28.39	14.72	20.44	23.14	29.42
LLaVA 1.6 34B	13.85	6.20	16.94	24.44	26.51	12.17	26.52	28.90	16.09	18.08	28.35	29.83
GPT 4V	22.67	16.27	-	-	29.24	-	26.89	30.86	-	-	-	31.82
Average	9.73	5.11	12.83	17.16	20.92	7.41	27.14	23.52	8.80	13.13	16.19	23.65

Model	Language											
	he	hi	hr	hu	id	it	ja	ko	mi	nl	no	pl
CogVLM	0.52	0.38	10.25	8.25	10.70	13.11	0.07	0.13	10.00	13.59	11.73	9.98
BakLLaVA	1.07	0.71	10.33	8.98	12.59	16.12	0.07	0.16	10.62	14.56	11.48	10.97
Qwen-VL	0.58	2.32	11.33	9.60	11.50	13.76	2.75	0.70	8.73	15.91	12.64	10.59
Yi-VL 6B	2.78	3.86	9.82	9.12	10.90	14.69	2.40	1.32	8.81	16.04	13.30	10.88
Yi-VL 34B	5.58	5.64	10.31	9.23	13.30	16.55	2.21	2.02	9.55	17.43	13.79	10.40
MiniCPM-V	4.86	2.36	11.96	10.91	16.94	19.06	2.92	0.39	10.49	18.47	14.27	11.51
Gemini Pro V	7.12	6.98	13.48	9.22	16.98	18.44	6.63	6.43	3.55	19.67	17.43	17.29
LLaVA 1.5 7B	3.76	6.29	13.05	11.69	19.33	20.73	3.48	3.93	10.10	23.30	19.79	16.10
mBliP mT0	11.16	12.08	10.26	14.59	17.39	17.92	5.79	6.00	11.88	24.20	19.97	14.49
InternVL V1.1	8.80	6.47	15.05	12.49	24.31	23.13	6.09	4.83	15.93	25.02	22.45	17.58
mBliP BloomZ	9.16	16.18	9.78	13.84	21.44	21.39	6.53	3.67	5.99	26.17	17.35	16.07
OmniLMM 12B	3.99	9.91	18.84	16.72	25.07	22.50	3.16	2.31	14.94	26.47	21.36	19.16
LLaVA 1.6 7B	10.61	10.26	16.52	18.26	24.05	24.71	6.66	6.09	13.12	25.07	20.49	19.38
LLaVA 1.5 13B	11.63	9.13	16.87	16.54	25.13	26.11	8.16	6.86	13.98	27.52	23.77	17.96
InternVL V1.2+	10.88	7.69	17.07	14.70	24.65	25.94	7.96	5.53	14.17	29.11	23.02	18.37
LLaVA 1.6 13B	12.54	11.00	19.99	19.52	26.15	26.66	8.27	6.95	13.73	27.15	21.19	21.03
LLaVA 1.6 34B	11.30	7.27	18.16	16.57	27.69	27.40	7.75	5.60	16.69	28.42	24.45	19.49
GPT 4V	-	17.16	-	-	33.24	-	11.46	-	-	-	-	-
Average	6.46	7.54	12.95	12.23	20.08	19.35	5.13	3.50	10.68	21.01	17.14	14.51

Model	Language												
	pt	quz	ro	ru	sv	sw	te	th	tr	uk	vi	zh	NEA
CogVLM	12.87	9.75	11.23	0.86	12.57	9.41	0.51	0.26	9.58	0.46	6.74	0.29	7.04
BakLLaVA	14.00	9.00	11.30	0.85	11.61	9.37	1.47	0.57	9.36	0.31	7.11	0.03	7.58
Qwen-VL	14.17	8.25	13.60	4.30	13.59	8.75	1.44	1.28	8.26	5.66	5.76	6.20	8.20
Yi-VL 6B	13.77	8.25	10.04	6.57	15.64	8.94	4.93	2.57	9.55	2.65	7.76	2.61	8.82
Yi-VL 34B	14.57	7.64	10.95	6.95	14.42	9.71	5.62	2.92	10.84	4.19	8.74	2.82	9.78
MiniCPM-V	18.21	7.21	14.94	3.69	15.36	11.16	1.83	2.24	13.47	1.74	8.88	2.46	10.30
Gemini Pro V	20.60	4.72	10.98	15.27	20.60	15.80	1.87	12.45	15.62	10.82	16.48	4.88	12.37
LLaVA 1.5 7B	21.57	9.55	12.38	10.08	20.68	9.59	2.23	5.51	11.78	5.84	14.34	3.87	12.44
mBliP mT0	19.35	7.70	13.05	14.63	20.66	14.45	12.42	14.76	14.13	13.60	18.73	2.59	14.80
InternVL V1.1	24.47	7.91	17.55	16.39	23.40	9.82	4.73	6.85	13.22	11.26	10.80	7.76	14.97
mBliP BloomZ	23.93	4.32	14.59	16.25	18.31	14.82	14.12	9.19	15.34	13.35	22.14	2.65	15.20
OmniLMM 12B	22.75	10.61	18.61	17.49	22.09	13.68	5.41	6.84	14.68	17.49	16.58	3.00	15.34
LLaVA 1.6 7B	23.42	10.04	15.55	15.18	21.42	11.69	4.60	9.62	14.81	11.40	15.54	5.58	15.46
LLaVA 1.5 13B	26.51	9.70	21.33	8.53	24.80	13.81	3.39	10.84	15.98	6.36	21.66	6.22	16.05
InternVL V1.2+	26.63	6.20	18.06	19.30	26.27	14.83	7.79	5.30	17.30	13.79	17.22	7.71	17.05
LLaVA 1.6 13B	25.07	10.60	21.96	14.86	21.01	14.80	5.18	11.11	17.03	14.03	21.44	6.02	17.44
LLaVA 1.6 34B	22.85	10.39	20.08	20.11	24.92	18.73	8.70	7.19	18.83	15.36	16.23	7.02	17.79
GPT 4V	30.13	-	25.41	-	-	-	-	-	25.70	-	-	-	24.91
Average	20.83	7.88	15.65	10.63	18.19	11.63	4.79	6.08	14.19	8.24	13.12	3.98	13.64

D.2 Language Fidelity Analysis

Table 17: Pearson correlation coefficients between language fidelity on xFlickrCO and Performance on other datasets.

Dataset	**Language**								
	Avg.	zh	en	de	id	ja	ru	es	tr
xFlickrCO	.91	.85	.65	0.86	.88	.91	.92	.90	.84
XM3600	.81	.74	.63	0.63	.69	.74	.76	.67	.82
MaXM	.55	.17	.43	-	-	-	-	-	-
XVNLI	.51	-	.46	-	-	-	.47	.20	-
MaRVL	.46	.21	.41	-	.50	-	-	-	.50
M5-VGR	.34	-	.11	0.15	-	-	.42	-	-
xGQA	.21	.35	.47	0.08	.37	-	-.04	-	-
M5-VLOD	.14	-	.44	0.20	-	-	.14	-	-

V

GIMMICK: Globally Inclusive Multimodal Multitask Cultural Knowledge Benchmarking

Bibliographic Entry

Florian Schneider, Carolin Holtermann, Chris Biemann, and Anne Lauscher. 2025b. GIMMICK – Globally Inclusive Multimodal Multitask Cultural Knowledge Benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2025,* in press. Vienna, Austria: Association for Computational Linguistics

GIMMICK
Globally Inclusive Multimodal Multitask Cultural Knowledge Benchmarking

Florian Schneider[1], Carolin Holtermann[2], Chris Biemann[1], Anne Lauscher[2]
[1]Language Technology Group, University of Hamburg
[2]Data Science Group, University of Hamburg
florian.schneider-1@uni-hamburg.de

Abstract

Large Vision-Language Models (LVLMs) have recently gained attention due to their distinctive performance and broad applicability. While it has been previously shown that their efficacy in usage scenarios involving non-Western contexts falls short, existing studies are limited in scope, covering just a narrow range of cultures, focusing exclusively on a small number of cultural aspects, or evaluating a limited selection of models on a single task only. Towards globally inclusive LVLM research, we introduce GIMMICK, an extensive multimodal benchmark designed to assess a broad spectrum of cultural knowledge across 144 countries representing six global macro-regions. GIMMICK comprises six tasks built upon three new datasets that span 728 unique cultural events or facets on which we evaluated 20 LVLMs and 11 LLMs, including five proprietary and 26 open-weight models of all sizes. We systematically examine (1) regional cultural biases, (2) the influence of model size, (3) input modalities, and (4) external cues. Our analyses reveal strong biases toward Western cultures across models and tasks and highlight strong correlations between model size and performance, as well as the effectiveness of multimodal input and external geographic cues. We further find that models have more knowledge of tangible than intangible aspects (e.g., *food* vs. *rituals*) and that they excel in recognizing broad cultural origins but struggle with a more nuanced understanding.[1]

1 Introduction

Recently, proprietary as well as open-weight Large Vision-Language Models (LVLMs) (OpenAI, 2023; Liu et al., 2023; Wang et al., 2024; Chen et al., 2023, *inter alia*) have attracted marked attention due to their broad applicability across various domains. Several large-scale holistic benchmarks (Duan et al., 2024; Yue et al., 2024; Fu et al., 2023) demonstrate LVLMs' remarkable performances in a wide range of multimodal tasks. However, most benchmarks concentrate on Western-centric English tasks, and multilingual benchmarks (Ahuja et al., 2024; Schneider and Sitaram, 2024) reveal a significant deterioration in performance on non-English tasks. While multilingualism is essential for globally equitable AI, *multiculturalism* (Gabriel, 2020; Adilazuarda et al., 2024) is equally crucial for models to reflect and respect the diverse cultural backgrounds of users worldwide. In this context, it has been shown that current LLMs (Myung et al., 2024; Chiu et al., 2024) and LVLMs suffer in tasks involving knowledge from non-Western cultures. However, the scope of existing multimodal cultural studies is still severely limited: Existing research often focuses only on specific concepts like food or dance (Winata et al., 2025; Burda-Lassen et al., 2025), covers a limited number of cultures (Urailertprasert et al., 2024; Baek et al., 2024), evaluates only a small selection of LVLM models (Cao et al., 2024; Nayak et al., 2024), or tests only a single combination of input modalities.

To address these gaps, we introduce GIMMICK, a comprehensive evaluation framework assessing 31 state-of-the-art models, ranging from proprietary LVLMs to open-weight LLMs and LVLMs of all sizes—from 500M to 78B parameters—across multiple model families. It comprises six tasks built on three novel datasets that contain 728 unique cultural events or facets (CEFs) from 144 countries in six global macro-regions and target both high-level and nuanced cultural knowledge through multimodal and unimodal tasks. Our VQA tasks span a total of 57 cultural aspects (see §B.2) Ultimately, GIMMICK enables us to answer four research questions:

(RQ1) ***Are there regional biases in LLMs' and LVLMs' cultural knowledge, and if so, which?*** For the most complex tasks, we observe consis-

[1]http://github.com/floschne/gimmick

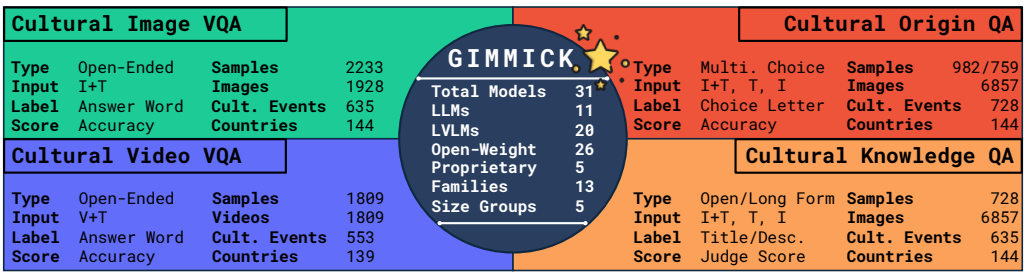


Figure 1: An overview of the GIMMICK benchmark and its tasks.

tent cultural regional biases (up to 14.72pp difference between instances targeting Western Europe & North America vs. Subsaharian Africa; §5.1) – even for the largest models. For less complex tasks, these differences flatten out.

(RQ2) ***To what degree does model size influence performance?*** We show that increasing the number of parameters significantly boosts performance on complex tasks, with larger models exhibiting less regional biases (§5.2). Still, even the largest models still struggle with nuanced cultural understanding.

(RQ3) ***How do input modalities affect cultural understanding?*** We observe that providing input in multiple modalities typically leads to the best results, as models leverage the cultural cues present in the visual inputs we provide (§5.3). Interestingly, on text-only tasks, LVLMs perform consistently worse than their LLM backbones, indicating a loss of cultural knowledge during integration training.

(RQ4) ***What is the influence of external cultural cues?*** We demonstrate that providing country information consistently guides the models towards better answers, especially for regions for which the models perform poorly (§5.4). Overall, with GIMMICK, we hope to encourage more research on culturally-aware and more globally-inclusive AI.

2 Related Work

Multicultural LLM Benchmarks. Naous et al. (2024) introduce CAMeL, a dataset that contrasts Arab and Western cultures to measure cultural biases in LLMs through extrinsic and intrinsic evaluations on core NLP tasks. With CultureAtlas, Fung et al. (2024) introduced an approach for massively multicultural knowledge acquisition and benchmarking of 5 LLMs from Wikipedia articles on cultural topics. BLEnD (Myung et al., 2024) is a large benchmark to evaluate LLMs' everyday knowledge across diverse cultures and from 16 countries in 13 different languages. (Mukherjee et al., 2024) test four popular LLMs with culturally sensitive and non-sensitive prompts on both sensitive and neutral datasets. Instead of assessing models' intrinsic cultural knowledge, (Bhatt and Diaz, 2024) focuses on the extrinsic evaluation of cultural competence, e.g., in user-interaction, in two text generation tasks, open-ended question answering, and story generation of 6 LLMs.

Benchmark	#M	#DS	#T	#S	#C	#R	Mods
SEA-VQA Urailertprasert et al. (2024)	2	1	1	1,999	8	1	**T+I**
WorldCuisines Winata et al. (2025)	18	1	2	1.15M	189	6	**T+I**
CROPE Nikandrou et al. (2025)	17	1	1	1,060	6	3	**T+I**
CulturalVQA Nayak et al. (2024)	8	1	1	2,378	11	5	**T+I**
Ananthram et al. (2025)	10	–	3	–	2	2	**T+I**
GlobalRG Bhatia et al. (2024)	12	2	2	3,591	51	6	**T+I**
MOSAIC-1.5K Burda-Lassen et al. (2025)	4	1	1	1,500	–	6	**T+I**
FoodieQA Li et al. (2024)	8	1	3	1,839	1	1	**T+I**
Cao et al. (2024)	1	–	3	–	5	3	**T+I**
K-VISCUIT Baek et al. (2024)	13	1	1	657	1	1	**T+I**
CVQA Romero et al. (2024)	8	1	1	10,374	30	6	**T+I**
CulturalBench Chiu et al. (2024)	30	2	1	6,135	45	6	**T+I**
GIMMICK **(ours)**	31	3	6	7,239	144	6	**T+I** **V+T** **T, I**

Table 1: A comparative overview of recent benchmarks assessing cultural knowledge of LVLMs. The abbreviations in the columns stand for the (combined) number of: (unique) **M**odels, **D**atasets, **T**asks, **S**amples, **C**ountries, or **R**egions contained. The **Mod**alities column lists the input modalities—**T**ext, **I**mage, **V**ideo—contained.

Multicultural LVLM Benchmarks. Bhatia et al. (2024) introduced the GlobalRG benchmark, which comprises two tasks: retrieving culturally diverse images for universal concepts from 50 countries and grounding culture-specific concepts within images from 15 countries. Karamolegkou et al. (2024) proposed a culture-centric evaluation benchmark investigating the reliability of LVLMs as visual as-

sistants for blind people in a culturally diverse setting. Using the CulturalVQA (Nayak et al., 2024), the authors assessed geo-diverse cultural understanding of nine "1st-Gen" LVLMs on a curated dataset of 2,378 VQA pairs representing cultures from 11 countries and five cultural aspects. CulturalBench (Chiu et al., 2024) is a dataset of 1,227 human-written and human-verified questions for evaluating LLMs' cultural knowledge, covering 45 global "regions". Nikandrou et al. (2025) propose CROPE, a VQA benchmark designed to probe the knowledge of culture-specific concepts and evaluate the capacity for cultural adaptation through contextual information featuring over 1M data points across 30 languages and dialects.See Table 1 for an overview and a comparison or related work with GIMMICK.

Multilingual Multicultural LVLM Benchmarks. Several studies evaluate the cultural awareness and capabilities of LVLMs in a multilingual setting. Geigle et al. (2025) extensively benchmarked state-of-the-art LVLMs across multiple multilingual and multicultural datasets, including MaRVL (Liu et al., 2021), XM3600 (Thapliyal et al., 2022) and MaXM(Changpinyo et al., 2023), M5B-VGR and M5B-VLOD (Schneider and Sitaram, 2024), CVQA (Romero et al., 2024) Winata et al. (2025) created WorldCuisines, a large-scale benchmark for multilingual and multicultural VQA on global cuisines. However, in GIMMICK, we focus on the English language, considering English performance as an upper bound.

3 The GIMMICK Benchmark

Cultural Benchmark Positioning Adilazuarda et al. (2024) surveyed 90+ recent papers on cultural awareness in LLMs and found that *none* explicitly define "culture". Instead, these studies evaluate models on datasets capturing only specific cultural aspects, which the authors organize into two dimensions: *demographic* and *semantic* proxies (with seven and five subsets, respectively). In GIMMICK, we adopt the proposed taxonomy by using countries and regions as *demographic* cultural proxies. Our tasks span all five *semantic* proxies: "emotions and values", "food and drink", "social and political relations", "basic actions and technology", and "names". We implement primarily "black-box" generative and discriminative probing approaches.

UNESCO Intangible Cultural Heritage. All tasks in GIMMICK are based on high-quality open-access data from the UNESCO Intangible Cultural Heritage (ICH) project[2], which aims to safeguard cultural traditions and practices vital to the identity and heritage of communities worldwide while honoring cultural diversity.Intangible cultural heritage encompasses oral traditions, performing arts, rituals, festive events, traditional craftsmanship, and cultural knowledge. The open-access dataset is structured as a knowledge graph, where most nodes represent cultural events or facets (CEFs; e.g., *Yuki-tsumugi*, a silk fabric production technique from Japan[3]), with additional nodes including countries, regions, case studies in which the CEFs occur. For GIMMICK, we extract the CEFs, each together with their title, description, associated macro-regions and countries, and several images depicting different aspects of the CEF. Moreover, each CEF is detailed in one or more YouTube videos. In total, GIMMICK contains 728 CEFs from 144 countries represented by 6,887 images and 993 videos[4]. While most CEFs (88.60%) are associated with one country, some are associated with two or more countries.The UNESCO ICH project groups the countries into six global macro-regions[5], which we adopt in this work. Throughout the paper—including all figures and tables—we use the region abbreviations listed in Table 2.

REGION	ABBRV.	#C	#CEF
Arab	A	18	76
Asia & Pacific	AP	35	226
Eastern Europe	E	25	150
Latin-America & Caribbean	LAC	28	98
Subsaharian Africa	SA	40	73
Western Europe & North America	W	23	149
Unique		144	728

Table 2: Regions within GIMMICK. #C and #CEF stand for the number of Countries and CEFs related to the respective region. Some CEFs may span multiple regions.

3.1 Datasets and Tasks

We created three novel multimodal datasets that serve as the foundation for six tasks designed to evaluate the cultural knowledge of models. See Figure 1 for an overview of the different tasks.[6]

[2] https://ich.unesco.org
[3] More examples including images are shown in §A.2.1
[4] We provide licensing details in §A.1
[5] We provide a comprehensive list in Table 4 in §A.3
[6] Sample counts per task & region are shown in §A.3.1

3.2 Cultural Image VQA

In the Cultural Image VQA (`CIVQA`) task, models are presented with an image depicting a CEF and a question that relates to a particular CEF aspect (see §B.1 for examples). Models are evaluated based on answer correctness. To create the data for `CIVQA`, we couple synthetic data generation with a two-stage annotation process.

Synthetic Data Generation. Building on the high-quality UNESCO ICH data, we applied synthetic data generation by prompting GPT-4o[7] to construct the basis for our dataset.Each VQA pair is related to a CEF and consists of an image depicting one aspect of the CEF, a question related to the CEF and the image, and an answer. Maximizing the quality of the generated silver data, we applied extensive prompt engineering combining techniques such as Few-Shot, Chain-of-Thought, ReAct (Wei et al., 2022; Zhang et al., 2023; Zheng et al., 2024; Sahoo et al., 2024) to craft the prompt. Key aspects of the prompt are a role description, a general task description, detailed annotation guidelines, a step-by-step strategy, an expected output format, few-shot examples, and the information of the target CEF (see §B.4 for the full prompt). We then generated silver VQA pairs for each of the 6,827 images contained in the ICH data source, which resulted in 17,369 pairs. Afterward, we automatically removed pairs where 1) the question contained words that introduce subjectiveness or ambiguity ("*could*", "*should*", "*maybe*", etc.); 2) the answer contained abstract words that are hard to depict visually; and 3) where the answer is not a substring of the description of the related CEF. This way, we obtained 9,900 silver VQA samples related to 5,517 images from all 728 CEFs.

Annotation Process. Opting for high-quality VQA pairs as well as cultural diversity, we devised a two-stage annotation process with 18 trained experts from various cultural backgrounds covering all six regions (see Table 8 in §B.5). Each silver pair was evaluated using two questionnaires—one with seven question-related requirements and another with four answer-related requirements. Questions had to target the CEF and image content directly, require cultural knowledge, and depend on visual evidence (Chen et al., 2024a). Answers needed to be clear, objective, concise, and depictable. For details on the annotation process, see §B.5.

In the first round, we annotated each sample once, resulting in 4,114 samples, of which 2,826 (68.69%) met all criteria. In the second round, five annotators re-evaluated these, retaining only samples with concordant approval. This process finally yielded 2,233 samples for 1,928 images from 728 CEFs across 144 countries in six global regions.

3.3 Cultural Video VQA

In this task, models are evaluated on questions relating to videos instead of single images, again employing accuracy as the metric. To this end, we extend `CIVQA` in two steps: synthetic data generation and quality annotation.

Synthetic Data Generation. First, we adjusted the `CIVQA` questions by replacing the term *"image"* with *"video"*. We then coupled the question with a short video clip, for which we started from the CEF's associated YouTube video. We ensured that the shortened clip contains relevant information for answering the question as follows: From each video, we extracted one frame per second, and computed image embeddings for both the frames and the `CIVQA` image, using DINOv2[8] (Oquab et al., 2024; Darcet et al., 2024). We then identified the frame that best matches the original image by calculating Cosine similarity. We selected this frame as the center (at $t = 0$) for a 10-second clip[9] (from $t = -5$ to $t = 5$). We only include clips with a best-matching frame similarity > 0.5, which we found to yield high-quality instances based on a manual inspection of random samples. Overall, this procedure resulted in 2,001 silver samples.

Annotation Process. For additional quality control, a trained expert annotated 20% of the silver data (400 samples). Each sample was evaluated using a three-item questionnaire[10] assessing whether (1) the video contained frames resembling the CEF image, (2) it clearly answered the question, or (3) neither condition was met. Overall, 95% of the annotated samples were accepted. For closer inspection, we stratified the annotated samples into four similarity bins, revealing that roughly 10% of those in the lower bins ($[0.5, 0.75[$) were rejected, while nearly all, i.e., 99% and 100%, in the higher bins ($[0.75, 1.0]$) were retained. The residual 5% label noise was considered acceptable based on further manual analysis. Notably, we found that of

[7] `gpt-4o-2024-08-06`

[8] `facebook/dinov2-with-registers-large`

[9] We do not include the audio stream in our clips.

[10] cf. §C for details.

the 20 rejected samples, only 9 were unanswerable based on the video, while the remaining 11 exhibited only a suboptimal frame match w.r.t. the CIVQA image. The final GIMMICK CVVQA dataset contains 1,809 samples (see §C.1 for examples) linked to 553 CEFs from 139 countries.

3.4 Cultural Origin QA

With Cultural Origin QA (COQA), we test a model's ability to capture coarse-grained cultural knowledge. Given a CEF's images, title, or both, the models must select its cultural origin (multiple-choice). We refer to the task as $\texttt{COQA}_\texttt{R}$ when the origin is a region and as $\texttt{COQA}_\texttt{C}$ when it is a country.

Dataset Construction. The COQA dataset contains all 728 CEFs from UNESCO ICH. To ensure that each instance corresponds to a unique origin, we replicate each CEF N times—where N represents the number of associated regions (for $\texttt{COQA}_\texttt{R}$) or countries (for $\texttt{COQA}_\texttt{C}$). For $\texttt{COQA}_\texttt{R}$, three negatives are randomly sampled from the remaining pool. Negatives for $\texttt{COQA}_\texttt{C}$ drawn from those within the same region as the target country.

Input Modalities and Prompts. The COQA tasks support multiple input configurations alongside the task prompt. In the text-only setting, only the title of the CEF is provided, whereas in the "image-only" setting, *all* images associated with the CEF are included. Both the title and the images are used in the text-image setting. Examples and complete prompts for all variations are shown in §D.2.

3.5 Cultural Knowledge QA

In GIMMICK Cultural Knowledge QA (CKQA), we evaluate whether current AI models capture fine-grained cultural knowledge. The dataset supports two open-answer tasks: naming ($\texttt{CKQA}_\texttt{N}$) and describing ($\texttt{CKQA}_\texttt{D}$). For $\texttt{CKQA}_\texttt{N}$, the ground truth corresponds to the title of the CEF, while for $\texttt{CKQA}_\texttt{D}$, it is the detailed description. For both tasks, we leverage all 728 CEFs from UNESCO ICH. As with COQA, CKQA supports multiple input configurations: text-only, "image-only", and text+image. We provide examples and prompts for all variations in §E.1.

4 Experimental Setup

Models and Inference. We evaluate a total of 31 models, including five proprietary LVLMs, 15 open-weight LVLMs, and 11 open-weight LLMs—the backbones of the respective LVLMs—covering 9 LVLM and 4 LLM model families. The sizes of the open-weight models vary, categorized as small, medium, large, and extra-large (see Table 3). A comprehensive list of models is provided in Table 6 in §A.4. For our experiments, we download open weights from the respective Huggingface (Wolf et al., 2020) repositories (see Table 6) and generate responses employing greedy decoding. For proprietary models, we use the official Python SDKs. More details are reported in §F.

GROUP	PARAMETERS (B)	LLMS	LVLMS
S	0.5 – 4	5	5
M	7 – 11	3	6
L	26 – 38	2	2
XL	72 – 78	1	2
Closed	unkown	0	5
Total		11	20

Table 3: The size groups we define for result aggregation according to models' number of parameters.

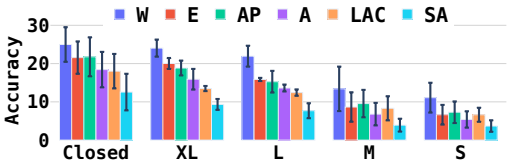
Metrics. For the CIVQA, CVVQA, and COQA tasks, we report relaxed answer accuracy, for which we consider a generated answer correct if it starts with the ground truth answer. For $\texttt{CKQA}_\texttt{D}$ and $\texttt{CKQA}_\texttt{N}$, due to their generative nature, we use GPT-4o[11] in an "LVLM-as-a-Judge" (Zheng et al., 2023; Xiong et al., 2024) setup to judge responses with a score $s \in [0, 100]$. Where $s = 0$, $s = 50$, and $s = 100$ indicate *completely incorrect or irrelevant*, *partially correct or relevant*, and *perfectly correct and complete* answers, respectively.

Video Processing. The 10-second video clips from CVVQA do not contain an audio stream, and we only use the visual information. Following established praxis (e.g., Wang et al., 2024), we extract one frame per second from the videos and provide them to the models as input alongside the textual prompt. Specifics about the image and video processing of the individual models are documented in the code.

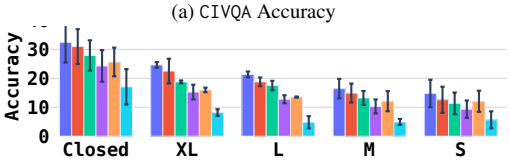
5 Results and Analyses

In this section, we present a series of in-depth analyses based on the outcomes of our benchmark. We show aggregated results: open-weight models are grouped and averaged by parameter size, and proprietary models are averaged together (see Table 3). We provide the complete numerical results for all tasks and models in § G. In the following, we use abbreviations for regions, as defined in Table 2.

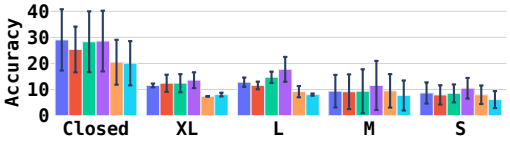
[11] gpt-4o-2024-11-20



(a) CIVQA Accuracy



(b) CVVQA Accuracy



(c) CKQA_{N} Accuracy

Figure 2: Aggregated results of the VQA tasks.

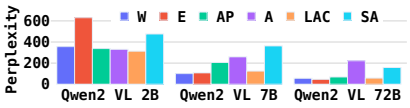


Figure 3: CIVQA ground-truth answer perplexity.

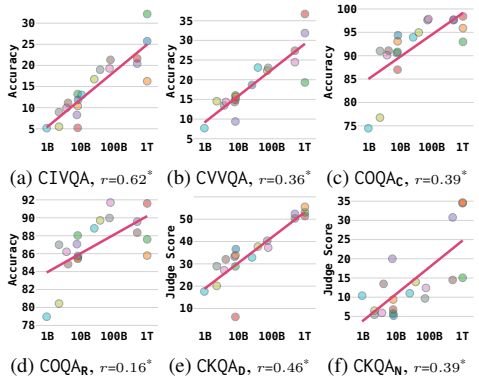
5.1 General Trends and Cultural Bias

We discuss general trends and investigate cultural bias across regions (Figures 2 and 3).

CIVQA & CVVQA. Figures 2a–c show clear regional performance disparities. Across all models—proprietary and open-weight, regardless of size—scores are highest for Western and Asian targets (**W**, **E**, and **AP**) and lowest for **SA**. XL models, e.g., reach 24.04 on **W** and 9.32 on **SA** on average. **A** and **LAC** fall in between, with model performance varying by size. Since CIVQA is an open-answer task, often with rare culturally specific terms, we also evaluated the task with GPT-4o as LVLM-as-a-Judge to account for imperfect naming or spelling. While this method yields higher scores, it confirms the same trend: models exhibit a strong bias toward Western contexts. However, even the best model (GPT-4o) scores only 31.58% on **W** and 25.44% on average, highlighting GIMMICK as a challenging benchmark and the lack of fine-grained cultural knowledge in current models. We supplement our analysis with a more fine-grained investigation of how well models "know" the cultural concepts discussed. Here, we focus on the QWENVL models on CIVQA and the compute perplexity of ground truth answers (conditioned on the input context) as a proxy of model cultural knowledge (details in §G.1.2). Figure 3 shows that for the 7B and 72B models, perplexity is consistently lower for **W**, **E**, and **AP** compared to **A** and **SA**, aligning with our performance findings. For the 2B model, however, **E** and **SA** yield the highest perplexities, which we attribute to the overall brittleness of the model. Moreover, we revisit the performance on questions about the prevalent cultural aspects in CIVQA (details in §G.1.2) and find that models perform notably better on tangible cultural aspects than on intangible ones. For instance, closed models achieve an accuracy of 30% for food-related questions and only 8% and 10% for questions concerning rituals or festivals. This highlights biases along the cultural dimension, which are particularly pronounced in non-Western contexts.

CKQA_{N} & CKQA_{D}. For CKQA_{N}, regional differences are minor, though proprietary models significantly outperform open-weight ones (see Figure 2c). The large error bars for closed models indicate inconsistent performance—particularly from GPT-4O MINI and GEMINI FLASH models, which perform similarly to large open-weight models. XL and L models perform worst on **SA** and **LAC** and best on **A** and **AP** with minor differences to **W** and **E**. For CKQA_{D} (Figure 6c), performance is $10\tilde{}20\%$ higher than on CKQA_{N}, likely because describing a CEF is easier than exactly naming it. However, regional biases are larger, with consistently higher scores on **W** than on **SA**, primarily for closed models like GPT-4o, which reaches 53.66 for **W** and 43.70 on **SA**.

COQA_{C} & COQA_{R}. Figure 6a shows minimal regional differences for COQA_{C}.Average accuracies range from close to or above 90% for closed, XL, and L models to 77.42% for S models. However, performance on COQA_{R} is lower than on COQA_{C}—85.02% vs. 81.17% on average over all models and regions— with models achieving the highest scores in **AP**. Notably, the regional ranking is mostly inverted compared to other tasks—**SA**, **A**,



(a) CIVQA, r=0.62* (b) CVVQA, r=0.36* (c) COQA$_C$, r=0.39*

(d) COQA$_R$, r=0.16* (e) CKQA$_D$, r=0.46* (f) CKQA$_N$, r=0.39*

Figure 4: Model size vs. performance on GIMMICK tasks. The x-axis is in log scale. The trend line was computed using OLS regression. We report the Pearson correlation coefficient r (* indicates statistical significance).

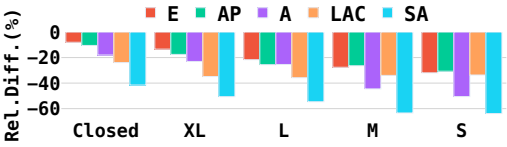


Figure 5: Relative Difference to W for CIVQA.

LAC, E, and AP score higher than W—suggesting more distinct visual and linguistic features in non-Western regions.

5.2 Influence of Model Size

We assess how model size impacts performance and whether it affects regions equally.

Figure 4 shows that model size[12] significantly influences performance, with moderate to strong Pearson correlations and steep regression lines across tasks except COQA$_R$, where the effect is minimal. Figure 5 shows that relative performance declines from the best-performing region (W) to others, particularly SA, varying by model size: the drops are -63.39 (S), -63.85 (M), -50.60 (L), -54.57 (XL), and -41.52 (Closed). We conclude that bigger sizes tend to result in smaller gaps without size presenting a strict ordering criterion.

5.3 Influence of Modalities

We explore how input modality—text-only, image-only, or text+image—affects performance on COQA$_C$, COQA$_R$, and CKQA$_D$. Further, we compare LVLMs to their LLM backbones to assess potential losses in cultural knowledge during multimodal training.

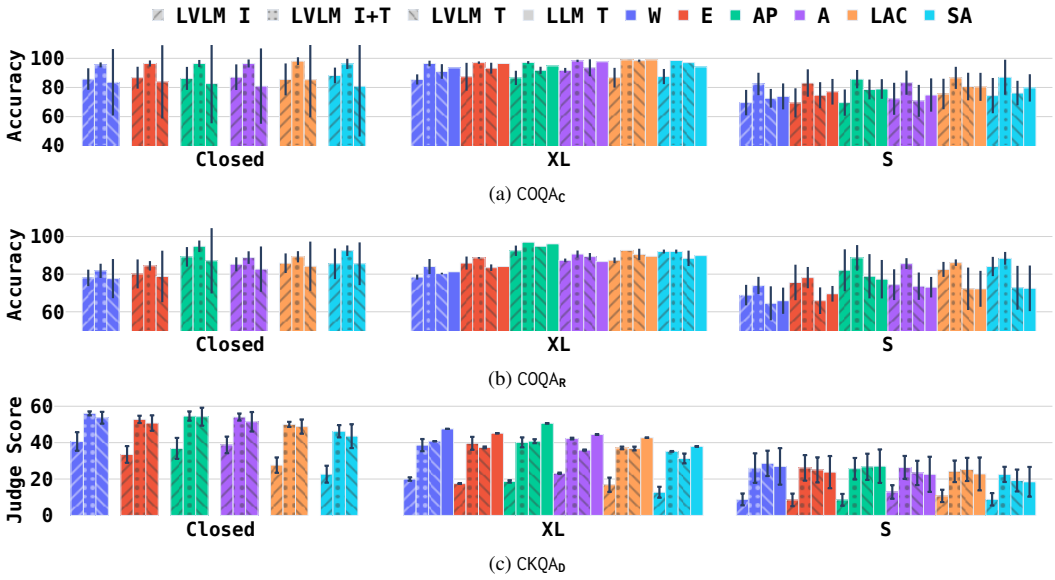
Input Modalities. Figure 6 shows that text+image (I+T) inputs consistently yield the highest performance across all tasks, confirming that textual and visual data provide complementary cultural cues. The gap between I+T and text-only (T) is slightly more prominent for COQA$_C$ than COQA$_R$, suggesting that visual information aids in inferring fine-grained, country-level details. In contrast, image-only (I) inputs perform poorly, indicating that textual information, such as CEF titles, carries more cultural context. The high variance in T results for the COQA tasks stems from the performance disparity between GEMINI PRO and CLAUDE 3.5 SONNET (e.g., 59.38 vs. 83.75 for W).

LVLM vs. LLM-Backbone. Comparing LVLMs with their LLM backbones reveals that multimodal training can impair the acquisition of detailed cultural knowledge (notably in CKQA$_D$) while having minimal impact on coarse-grained cultural understanding (COQA). For large models, significant performance gaps—50.62 for QWEN2.5 72B vs. 40.02 for QWEN2VL 72B on AP—on the CKQA$_D$ task between the LVLMS and their LLM backbones can be observed, whereas, for smaller models, the effect is subtle. Overall, our findings highlight that while images complement text for culturally grounded tasks, it is ultimately the synergy between both modalities that leads to robust and broad cultural understanding.

5.4 Influence of External Cues

We examine how external hints, i.e., informing a model about the country or region of a CEF, affect VQA performance. For CIVQA (Figure 7a), country hints consistently boost performance across model sizes and regions, while regional cues yield only modest—or even slightly adverse—effects in larger models. Gains from country hints are around 50% for most regions, but in SA, improvements nearly double (e.g., 97.48% for INTERNVL 2.5 78B and 97.13% for INTERNVL 2.5 38B). A similar pattern emerges for CVVQA (Figure 7b). Hints generally enhance performance across regions

[12]For closed source models, we manually set the number of parameters to 1T, except for Gemini Flash and GPT-4o mini, for which we set the number to 500B.

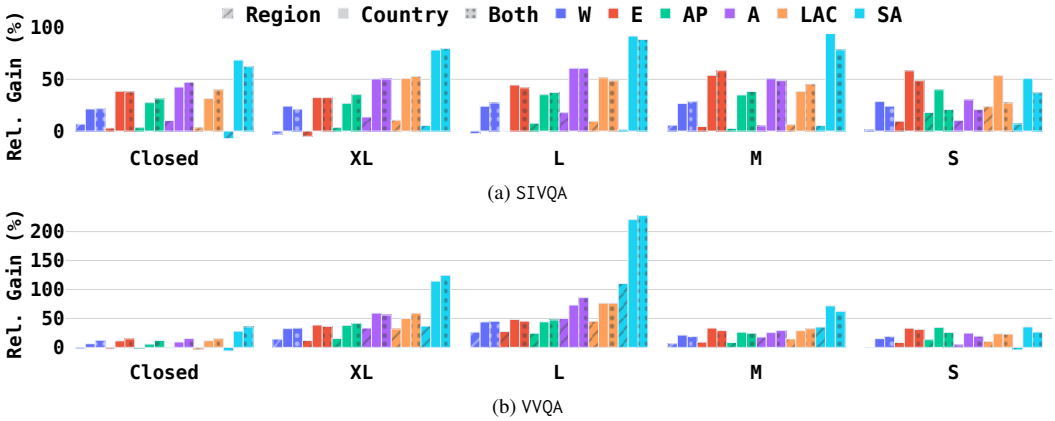


(a) COQA$_C$

(b) COQA$_R$

(c) CKQA$_D$

Figure 6: Aggregated results including multimodal input variations: **T**ext-only, **I**mage-only, **T**ext+**I**mage.



(a) SIVQA

(b) VVQA

Figure 7: Relative gains on VQA tasks from providing external geographical hints.

and models, with **SA** showing the most significant gains. Proprietary and small models exhibit subtle improvements, whereas L and XL models see much higher relative gains—up to 240.7% for INTERN VL 38B. Notably, regional cues have a more positive impact on CVVQA than on CIVQA.

6 Conclusion

We introduce GIMMICK, a comprehensive benchmark to assess various aspects of cultural knowledge of current LVLMs and LLMs and introduce six tasks built upon three novel datasets, which span 728 unique cultural events or facets (CEFs) from 144 countries grouped into six global macro-regions. Through extensive analyses, we study general cultural biases and the influence of model size, input modalities, and external cues. Our results consistently reveal a prominent bias toward Western cultures across all models. Interestingly, when only coarse cultural knowledge is required—such as regional origins—models performed remarkably better. Across all tasks, significant correlations between a model's performance and its size are evident,

with a substantial gap between proprietary and open-weight models. Our analyses show that while models grasp broad cultural categories, they struggle with nuanced understanding. This suggests that GIMMICK poses a challenging benchmark and highlights the need for further advances in modeling broad cultural awareness.

Limitations

English-Only Benchmark Although we consider the performance on tasks requiring cultural understanding in English as an upper bound for the majority of models, it is yet to be tested if that hypothesis generally holds across tasks, model size, and model family. Especially for models like QwenVL and InternVL, which were pretrained on large portions of Chinese textual data, Chinese could be pivotal instead of English. Moreover, some cultural nuances might not be translatable to other languages.

Open-Ended VQA. CIVQA and CVVQA comprise open-ended answers to their questions, imposing challenges for adequate evaluation, especially when employing binary metrics like accuracy. This is especially true for rare, culturally specific answer terms, such as in our tasks, which are prone to spelling inaccuracies or might have different names in different cultures or languages. Although we alleviate this issue by computing scores using GPT-4o in an LVLM-as-a-Judge setting and thereby confirm our findings, this requires additional computational and financial resources. A typical solution for this is transforming the questions into multiple-choice questions, which, however, requires culturally expert annotators, who are challenging to find or train and expensive if hired via professional annotation companies.

Small Number of Samples. With a total of 7239 unique samples across all tasks in GIMMICK—2233 (CIVQA), 1809 (CVVQA), 982 (COQA_C), 759 (COQA_R), 728 (CKQA_D), and 728 (CKQA_N)—, the benchmark itself has the third most samples compared to other recent benchmarks. However, the per-task number falls relatively low, leading to even fewer counts per country or culture, making judgments about single countries not informative.

To increase the number of samples, we consider two main options: 1) By expanding the number of annotations by employing expert annotators for an additional period of time and/or increasing the amount of silver data as described in §B.4, which would lead to an increase of samples for the CIVQA and CVVQA datasets. 2) By incorporating the newly released UNESCO data every year, as well as leveraging other high-quality sources such as UNESCO World Heritage[13], the European Commission[14], the Southeast Asian Cultural Heritage Alliance (SEACHA)[15], the Journal of African Cultural Heritage[16], or ICH Links[17]

Ethical Considerations

Country and Region Definitions. GIMMICK adopts the country and region classifications from the UNESCO ICH dataset. While these classifications are widely used, we recognize the potential for differing interpretations.

Potentially Offensive Questions. We employed semi-automatic data generation strategies to create the CIVQA dataset. Here, the silver data was generated using GPT-4o, which we showed displays significant cultural biases towards Western contexts. Although we provided the model with high-quality ground-truth information from the UNESCO ICH project and trained expert annotators with diverse cultural backgrounds to filter low-quality VQA samples, certain questions or their answers might still be offensive to people with certain cultural origins. Since this is subjective, we need to accept it as is for now. Nevertheless, we encourage contacting us if any offensive or otherwise harmful sample raises someone's attention.

Acknowledgements

We thank our annotators for the CIVQA and CVVQA tasks with special thanks to Timm Dill, Narges Baba Ahmadi, Niloufar Baba Ahmadi,

[13] https://www.unesco.org/world-heritage
[14] https://culture.ec.europa.eu/cultural-heritage
[15] https://seacha.org/
[16] https://jachs.org/
[17] https://www.ichlinks.com

and Abdullah Abdelhafez for their extra efforts. The work of Carolin Holtermann and Anne Lauscher is funded by the Excellence Strategy of the German Federal Government and the Federal States.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, and 68 others. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *CoRR*, abs/2404.14219.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards Measuring and Modeling "Culture" in LLMs: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15763–15784, Miami, Florida, USA.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 2598–2637, Mexico City, Mexico.

Meta AI. 2024. Llama 3.2: Revolutionizing Edge AI and Vision with Open, Customizable Models.

Amith Ananthram, Elias Stengel-Eskin, Carl Vondrick, Mohit Bansal, and Kathleen R. McKeown. 2025. See It from My Perspective: Diagnosing the Western Cultural Bias of Large Vision-Language Models in Image Understanding. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, Singapore, Republic of Singapore.

AI Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.

Yujin Baek, ChaeHun Park, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. Evaluating Visual and Cultural Interpretation: The K-Viscuit Benchmark with Human-VLM Collaboration. *CoRR*, abs/2406.16469.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. 2024. From Local Concepts to Universals: Evaluating the Multicultural Understanding of Vision-Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6763–6782, Miami, FL, USA.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic Evaluation of Cultural Competence in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 16055–16074, Miami, FL, USA.

Olena Burda-Lassen, Aman Chadha, Shashank Goswami, and Vinija Jain. 2025. How Culturally Aware Are Vision-Language Models? In *2025 IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS)*, volume CFP2540Z-ART, pages 1–6, Lyon, France.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 52 others. 2024. InternLM2 Technical Report. *CoRR*, abs/2403.17297.

Yong Cao, Wenyan Li, Jiaang Li, Yifei Yuan, Antonia Karamolegkou, and Daniel Hershcovich. 2024. Exploring Visual Culture Awareness in GPT-4V: A Comprehensive Probing. *CoRR*, abs/2402.06015.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. MaXM: Towards Multilingual Visual

Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2667–2682, Singapore.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024a. Are We on the Right Way for Evaluating Large Vision-Language Models? In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 21 others. 2024b. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *CoRR*, abs/2412.05271.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, Seattle, WA, USA.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. CulturalBench: a Robust, Diverse and Challenging Benchmark on Measuring the (Lack of) Cultural Knowledge of LLMs. *CoRR*, abs/2410.02677.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier. *CoRR*, abs/2412.04261.

Tri Dao. 2024. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2024. Vision Transformers Need Registers. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 2024. VLMEvalKit: An Open-Source ToolKit for Evaluating Large Multi-Modality Models. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pages 11198–11201, Melbourne, VIC, Australia.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *CoRR*, abs/2306.13394.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively Multi-Cultural Knowledge Acquisition & LM Benchmarking. *CoRR*, abs/2402.09369.

Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437.

Gregor Geigle, Florian Schneider, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, and Goran Glavaš. 2025. Centurio: On Drivers of Multilingual Ability of Large Vision-Language Model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Vienna, Austria.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. GPT-4o System Card. *CoRR*, abs/2410.21276.

Antonia Karamolegkou, Phillip Rust, Yong Cao, Ruixiang Cui, Anders Søgaard, and Daniel Hershcovich. 2024. Vision-Language Models under Cultural and Inclusive Considerations. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 53–66, Bangkok, Thailand.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024. FoodieQA: A Multimodal Dataset for Fine-Grained Understanding of Chinese Food Culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 19077–19095, Miami, Florida, USA.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually Grounded Reasoning across Languages and Cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10467–10485, Punta Cana, Dominican Republic.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 34892–34916, New Orleans, LA, USA.

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024. Cultural Conditioning or Placebo? On the Effectiveness of Socio-Demographic Prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15811–15837, Miami, FL, USA.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:78104–78146.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 16366–16393, Bangkok, Thailand.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking Vision Language Models for Cultural Understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5769–5790, Miami, FL, USA.

Malvina Nikandrou, Georgios Pantazopoulos, Nikolas Vitsakis, Ioannis Konstas, and Alessandro Suglia. 2025. CROPE: Evaluating In-Context Adaptation of Vision and Language Models to Culture-Specific Concepts. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pages 7917–7936, Albuquerque, New Mexico.

OpenAI. 2023. GPT-4 Vision System Card.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesús-Germán Ortiz-Barajas, Emilio Villa-Cueva, Jinheon Baek, Soyeong Jeong, Injy Hamed, Zheng Xin Yong, Zheng Wei Lim, Paula Mónica Silva, Jocelyn Dunstan, Mélanie Jouitteau, David Le Meur, Joan Nwatu, Ganzorig Batnasan, and 57 others. 2024. CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *CoRR*, abs/2402.07927.

Florian Schneider and Sunayana Sitaram. 2024. M5 - A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks. In *Findings of the Association for Computational Linguistics: (EMNLP)*, pages 4309–4345, Miami, Florida, USA.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context. *arXiv preprint arXiv:2403.05530*.

Ashish V. Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 715–729, Abu Dhabi, United Arab Emirates.

Norawit Urailertprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. SEA-VQA: Southeast Asian Cultural Context Dataset For Visual Question Answering. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 173–185, Bangkok, Thailand.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *CoRR*, abs/2409.12191.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Ching Lam Cheng, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025. WorldCuisines: A Massive-Scale Benchmark for Multilingual and Multicultural Visual Question Answering on Global Cuisines. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pages 3242–3264, Albuquerque, New Mexico.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 38–45, Online.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. LLaVA-Critic: Learning to Evaluate Multimodal Models. *CoRR*, abs/2410.02712.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 Technical Report. *CoRR*, abs/2412.15115.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *CoRR*, abs/2408.01800.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, Seattle, WA, USA.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic Chain of Thought Prompting in Large Language Models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, Kigali, Rwanda.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA.

Appendix Overview

Due to the number of experiments, the general density of our work, and our aim to be as transparent as possible in the sense of open science, the following appendix is extensive. Hence, we provide a brief outline of its content to ease navigation and to get an overview quickly.

A GIMMICK Benchmark Details
Details on license, examples, regions, models.
B CIVQA Details
Details on examples, synthetic data generation, and the annotation project.
C VVQA Details
Details on examples and the annotation project.
D COQA Details
Details on prompts, and examples.
E CKQA Details
Details on prompts.
F Experimental Setup
Details on prompts hyperparameters.
G Results and Analyses
Details on complete results of all models and datasets and additional analyses.

A GIMMICK Benchmark Details

A.1 Data License

GIMMICK is built upon the open-access data from the UNESCO Intangible Cultural Heritage (ICH) project, which is organized as a knowledge graph. The graph can be downloaded in English, French, and Spanish on the ICH project website: https://ich.unesco.org/en/open-access-to-dive-data-01218, with details about its structure and subsets also provided. In GIMMICK, we work with the English graph only. The open-access license of the knowledge graph is defined on the UNESCO website[18] as follows:

> *By 'open access' to the literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself.*

The images and videos within the data are shared via URLs and hosted by UNESCO or on YouTube, respectively. Further, each image and video node in the knowledge graph has individual copyright information attached. However, the licenses themselves are not discussed, and merely the name of the photographer or institution or UNESCO itself is stated. Unfortunately, we did not receive an answer to multiple emails in which we asked for clarification. Hence, we assume that the image and video content also fall under the definition of "open access". If you are a copyright holder of any of the images or videos and do not want your intellectual property to be used or shared by us, please reach out via email: florian.schneider-1@uni-hamburg.de.

A.2 Cultural Event or Facets (CEFs)

A.2.1 Examples

In the following, we provide one example of CEFs per region from the UNESCO ICH project. We also use the same information for the $\texttt{CKQA}_{\texttt{N}}$ and $\texttt{CKQA}_{\texttt{D}}$ tasks.

[18] https://www.unesco.org/en/open-access

Western Europe (■W)

Title: The skills related to perfume in Pays de Grasse: the cultivation of perfume plants, the knowledge and processing of natural raw materials, and the art of perfume composition
Countries: France
Regions: Western European and North American States
Description:
The skills related to perfume in Pays de Grasse cover three different aspects: the cultivation of perfume plants; the knowledge and processing of natural raw materials; and the art of perfume composition. The practice involves a wide range of communities and groups, brought together under the Association du Patrimoine Vivant du Pays de Grasse (Living Heritage Association of the Region of Grasse). Since at least the sixteenth century, the practices of growing and processing perfume plants and creating fragrant blends have been developed in Pays de Grasse, in a craft industry long dominated by leather tanning. Perfume plant cultivation involves a wide range of skills and knowledge, for instance pertaining to nature, soil, weather, biology, plant physiology and horticultural practices, as well as specific techniques such as extraction and hydraulic distillation methods. The inhabitants of Grasse have made these techniques their own and helped improve them. In addition to technical skills, however, the art also calls for imagination, memory and creativity. Perfume forges social bonds and provides an important source of seasonal labour. Related knowledge is mostly transmitted informally through a long learning process that still takes place primarily in perfumeries. In recent decades, however, there has been a growing interest in standardizing learning through formalized teaching.

UNESCO ICH URL: https://ich.unesco.org/en/RL/the-skills-related-to-perfume-i...

Copyrigth: JM. Ghibaudo APVPG 2011

Copyrigth: Musées de Grasse 2011

Copyrigth: N. Bédar APVPG 2015

Copyrigth: C. Barbiero/Musées de Grasse 2010

Copyrigth: Daniel, Serre, M. Roudnitska APVPG 2014

Copyrigth: Musées de Grasse 2012

Copyrigth: G. Voinot/Université Sophia Antiopolis 2011

Copyrigth: Esat Les Restanques 2013

Copyrigth: Forum des Associations Pays de Grasse 2014

Copyrigth: PH. Massé APVPG 2014

Eastern Europe (■E)

Title: Cultural Heritage of Boka Navy Kotor: a festive representation of a memory and cultural identity
Countries: Montenegro
Regions: Eastern European States
Description:
Boka Navy is a traditional, non-governmental maritime organization founded in Kotor, Montenegro in 809. Its origin is linked to the arrival of the relics of St. Tryphon, the patron saint of the city of Kotor. Comprised of a community of seafarers with military, economic, educational and humanitarian functions, Boka Navy has played a memorial role for two centuries, preserving and promoting maritime history and tradition. Membership is voluntary and open to men, women and children of all ages. The organization is founded on the respect of human rights and of religious, national and cultural diversity. During formal celebrations, members wear colourful traditional uniforms, carry historic weapons and perform the traditional circle kolo dance. Boka Navy is the backbone of the annual St. Tryphon festivities, which take place from 13 January through 3 February and include a procession and a series of rituals in the cathedral. The external festivities begin with the Boka Navy's traditional kolo circle dance and are followed by a procession carrying the relics of St. Tryphon through the main town squares and streets. Thousands of spectators attend the processions in the historic centre and observe the festive events. Hundreds of women, men and children also participate in preparations of the activities.

UNESCO ICH URL: https://ich.unesco.org/en/RL/cultural-heritage-of-boka-navy-...

Copyrigth: Ministry of Culture of Montenegro

Copyrigth: Ministry of Culture of Montenegro

Copyrigth: Ministry of Culture of Montenegro

Copyrigth: Ministry of Culture of Montenegro

Copyrigth: Ministry of Culture of Montenegro

Copyrigth: Ministry of Culture of Montenegro

Copyrigth: Ministry of Culture of Montenegro

Copyrigth: Ministry of Culture of Montenegro

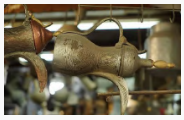
Copyrigth: Ministry of Culture of Montenegro

Copyrigth: Ministry of Culture of Montenegro

Arab (■A)

Title: Arts, skills and practices associated with engraving on metals (gold, silver and copper)
Countries: Algeria, Saudi Arabia, Egypt, Iraq, Morocco, Mauritania, Palestine, Sudan, Tunisia, Yemen
Regions: Arab States
Description:
Engraving on metals such as gold, silver and copper is a centuries-old practice that entails manually cutting words, symbols or patterns into the surfaces of decorative, utilitarian, religious or ceremonial objects. The craftsperson uses different tools to manually cut symbols, names, Quran verses, prayers and geometric patterns into the objects. Engravings can be concave (recessed) or convex (elevated), or the result of a combination of different types of metals, such as gold and silver. Their social and symbolic meanings and functions vary according to the communities concerned. Engraved objects, such as jewelry or household objects, are often presented as traditional gifts for weddings or used in religious rituals and alternative medicine. For instance, certain types of metals are believed to have healing properties. Engraving on metals is transmitted within families, through observation and hands-on practice. It is also transmitted through workshops organized by training centres, organizations and universities, among others. Publications, cultural events and social media further contribute to the transmission of the related knowledge and skills. Practised by people of all ages and genders, metal engraving and the use of engraved objects are means of expressing the cultural, religious and geographical identity and the socioeconomic status of the communities concerned.

UNESCO ICH URL: https://ich.unesco.org/en/RL/arts-skills-and-practices-assoc...



Copyrigth: Huzaifa Ayad Bahaa El Din, Iraq, 2021

Copyrigth: Huzaifa Ayad Bahaa El Din, Iraq, 2021

Copyrigth: Huzaifa Ayad Bahaa El Din, Iraq, 2021

Copyrigth: Zahia Benabdallah, Algeria, 2021

Copyrigth: Azza Fahmi, Egypt, 2021

Copyrigth: Mustafa Kamil, Egypt, 2021

Copyrigth: National Heritage Preservation, Ministry of Culture, Youth and Sport and Relations with the Parliament, Egypt, 2022

Copyrigth: Direction du Patrimoine Culturel, Morocco, 2021

Copyrigth: Direction du Patrimoine Culturel, Morocco, 2021

Copyrigth: Ministry of Culture, Palestine, 2021

Asia and Pacific (■AP)

Title: Tugging rituals and games
Countries: Cambodia, Korea, Philippines, Vietnam
Regions: Asian and Pacific States
Description:
Tugging rituals and games in the rice-farming cultures of East Asia and Southeast Asia are enacted among communities to ensure abundant harvests and prosperity. They promote social solidarity, provide entertainment and mark the start of a new agricultural cycle. Many tugging rituals and games also have profound religious significance. Most variations include two teams, each of which pulls one end of a rope attempting to tug it from the other. The intentionally uncompetitive nature of the event removes the emphasis on winning or losing, affirming that these traditions are performed to promote the well-being of the community, and reminding members of the importance of cooperation. Many tugging games bear the traces of agricultural rituals, symbolizing the strength of natural forces, such as the sun and rain while also incorporating mythological elements or purification rites. Tugging rituals and games are often organized in front of a village's communal house or shrine, preceded by commemorative rites to local protective deities. Village elders play active roles in leading and organizing younger people in playing the game and holding accompanying rituals. Tugging rituals and games also serve to strengthen unity and solidarity and sense of belonging and identity among community members.

UNESCO ICH URL: https://ich.unesco.org/en/RL/tugging-rituals-and-games-01080...

Copyrigth: Siyonn Sophearith, 2013

Copyrigth: Siyonn Sophearith, 2013

Copyrigth: Siyonn Sophearith, 2013

Copyrigth: Renato S. Rastrollo, NCCA

Copyrigth: Renato S. Rastrollo, NCCA

Copyrigth: Vietnam Institute of Culture and Arts Studies, 2013

Copyrigth: Vietnam Institute of Culture and Arts Studies, 2013

Copyrigth: Joo Byung Soo, 2006

Copyrigth: Joo Byung Soo, 2006

Latin America & Caribbean (LAC)

Title: Ancestral system of knowledge of the four indigenous peoples, Arhuaco, Kankuamo, Kogui and Wiwa of the Sierra Nevada de Santa Marta
Countries: Colombia
Regions: Latin-American and Caribbean States
Description:
The Ancestral System of Knowledge of the Arhuaco, Kankuamo, Kogui and Wiwa peoples of the Sierra Nevada de Santa Marta is comprised of sacred mandates that keep the existence of the four peoples in harmony with the physical and spiritual universe. Through many years of dedication, the knowledgeable men (Mamos) and women (Sagas) acquire the necessary skills and sensitivity to communicate with the snow-capped peaks, connect with the knowledge of the rivers and decipher the messages of nature. Based on the Law of Origin, a philosophy that governs human relationships to nature and the universe, the Ancestral System of Knowledge entails caring for sacred sites and partaking in baptism rituals, marriage rites, traditional dances and songs, and retributions or offerings to spiritual powers. This ancestral wisdom is believed to play a fundamental role in protecting the Sierra Nevada ecosystem and avoiding the loss of the cultural identity of the four peoples of the region. The Ancestral System of Knowledge is transmitted from generation to generation through cultural practice, community activities, the use of the indigenous language and the implementation of the sacred mandates. The transmission process includes the understanding of physical and spiritual relationships with Mother Nature and sacred sites.

UNESCO ICH URL: https://ich.unesco.org/en/RL/ancestral-system-of-knowledge-o...

Copyrigth: William Diaz, 2021

Copyrigth: Jorge Mario Suarez/Government of Magdalena, 2017

Copyrigth: Jorge Mario Suarez/Government of Magdalena, 2017

Copyrigth: William Diaz, 2021

Copyrigth: Jorge Mario Suarez/Government of Magdalena, 2017

Copyrigth: Jorge Mario Suarez/Government of Magdalena, 2017

Copyrigth: Jorge Mario Suarez/Government of Magdalena, 2017

Copyrigth: Jorge Mario Suarez/Government of Magdalena, 2017

Copyrigth: Jorge Mario Suarez/Government of Magdalena, 2017

Copyrigth: William Diaz, 2021

Subsaharian Africa (SA)

Title: Gada system, an indigenous democratic socio-political system of the Oromo
Countries: Ethiopia
Regions: Subsaharian African States
Description:
Gada is a traditional system of governance used by the Oromo people in Ethiopia developed from knowledge gained by community experience over generations. The system regulates political, economic, social and religious activities of the community dealing with issues such as conflict resolution, reparation and protecting women's rights. It serves as a mechanism for enforcing moral conduct, building social cohesion, and expressing forms of community culture. Gada is organized into five classes with one of these functioning as the ruling class consisting of a chairperson, officials and an assembly. Each class progresses through a series of grades before it can function in authority with the leadership changing on a rotational basis every eight years. Class membership is open to men, whose fathers are already members, while women are consulted for decision-making on protecting women's rights. The classes are taught by oral historians covering history, laws, rituals, time reckoning, cosmology, myths, rules of conduct, and the function of the Gada system. Meetings and ceremonies take place under a sycamore tree (considered the Gada symbol) while major clans have established Gada centres and ceremonial spaces according to territory. Knowledge about the Gada system is transmitted to children in the home and at school.

UNESCO ICH URL: https://ich.unesco.org/en/RL/gada-system-an-indigenous-democ...

Copyrigth: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014

Copyrigth: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014

Copyrigth: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014

Copyrigth: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014

Copyrigth: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014

Copyrigth: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014

Copyrigth: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014

Copyrigth: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014

Copyrigth: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014

Copyrigth: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014

A.2.2 CEFs as Python a `dataclass`

Listing 1 presents a CEF implemented as a Python dataclass.

```python
from dataclasses import dataclass

@dataclass
class CEF:
    title: str
    description: str
    countries: list[str]
    regions: list[str]
    images: list[str]  # URLs
    videos: list[str]  # URLs
```

Listing 1: Python pseudo-code for a dataclass representing a CEF.

Region	Abbrv.	Countries	Countries
Arab	A	18	Algeria, Bahrain, Egypt, Iraq, Jordan, Kuwait, Lebanon, Mauritania, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, United Arab Emirates, Yemen
Asia & Pacific	AP	35	Lao People's Democratic Republic, Afghanistan, Australia, Bangladesh, Bhutan, Cambodia, China, Cook Islands, Democratic People's Republic of Korea, Fiji, India, Indonesia, Iran, Japan, Kazakhstan, Korea, Kyrgyzstan, Malaysia, Micronesia, Mongolia, Myanmar, Nepal, New Zealand, Pakistan, Papua New Guinea, Philippines, Samoa, Singapore, Sri Lanka, Thailand, Timor-Leste, Tonga, Turkmenistan, Vanuatu, Vietnam
Eastern Europe	E	25	Albania, Armenia, Azerbaijan, Belarus, Bosnia and Herzegovina, Bulgaria, Croatia, Czechia, Estonia, Georgia, Hungary, Latvia, Lithuania, Moldova, Montenegro, North Macedonia, Poland, Romania, Russia, Serbia, Slovakia, Slovenia, Tajikistan, Ukraine, Uzbekistan
Latin-America & Caribbean	LAC	28	Antigua and Barbuda, Argentina, Bahamas, Belize, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Curaçao, Dominican Republic, Ecuador, El Salvador, Grenada, Guatemala, Haiti, Honduras, Jamaica, Mexico, Nicaragua, Panama, Paraguay, Peru, Saint Kitts and Nevis, Saint Vincent and the Grenadines, Uruguay, Venezuela
Subsaharian Africa	SA	40	Côte d'Ivoire, Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Congo, Democratic Republic of the Congo, Djibouti, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Kenya, Lesotho, Madagascar, Malawi, Mali, Mauritius, Mozambique, Namibia, Niger, Nigeria, Rwanda, Senegal, Seychelles, Somalia, South Africa, South Sudan, Togo, Uganda, Zambia, Zimbabwe
Western Europe & North America	W	23	Andorra, Austria, Belgium, Canada, Cyprus, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Malta, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, Türkiye, United Kingdom of Great Britain and Northern Ireland

Table 4: Caption

A.3 Regions

A.3.1 Number of Samples per Task per Region

A.4 Models

We present the comprehensive list of all 31 models evaluated in GIMMICK in Table 6.

B CIVQA Details

REGION	CIVQA	CVVQA	COQA_{R}	COQA_{C}	CKQA_{D}	CKQA_{N}
A	375	296	71	127	71	71
A AP	4	4	2	2	1	1
A AP E W	5	5	0	36	2	2
A E W	1	0	3	7	1	1
A SA	8	0	2	3	1	1
AP	444	407	211	222	211	211
AP E	7	7	6	6	3	3
AP E LAC SA W	1	1	0	8	1	1
AP E W	10	7	21	35	7	7
AP W	4	3	2	3	1	1
E	302	242	125	136	125	125
E W	21	20	22	56	11	11
LAC	420	341	96	106	96	96
LAC W	2	2	2	2	1	1
SA	388	299	71	80	71	71
W	241	175	125	153	125	125

Table 5: Number of samples per region(s) in GIMMICK tasks.

Model ID	Paper Name	Open-Weight	Size Group	Image Input	Video Input	Text Input	LLM Backbone
claude-3-5-sonnet-20241022	Claude 3.5 Sonnet (Anthropic, 2024)	No	A	Yes	Yes	Yes	–
gemini-1.5-pro-002	Gemini Pro (Team et al., 2024)	No	A	Yes	Yes	Yes	–
gemini-1.5-flash-002	Gemini Flash (Team et al., 2024)	No	A	Yes	Yes	Yes	–
gpt-4o-2024-11-20	GPT-4o (Hurst et al., 2024)	No	A	Yes	Yes	Yes	–
gpt-4o-mini-2024-07-18	GPT-4o Mini (Hurst et al., 2024)	No	A	Yes	Yes	Yes	–
opengvlab/internvl2_5-78b	InternVL2.5 78B (Chen et al., 2024b)	Yes	XL	Yes	Yes	Yes	qwen/qwen2.5-72b-instruct
qwen/qwen2-vl-72b-instruct	Qwen2 VL 72B (Wang et al., 2024)	Yes	XL	Yes	Yes	Yes	qwen/qwen2.5-72b-instruct
opengvlab/internvl2_5-26b	InternVL2.5 26B (Chen et al., 2024b)	Yes	L	Yes	Yes	Yes	internlm/internlm2_5-20b-chat
opengvlab/internvl2_5-38b	InternVL2.5 38B (Chen et al., 2024b)	Yes	L	Yes	Yes	Yes	qwen/qwen2.5-32b-instruct
meta-llama/llama-3.2-11b-vision-instruct	Llama 3.2 11B Vision (AI, 2024)	Yes	M	Yes	Yes	Yes	–
qwen/qwen2-vl-7b-instruct	Qwen2 VL 7B (Wang et al., 2024)	Yes	M	Yes	Yes	Yes	qwen/qwen2.5-7b-instruct
openbmb/minicpm-v-2_6	MiniCPM V 2.6 (Yao et al., 2024)	Yes	M	Yes	Yes	Yes	–
wuenlp/centurio_aya	Centurio Aya (Geigle et al., 2025)	Yes	M	Yes	Yes	Yes	cohereforai/aya-expanse-8b
opengvlab/internvl2_5-8b	InternVL2.5 8B (Chen et al., 2024b)	Yes	M	Yes	Yes	Yes	internlm/internlm2_5-7b-chat
wuenlp/centurio_qwen	Centurio Qwen (Geigle et al., 2025)	Yes	M	Yes	Yes	Yes	qwen/qwen2.5-7b-instruct
qwen/qwen2-vl-2b-instruct	Qwen2 VL 2B (Wang et al., 2024)	Yes	S	Yes	Yes	Yes	qwen/qwen2.5-1.5b-instruct
microsoft/phi-3.5-vision-instruct	Phi 3.5 Vision (Abdin et al., 2024)	Yes	S	Yes	Yes	Yes	microsoft/phi-3.5-mini-instruct
opengvlab/internvl2_5-4b	InternVL2.5 4B (Chen et al., 2024b)	Yes	S	Yes	Yes	Yes	qwen/qwen2.5-3b-instruct
opengvlab/internvl2_5-1b	InternVL2.5 1B (Chen et al., 2024b)	Yes	S	Yes	Yes	Yes	qwen/qwen2.5-0.5b-instruct
opengvlab/internvl2_5-2b	InternVL2.5 2B (Chen et al., 2024b)	Yes	S	Yes	Yes	Yes	internlm/internlm2_5-1_8b-chat
qwen/qwen2.5-72b-instruct	Qwen2.5 72B (Yang et al., 2024)	Yes	XL	No	No	Yes	–
qwen/qwen2.5-32b-instruct	Qwen2.5 32B (Yang et al., 2024)	Yes	L	No	No	Yes	–
internlm/internlm2_5-20b-chat	InternLM2.5 20B (Cai et al., 2024)	Yes	L	No	No	Yes	–
cohereforai/aya-expanse-8b	Aya Expanse 8B (Dang et al., 2024)	Yes	M	No	No	Yes	–
internlm/internlm2_5-7b-chat	InternLM2.5 7B (Cai et al., 2024)	Yes	M	No	No	Yes	–
qwen/qwen2.5-7b-instruct	Qwen2.5 7B (Yang et al., 2024)	Yes	M	No	No	Yes	–
qwen/qwen2.5-0.5b-instruct	Qwen2.5 0.5B (Yang et al., 2024)	Yes	S	No	No	Yes	–
qwen/qwen2.5-3b-instruct	Qwen2.5 3B (Yang et al., 2024)	Yes	S	No	No	Yes	–
qwen/qwen2.5-1.5b-instruct	Qwen2.5 1.5B (Yang et al., 2024)	Yes	S	No	No	Yes	–
internlm/internlm2_5-1_8b-chat	InternLM2.5 1.8B (Cai et al., 2024)	Yes	S	No	No	Yes	–
microsoft/phi-3.5-mini-instruct	Phi 3.5 Mini (Abdin et al., 2024)	Yes	S	No	No	Yes	–

Table 6: Details about the models evaluated within the GIMMICK benchmark. The size "A" indicates that the model is a proprietary API model with unknown size.

B.1 Examples

In the following, we provide one random sample per region for the CIVQA task. Note that the lower part of the examples, where the related CEF is provided, is *not* part of the actual sample.

■ A

Copyrigth: Conseil municipal de Sefrou, 2010

Question: What title is given to the woman wearing the sash in the image?
Answer: Cherry Queen

Related Cultural Event or Facet

Title: Cherry festival in Sefrou
Countries: Morocco
Regions: Arab States
Description:
For three days in June each year, the local population of Sefrou celebrates the natural and cultural beauty of the region, symbolized by the cherry fruit and that year's newly chosen Cherry Queen, selected during a pageant that draws competitors from the region and entire country. The highlight of the festival is a parade with performing troupes, rural and urban music, majorettes and bands, and floats featuring local producers. At the centre is the Cherry Queen, who offers cherries to onlookers while dressed ornately and surrounded by attendants. The whole population contributes to the success of the festival: craftswomen make silk buttons for traditional dresses, fruit growers supply cherries, local sports clubs participate in competitions, and music and dancing troupes animate the entire festival. The cherry festival provides an opportunity for the entire city to present its activities and achievements. The younger generation are also integrated into festival activities to ensure their sustainability. The festival is a source of pride and belonging that enhances the self-esteem of the city and its people and constitutes a fundamental contribution to their local identity.

UNESCO ICH URL: https://ich.unesco.org/en/RL/cherry-festival-in-sefrou-00641...

AP

Copyrigth: 2010 by Centre for Research and Development of Culture, Indonesia

Question: What traditional dance are the performers engaging in, as seen in the image?
Answer: Saman dance

Related Cultural Event or Facet

Title: Saman dance
Countries: Indonesia
Regions: Asian and Pacific States
Description:
The Saman dance is part of the cultural heritage of the Gayo people of Aceh province in Sumatra. Boys and young men perform the Saman sitting on their heels or kneeling in tight rows. Each wears a black costume embroidered with colourful Gayo motifs symbolizing nature and noble values. The leader sits in the middle of the row and leads the singing of verses, mostly in the Gayo language. These offer guidance and can be religious, romantic or humorous in tone. Dancers clap their hands, slap their chests, thighs and the ground, click their fingers, and sway and twist their bodies and heads in time with the shifting rhythm – in unison or alternating with the moves of opposing dancers. These movements symbolize the daily lives of the Gayo people and their natural environment. The Saman is performed to celebrate national and religious holidays, cementing relationships between village groups who invite each other for performances. The frequency of Saman performances and its transmission are decreasing, however. Many leaders with knowledge of the Saman are now elderly and without successors. Other forms of entertainment and new games are replacing informal transmission, and many young people now emigrate to further their education. Lack of funds is also a constraint, as Saman costumes and performances involve considerable expense.

UNESCO ICH URL: https://ich.unesco.org/en/USL/saman-dance-00509...

E

Copyrigth: 2010 by M.Rahimov/Ministry of Culture and Tourism

Question: What is the name of the musical instrument observed by the man in the image?
Answer: Tar

Related Cultural Event or Facet

Title: Craftsmanship and performance art of the Tar, a long-necked string musical instrument
Countries: Azerbaijan
Regions: Eastern European States
Description:
The Tar is a long-necked plucked lute, traditionally crafted and performed in communities throughout Azerbaijan. Considered by many to be the country's leading musical instrument, it features alone or with other instruments in numerous traditional musical styles. Tar makers transmit their skills to apprentices, often within the family. Craftsmanship begins with careful selection of materials for the instrument: mulberry wood for the body, nut wood for the neck, and pear wood for the tuning pegs. Using various tools, crafters create a hollow body in the form of a figure eight, which is then covered with the thin pericardium of an ox. The fretted neck is affixed, metal strings are added and the body is inlaid with mother-of-pearl. Performers hold the instrument horizontally against the chest and pluck the strings with a plectrum, while using trills and a variety of techniques and strokes to add colour. Tar performance has an essential place in weddings and different social gatherings, festive events and public concerts. Players transmit their skills to young people within their community by word of mouth and demonstration, and at educational musical institutions. Craftsmanship and performance of the tar and the skills related to this tradition play a significant role in shaping the cultural identity of Azerbaijanis.

UNESCO ICH URL: https://ich.unesco.org/en/RL/craftsmanship-and-performance-a...

LAC

Copyrigth: Py, 2019

Question: What traditional tool from the Guaraní culture is depicted in the image for drinking Terere?
Answer: Bombilla

Related Cultural Event or Facet

Title: Practices and traditional knowledge of Terere in the culture of Pohã Ñana, Guaraní ancestral drink in Paraguay
Countries: Paraguay
Regions: Latin-American and Caribbean States
Description:
The practices and traditional knowledge of Terere in the culture of Pohã Ñana, Guaraní ancestral drink in Paraguay, are widespread in the Paraguayan territory and involve a variety of bearers. Terere is a traditional drink prepared in a jug or thermos, in which cold water is mixed with Pohã Ñana crushed in a mortar. It is served in a glass pre-filled with yerba mate and sucked with a bombilla (metal or cane straw). Preparing the Terere is an intimate ritual involving a series of pre-established codes and each Pohã Ñana herb has health benefits linked to popular wisdom passed down through the generations. Terere practices in the culture of Pohã Ñana have been transmitted in Paraguayan families since approximately the sixteenth century. Traditional knowledge about the healing attributes of the medicinal herbs that make up the Pohã Ñana and their correct use are also transmitted spontaneously within the family. In recent years, the figure of apprentices has risen, but family transmission remains the main mode of transmission. The practice of the Terere in the culture of Pohã Ñana fosters social cohesion as the time and space dedicated to preparing and consuming the Terere promote inclusion, friendship, dialogue, respect and solidarity. The practice also strengthens new generations' appreciation of the rich cultural and botanical heritage of Guaraní origin.

UNESCO ICH URL: https://ich.unesco.org/en/RL/practices-and-traditional-knowl...

SA

Copyrigth: The Authority for Research and Conservation of Cultural Heritage (ARCCH), 2013

Question: What festival are the people in the image celebrating?
Answer: Fichee-Chambalaalla

Related Cultural Event or Facet

Title: Fichee-Chambalaalla, New Year festival of the Sidama people
Countries: Ethiopia
Regions: Subsaharian African States
Description:
Fichee-Chambalaalla is a New Year festival celebrated among the Sidama people. According to the oral tradition, Fichee commemorates a Sidama woman who visited her parents and relatives once a year after her marriage, bringing ”buurisame”, a meal prepared from false banana, milk and butter, which was shared with neighbours. Fichee has since become a unifying symbol of the Sidama people. Each year, astrologers determine the correct date for the festival, which is then announced to the clans. Communal events take place throughout the festival, including traditional songs and dances. Every member participates irrespective of age, gender and social status. On the first day, children go from house to house to greet their neighbours, who serve them ”buurisame”. During the festival, clan leaders advise the Sidama people to work hard, respect and support the elders, and abstain from cutting down indigenous trees, begging, indolence, false testimony and theft. The festival therefore enhances equity, good governance, social cohesion, peaceful co-existence and integration among Sidama clans and the diverse ethnic groups in Ethiopia. Parents transmit the tradition to their children orally and through participation in events during the celebration. Women in particular, transfer knowledge and skills associated with hairdressing and preparation of ”buurisame” to their daughters and other girls in their respective villages.

UNESCO ICH URL: https://ich.unesco.org/en/RL/fichee-chambalaalla-new-year-fe...

W

Copyrigth: Município de Estremoz, 2015

Question: What specific region's attire is represented by the figures in the image?
Answer: Alentejo

Related Cultural Event or Facet

Title: Craftmanship of Estremoz clay figures
Countries: Portugal
Regions: Western European and North American States
Description:
The Craftsmanship of Estremoz Clay Figures involves a production process lasting several days: the elements of the figures are assembled before being fired in an electric oven and then painted by the artisan and covered with a colourless varnish. The clay figures are dressed in the regional attires of Alentejo or the clothing of religious Christian iconography, and follow specific themes. The production of clay figures in Estremoz dates back to the seventeenth century, and the very characteristic aesthetic features of the figures make them immediately identifiable. The craft is strongly attached to the Alentejo region, since the vast majority of the figures depict natural elements, local trades and events, popular traditions and devotions. The viability and recognition of the craft are ensured through non-formal education workshops and pedagogical initiatives by the artisans, as well as by the Centre for the Appreciation and Safeguarding of the Estremoz Clay Figure. Fairs are organized at the local, national and international levels. Knowledge and skills are transmitted both in family workshops and professional contexts, and artisans teach the basics of their craft through non-formal training initiatives. Artisans are actively involved in awareness-raising activities organized in schools, museums, fairs and other events.

UNESCO ICH URL: https://ich.unesco.org/en/RL/craftmanship-of-estremoz-clay-f...

B.2 Cultural Aspects

During the synthetic data generation phase of the CIVQA, we also obtained a "target aspect" per question (see §B.4 and §B.4.1). We report these aspects in the following.

B.3 External Hint Variations

For the CIVQA (and CVVQA) task, we ablate the effect of external cues or hints on the task performance of models. In the following, we provide the Python pseudo-code snippet to generate the prompt for a given sample.

B.4 Synthetic Data Generation

Aspect	Questions	Aspect	Questions	Aspect	Questions
traditions	390	education	3	jewelry	1
rituals	241	culture	3	objects	1
art	233	games	3	animal	1
music	210	performing arts	3	plants	1
craftsmanship	177	language	3	process	1
instruments	155	performance	3	agriculture	1
festivals	151	characters	2	celebrations	1
dance	150	practices	2	details	1
tools	108	skills	2	historical	1
food	96	origin	2	function or usage	1
clothing	93	cultural identity	2	symbolism	1
architecture	52	technology	1	healthcare	1
sports	38	people	1	knowledge	1
location	28	community	1	social status	1
symbols	19	identity	1	religion	1
drinks	14	environment	1	cultural space	1
customs	13	traditional medicine	1	social space	1
cultural significance	6	nature	1	cultural practice	1
theatre	4	communication	1	unknown	1

Table 7: Cultural aspects targeted by the questions within the CIVQA task.

Python Pseudo-Code for the external cue settings of the CIVQA and CVVQA tasks.

```python
def apply_gimmick_prompt_template(
    sample: dict[str, Any],
    regions_hint: bool,
    countries_hint: bool,
) -> str:

    prompt_template = "{QUESTION}\n{HINTS}\n"
    hints = ""

    if regions_hint:
        hints += (
            "Hint: The question is related to a cultural event or facet from the following
            ↪  region(s): "
            f"{', '.join(sample['regions'])}\n"
        )

    if countries_hint:
        hints += (
            "Hint: The question is related to a cultural event or facet from the following
            ↪  country or countries: "
            f"{', '.join(sample['countries'])}\n"
        )

    return prompt_template.format(
        QUESTION=sample["prompt"],
        HINTS=hints,
    )
```

Figure 73: Python Pseudo-Code to generate the prompt for a given CIVQA (or CVVQA) sample for the external cues settings.

B.4.1 System Prompt

```
# Your Role

You are a professional annotator specialized in creating VQA samples based on a provided
↪  intangible cultural heritage(ICH) item. You will be given the following information
↪  related to the item:

- Image: An image representing one aspect of the ICH item.
- Countries of Origin: The country or countries where this ICH is recognized.
- Regions of Origin: The country or countries where this ICH is recognized.
- Title: The official title of the ICH item.
- Description: A detailed description of the ICH item, including relevant details.

# Your Task

Your task is it to generate high-quality question-answer pairs in a VQA style to assess the
↪  cultural knowledge of the intangible cultural heritage (ICH) item of state-of-the-art
↪  multimodal AI models. Be sure to follow the annotation guidelines provided below to ensure
↪  the quality and relevance of the question-answer pairs.

# Annotation Guidelines

## Question Requirements

Make sure the question meets all of the following requirements:

1. Clear and Concise
    The question is clear and concise and no longer than a single sentence.
2. Directly related to the ICH item
    The question is directly related to the ICH item.
3. Directly related to the visible content
    The question is directly related to the visible content in the image and requires visual
    ↪  analysis to answer.
4. Does not (partially) contain the answer
    The question does not contain any hints or clues to or parts of the answer that would make
    ↪  the answer obvious.
5. Does not contain subjective words
    The question does not contain subjective words like 'likely', 'possibly', 'probably',
    ↪  'eventually', 'might', 'could', 'should', etc., which could introduce ambiguity.
6. Requires both image and cultural knowledge to answer
    The question requires both image and cultural knowledge to answer and is not answerable by
    ↪  looking only at the image or only knowing about the ICH item or reading the textual
    ↪  description.
7. (optional) Includes specific cultural terms
    The answer includes specific cultural terms, names, or phrases related to the ICH item.
    ↪  E.g., particular names mentioned in the description or parts of the title.

## Answer Requirements

Make sure the answer meets all of the following requirements:

1. Single Word or Multiword Expression
    The answer is a single word or multiword expression.
2. Clear, Objective, and Correct
    The answer is clear, objective, and unambiguously correct.
3. Directly Related to Visual Content
    The answer is directly related to the visual content of the image.
4. No General or Abstract Words
    The answer does not contain general, abstract, or non-depictable words like "Traditional",
    ↪  "Cooperation", "Gathering", "Solidarity", "Community", "Indoor", "Outdoor", "Urban",
    ↪  "Rural", etc.
5. Verifiable by Text and Image
    The answer is unambiguously verifiable by reading the textual information and inspecting
    ↪  the image.
6. (optional) Includes specific cultural terms
    The answer includes specific cultural terms, names, or phrases related to the ICH item.
    ↪  E.g., particular names mentioned in the description or parts of the title.
```

```
## Question Characteristics

### Target Aspects

Make sure the question targets different aspects of the ICH item, such as:

- Food
- Drinks
- Clothing
- Art
- Tools
- Sports
- Instruments
- Dance
- Music
- Rituals
- Traditions
- Festivals
- Customs
- Symbols
- Architecture
- Other

### Question Categories

Make sure the question falls into different categories, such as:

- Identification
    Questions that ask for the identification of objects, people, or elements in the image.
    ↪  E.g.: What is the name of the instrument shown in the image?
- Origin
    Questions that inquire about the origin or source of the CEF. E.g.: Which culture or
    ↪  country does this artifact belong to?
- Cultural Significance
    Questions that explore the cultural or religious significance of the depicted element. E.g.:
    ↪  What cultural or religious significance does this item hold in its native context?
- Function or Usage
    Questions that ask about the traditional or historical function or usage of the depicted
    ↪  element. E.g.: What was this object traditionally used for?
- Material and Craftsmanship
    Questions that focus on the materials used and the craftsmanship involved in creating the
    ↪  depicted element. E.g.: What material is used to construct this artifact?
- Location
    Questions that ask about the geographical location where the cultural event or facet takes
    ↪  place. E.g.: In which place does this dance take place?
- Symbolism
    Questions that delve into the symbolic meanings associated with the depicted element. E.g.:
    ↪  What does the color red symbolize in this cultural context?
- Historical
    Questions that relate to historical events or contexts depicted in the image. E.g.: What
    ↪  historical event is depicted in this image?
- Details
    Questions that ask for specific details about the formation, arrangement, or other aspects
    ↪  of the depicted element. E.g.: What formation are the dancers in?
- Other
    Questions that do not fall into the above categories but are relevant to the ICH item.


# Task Strategy

Before generating a question-answer pair, first think step-by-step and analyse the image:

1. What is visible in the image? Generate a highly detailed description of the key elements,
↪  objects, or people in the image. Take into account the textual description provided to
↪  identify details.
2. How does the visible content relate to the intangible cultural heritage item? Identify the
↪  connection between the contents of the image and the intangible cultural heritage item.
```

````
Then, think step-by-step about potential questions:

1. What can be asked about the image that is directly related to the visible content and the intangible cultural heritage item?
2. Can a concise and clear answer to the questions be inferred from the image and the provided information?

Finally, think step-by-step before generating the final question-answer pairs:

1. Does the question-answer pair strictly adhere to the guidelines provided above? Percisly check every part of the guidelines and drop the question-answer pair if it does not meet the criteria.
2. What aspect of the intangible cultural heritage item is targeted with the question?
3. What category does the question fall into?

# Output Format

For each question-answer pair, provide the following information in the following format:
```xml
<vqa-task>
    <image-analysis>
        <description>
            <!-- PUT YOUR DETAILED DESCRIPTION OF THE IMAGE HERE -->
        </description>
        <cultural-relatetness>
            <!-- PUT YOUR ANALYSIS OF HOW THE CONTENTS OF THE IMAGE RELATE TO THE INTANGIBLE CULTURAL HERITAGE ITEM HERE -->
        </cultural-relatetness>
    </image-analysis>
    <potential-questions>
        <qa-candidate>
            <question>
                <!-- PUT YOUR QUESTION HERE -->
            </question>
            <answer>
                <!-- PUT YOUR ANSWER HERE -->
            </answer>
            <guideline-adherence>
                <question-requirments>
                    <clear-and-concise>
                        <!-- YES OR NO -->
                    </clear-and-concise>
                    <directly-related-to-ich>
                        <!-- YES OR NO -->
                    </directly-related-to-ich>
                    <directly-related-to-visual-content>
                        <!-- YES OR NO -->
                    </directly-related-to-visual-content>
                    <does-not-contain-answer>
                        <!-- YES OR NO -->
                    </does-not-contain-answer>
                    <does-not-contain-subjective-words>
                        <!-- YES OR NO -->
                    </does-not-contain-subjective-words>
                    <requires-both-image-and-cultural-knowledge>
                        <!-- YES OR NO -->
                    </requires-both-image-and-cultural-knowledge>
                    <includes-specific-cultural-terms>
                        <!-- YES OR NO -->
                    </includes-specific-cultural-terms>
                </question-requirments>
                <answer-requirments>
                    <single-word-or-multiword-expression>
                        <!-- YES OR NO -->
                    </single-word-or-multiword-expression>
                    <clear-objective-and-correct>
                        <!-- YES OR NO -->
````
````

```
                    </clear-objective-and-correct>
                    <directly-related-to-visual-content>
                        <!-- YES OR NO -->
                    </directly-related-to-visual-content>
                    <no-general-or-abstract-words>
                        <!-- YES OR NO -->
                    </no-general-or-abstract-words>
                    <verifiable-by-text-and-image>
                        <!-- YES OR NO -->
                    </verifiable-by-text-and-image>
                    <includes-specific-cultural-terms>
                        <!-- YES OR NO -->
                    </includes-specific-cultural-terms>
                </answer-requirments>
            </guideline-adherence>
        </qa-candidate>
        ...
    </potential-questions>
    <final-qa-pairs>
        <!-- PUT ALL QA PAIRS THAT MEET ALL MANDATORY REQUIREMENTS HERE -->
        <qa-pair>
            <meets-requirements>
                <!-- DOES YOUR QUESTION-ANSWER PAIR MEET ALL MANDATORY REQUIREMENTS? YES OR NO
                ↪ -->
            </meets-requirements>
            <final-result-json>
                <!-- PUT YOUR FINAL RESULT AS JSON HERE -->
                {
                    "question": <insert question here>,
                    "answer": <insert answer here>,
                    "target_aspect": <insert target aspect here>
                    "question_category": <insert question category here>
                }
            </final-result-json>
        </qa-pair>
        ...
    </final-qa-pairs>
</vqa-task>
```

B.4.2 User Prompt Template

```
# Intangible Cultural Heritage Item

### Image

{IMAGE_PLACEHOLDER}

### Countries of Origin:

{LIST_OF_COUNTRIES}

### Regions of Origin

{LIST_OF_REGIONS}

### Title

{TITLE}

### Description

{DESCRIPTION}
```

B.5 Annotation Project Details

We first conducted several internal pilot studies to iteratively create a straightforward annotation task, guidelines, and an intuitive interface for the final annotation project. To find annotators, we advertised the task in our faculty research network, emphasizing our goal of creating a culturally diverse benchmark for assessing the cultural awareness of current AI models. Therefore, we targeted primarily individuals from non-Western cultural backgrounds. We found 18 volunteers who have spent most of their lives in 10 different countries from all six regions and thus cover diverse cultural backgrounds (see Table 8). To train the annotators, we provided detailed annotation guidelines, followed by an oral introduction to the task. For more details, refer to the (anonymized) original annotation guidelines we shared here.

For the second annotation round, we hired 5 of the previous volunteering annotators (0, 1, 8, 15, 17) who assessed the kept samples from the first round to obtain two annotations (from distinct annotators) per sample. We paid the second-round annotators a salary of roughly 12.5€ per hour.

ID	Age	Pronouns	Education	Country	Region	Round(s)
0	23	she/her	Bachelor	Iran	AP	1, 2
1	23	she/her	Bachelor	Iran	AP	1, 2
2	28	she/her	PhD	Russia	E	1
3	35	he/him	Master	Germany	W	1
5	29	he/him	Bachelor	Guatemala	LAC	1
6	29	he/him	Master	Germany	W	1
7	42	he/him	PhD	Ethiopia	SA	1
8	23	he/him	Bachelor	Egypt	A	1, 2
9	33	she/her	Master	Iran	AP	1
10	29	she/her	Bachelor	Afghanistan	AP	1
11	23	she/her	Bachelor	India	AP	1
12	33	he/him	Bachelor	Germany	W	1
13	22	she/her	Bachelor	Pakistan	AP	1
14	27	he/him	Master	China	AP	1
15	29	she/her	High School	Germany	W	1, 2
16	22	she/her	Bachelor	China	AP	1
17	26	he/him	High School	Germany	W	1, 2, 3

Table 8: Demographics of the annotators who participated in our VQA annotation project. For the country, we asked the question, "*Where did you spend most of your life?*". The Round(s) column indicates which annotation rounds the annotator participated in.

B.5.1 `CIVQA` Annotation Interface

For the annotation project, we used a self-hosted Label Studio[19] instance with a custom labeling interface (see Figure 74) for all annotation projects.

[19] https://labelstud.io/

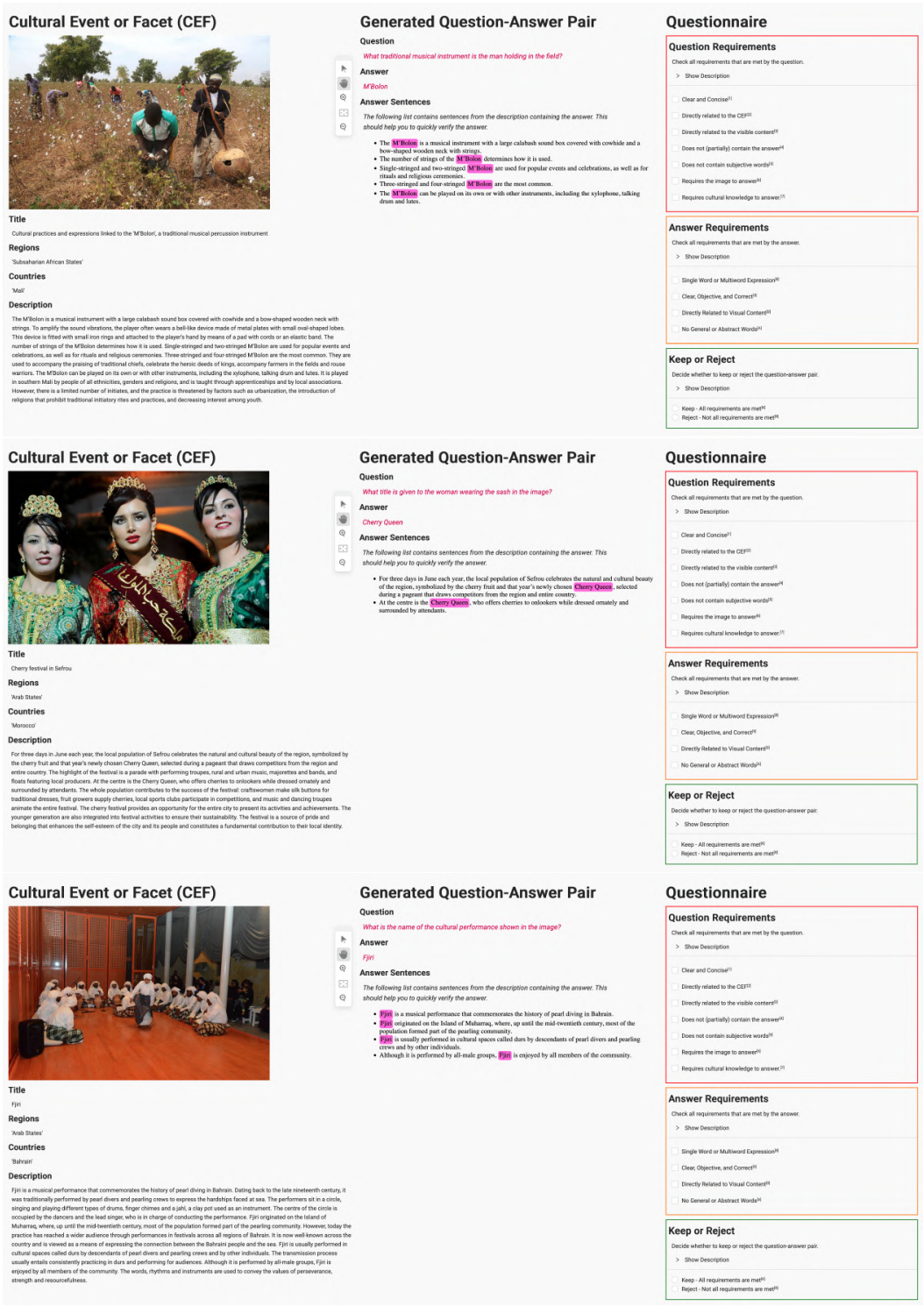


Figure 74: Three screenshots showing examples of the Label Studio interface used in our CIVQA annotation tasks.

B.5.2 First Annotation Round Statistics

Country	Count	Country	Count
United Arab Emirates	101	Nicaragua	18
China	98	Chile	17
Oman	91	Serbia	17
Saudi Arabia	87	Cambodia	17
France	86	Bangladesh	17
Croatia	84	Bulgaria	17
Algeria	82	Qatar	17
Morocco	81	Ireland	17
Türkiye	78	Panama	16
Peru	75	Ukraine	16
Spain	74	Malaysia	16
Azerbaijan	69	Namibia	16
Colombia	68	Philippines	15
Islamic Republic of Iran	66	Bosnia and Herzegovina	15
Mali	65	Niger	15
Mexico	64	Estonia	14
Republic of Korea	62	Netherlands	14
Egypt	62	Zimbabwe	14
Tunisia	56	Senegal	14
Iraq	54	Madagascar	14
Japan	52	Belarus	13
Brazil	50	Luxembourg	13
Italy	50	Togo	12
Belgium	50	Burundi	12
Plurinational State of Bolivia	49	Dominican Republic	12
Mauritania	49	Congo	11
Bolivarian Republic of Venezuela	47	Democratic Republic of the Congo	11
Nigeria	46	Benin	11
India	45	Finland	11
Malawi	43	Angola	10
Palestine	40	Afghanistan	10
Greece	38	Seychelles	10
Uzbekistan	37	Democratic People's Republic of Korea	10
Kuwait	37	Norway	9
Kyrgyzstan	36	Lao Peoples Democratic Republic	9
Cuba	35	Burkina Faso	9
Mauritius	34	Sweden	9
Mongolia	34	Bahamas	9
Czechia	34	Georgia	9
Jordan	32	Albania	9
Zambia	31	Republic of Moldova	9
Côte d'Ivoire	31	Cabo Verde	8
Syrian Arab Republic	31	North Macedonia	8
Kazakhstan	30	Jamaica	8
Portugal	29	Honduras	7
Switzerland	29	Latvia	7
Uganda	29	Denmark	7
Ethiopia	29	Pakistan	7
Botswana	28	Belize	7
Viet Nam	28	Uruguay	7
Argentina	28	Timor-Leste	6
Armenia	28	Montenegro	6
Yemen	28	Sri Lanka	6
Turkmenistan	26	Thailand	6
Sudan	26	Guinea	6
Bahrain	26	Malta	5
Indonesia	26	Andorra	5
Ecuador	25	Russian Federation	5
Mozambique	25	Lithuania	5
Tajikistan	25	Tonga	4
Austria	24	Costa Rica	4
Hungary	24	Cameroon	4
Slovakia	23	Vanuatu	3
Lebanon	23	Singapore	3
Cyprus	22	Gambia	3
Slovenia	22	Iceland	3
Paraguay	21	Federated States of Micronesia	2
Germany	21	Grenada	2
Romania	21	Samoa	2
Guatemala	20	Bhutan	1
Kenya	20	Djibouti	1
Poland	20	Central African Republic	1

Table 9: The number of countries related to the QA pairs collected in the first annotation round for CIVQA.

C VVQA Details

C.1 Examples

In the following, we provide one random sample per region for the CVVQA task. Note that the lower part of the examples, where the related CEF is provided, is *not* part of the actual sample.

■ A

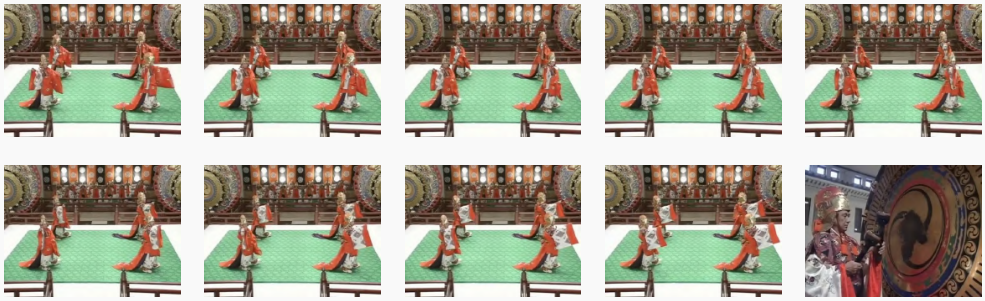
Question: What event are the women in the video participating in?
Answer: Moussem of Tan-Tan

Related Cultural Event or Facet

Title: Moussem of Tan-Tan
Countries: Morocco
Regions: Arab States
Description:
The Moussem of Tan-Tan in southwest Morocco is an annual gathering of nomadic peoples of the Sahara that brings together more than thirty tribes from southern Morocco and other parts of northwest Africa. Originally this was an annual event around the month of May. Part of the agricultural and herding calendar of the nomads, these gatherings were an opportunity to group together, buy, sell and exchange foodstuffs and other products, organize camel and horse-breeding competitions, celebrate weddings and consult herbalists. The Moussem also included a range of cultural expressions such as musical performances, popular chanting, games, poetry contests and other Hassanie oral traditions.
These gatherings took the form of a Moussem (a type of annual fair with economic, cultural and social functions) in 1963 when the first Moussem of Tan-Tan was organized to promote local traditions and provide a place for exchange, meeting and celebration. The Moussem is said to have been initially associated with Mohamed Laghdaf, who resisted the Franco-Spanish occupation. He died in 1960, and his tomb lies near the town. However, between 1979 and 2004 it was not possible to hold the Moussem because of security problems in the region.
Today, the nomadic populations are particularly concerned to protect their way of life. Economic and technical upheavals in the region have profoundly altered the lifestyle of the nomadic Bedouin communities, forcing many of them to settle. Moreover, urbanization and rural exodus have contributed to the loss of many aspects of the traditional culture of these populations, such as crafts and poetry. Because of these risks, Bedouin communities rely strongly on the renewed Moussem of Tan-Tan to assist them in ensuring the survival of their know-how and traditions.

UNESCO ICH URL: https://ich.unesco.org/en/RL/moussem-of-tan-tan-00168...

AP



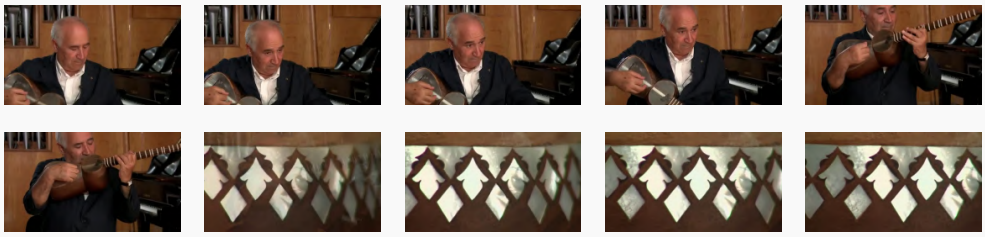
Question: What traditional Japanese performance art is depicted by the performers in the video?
Answer: Gagaku

Related Cultural Event or Facet

Title: Gagaku
Countries: Japan
Regions: Asian and Pacific States
Description:
Gagaku, characterized by long, slow songs and dance-like movements, is the oldest of the Japanese traditional performing arts. It is performed at banquets and ceremonies in the Imperial Palace and in theatres throughout the country, and encompasses three distinct arts. The first, Kuniburi no Utamai, features ancient Japanese songs, partial accompaniment by harp and flute and simple choreography. The second consists of instrumental music (especially wind instruments) and a ceremonial dance developed on the Asian continent and subsequently adapted by Japanese artists. The third, Utamono, is danced to vocal music whose texts include Japanese folk songs and Chinese poems. Influenced by the politics and culture of different periods over its long evolution, Gagaku continues to be transmitted to apprentices by masters in the Music Department of the Imperial Household Agency, many of whom are the descendants of families with deep roots in the art. It is not only an important cultural tool in confirming Japanese identity and a crystallization of the history of Japanese society, but also a demonstration of how multiple cultural traditions can be fused into a unique heritage through constant recreation over time.

UNESCO ICH URL: https://ich.unesco.org/en/RL/gagaku-00265...

E



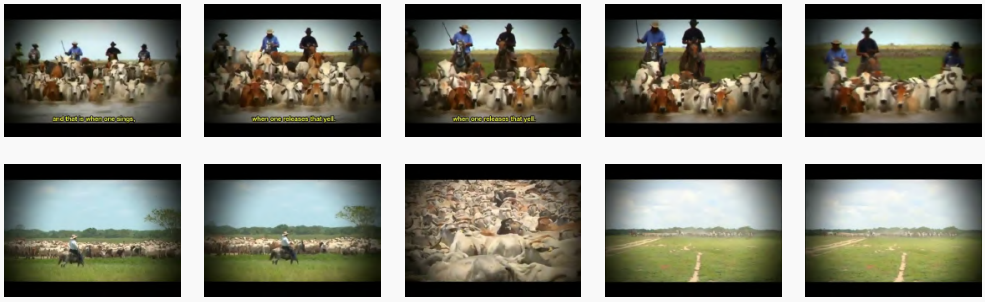
Question: What instrument is the individual playing in the video?
Answer: Tar

Related Cultural Event or Facet

Title: Craftsmanship and performance art of the Tar, a long-necked string musical instrument
Countries: Azerbaijan
Regions: Eastern European States
Description:
The Tar is a long-necked plucked lute, traditionally crafted and performed in communities throughout Azerbaijan. Considered by many to be the country's leading musical instrument, it features alone or with other instruments in numerous traditional musical styles. Tar makers transmit their skills to apprentices, often within the family. Craftsmanship begins with careful selection of materials for the instrument: mulberry wood for the body, nut wood for the neck, and pear wood for the tuning pegs. Using various tools, crafters create a hollow body in the form of a figure eight, which is then covered with the thin pericardium of an ox. The fretted neck is affixed, metal strings are added and the body is inlaid with mother-of-pearl. Performers hold the instrument horizontally against the chest and pluck the strings with a plectrum, while using trills and a variety of techniques and strokes to add colour. Tar performance has an essential place in weddings and different social gatherings, festive events and public concerts. Players transmit their skills to young people within their community by word of mouth and demonstration, and at educational musical institutions. Craftsmanship and performance of the tar and the skills related to this tradition play a significant role in shaping the cultural identity of Azerbaijanis.

UNESCO ICH URL: https://ich.unesco.org/en/RL/craftsmanship-and-performance-a...

LAC



Question: In which environment do the cultural practices depicted in the video typically occur?
Answer: Llanos

Related Cultural Event or Facet

Title: Colombian-Venezuelan llano work songs
Countries: Colombia, Venezuela (Bolivarian Republic of)
Regions: Latin-American and Caribbean States
Description:
Colombian-Venezuelan llano work songs are a practice of vocal communication consisting of tunes sung individually, a capella, on the themes of herding and milking. The practice emerged from the close relationship between human communities and cattle and horses and is in harmony with the environmental conditions and the dynamics of nature, forming part of the traditional animal husbandry system of the Llanos. Transmitted orally from childhood, the songs are repositories of the individual and collective stories of the llaneros. Llano work songs have been gradually affected by economic, political and social processes that, modifying the llanero cultural universe, have significantly weakened the practice. For example, ambitious government plans conceived from a developmental perspective have led to profound changes in the use of the land and in ownership systems, and the modification of the social, cultural and natural sites of the songs have resulted in a loss of interest in the values and techniques of llano work. Llanero work songs thus face various threats to their viability. Efforts to safeguard the element are nonetheless widespread, including a pedagogical strategy involving more than twenty meetings for bearers and young people in the region, training projects for schoolteachers and a proliferation of festivals.

UNESCO ICH URL: https://ich.unesco.org/en/USL/colombian-venezuelan-llano-wor...

SA

Question: What type of theatre is depicted in the video, known for using elaborate costumes and performances?
Answer: Kwagh-Hir

Related Cultural Event or Facet

Title: Kwagh-Hir theatrical performance
Countries: Nigeria
Regions: Subsaharian African States
Description:
Kwagh-Hir theatrical performance is a composite art form encompassing a spectacle that is both visually stimulating and culturally edifying. Kwagh-hir has its roots in the story-telling tradition of the Tiv people called 'kwagh-alom', a practice where the family was treated to a storytelling session by creative storytellers, usually in the early hours of the night after the day's farming work. With time, creative storytellers began to dramatize these stories, culminating in the present stage and status of Kwagh-hir. The practice is a social performance with the potential to entertain and teach moral lessons through the dramatization and performance of past and current social realities. As a form of total theatre, Kwagh-hir incorporates puppetry, masquerading, poetry, music, dance and animated narratives in articulating the reality of the Tiv people. People's daily struggles, aspirations, successes and failures are all given expression through creative dramatization. Khwagh-hir theatre is owned by the community, with knowledge and skills being transmitted through apprenticeship. People who indicate an interest in the troupe's activities are trained and mentored until they reach a certain level of proficiency; they are then accepted into the troupe. Regular performances are held to ensure the art is kept alive and that the younger generation continues to identify with it.

UNESCO ICH URL: https://ich.unesco.org/en/RL/kwagh-hir-theatrical-performanc...

W

Question: What traditional practice is depicted with the herders and sheep in the video?
Answer: Transhumance

Related Cultural Event or Facet

Title: Transhumance, the seasonal droving of livestock
Countries: Albania, Andorra, Austria, Croatia, Spain, France, Greece, Italy, Luxembourg, Romania
Regions: Western European and North American States, Eastern European States
Description:
Transhumance refers to the seasonal movement of people with their livestock between geographical or climatic regions. Each year, in spring and autumn, men and women herders organise the movement of thousands of animals along traditional pastoral paths. They move on foot or horseback, leading with their dogs and sometimes accompanied by their families. An ancestral practice, transhumance stems from a deep knowledge about the environment and entails social practices and rituals related to the care, breeding and training of animals and the management of natural resources. An entire socio-economic system has been developed around transhumance, from gastronomy to local handicrafts and festivities marking the beginning and end of a season. Families have been enacting and transmitting transhumance through observation and practice for many generations. Communities living along transhumance routes also play an important role in its transmission, such as by celebrating herd crossings and organising festivals. The practice is also transmitted through workshops organised by local communities, associations and networks of herders and farmers, as well as through universities and research institutes. Transhumance thus contributes to social inclusion, strengthening cultural identity and ties between families, communities and territories while counteracting the effects of rural depopulation.

UNESCO ICH URL: https://ich.unesco.org/en/RL/transhumance-the-seasonal-drovi...

C.2 Annotation Project Details

The expert who annotated the samples was Annotator 17 from Table 8. As for the CIVQA task, we used a self-hosted Label Studio instance with a custom labeling interface. The UI is depicted in Figure 75.

D COQA Details

D.1 Prompts

In the following, the prompts for the COQA_{R} and COQA_{C} tasks are provided. For the variations involving images, the image placeholder gets replaced N times, where N is the number of images related to the target CEF.

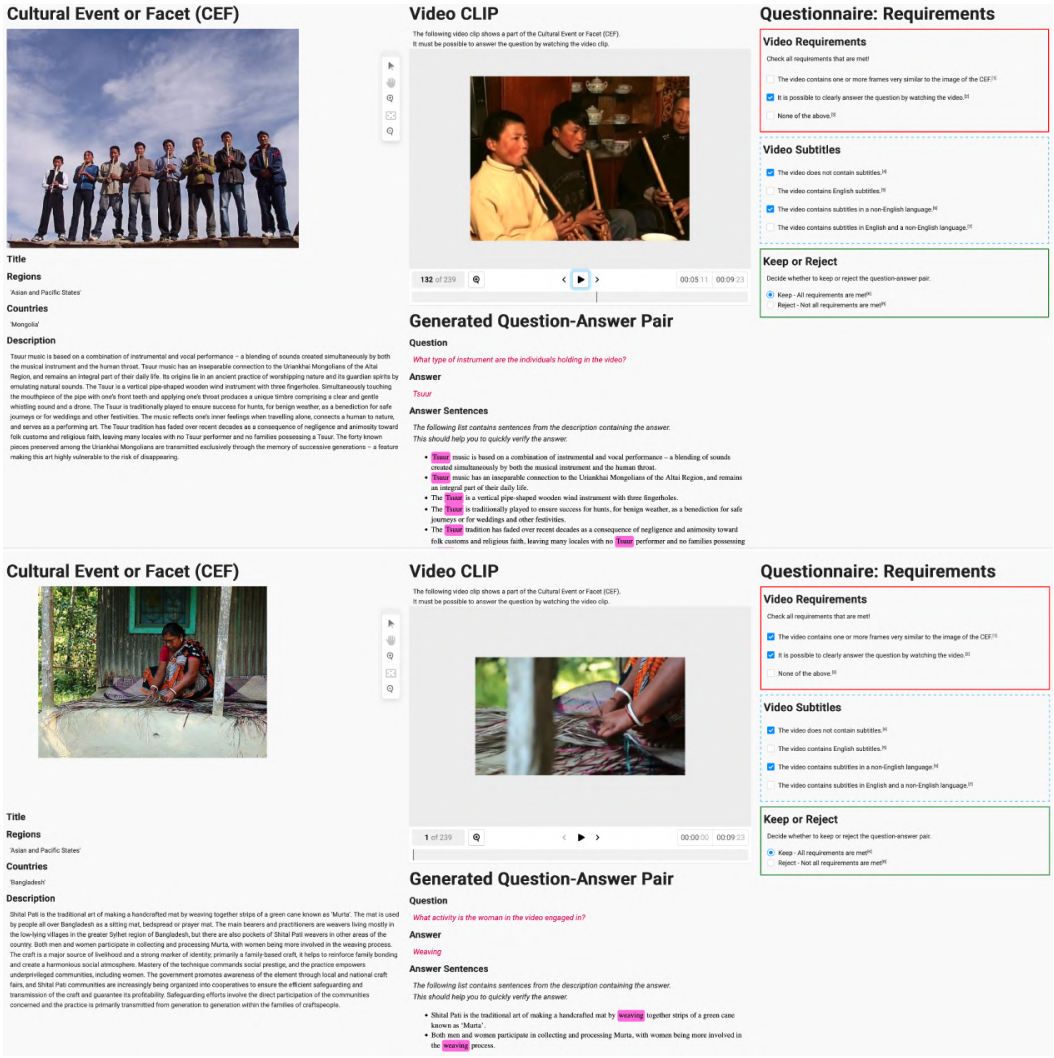


Figure 75: Two screenshots showing examples of the Label Studio interface used in our VVQA annotation tasks.

Region — Text-Only

```
From which of the following regions does the cultural event or facet with the title `{TITLE}`
↪  originate?
Choose from the following options and output only the corresponding letter.

A. {REGION_OPTION_A}
B. {REGION_OPTION_B}
C. {REGION_OPTION_C}
D. {REGION_OPTION_D}

Your answer letter:
```

Region — Image-Only

```
<IMAGE_PLACEHOLDER>

From which of the following countries does the cultural event or facet shown in the images
↪  originate?
Choose from the following options and output only the corresponding letter.

A. {REGION_OPTION_A}
B. {REGION_OPTION_B}
C. {REGION_OPTION_C}
D. {REGION_OPTION_D}

Your answer letter:
```

Region — Text-Image

```
<IMAGE_PLACEHOLDER>

From which of the following regions does the cultural event or facet with the title `{TITLE}`
↪  shown in the images originate?
Choose from the following options and output only the corresponding letter.

A. {REGION_OPTION_A}
B. {REGION_OPTION_B}
C. {REGION_OPTION_C}
D. {REGION_OPTION_D}

Your answer letter:
```

Figure 76: Prompts for the $\texttt{COQA}_{\texttt{R}}$ task.

Country — Text-Only

```
From which of the following countries does the cultural event or facet with the title
↪  `{TITLE}` originate?
Choose from the following options and output only the corresponding letter.

A. {COUNTRY_OPTION_A}
B. {COUNTRY_OPTION_B}
C. {COUNTRY_OPTION_C}
D. {COUNTRY_OPTION_D}

Your answer letter:
```

Country — Image-Only

```
<IMAGE_PLACEHOLDER>

From which of the following countries does the cultural event or facet with the title
↪  `{TITLE}` originate?
Choose from the following options and output only the corresponding letter.

A. {COUNTRY_OPTION_A}
B. {COUNTRY_OPTION_B}
C. {COUNTRY_OPTION_C}
D. {COUNTRY_OPTION_D}

Your answer letter:
```

Country — Text-Image

```
<IMAGE_PLACEHOLDER>

From which of the following countries does the cultural event or facet with the title
↪  `{TITLE}` shown in the images originate?
Choose from the following options and output only the corresponding letter.

A. {COUNTRY_OPTION_A}
B. {COUNTRY_OPTION_B}
C. {COUNTRY_OPTION_C}
D. {COUNTRY_OPTION_D}

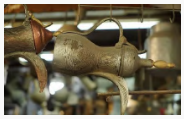
Your answer letter:
```

Figure 77: Prompts for the COQA_{C} task.

D.2 Examples

In the following, we provide one random sample per region for the $\texttt{COQA}_\texttt{C}$ task in the image-only setting. For the other settings and the COQA tasks, the same pattern applies using the respective prompts from above. Note that the lower part of the examples, where the related CEF is provided, is *not* part of the actual sample.

A



Copyrigth: Huzaifa Ayad Bahaa El Din, Iraq, 2021

Copyrigth: Huzaifa Ayad Bahaa El Din, Iraq, 2021

Copyrigth: Huzaifa Ayad Bahaa El Din, Iraq, 2021

Copyrigth: Zahia Benabdallah, Algeria, 2021

Copyrigth: Azza Fahmi, Egypt, 2021

Copyrigth: Mustafa Kamil, Egypt, 2021

Copyrigth: National Heritage Preservation, Ministry of Culture, Youth and Sport and Relations with the Parliament, Egypt, 2022

Copyrigth: Direction du Patrimoine Culturel, Morocco, 2021

Copyrigth: Direction du Patrimoine Culturel, Morocco, 2021

Copyrigth: Ministry of Culture, Palestine, 2021

Question: In which of the following countries does the event shown in the images take place? Choose from the following options and output only the corresponding letter.

A. Kuwait

B. Jordan

C. Egypt

D. United Arab Emirates

Your answer letter:

Answer: C

Related Cultural Event or Facet

Title: Arts, skills and practices associated with engraving on metals (gold, silver and copper)

Countries: Algeria, Saudi Arabia, Egypt, Iraq, Morocco, Mauritania, Palestine, Sudan, Tunisia, Yemen

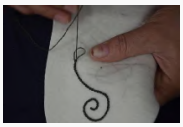
Regions: Arab States

Description:

Engraving on metals such as gold, silver and copper is a centuries-old practice that entails manually cutting words, symbols or patterns into the surfaces of decorative, utilitarian, religious or ceremonial objects. The craftsperson uses different tools to manually cut symbols, names, Quran verses, prayers and geometric patterns into the objects. Engravings can be concave (recessed) or convex (elevated), or the result of a combination of different types of metals, such as gold and silver. Their social and symbolic meanings and functions vary according to the communities concerned. Engraved objects, such as jewelry or household objects, are often presented as traditional gifts for weddings or used in religious rituals and alternative medicine. For instance, certain types of metals are believed to have healing properties. Engraving on metals is transmitted within families, through observation and hands-on practice. It is also transmitted through workshops organized by training centres, organizations and universities, among others. Publications, cultural events and social media further contribute to the transmission of the related knowledge and skills. Practised by people of all ages and genders, metal engraving and the use of engraved objects are means of expressing the cultural, religious and geographical identity and the socioeconomic status of the communities concerned.

UNESCO ICH URL: https://ich.unesco.org/en/RL/arts-skills-and-practices-assoc...

AP



Question: In which of the following countries does the event shown in the images take place? Choose from the following options and output only the corresponding letter.

A. Kyrgyzstan
B. Timor-Leste
C. Thailand
D. Turkmenistan

Your answer letter:
Answer: A

Related Cultural Event or Facet

Title: Ak-kalpak craftsmanship, traditional knowledge and skills in making and wearing Kyrgyz men's headwear
Countries: Kyrgyzstan
Regions: Asian and Pacific States
Description:
Ak-kalpak craftsmanship is a traditional Kyrgyz handicraft. The Ak-kalpak is a traditional male hat made with white felt, which bears deep sacral meanings. Ak-kalpak craftsmanship is a cumulative, ever-evolving body of knowledge and skills passed down by craftswomen in the communities concerned comprising felting, cutting and sewing and pattern embroidery. Related knowledge and skills are transmitted via oral coaching, hands-on training and joint making in workshops. More than eighty kinds of Ak-kalpak can be distinguished, decorated with various patterns bearing a sacred meaning and history. Environmentally friendly and comfortable, the Ak-kalpak resembles a snow peak, with four sides representing the four elements: air, water, fire and earth. The four edging lines symbolize life, with the tassels on the top symbolizing ancestors' posterity and memory, and the pattern symbolizing the family tree. Ak-kalpak unites different Kyrgyz tribes and communities and makes Kyrgyz people recognizable to other ethnic groups. It also fosters inclusivity when representatives of other ethnic groups wear it on holidays or days of mourning to express unity and sympathy. There are workshops all over the country where related knowledge and skills are passed down, and in 2013 a project entitled 'From generation to generation' was conducted on traditional Ak-kalpak-making techniques nationwide, resulting in an exhibition and published book.

UNESCO ICH URL: https://ich.unesco.org/en/RL/ak-kalpak-craftsmanship-traditi...

E

Copyrigth: Lithuanian National Culture Centre, Archive, 2021

Copyrigth: Vilnius Ethnic Culture Centre, Archive, 2021

Copyrigth: Vilnius Ethnic Culture Centre, Archive, 2021

Copyrigth: Lithuanian National Culture Centre, Archive, 2021

Copyrigth: Vilnius Ethnic Culture Centre, Archive, 2021

Copyrigth: Vilnius Ethnic Culture Centre, Archive, 2021

Copyrigth: Vilnius Ethnic Culture Centre, Archive, 2021

Copyrigth: Lithuanian National Culture Centre, Archive, 2021

Copyrigth: Marija Liugienė, Archive, 2003

Copyrigth: Lithuanian National Culture Centre, Archive, 2021

Question: In which of the following countries does the event shown in the images take place? Choose from the following options and output only the corresponding letter.

A. Lithuania
B. Bosnia and Herzegovina
C. Russia
D. Poland

Your answer letter:
Answer: A

Related Cultural Event or Facet

Title: Sodai straw garden making in Lithuania
Countries: Lithuania
Regions: Eastern European States
Description:
Sodai straw gardens are hanging ornaments made from the stalks of grains. This practice involves the cultivation of grain (typically rye), the treatment of straw and the creation of geometric structures of varying sizes. The structures are then decorated with details symbolizing fertility and prosperity. Sodai gardens are believed to reflect the pattern of the universe and are associated with well-being and spirituality. They are hung over the cradles of babies and over a wedding or family table to wish happiness to newborns, fertility to newlyweds or harmony to the family. Lithuanian homes are also frequently decorated with sodai gardens for Easter and Christmas. Some sodai-making families have been practising the tradition for generations. Although most of the practitioners are women, workshops exist and are open to people of all ages and genders. The practice is passed on informally within families or during events such as festivals, exhibitions, conferences and summer camps. An integral part of traditional wooden home interiors, sodai gardens are viewed as spiritual gifts. They provide a sense of shared cultural heritage and continuity to the practising communities while strengthening communal partnerships, intergenerational bonds and cultural diversity.

UNESCO ICH URL: https://ich.unesco.org/en/RL/sodai-straw-garden-making-in-li...

LAC

Copyrigth: Gerson Fonseca/Ministry of Culture of Colombia, 2018

Copyrigth: Gerson Fonseca/Ministry of Culture of Colombia, 2018

Copyrigth: Gerson Fonseca/Ministry of Culture of Colombia, 2018

Copyrigth: Gerson Fonseca/Ministry of Culture of Colombia, 2018

Copyrigth: Gerson Fonseca/Ministry of Culture of Colombia, 2018

Copyrigth: Gerson Fonseca/Ministry of Culture of Colombia, 2018

Copyrigth: Gerson Fonseca/Ministry of Culture of Colombia, 2018

Copyrigth: Gerson Fonseca/Ministry of Culture of Colombia, 2018

Copyrigth: Gerson Fonseca/Ministry of Culture of Colombia, 2018

Copyrigth: Gerson Fonseca/Ministry of Culture of Colombia, 2018

Question: In which of the following countries does the event shown in the images take place? Choose from the following options and output only the corresponding letter.

A. Dominican Republic

B. Chile

C. Colombia

D. Grenada

Your answer letter:

Answer: C

Related Cultural Event or Facet

Title: Safeguarding strategy of traditional crafts for peace building
Countries: Colombia
Regions: Latin-American and Caribbean States
Description:
The safeguarding strategy of traditional crafts for peace building addresses the weakening of traditional crafts through a system of intergenerational transmission of knowledge between master and apprentice based on the non-formal 'learning by doing' method. The safeguarding strategy aims to train different sectors of the population, create labour connections and foster cultural entrepreneurship. It establishes a link between bearers of traditional crafts and skills who are recognized by their communities for their empirical knowledge of the peculiarities of their region and apprentices aged between fourteen and thirty-five who become builders of peace by learning a skill or craft, seeking to transform their situation of vulnerability. The safeguarding strategy is therefore geared at: allowing for the qualification of traditional crafts, thereby improving employment opportunities; implementing a Traditional Crafts Policy to guide and ensure continuity in the transmission and practice of these crafts; and enhancing the Workshop Schools Programme. Priority is accorded to young people who are exposed to the effects of armed conflict, a lack of opportunities, school desertion and unemployment. Training is also combined with work, guaranteeing apprentices' future employability. The strategy thus aims to foster the safeguarding of traditional crafts as a tool for social inclusion, employment and cultural entrepreneurship. In turn, the community can recognize the cultural and societal value of safeguarding different traditional skills and crafts.

UNESCO ICH URL: https://ich.unesco.org/en/BSP/safeguarding-strategy-of-tradi...

SA

Copyrigth: Etienne Kokolo, Kinshasa, République du Congo, 2018

Copyrigth: Etienne Kokolo, Kinshasa, République du Congo, 2019

Copyrigth: Etienne Kokolo, Kinshasa, République du Congo, 2018

Copyrigth: Etienne Kokolo, Kinshasa, République du Congo, 2018

Copyrigth: Etienne Kokolo, Kinshasa, République du Congo, 2018

Copyrigth: Etienne Kokolo, Kinshasa, République du Congo, 2017

Copyrigth: Etienne Kokolo, Kinshasa, République du Congo, 2018

Copyrigth: Etienne Kokolo, Kinshasa, République du Congo, 2020

Copyrigth: Etienne Kokolo, Kinshasa, République du Congo, 2017

Copyrigth: Etienne Kokolo, Kinshasa, République du Congo, 2020

Question: In which of the following countries does the event shown in the images take place? Choose from the following options and output only the corresponding letter.

A. Congo
B. Togo
C. Namibia
D. Nigeria

Your answer letter:
Answer: A

Related Cultural Event or Facet

Title: Congolese rumba
Countries: Congo, Democratic Republic of the Congo
Regions: Subsaharian African States
Description:
Congolese rumba is a musical genre and a dance common in urban areas of the Democratic Republic of the Congo and the Republic of the Congo. Generally danced by a male-female couple, it is a multicultural form of expression originating from an ancient dance called nkumba (meaning 'waist' in Kikongo). The rumba is used for celebration and mourning, in private, public and religious spaces. It is performed by professional and amateur orchestras, choirs, dancers and individual musicians, and women have played a predominant role in the development of religious and romantic styles. The tradition of Congolese rumba is passed down to younger generations through neighbourhood clubs, formal training schools and community organisations. For instance, rumba musicians maintain clubs and apprentice artists to carry on the practice and the manufacture of instruments. The rumba also plays an important economic role, as orchestras are increasingly developing cultural entrepreneurship aimed at reducing poverty. The rumba is considered an essential and representative part of the identity of Congolese people and its diaspora. It is perceived as a means of conveying the social and cultural values of the region and of promoting intergenerational and social cohesion and solidarity.

UNESCO ICH URL: https://ich.unesco.org/en/RL/congolese-rumba-01711...

W

Copyrigth: Servicio de Patrimonio Histórico de la Región de Murcia, 2005

Copyrigth: Generalitat Valenciana, 2005

Copyrigth: Servicio de Patrimonio Histórico de la Región de Murcia, 2005

Copyrigth: Generalitat Valenciana, 2005

Copyrigth: Servicio de Patrimonio Histórico de la Región de Murcia, 2005

Copyrigth: Servicio de Patrimonio Histórico de la Región de Murcia, 2005

Copyrigth: Servicio de Patrimonio Histórico de la Región de Murcia, 2005

Copyrigth: Servicio de Patrimonio Histórico de la Región de Murcia, 2005

Copyrigth: Servicio de Patrimonio Histórico de la Región de Murcia, 2005

Copyrigth: Servicio de Patrimonio Histórico de la Región de Murcia, 2005

Question: In which of the following countries does the event shown in the images take place? Choose from the following options and output only the corresponding letter.

A. Austria

B. Spain

C. Cyprus

D. United Kingdom of Great Britain and Northern Ireland

Your answer letter:

Answer: B

Related Cultural Event or Facet

Title: Irrigators' tribunals of the Spanish Mediterranean coast: the Council of Wise Men of the plain of Murcia and the Water Tribunal of the plain of Valencia

Countries: Spain

Regions: Western European and North American States

Description:

The irrigators' tribunals of the Spanish Mediterranean coast are traditional law courts for water management that date back to the al-Andalus period (ninth to thirteenth centuries). The two main tribunals – the Council of Wise Men of the Plain of Murcia and the Water Tribunal of the Plain of Valencia – are recognized under Spanish law. Inspiring authority and respect, these two courts, whose members are elected democratically, settle disputes orally in a swift, transparent and impartial manner. The Council of Wise Men has seven geographically representative members, and has jurisdiction over a landowners' assembly of 23,313 members. The Water Tribunal comprises eight elected administrators representing a total of 11,691 members from nine communities. In addition to their legal role the irrigators' tribunals play a key part in the communities of which they are a visible symbol, as apparent from the rites performed when judgments are handed down and the fact that the tribunals often feature in local iconography. They provide cohesion among traditional communities and synergy between occupations (wardens, inspectors, pruners, etc.), contribute to the oral transmission of knowledge derived from centuries-old cultural exchanges, and have their own specialist vocabulary peppered with Arabic borrowings. In short, the courts are long-standing repositories of local and regional identity and are of special significance to local inhabitants.

UNESCO ICH URL: https://ich.unesco.org/en/RL/irrigators-tribunals-of-the-spa...

E CKQA Details

E.1 Prompts

In the following, the prompts for the $\texttt{CKQA}_{\texttt{N}}$ and $\texttt{CKQA}_{\texttt{D}}$ tasks are provided. For the variations involving images, the image placeholder gets replaced N times, where N is the number of images related to the target CEF. Examples without the respective prompts, i.e., only the related CEFs, are provided in §A.2.1.

Naming — Image-Only

```
Name the cultural event or facet depicted by the following images. Answer briefly and
↪  concisely.

<IMAGE_PLACEHOLDER>

Your answer:
```

Figure 138: Prompt for the $\texttt{CKQA}_{\texttt{N}}$ task.

Describing — Text-Only

```
Write a brief essay about the cultural event or facet with the title `{TITLE}`.

Your answer:
```

Describing — Image-Only

```
Write a brief essay about the cultural event or facet depicted by the following images.

<IMAGE_PLACEHOLDER>

Your answer:
```

Describing — Text-Image

```
Write a brief essay about the cultural event or facet depicted by the following images. It has
↪  the title `{TITLE}`.

<IMAGE_PLACEHOLDER>

Your answer:
```

Figure 139: Prompts for the $\texttt{CKQA}_{\texttt{D}}$ task.

F Experimental Setup

For inference, we load all models using the *transformers* library (`v.4.48.0`) in 16-bit with Flash Attention 2 (Dao et al., 2022; Dao, 2024) (`v.2.7.3`), PyTorch (`v.2.4.0`), and CUDA (`v12.1`). We used A40 (46GB) GPUs for models up to 26B parameters, A100 (80GB) GPUs for models up to 38B parameters, and two H100 (96GB) GPUs for 70B+ models in a multi-GPU setup. To generate responses, we use greedy decoding, i.e., we use the following arguments for the generation method:

```
generation_kwargs = {
    "max_new_tokens": 512,
    "do_sample": False,
    "temperature": None,
    "top_p": None,
    "top_k": None,
}
```

More details and exact hyperparameters are documented in the code base: https://github.com/floschne/gimmick.

G Results and Analyses

G.1 CIVQA

G.1.1 Results

Relaxed Accuracy

Model	West EU & North America				Asia & Pacific				Subsaharian Africa				Arab				East EU				Latin-America & Caribbean				Average			
	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B
GPT-4o	31.58	34.39	41.05	40.70	29.89	31.37	36.63	37.68	17.38	17.63	32.49	31.74	25.70	30.53	39.19	39.95	26.80	32.56	42.94	41.79	23.17	25.77	30.26	32.39	25.44	28.17	36.59	37.08
Gemini Pro	27.02	30.53	31.23	32.28	22.53	26.11	31.16	29.68	16.84	14.61	26.20	24.18	19.85	22.39	28.50	28.50	25.07	25.94	31.12	32.56	22.46	19.86	24.35	27.19	21.50	22.84	28.30	28.30
GPT-4o Mini	23.86	25.26	30.18	29.82	21.05	21.89	26.74	26.53	9.32	10.58	16.12	15.37	17.30	19.85	25.45	25.95	19.02	19.02	28.53	28.24	16.31	17.73	23.64	22.93	17.38	18.54	24.81	24.59
Gemini Flash	22.81	25.96	27.02	24.91	18.95	20.21	26.11	25.89	12.91	10.83	20.40	18.89	15.27	17.56	20.36	20.61	20.17	19.31	24.78	24.50	14.66	16.55	22.22	20.57	16.85	18.00	23.29	22.44
InternVL2.5 78B	25.61	23.86	29.82	29.12	20.21	19.79	26.32	27.58	10.33	11.08	20.40	20.40	17.81	19.85	27.74	27.99	19.02	17.58	24.50	23.63	13.95	15.13	20.80	21.51	16.75	16.97	24.45	24.72
Qwen2 VL 72B	22.46	22.81	29.82	29.12	17.47	19.16	21.47	23.37	8.31	8.56	12.85	13.10	13.99	16.28	20.10	19.85	21.04	20.46	28.53	29.39	13.00	14.66	19.86	19.62	15.32	16.26	21.45	21.59
InternVL2.5 38B	23.86	23.16	28.77	29.82	17.26	17.89	22.32	23.16	9.07	8.82	17.88	16.62	14.25	17.30	23.16	22.65	16.14	17.29	24.78	23.92	11.82	12.29	17.97	17.49	14.55	15.41	21.99	21.63
Claude 3.5 Sonnet	19.65	17.19	22.11	24.21	16.42	12.84	18.11	22.95	6.30	4.53	10.58	11.59	13.99	11.20	17.81	20.61	16.71	14.12	21.90	21.61	13.48	13.00	17.97	22.93	14.02	11.64	17.60	20.24
InternVL2.5 26B	20.00	19.65	25.61	25.96	13.26	14.95	18.95	18.74	6.30	6.80	11.59	12.34	12.98	14.76	20.61	21.12	15.56	14.41	21.04	21.04	13.00	14.89	19.62	19.39	13.03	14.15	19.44	19.61
Llama 3.2 11B Vision	16.49	18.95	20.70	20.35	13.26	12.84	15.79	16.84	5.29	5.54	9.07	8.31	7.89	7.89	10.18	10.18	12.68	13.54	17.29	19.02	11.82	11.82	13.95	14.66	10.61	11.06	13.97	14.20
InternVL2.5 8B	19.30	17.89	23.16	23.51	11.79	12.00	16.42	16.84	5.04	6.30	10.58	9.57	9.41	9.67	14.50	14.25	9.80	9.80	15.27	15.56	9.46	9.69	13.71	14.89	10.34	10.39	15.41	15.41
Qwen2 VL 7B	17.19	17.19	20.35	18.95	9.47	9.47	12.00	11.37	5.79	6.30	8.56	8.31	8.91	9.92	11.45	11.45	10.95	12.10	15.56	14.41	9.69	11.11	13.00	13.24	9.63	10.26	12.76	12.36
Phi 3.5 Vision	14.39	12.63	20.00	18.95	8.84	10.74	13.89	13.47	6.05	6.05	8.31	8.82	6.62	8.14	9.41	9.92	8.93	8.65	14.70	14.70	8.27	9.93	13.71	12.77	8.55	9.18	12.99	12.85
MiniCPM V 2.6	12.98	14.39	14.39	17.19	10.74	10.32	13.68	14.74	2.52	3.27	6.55	6.05	6.36	6.36	9.67	9.67	10.09	9.80	13.26	14.70	9.46	9.22	12.77	13.24	8.11	8.15	11.60	11.96
InternVL2.5 4B	14.04	16.49	16.84	14.39	9.47	14.53	13.47	9.05	3.53	7.05	7.56	4.03	7.89	9.16	9.16	6.87	8.07	11.53	10.66	8.07	8.04	11.58	11.82	7.33	7.97	11.42	11.29	7.79
Qwen2 VL 2B	13.33	12.28	13.68	14.39	9.68	9.47	11.79	10.95	4.03	3.78	5.54	4.28	6.11	5.09	6.11	6.11	8.36	8.93	12.97	12.10	7.33	8.27	10.40	9.69	7.97	7.88	9.94	9.49
Centurio Qwen	11.23	9.12	14.39	14.39	9.05	8.42	10.32	9.68	3.02	1.76	6.05	6.05	6.87	5.34	9.92	8.91	6.34	5.48	11.24	11.24	6.62	5.67	9.22	9.46	6.81	5.69	9.85	9.76
InternVL2.5 2B	6.67	7.37	10.18	9.47	4.21	4.63	6.95	5.89	2.27	2.02	3.53	5.29	2.80	3.56	5.34	5.60	3.17	3.75	7.49	6.92	5.44	5.44	8.04	6.15	4.03	4.39	6.85	6.45
InternVL2.5 1B	7.02	7.37	10.53	11.58	4.21	3.58	4.84	4.63	2.52	0.76	2.77	2.77	3.56	3.82	5.09	4.07	4.61	3.46	6.63	7.49	4.02	5.67	6.86	6.15	4.03	4.03	5.96	5.87
Centurio Aya	3.16	7.37	8.77	8.77	2.95	5.68	9.05	9.68	1.76	1.51	4.79	3.53	1.27	3.56	5.60	6.11	2.02	3.46	7.20	7.20	2.84	5.44	6.38	7.09	2.24	4.39	6.99	7.17
Average X-Large	24.04	23.33	29.82	29.12	18.84	19.47	23.89	25.47	9.32	9.82	16.62	16.75	15.90	18.07	23.92	23.92	20.03	19.02	26.51	26.51	13.48	14.89	20.33	20.57	16.03	16.61	22.95	23.15
Average Large	21.93	21.40	27.19	27.89	15.26	16.42	20.63	20.95	7.68	7.81	14.74	14.48	13.61	16.03	21.88	21.88	15.85	15.85	22.91	22.48	12.41	13.59	18.79	18.44	13.79	14.78	20.71	20.62
Average Medium	13.39	14.15	16.96	17.19	9.54	9.79	12.88	13.19	3.90	4.11	7.60	6.97	6.79	7.12	10.22	10.09	8.65	9.03	13.30	13.69	8.31	8.83	11.51	12.10	7.96	8.32	11.76	11.81
Average Small	11.09	11.23	14.25	13.75	7.28	8.59	10.19	8.80	3.68	3.93	5.54	5.04	5.39	5.95	7.02	6.51	6.63	7.26	10.49	9.86	6.62	8.18	10.17	8.42	6.51	7.38	9.40	8.49
Average Open	15.18	15.37	19.13	19.06	10.79	11.56	14.48	14.40	5.05	5.31	9.07	8.63	8.45	9.38	12.54	12.32	10.45	10.68	15.41	15.29	8.98	10.06	13.21	12.84	9.33	9.97	13.66	13.39
Average Proprietary	24.98	26.67	30.32	30.39	21.77	22.48	27.75	28.55	12.55	11.64	21.16	20.35	18.42	20.31	26.26	27.12	21.56	22.19	29.86	29.74	18.01	18.58	23.69	25.20	19.04	19.84	26.12	26.53
Average	17.63	18.19	21.93	21.89	13.54	14.29	17.80	17.94	6.93	6.89	12.09	11.56	10.94	12.11	15.97	16.02	13.23	13.56	19.02	18.90	11.24	12.19	15.83	15.93	11.76	12.44	16.78	16.67

Table 10: Cultural Image Visual Question Answering (`CIVQA`) scores. The reported score is relaxed accuracy. The columns **N**, **R**, **C**, and **B** stand for the hints **"None"**, **"Region"**, **"Country"**, and **"Both"**, respectively.

Judge Score

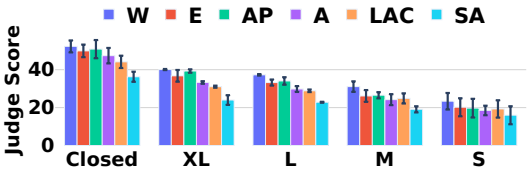


Figure 140: An overview of aggregated `CIVQA` Judge Score results.

Model	West EU & North America				Asia & Pacific				Subsaharian Africa				Arab				East EU				Latin-America & Caribbean				Average			
	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B
GPT-4o	55.86	55.46	64.33	64.49	56.91	56.42	63.09	63.20	39.13	35.97	48.90	47.93	51.89	56.53	65.71	67.59	54.26	55.62	67.56	67.33	48.32	46.69	53.05	54.09	51.06	51.12	60.44	60.77
Gemini Pro	55.44	56.30	63.15	62.06	54.23	54.15	59.84	59.14	38.74	39.51	47.63	46.74	49.83	52.96	58.10	57.65	51.67	52.23	62.48	63.77	46.54	45.32	51.95	52.03	49.41	50.08	57.19	56.90
Claude 3.5 Sonnet	50.65	51.35	65.33	64.14	50.11	51.07	63.70	63.59	35.06	39.90	57.56	56.52	48.99	55.34	67.49	67.86	50.42	51.80	70.12	70.20	41.49	46.10	57.85	59.53	46.12	49.26	63.68	63.64
Gemini Flash	50.49	49.93	56.29	56.11	48.19	47.42	53.20	53.11	35.60	31.32	39.58	38.54	44.04	44.25	50.85	49.03	46.66	47.34	57.24	55.22	43.59	41.58	47.74	46.88	44.76	43.64	50.82	49.81
GPT-4o Mini	48.98	48.47	53.10	54.28	44.83	44.24	49.96	49.05	32.78	35.06	35.49	35.98	42.24	43.83	47.98	48.58	46.57	43.40	55.06	53.76	40.74	38.24	44.21	44.19	42.69	42.21	47.63	47.64
Qwen2 VL 72B	40.28	40.05	48.04	46.90	38.65	39.25	43.02	44.15	25.79	26.30	31.52	30.57	32.77	35.27	42.89	41.64	38.95	39.44	50.55	49.74	30.64	33.06	39.66	40.27	34.51	35.56	42.61	42.21
InternVL2.5 78B	39.88	39.52	46.43	47.01	39.93	38.01	47.78	49.26	22.18	20.79	30.15	30.42	33.72	35.80	46.82	47.86	34.57	32.63	41.89	40.73	31.42	30.40	39.09	38.84	33.62	32.86	42.03	42.35
InternVL2.5 26B	37.00	34.65	39.75	41.00	32.64	32.97	39.10	39.47	22.63	21.71	29.40	27.22	30.89	31.39	38.10	38.81	34.34	32.38	41.14	41.53	29.34	29.69	37.05	37.78	31.14	30.47	37.42	37.64
InternVL2.5 38B	37.55	37.51	45.58	45.49	35.45	36.26	42.65	43.88	22.98	22.52	29.11	28.35	28.71	31.78	38.96	38.63	32.08	31.69	41.98	41.21	28.39	29.15	36.46	35.18	30.86	31.48	39.12	38.79
Qwen2 VL 7B	33.36	34.84	38.64	38.12	28.19	28.97	31.23	31.13	21.31	25.25	25.09	26.26	28.72	28.45	32.00	32.28	29.19	31.13	35.53	37.11	27.84	28.61	31.45	32.87	28.10	29.54	32.33	32.96
Llama 3.2 11B Vision	35.16	36.56	37.22	37.81	27.06	27.59	31.14	33.09	19.24	17.97	24.38	26.42	25.09	26.53	31.43	30.47	28.34	27.88	33.96	36.73	26.89	28.82	32.14	32.88	26.96	27.56	31.71	32.90
MiniCPM V 2.6	30.61	32.35	32.73	35.48	27.73	25.88	31.13	33.25	20.29	18.92	25.31	24.58	24.52	24.57	28.47	28.19	28.13	25.07	34.31	36.27	26.74	26.04	29.16	30.37	26.34	25.47	30.18	31.36
Qwen2 VL 2B	28.86	28.30	28.94	30.85	25.18	24.06	26.32	26.31	21.02	19.32	23.06	21.94	20.92	20.32	22.98	23.30	25.10	26.01	32.90	31.34	23.91	24.07	26.73	25.85	24.16	23.68	26.82	26.60
Centurio Aya	29.84	30.21	30.67	32.31	26.64	25.51	28.70	28.81	18.81	17.87	21.23	20.97	19.75	20.43	24.02	24.01	25.42	24.93	28.79	30.72	23.58	24.65	25.66	26.68	24.01	23.93	26.51	27.25
InternVL2.5 8B	30.12	32.19	35.35	36.47	23.62	23.93	29.75	29.92	16.70	17.20	20.54	21.81	23.61	23.46	30.65	29.92	24.94	24.67	32.73	33.66	22.80	22.13	26.78	27.99	23.63	23.93	29.30	29.96
Phi 3.5 Vision	24.84	26.93	33.43	33.45	23.46	25.36	29.18	29.02	21.28	21.65	23.63	25.92	21.06	23.26	26.18	26.48	24.70	24.88	31.32	31.82	24.47	25.73	29.67	30.56	23.30	24.64	28.90	29.54
Centurio Qwen	27.32	26.32	29.21	30.42	25.84	26.54	26.88	28.63	18.12	17.91	20.14	22.69	23.46	22.32	27.19	27.72	20.84	20.56	26.53	28.83	21.21	21.74	23.19	23.82	22.80	22.56	25.52	27.02
InternVL2.5 4B	25.18	26.06	26.71	29.67	20.67	22.04	26.29	27.53	12.32	14.45	14.99	17.91	18.42	21.62	24.43	25.80	20.22	22.07	26.43	28.43	17.56	20.93	23.45	24.29	19.06	21.19	23.72	25.61
InternVL2.5 1B	19.67	20.32	20.90	23.91	14.46	13.92	14.95	17.03	12.05	13.50	16.60	15.86	16.48	16.31	16.88	17.42	16.10	14.90	18.27	20.82	14.94	15.64	16.75	17.59	15.62	15.76	17.39	18.77
InternVL2.5 2B	18.19	19.35	20.25	21.95	14.75	15.99	16.42	18.36	13.14	10.52	12.88	14.96	15.55	14.08	16.69	18.03	14.77	13.73	17.43	18.32	15.57	15.86	18.04	18.77	15.33	14.92	16.95	18.40
Average X-Large	40.08	39.79	47.23	46.95	39.29	38.63	45.40	46.71	23.99	23.55	30.83	30.49	33.25	35.54	44.86	44.75	36.76	36.03	46.22	45.24	31.03	31.73	39.37	39.56	34.06	34.21	42.32	42.28
Average Large	37.28	36.08	42.67	43.25	34.05	34.61	40.88	41.68	22.81	22.12	29.25	27.79	29.80	31.58	38.53	38.72	33.21	32.04	41.56	41.37	28.86	29.42	36.75	36.48	31.00	30.98	38.27	38.22
Average Medium	31.07	32.08	33.97	35.10	26.51	26.40	29.80	30.80	19.08	19.19	22.78	23.79	24.19	24.29	28.96	28.77	26.14	25.71	31.98	33.89	24.84	25.33	28.06	29.10	25.31	25.50	29.26	30.24
Average Small	23.35	24.19	26.05	27.96	19.70	20.27	22.63	23.65	15.96	15.89	18.23	19.32	18.48	19.12	21.43	22.21	20.18	20.32	25.27	26.15	19.29	20.45	22.93	23.41	19.49	20.04	22.76	23.78
Average Open	30.52	31.01	34.26	35.39	26.95	27.08	30.97	31.99	19.19	19.06	23.20	23.73	24.24	25.04	29.85	30.04	26.51	26.13	32.92	33.82	24.35	25.10	29.02	29.58	25.30	25.57	30.03	30.76
Average Proprietary	52.28	52.30	60.44	60.22	50.85	50.66	57.96	57.62	36.26	36.35	45.83	45.14	47.40	50.58	58.03	58.14	49.92	50.08	62.49	62.05	44.14	43.59	50.96	51.34	46.81	47.26	55.95	55.75
Average	35.96	36.33	40.80	41.59	32.93	32.98	37.72	38.40	23.46	23.38	28.86	29.08	30.03	31.42	36.89	37.06	32.36	32.12	40.31	40.88	29.30	29.72	34.50	35.02	30.67	30.99	36.51	37.01

Table 11: Cultural Image Visual Question Answering (CIVQA) scores. The reported score is the average judge score. The columns **N**, **R**, **C**, and **B** stand for the hints **"None"**, **"Region"**, **"Country"**, and **"Both"**, respectively.

G.1.2 Ground-Truth Answer Perplexity

The perplexity for every sample is computed as follows:

$$\mathrm{PPL}(y \mid x) = \exp\left(-\frac{1}{N}\sum_{t=0}^{N}\log p\left(y_t \mid y_{t-1},\, x\right)\right) \tag{1}$$

where $x = \{s, v\}$ are the textual (s) and visual (v) prompt (prefix) tokens and y are the N ground-truth answer tokens.

Results Per Cultural Aspect

We computed the average accuracy for questions targeting one of the ten most frequent cultural aspects (see §B.2), grouped by model size and region. For better interpretation, Table 12 reports the counts of questions associated with each cultural aspect per region. As shown in Table 13, our results reveal a consistent trend: models perform significantly better on tangible cultural aspects (e.g., food) than on intangible ones. For instance, across all regions, closed models achieve an average accuracy of 30% for food-related questions, compared to only 8% and 10% for questions concerning rituals and festivals, respectively. These findings highlight not only regional biases but also biases along the cultural dimension, the latter being particularly pronounced in non-Western contexts.

aspect	art	craftsmanship	dance	festivals	food	instruments	music	rituals	tools	traditions
A	45	32	20	6	33	20	37	32	30	76
AP	57	44	31	14	12	25	32	53	22	68
E	53	36	18	19	10	19	26	18	20	49
LAC	31	22	31	66	6	13	51	47	12	78
SA	14	16	40	16	22	64	41	73	7	70
W	33	27	10	30	13	14	23	18	17	49

Table 12: Number of questions targeting one of the top-10 cultural aspects per region in CIVQA.

G.2 CVVQA

G.2.1 Results

	AP					A					SA					W					E					LAC					OVERALL				
	A	XL	L	M	S	A	XL	L	M	S	A	XL	L	M	S	A	XL	L	M	S	A	XL	L	M	S	A	XL	L	M	S	A	XL	L	M	S
food	0.28	0.23	0.21	0.07	0.10	0.28	0.35	0.31	0.06	0.09	0.21	0.16	0.07	0.03	0.04	0.18	0.12	0.19	0.07	0.09	0.18	0.36	0.31	0.10	0.12	0.68	0.54	0.71	0.31	0.39	0.30	0.29	0.30	0.11	0.14
instruments	0.29	0.27	0.25	0.05	0.07	0.20	0.15	0.12	0.03	0.04	0.16	0.15	0.16	0.03	0.04	0.26	0.37	0.32	0.05	0.11	0.32	0.31	0.26	0.04	0.05	0.45	0.44	0.31	0.15	0.20	0.28	0.28	0.24	0.06	0.08
craftsmanship	0.15	0.17	0.15	0.04	0.07	0.18	0.24	0.22	0.08	0.08	0.11	0.09	0.08	0.04	0.02	0.14	0.19	0.12	0.10	0.08	0.26	0.28	0.27	0.13	0.17	0.12	0.12	0.12	0.06	0.07	0.16	0.18	0.16	0.08	0.08
music	0.20	0.32	0.26	0.07	0.09	0.10	0.09	0.12	0.03	0.04	0.13	0.11	0.13	0.03	0.03	0.25	0.27	0.28	0.10	0.13	0.10	0.10	0.05	0.02	0.02	0.19	0.22	0.22	0.08	0.12	0.16	0.19	0.18	0.06	0.07
tools	0.19	0.29	0.18	0.09	0.11	0.18	0.17	0.18	0.05	0.06	0.00	0.05	0.04	0.00	0.04	0.22	0.15	0.19	0.09	0.11	0.14	0.17	0.17	0.04	0.04	0.23	0.17	0.30	0.05	0.05	0.16	0.17	0.18	0.05	0.07
traditions	0.19	0.18	0.15	0.06	0.07	0.11	0.09	0.09	0.04	0.05	0.06	0.07	0.06	0.04	0.04	0.21	0.25	0.21	0.09	0.09	0.16	0.16	0.15	0.06	0.08	0.14	0.16	0.13	0.06	0.08	0.14	0.15	0.13	0.06	0.07
art	0.15	0.17	0.13	0.07	0.07	0.18	0.13	0.14	0.04	0.04	0.07	0.09	0.06	0.01	0.00	0.16	0.27	0.23	0.10	0.13	0.13	0.20	0.12	0.06	0.07	0.10	0.11	0.11	0.06	0.09	0.13	0.16	0.13	0.06	0.07
dance	0.07	0.07	0.04	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.05	0.04	0.02	0.01	0.01	0.26	0.39	0.35	0.17	0.11	0.22	0.23	0.17	0.04	0.03	0.14	0.11	0.08	0.04	0.03	0.13	0.14	0.11	0.05	0.03
festivals	0.18	0.20	0.18	0.09	0.08	0.02	0.06	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.10	0.10	0.10	0.04	0.06	0.13	0.11	0.05	0.02	0.01	0.13	0.11	0.11	0.04	0.04	0.10	0.10	0.07	0.03	0.03
rituals	0.09	0.08	0.09	0.02	0.03	0.06	0.06	0.05	0.02	0.03	0.06	0.07	0.04	0.02	0.04	0.12	0.10	0.08	0.02	0.02	0.13	0.16	0.08	0.01	0.01	0.04	0.06	0.03	0.01	0.00	0.08	0.09	0.06	0.02	0.02
Average	0.18	0.20	0.16	0.06	0.07	0.13	0.13	0.12	0.03	0.04	0.09	0.08	0.07	0.02	0.03	0.19	0.22	0.21	0.08	0.09	0.18	0.21	0.16	0.05	0.06	0.22	0.20	0.21	0.09	0.11	0.16	0.17	0.16	0.06	0.07

Table 13: The averaged accuracy per region per model size group (A, XL, L, M, S) per target cultural aspect for samples in the CIVQA task.

Model	West EU & North America				Asia & Pacific				Subsaharian Africa				Arab				East EU				Latin-America & Caribbean				Average			
	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B
GPT-4o	38.97	39.91	41.31	44.13	34.56	36.41	35.71	39.17	23.67	26.33	36.67	36.00	29.18	32.46	36.72	36.39	37.59	40.43	47.16	45.74	31.98	32.56	38.95	39.24	32.67	34.49	39.19	39.97
GPT-4o Mini	38.06	31.58	34.01	38.87	29.45	25.64	25.64	29.66	20.32	13.33	15.56	20.63	28.61	24.40	25.60	29.52	35.37	29.27	30.18	38.72	25.13	21.20	23.04	25.65	28.69	23.89	24.84	29.69
Gemini Pro	33.80	37.09	40.85	39.91	30.41	31.34	34.10	34.79	20.07	22.33	28.67	28.67	26.56	28.85	32.13	32.13	32.27	33.33	36.52	36.88	28.78	30.52	33.43	32.85	28.32	29.91	33.67	33.78
Gemini Flash	29.55	29.96	30.36	34.82	22.67	24.36	26.69	26.69	12.06	12.06	15.87	19.05	20.18	20.78	21.39	23.49	26.52	27.74	32.01	31.71	23.30	24.61	26.18	27.49	21.64	22.59	24.89	26.29
Claude 3.5 Sonnet	21.86	19.84	25.91	24.29	22.46	19.92	25.21	25.85	9.21	6.03	12.38	11.11	16.87	14.16	16.87	18.37	23.17	20.12	26.52	24.70	19.11	15.45	21.47	22.25	18.74	15.89	21.44	21.24
Qwen2 VL 72B	25.35	27.23	33.33	34.27	18.43	19.12	23.73	23.73	9.00	10.00	16.33	17.00	17.05	18.36	22.62	21.64	25.53	25.53	32.27	31.91	16.57	20.93	23.55	24.42	18.13	19.62	24.27	24.65
InternVL2.5 78B	23.94	29.11	31.92	31.46	19.12	24.19	28.11	29.49	7.33	12.33	18.67	19.67	13.44	22.30	25.90	25.90	19.50	24.82	30.14	29.43	15.41	21.22	24.42	26.45	15.75	21.56	25.98	26.70
InternVL2.5 38B	22.07	28.64	32.86	32.86	18.66	24.19	27.19	26.73	6.33	13.00	21.67	21.33	13.77	22.95	24.59	27.54	19.86	26.24	30.85	30.50	13.37	20.93	26.16	24.71	14.98	21.78	26.31	26.37
Qwen2 VL 2B	19.72	18.78	21.13	23.00	13.13	14.75	16.59	15.67	6.67	4.67	7.00	6.67	13.11	11.15	13.44	12.46	16.67	16.67	17.38	17.38	15.70	15.12	16.28	16.28	13.88	13.27	15.26	14.98
Qwen2 VL 7B	18.78	18.78	22.07	21.60	14.06	14.06	17.05	16.59	5.00	6.00	7.67	7.33	13.11	15.08	17.05	17.70	15.25	17.73	18.79	19.50	15.70	18.60	19.48	20.06	13.54	14.76	16.86	16.92
InternVL2.5 26B	20.66	25.35	28.64	29.11	16.36	19.35	23.27	24.88	3.33	7.33	9.33	10.33	11.80	15.41	19.67	20.00	17.73	21.63	24.82	24.11	13.66	18.31	21.51	22.97	13.32	17.30	20.78	21.61
MiniCPM V 2.6	16.90	19.25	18.31	19.72	14.75	16.82	17.28	18.43	5.67	10.00	11.33	11.00	12.13	13.44	14.75	14.75	19.15	20.21	22.34	21.99	15.41	17.44	16.28	19.19	13.16	15.37	16.14	17.03
Phi 3.5 Vision	16.43	14.55	16.90	16.90	13.82	14.06	17.51	17.28	8.67	8.33	10.67	10.33	9.84	10.16	11.15	10.82	15.60	15.25	19.15	19.86	13.95	15.12	18.60	18.90	12.82	12.88	16.09	15.87
Centurio Qwen	20.19	17.84	23.00	21.13	15.67	15.44	18.43	17.74	6.00	6.33	7.33	7.33	9.51	10.82	10.16	10.49	14.89	15.96	22.34	20.92	11.63	11.92	15.70	14.53	12.38	12.55	15.70	15.15
InternVL2.5 8B	14.55	19.25	20.66	23.00	11.98	15.44	18.43	18.43	3.33	6.33	9.33	9.00	9.84	13.77	15.08	16.07	15.25	17.73	23.05	23.05	10.17	12.21	16.28	16.57	10.61	13.82	16.92	17.36
InternVL2.5 4B	14.55	15.96	18.78	18.31	12.67	14.29	17.74	16.59	5.67	6.33	9.00	9.00	8.52	9.84	13.11	12.46	11.35	15.25	19.50	18.09	11.34	14.24	15.70	15.12	10.45	12.38	15.70	14.70
Centurio Aya	11.74	12.21	15.49	12.21	9.68	9.91	12.21	11.06	4.67	4.67	6.67	5.33	6.89	7.54	7.54	7.54	9.93	9.57	12.77	10.64	7.56	9.01	10.17	9.59	8.46	8.96	10.95	9.62
InternVL2.5 1B	8.45	9.86	11.27	12.21	5.76	8.29	9.22	7.60	1.67	2.33	4.00	2.67	5.90	7.87	8.85	8.85	6.74	7.45	10.99	10.64	7.27	8.72	9.01	9.01	5.86	7.46	8.90	8.35
Average X-Large	24.65	28.17	32.63	32.86	18.78	21.66	25.92	26.61	8.17	11.17	17.50	18.33	15.25	20.33	24.26	23.77	22.52	25.18	31.21	30.67	15.99	21.08	23.98	25.44	16.94	20.59	25.12	25.68
Average Large	21.36	27.00	30.75	30.99	17.51	21.77	25.23	25.81	4.83	10.17	15.50	15.83	12.79	19.18	22.13	23.77	18.79	23.94	27.84	27.30	13.52	19.62	23.84	23.84	14.15	19.54	23.55	23.99
Average Medium	16.43	17.46	19.91	19.53	13.23	14.33	16.68	16.45	4.93	6.67	8.47	8.00	10.30	12.13	12.92	13.31	14.89	16.24	19.86	19.22	12.09	13.84	15.58	15.99	11.63	13.09	15.31	15.21
Average Small	14.79	14.79	17.02	17.61	11.35	12.85	15.26	14.29	5.67	5.42	7.67	7.17	9.34	9.75	11.64	11.15	12.59	13.65	16.76	16.49	12.06	13.30	14.90	14.83	10.75	11.50	13.99	13.47
Average Open	17.95	19.75	22.64	22.75	14.16	16.15	18.98	18.79	5.64	7.51	10.69	10.54	11.15	13.75	15.69	15.86	15.96	18.00	21.88	21.39	12.90	15.68	17.93	18.29	12.57	14.75	17.68	17.64
Average Proprietary	32.45	31.67	34.49	36.40	27.91	27.53	29.47	31.23	17.06	16.02	21.83	23.09	24.28	24.13	26.54	27.98	30.98	30.18	34.48	35.55	25.66	24.87	28.61	29.50	26.01	25.35	28.80	30.19
Average	21.98	23.07	25.93	26.54	17.98	19.31	21.90	22.24	8.81	9.88	13.79	14.03	14.80	16.63	18.70	19.23	20.13	21.39	25.38	25.32	16.45	18.23	20.90	21.40	16.30	17.69	20.77	21.13

Table 14: GIMMICK Video Visual Question Answering (VVQA) results. The reported score is relaxed accuracy. The columns **N**, **R**, **C**, and **B** stand for the hints **"None"**, **"Region"**, **"Country"**, and **"Both"**, respectively.

G.3 COQA Details

G.3.1 Results

	West EU & North Am.			East EU			Asia & Pacific			Lat. Am. & Carib.			Arab			Subs. Africa			Average			
	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	Avg.
GPT-4o	82.50	83.75	85.00	85.89	90.18	88.34	94.37	96.54	97.40	93.68	92.63	92.63	88.00	88.00	92.00	91.30	91.30	94.20	89.29	90.40	91.60	90.43
Claude 3.5 Sonnet	72.50	83.75	81.25	76.69	85.89	82.21	87.88	95.67	95.67	83.16	89.47	87.37	84.00	90.67	90.67	82.61	89.86	88.41	81.14	89.22	87.60	85.98
InternVL2.5 78B	77.50	80.00	86.88	83.44	82.21	88.96	94.37	94.81	96.97	88.42	92.63	92.63	88.00	90.67	92.00	92.75	91.30	92.75	87.41	88.60	91.70	89.24
Qwen2.5 72B	–	81.25	–	–	84.05	–	–	96.10	–	–	89.47	–	–	86.67	–	–	89.86	–	–	87.90	–	87.90
GPT-4o Mini	76.25	82.50	86.25	84.66	82.21	84.66	94.37	95.67	96.54	87.37	90.53	90.53	85.33	86.67	86.67	91.30	89.86	92.75	86.55	87.90	89.57	88.01
InternVL2.5 38B	81.25	81.25	84.38	85.89	84.66	85.28	90.04	95.24	92.64	86.32	86.32	92.63	89.33	90.67	92.00	89.86	86.96	91.30	87.11	87.51	89.70	88.11
Qwen2 VL 72B	79.38	80.62	81.25	88.34	84.66	88.34	90.48	94.81	96.97	86.32	88.42	92.63	86.67	88.00	89.33	91.30	85.51	91.30	87.08	87.00	89.97	88.02
Gemini Flash	82.50	78.75	78.13	85.28	80.37	84.66	87.01	91.34	94.81	85.11	87.37	90.53	89.19	86.67	90.67	89.86	91.30	91.30	86.49	85.97	88.35	86.94
Qwen2.5 32B	–	76.88	–	–	79.75	–	–	94.37	–	–	87.37	–	–	84.00	–	–	89.86	–	–	85.37	–	85.37
Qwen2 VL 7B	71.25	74.38	76.25	82.82	80.37	84.05	92.64	93.51	93.51	85.26	88.42	92.63	80.00	82.67	84.00	86.96	85.51	84.06	83.16	84.14	85.75	84.35
MiniCPM V 2.6	72.50	72.50	75.00	81.60	79.14	80.37	88.74	90.48	93.07	80.00	87.37	90.53	80.00	77.33	86.67	88.41	85.51	86.96	81.87	82.05	85.43	83.12
InternVL2.5 26B	77.50	74.38	80.62	87.12	75.46	87.12	91.77	91.77	96.54	88.42	84.21	93.68	84.00	85.33	88.00	91.30	79.71	86.96	86.69	81.81	88.82	85.77
Phi 3.5 Mini	–	74.38	–	–	72.39	–	–	88.31	–	–	83.16	–	–	81.33	–	–	86.96	–	–	81.09	–	81.09
InternLM2.5 7B	–	74.38	–	–	76.69	–	–	90.48	–	–	80.00	–	–	78.67	–	–	85.51	–	–	80.95	–	80.95
Centurio Qwen	75.63	74.38	80.00	79.75	76.69	82.82	86.58	92.64	92.21	83.16	86.32	89.47	78.67	77.33	88.00	86.96	76.81	89.86	81.79	80.69	87.06	83.18
InternLM2.5 20B	–	74.38	–	–	75.46	–	–	89.18	–	–	86.32	–	–	76.00	–	–	82.61	–	–	80.66	–	80.66
Qwen2.5 7B	–	71.88	–	–	72.39	–	–	93.51	–	–	85.26	–	–	77.33	–	–	81.16	–	–	80.26	–	80.26
Aya Expanse 8B	–	68.12	–	–	77.30	–	–	91.77	–	–	81.05	–	–	80.00	–	–	81.16	–	–	79.90	–	79.90
InternVL2.5 8B	68.12	72.50	75.63	83.44	76.07	83.44	87.88	89.61	94.37	84.21	83.16	92.63	84.00	73.33	89.33	88.41	81.16	92.75	82.68	79.31	88.03	83.34
Centurio Aya	80.62	68.12	78.75	82.21	75.46	80.37	90.91	85.71	92.21	84.21	82.11	85.26	81.33	82.67	85.33	85.51	81.16	91.30	84.13	79.21	85.54	82.96
Phi 3.5 Vision	65.62	72.50	75.63	69.94	70.55	76.69	89.18	91.34	95.24	80.00	81.05	86.32	72.00	80.00	86.67	85.51	79.71	88.41	77.04	79.19	84.82	80.35
InternVL2.5 4B	66.88	66.88	76.25	84.66	75.46	84.05	87.01	86.15	93.07	83.16	78.95	87.37	80.00	82.67	86.67	86.96	84.06	89.86	81.44	79.03	86.21	82.23
Qwen2 VL 2B	77.50	72.50	78.75	84.05	64.42	84.05	91.77	82.68	92.21	88.42	81.05	86.32	84.00	70.67	89.33	88.41	79.71	91.30	85.69	75.17	86.99	82.62
Qwen2.5 3B	–	68.75	–	–	73.01	–	–	83.12	–	–	73.68	–	–	74.67	–	–	75.36	–	–	74.76	–	74.76
Qwen2.5 1.5B	–	61.88	–	–	65.03	–	–	82.25	–	–	78.95	–	–	72.00	–	–	78.26	–	–	73.06	–	73.06
Qwen2.5 0.5B	–	68.12	–	–	72.39	–	–	67.53	–	–	65.26	–	–	70.67	–	–	55.07	–	–	66.51	–	66.51
InternLM2.5 1.8B	–	56.25	–	–	65.03	–	–	65.37	–	–	60.00	–	–	66.67	–	–	66.67	–	–	63.33	–	63.33
InternVL2.5 2B	70.62	51.88	72.50	76.69	58.28	71.78	77.92	72.29	82.68	83.16	62.11	83.16	73.33	66.67	82.67	84.06	60.87	89.86	77.63	62.02	80.44	73.36
InternVL2.5 1B	63.75	58.75	66.88	62.58	60.74	74.23	64.50	61.90	80.09	77.89	57.89	87.37	62.67	68.00	82.67	75.36	59.42	82.61	67.79	61.12	78.97	69.29
Gemini Pro	76.25	59.38	78.13	68.10	55.21	82.21	82.25	56.28	89.61	79.79	61.05	85.11	79.73	61.33	84.00	72.46	65.22	95.65	76.43	59.75	85.78	73.99
Average X-Large LVLMs	78.44	80.31	84.06	85.89	83.44	88.65	92.42	94.81	96.97	87.37	90.53	92.63	87.33	89.33	90.67	92.03	88.41	92.03	87.24	87.80	90.84	88.63
Average Large LVLMs	79.38	77.81	82.50	86.50	80.06	86.20	90.91	93.51	94.59	87.37	85.26	93.16	86.67	88.00	90.00	90.58	83.33	89.13	86.90	84.66	89.26	86.94
Average Medium LVLMs	73.62	72.38	77.12	81.96	77.55	82.21	89.35	90.39	93.07	83.37	85.47	90.11	80.80	78.67	86.67	87.25	82.03	88.99	82.73	81.08	86.36	83.39
Average Small LVLMs	68.88	64.50	74.00	75.58	65.89	78.16	82.08	78.87	88.66	82.53	72.21	86.11	74.40	73.60	85.60	84.06	72.75	88.41	77.92	71.31	83.49	77.57
Average LVLMs	73.44	71.47	77.77	80.89	74.58	82.25	87.41	87.35	92.27	84.21	81.43	89.47	80.29	79.71	87.33	87.27	79.81	89.23	82.25	79.06	86.39	82.57
Average X-Large LLMs	–	81.25	–	–	84.05	–	–	96.10	–	–	89.47	–	–	86.67	–	–	89.86	–	–	87.90	–	87.90
Average Large LLMs	–	75.62	–	–	77.61	–	–	91.77	–	–	86.84	–	–	80.00	–	–	86.23	–	–	83.02	–	83.02
Average Medium LLMs	–	71.46	–	–	75.46	–	–	91.92	–	–	82.11	–	–	78.67	–	–	82.61	–	–	80.37	–	80.37
Average Small LLMs	–	65.88	–	–	69.57	–	–	77.32	–	–	72.21	–	–	73.07	–	–	72.46	–	–	71.75	–	71.75
Average LLMs	–	70.57	–	–	73.95	–	–	85.64	–	–	79.14	–	–	77.09	–	–	79.31	–	–	77.62	–	77.62
Average X-Large	78.44	80.62	84.06	85.89	83.64	88.65	92.42	95.24	96.97	87.37	90.18	92.63	87.33	88.44	90.67	92.03	88.89	92.03	87.24	87.83	90.84	88.39
Average Large	79.38	76.72	82.50	86.50	78.83	86.20	90.91	92.64	94.59	87.37	86.05	93.16	86.67	84.00	90.00	90.58	84.78	89.13	86.90	83.84	89.26	84.98
Average Medium	73.62	72.03	77.12	81.96	76.76	82.21	89.35	90.96	93.07	83.37	84.21	90.11	80.80	78.67	86.67	87.25	82.25	88.99	82.73	80.81	86.36	82.26
Average Small	68.88	65.19	74.00	75.58	67.73	78.16	82.08	78.10	88.66	82.53	72.21	86.11	74.40	73.33	85.60	84.06	72.61	88.41	77.92	71.53	83.49	74.66
Average Open	73.44	71.08	77.77	80.89	74.31	82.25	87.41	86.60	92.27	84.21	80.42	89.47	80.29	78.56	87.33	87.27	79.59	89.23	82.25	78.43	86.39	80.39
Average Proprietary	78.00	77.62	81.75	80.12	78.77	84.42	89.18	87.10	94.81	85.82	84.21	89.23	85.25	82.67	88.80	85.51	85.51	92.46	83.98	82.65	88.58	85.07
Average	74.64	72.17	78.82	80.69	75.05	82.82	87.88	86.68	92.94	84.63	81.05	89.41	81.59	79.24	87.72	86.80	80.58	90.08	82.71	79.13	86.96	81.17

Table 15: GIMMICK Cultural Origin Question Answering – Regions ($\texttt{COQA}_{\texttt{R}}$) results. The reported score is relaxed accuracy. The columns **I** and **T** stand for **image-only** and **text-only** inputs to the model.

	West EU & North Am.			East EU			Asia & Pacific			Lat. Am. & Carib.			Arab			Subs. Africa			Average			
	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	Avg.
Claude 3.5 Sonnet	79.23	96.72	95.63	82.35	97.65	96.47	76.62	97.84	95.67	70.21	98.94	100.00	76.47	97.65	96.47	83.82	97.06	91.18	78.12	97.64	95.90	90.55
GPT-4o	93.44	95.08	96.17	94.71	98.24	98.24	93.51	97.40	98.27	97.87	98.94	98.94	95.29	95.29	98.82	95.59	97.06	100.00	95.07	97.00	98.41	96.83
InternVL2.5 78B	83.06	94.54	97.81	80.59	95.88	97.65	83.12	93.51	96.54	81.91	98.94	98.94	90.59	97.65	97.65	83.82	97.06	98.53	83.85	96.26	97.85	92.65
Qwen2.5 72B	–	93.44	–	–	96.47	–	–	94.81	–	–	98.94	–	–	97.65	–	–	94.12	–	–	95.90	–	95.90
GPT-4o Mini	89.07	93.99	95.63	90.00	95.29	97.65	90.48	93.51	96.97	90.43	95.74	100.00	94.12	88.24	97.65	91.18	95.59	98.53	90.88	93.73	97.74	94.11
Qwen2.5 32B	–	91.26	–	–	93.53	–	–	91.77	–	–	94.68	–	–	95.29	–	–	92.65	–	–	93.20	–	93.20
InternVL2.5 38B	78.69	91.80	92.35	77.06	91.18	92.94	77.49	93.07	93.94	79.79	95.74	96.81	84.71	94.12	95.29	88.24	92.65	98.53	80.99	93.09	94.98	89.69
Qwen2 VL 72B	87.98	87.43	95.08	94.12	90.59	96.47	90.04	90.04	97.84	91.49	97.87	98.94	92.94	89.41	98.82	91.18	97.06	98.53	91.29	92.07	97.61	93.66
Gemini Flash	90.56	89.01	97.27	90.59	88.82	97.06	91.77	90.48	98.70	90.43	93.62	97.87	90.59	88.24	97.65	88.24	95.59	97.06	90.36	90.96	97.60	92.97
InternVL2.5 26B	78.14	87.98	92.90	78.24	88.24	94.71	76.19	90.48	93.94	80.85	94.68	94.68	81.18	91.76	91.76	80.88	91.18	95.59	79.25	90.72	93.93	87.96
Qwen2.5 7B	–	86.34	–	–	88.24	–	–	85.28	–	–	95.74	–	–	90.59	–	–	94.12	–	–	90.05	–	90.05
Aya Expanse 8B	–	87.43	–	–	88.24	–	–	90.04	–	–	93.62	–	–	88.24	–	–	89.71	–	–	89.54	–	89.54
InternLM2.5 20B	–	86.89	–	–	87.06	–	–	90.91	–	–	90.43	–	–	85.88	–	–	89.71	–	–	88.48	–	88.48
MiniCPM V 2.6	81.97	84.70	90.16	81.18	86.47	92.94	78.79	87.45	92.21	86.17	87.23	92.55	82.35	89.41	96.47	88.24	92.65	94.12	83.12	87.98	93.08	88.06
Qwen2 VL 7B	87.43	83.61	90.71	82.35	85.29	92.94	87.01	84.85	94.37	91.49	88.30	94.68	84.71	88.24	96.47	92.65	94.12	97.06	87.61	87.40	94.37	89.79
Qwen2.5 3B	–	81.42	–	–	85.88	–	–	84.85	–	–	92.55	–	–	88.24	–	–	86.76	–	–	86.62	–	86.62
InternLM2.5 7B	–	83.61	–	–	85.88	–	–	85.71	–	–	90.43	–	–	77.65	–	–	88.24	–	–	85.25	–	85.25
Centurio Qwen	78.69	82.51	89.07	78.82	82.94	89.41	78.79	84.42	92.64	76.60	85.11	91.49	80.00	83.53	88.24	79.41	91.18	92.65	78.72	84.95	90.58	84.75
Centurio Aya	65.57	83.61	85.79	72.35	81.76	85.88	75.76	85.71	88.31	74.47	87.23	82.98	70.59	80.00	89.41	66.18	89.71	89.71	70.82	84.67	87.01	80.83
InternVL2.5 8B	68.31	82.51	88.52	70.59	84.71	90.00	75.32	86.58	91.34	75.53	87.23	94.68	76.47	83.53	90.59	82.35	82.35	89.71	74.76	84.49	90.81	83.35
Phi 3.5 Mini	–	80.87	–	–	82.94	–	–	83.98	–	–	84.04	–	–	82.35	–	–	88.24	–	–	83.74	–	83.74
InternVL2.5 4B	68.85	77.05	89.62	72.35	82.94	89.41	71.43	86.15	90.48	76.60	87.23	89.36	72.94	81.18	84.71	76.47	82.35	97.06	73.11	82.82	90.11	82.01
Phi 3.5 Vision	72.13	79.78	86.89	68.82	82.94	92.35	69.70	81.82	89.61	74.47	91.49	91.49	81.18	77.65	90.59	76.47	82.35	95.59	73.79	82.67	91.09	82.52
Qwen2.5 1.5B	–	78.69	–	–	81.18	–	–	82.68	–	–	82.98	–	–	75.29	–	–	80.88	–	–	80.28	–	80.28
Qwen2 VL 2B	83.06	74.32	87.43	84.71	77.06	87.65	83.55	80.95	90.48	92.55	81.91	94.68	83.53	76.47	91.76	89.71	80.88	94.12	86.18	78.60	91.02	85.27
Qwen2.5 0.5B	–	65.03	–	–	68.82	–	–	72.29	–	–	75.53	–	–	69.41	–	–	77.94	–	–	71.51	–	71.51
InternVL2.5 1B	61.20	66.12	73.77	59.41	65.88	73.53	62.34	75.76	77.06	67.02	74.47	76.60	56.47	63.53	75.29	55.88	72.06	70.59	60.39	69.64	74.47	68.17
InternLM2.5 1.8B	–	63.39	–	–	66.47	–	–	71.00	–	–	67.02	–	–	58.82	–	–	64.71	–	–	65.23	–	65.23
InternVL2.5 2B	62.84	65.57	74.32	61.76	64.71	72.35	61.04	68.40	80.95	67.02	68.09	80.85	67.06	55.29	74.12	73.53	63.24	77.94	65.54	64.22	76.76	68.84
Gemini Pro	76.67	43.17	92.70	75.88	39.41	92.94	78.79	34.20	92.64	78.72	39.36	93.62	78.82	35.29	91.76	82.35	19.12	94.12	78.54	35.09	92.96	68.86
Average X-Large LVLMs	85.52	90.98	96.45	87.35	93.24	97.06	86.58	91.77	97.19	86.70	98.40	98.94	91.76	93.53	98.24	87.50	97.06	98.53	87.57	94.16	97.73	93.16
Average Large LVLMs	78.42	89.89	92.62	77.65	89.71	93.82	76.84	91.77	93.94	80.32	95.21	95.74	82.94	92.94	93.53	84.56	91.91	97.06	80.12	91.90	94.46	88.82
Average Medium LVLMs	76.39	83.39	88.85	77.06	84.24	90.24	79.13	85.80	91.77	80.85	87.02	91.28	78.82	84.94	92.24	81.76	90.00	92.65	79.01	85.90	91.17	85.36
Average Small LVLMs	69.62	72.57	82.40	69.41	74.71	83.06	69.61	78.61	85.71	75.53	80.64	86.60	72.24	70.82	83.29	74.41	76.18	87.06	71.80	75.59	84.69	77.36
Average LVLMs	75.57	81.54	88.17	75.88	82.90	89.16	76.47	84.94	90.69	79.71	87.54	91.34	78.91	82.27	90.08	80.36	86.34	92.12	77.82	84.26	90.26	84.11
Average X-Large LLMs	–	93.44	–	–	96.47	–	–	94.81	–	–	98.94	–	–	97.65	–	–	94.12	–	–	95.90	–	95.90
Average Large LLMs	–	89.07	–	–	90.29	–	–	91.34	–	–	92.55	–	–	90.59	–	–	91.18	–	–	90.84	–	90.84
Average Medium LLMs	–	85.79	–	–	87.45	–	–	87.01	–	–	93.26	–	–	85.49	–	–	90.69	–	–	88.28	–	88.28
Average Small LLMs	–	73.88	–	–	77.06	–	–	78.96	–	–	80.43	–	–	74.82	–	–	79.71	–	–	77.48	–	77.48
Average LLMs	–	81.67	–	–	84.06	–	–	84.85	–	–	87.81	–	–	82.67	–	–	86.10	–	–	84.53	–	84.53
Average X-Large	85.52	91.80	96.45	87.35	94.31	97.06	86.58	92.78	97.19	86.70	98.58	98.94	91.76	94.90	98.24	87.50	96.08	98.53	87.57	94.74	97.73	94.07
Average Large	78.42	89.48	92.62	77.65	90.00	93.82	76.84	91.56	93.94	80.32	93.88	95.74	82.94	91.76	93.53	84.56	91.54	97.06	80.12	91.37	94.46	89.83
Average Medium	76.39	84.29	88.85	77.06	85.44	90.24	79.13	86.26	91.77	80.85	89.36	91.28	78.82	85.15	92.24	81.76	90.26	92.65	79.01	86.79	91.17	86.45
Average Small	69.62	73.22	82.40	69.41	75.88	83.06	69.61	78.79	85.71	75.53	80.53	86.60	72.24	72.82	83.29	74.41	77.94	87.06	71.80	76.53	84.69	77.42
Average Open	75.57	81.60	88.17	75.88	83.41	89.16	76.47	84.90	90.69	79.71	87.66	91.34	78.91	82.45	90.08	80.36	86.24	92.12	77.82	84.38	90.26	84.29
Average Proprietary	85.79	83.59	95.48	86.71	83.88	96.47	86.23	82.68	96.45	85.53	85.32	98.09	87.06	80.94	96.47	88.24	80.88	96.18	86.59	82.88	96.52	88.66
Average	78.26	81.93	90.10	78.73	83.49	91.08	79.04	84.53	92.21	81.24	87.27	93.11	81.05	82.20	91.76	82.43	85.34	93.19	80.13	84.13	91.91	85.02

Table 16: GIMMICK Cultural Origin Question Answering – Country ($\texttt{COQA}_\texttt{C}$) results. The reported score is relaxed accuracy. The columns **I** and **T** stand for **image-only** and **text-only** inputs to the model.

G.4 CKQA

G.4.1 LLM-as-a-Judge Evaluation

To evaluate the $\texttt{CKQA}_\texttt{D}$ and $\texttt{CKQA}_\texttt{N}$ tasks, we used GPT-4o (`gpt-4o-2024-11-20`) as a judge using the prompts shown in the next section. For each sample, we used the same system prompt and generated user prompts per sample individually.

System Prompt

```
# Your Role

You are an impartial judge who excels at critical and analytical thinking.

# Your Task

Your task is it to thoroughly analyze and evaluate the correctness of a generated answer to a
↪  Cultural Knowledge Test.
1. Carefully analyze the ground truth and the generated answer.
2. Provide a brief summary (1 - 3 sentences) of your analysis, covering the accuracy,
↪  relevance, and completeness of the generated answer.
3. Provide a one or two-sentence explanation justifying your final score. Ensure that your
↪  explanation and score are consistent with each other and accurately reflect the quality of
↪  the generated answer in relation to the ground truth.
4. Provide a single number from 0 to 100 representing the correctness of the generated answer,
↪  where:
    0 = Completely incorrect or irrelevant.
    25 = Mostly incorrect or irrelevant.
    50 = Partially correct or relevant.
    75 = Mostly correct and relevant.
    100 = Perfectly correct and complete.

    You may use any whole number within this range to reflect nuanced judgments.

# Output Format

Provide your evaluation in the following format:

```xml
<evaluation>
<analysis>
<!-- Put your analysis summary here -->
</analysis>
<explanation>
<!-- Put your explanation here -->
</explanation>
<score>
<!-- Put your score here -->
</score>
</evaluation>
```
```

User Prompt Template

```
Evaluate the correctness of the generated answer with respect to Ground Truth.

# Ground Truth

```
{GROUND_TRUTH}
```

Generated Answer

```
{GENERATED_ANSWER}
```

Evaluation
```

G.4.2 Results

	West EU & North Am.			East EU			Asia & Pacific			Lat. Am. & Carib.			Arab			Subs. Africa			Average			
	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	Avg.
GPT-4o	46.98	56.78	57.21	38.20	54.30	54.87	44.71	59.20	58.56	34.08	51.53	52.45	44.41	57.76	56.91	29.04	50.68	51.37	39.57	55.04	55.23	49.95
Claude 3.5 Sonnet	43.05	56.64	55.60	35.20	55.97	50.07	39.54	59.67	54.05	26.84	53.32	49.23	41.45	56.78	53.68	24.73	50.34	44.79	35.14	55.45	51.24	47.28
Gemini Pro	42.28	53.29	57.21	36.80	50.07	53.67	37.47	52.94	55.18	29.18	49.44	50.00	38.68	48.68	54.08	22.05	41.23	46.23	34.41	49.28	52.73	45.47
Qwen2.5 72B	–	47.55	–	–	45.17	–	–	50.62	–	–	42.70	–	–	44.47	–	–	37.95	–	–	44.74	–	44.74
Qwen2.5 32B	–	47.89	–	–	43.73	–	–	48.45	–	–	40.71	–	–	42.17	–	–	39.25	–	–	43.70	–	43.70
GPT-4o Mini	34.36	48.89	55.70	27.70	46.63	54.00	30.73	49.05	53.61	24.95	43.72	49.44	36.84	47.50	54.21	21.03	39.25	46.78	29.27	45.84	52.29	42.47
Gemini Flash	36.54	52.75	54.70	29.67	46.87	51.30	31.78	50.40	51.62	23.20	46.07	49.07	32.43	46.64	51.45	16.44	36.37	42.12	28.34	46.52	50.04	41.63
Phi 3.5 Mini	–	40.40	–	–	35.23	–	–	38.27	–	–	34.80	–	–	34.87	–	–	30.00	–	–	35.60	–	35.60
Aya Expanse 8B	–	40.17	–	–	36.13	–	–	39.42	–	–	34.18	–	–	36.32	–	–	26.71	–	–	35.49	–	35.49
Qwen2.5 7B	–	38.39	–	–	36.50	–	–	38.78	–	–	34.23	–	–	34.01	–	–	29.04	–	–	35.16	–	35.16
InternLM2.5 20B	–	37.01	–	–	34.13	–	–	36.59	–	–	31.17	–	–	32.83	–	–	27.53	–	–	33.21	–	33.21
Llama 3.2 11B Vision	–	36.44	–	–	32.77	–	–	35.75	–	–	30.00	–	–	33.68	–	–	27.40	–	–	32.67	–	32.67
InternVL2.5 38B	23.72	41.21	37.62	18.63	38.80	37.03	20.51	41.55	39.96	23.72	33.32	38.72	24.08	35.46	39.67	15.96	32.47	33.49	21.10	37.14	37.75	32.00
InternVL2.5 78B	19.33	40.84	36.28	17.63	37.73	37.10	19.16	41.57	38.23	19.64	37.50	36.58	22.89	35.66	42.57	14.86	33.22	35.34	18.92	37.75	37.68	31.45
Qwen2 VL 72B	20.67	40.81	41.01	17.37	37.03	42.13	18.23	40.02	42.10	14.08	36.02	37.60	23.42	36.12	41.91	10.62	29.38	35.14	17.40	36.56	39.98	31.31
InternLM2.5 7B	–	34.33	–	–	32.30	–	–	34.62	–	–	31.17	–	–	29.93	–	–	23.49	–	–	30.97	–	30.97
Qwen2.5 3B	–	32.75	–	–	28.90	–	–	33.05	–	–	28.47	–	–	26.58	–	–	22.81	–	–	28.76	–	28.76
InternVL2.5 26B	11.91	39.97	34.43	12.63	36.10	34.07	13.76	38.92	34.00	13.11	34.18	28.98	15.33	34.01	36.38	9.38	29.59	27.95	12.69	35.46	32.64	26.93
Qwen2 VL 7B	14.09	32.82	38.09	14.27	29.53	37.17	12.72	33.12	37.43	17.09	28.32	35.97	17.17	29.28	35.46	9.38	20.55	30.96	14.12	28.94	35.85	26.30
MiniCPM V 2.6	18.49	34.70	36.28	13.60	30.47	34.60	15.88	33.76	34.96	18.67	29.03	34.34	17.50	30.20	36.84	9.93	18.42	24.52	15.68	29.43	33.59	26.23
Centurio Qwen	14.87	31.38	32.05	14.47	29.23	29.97	15.07	31.57	34.76	16.38	27.40	35.77	18.62	27.30	36.32	13.01	20.41	32.60	15.40	27.88	33.58	25.62
Phi 3.5 Vision	12.08	36.64	35.03	12.43	32.10	35.23	10.09	33.36	32.30	13.78	29.74	29.80	16.97	31.97	33.42	11.99	25.75	26.78	12.89	31.59	32.09	25.53
InternVL2.5 8B	6.81	36.28	29.97	6.80	31.33	29.13	9.07	33.72	30.49	6.58	29.74	30.36	12.11	30.53	30.26	2.53	22.67	20.96	7.32	30.71	28.53	22.19
InternVL2.5 4B	5.44	35.81	27.89	5.40	33.07	27.80	5.97	35.71	28.08	7.14	34.29	27.24	9.01	28.62	29.54	4.79	25.68	23.63	6.29	32.20	27.36	21.95
Qwen2.5 1.5B	–	24.03	–	–	20.77	–	–	26.66	–	–	20.87	–	–	21.45	–	–	16.23	–	–	21.67	–	21.67
InternLM2.5 1.8B	–	23.56	–	–	22.30	–	–	22.94	–	–	18.52	–	–	21.78	–	–	14.66	–	–	20.63	–	20.63
Qwen2 VL 2B	11.41	23.29	31.95	11.57	20.03	28.90	12.48	20.97	29.00	14.18	20.51	28.16	16.32	18.29	29.47	11.92	13.63	25.96	12.98	19.45	28.91	20.45
InternVL2.5 1B	9.87	24.09	16.51	8.50	20.43	16.83	9.78	21.28	18.50	11.22	20.41	16.07	12.83	16.51	18.75	9.93	14.93	17.60	10.35	19.61	17.38	15.78
InternVL2.5 2B	5.30	23.26	18.72	4.80	19.50	21.90	4.47	22.26	20.58	7.30	21.48	19.95	9.34	21.38	20.72	5.68	15.96	18.77	6.15	20.64	20.11	15.63
Centurio Aya	4.80	29.33	5.94	5.30	25.50	7.53	2.94	28.85	5.02	7.50	24.23	8.47	4.28	24.21	4.21	4.45	19.38	5.89	4.88	25.25	6.18	12.10
Qwen2.5 0.5B	–	13.96	–	–	11.77	–	–	14.40	–	–	11.43	–	–	8.29	–	–	8.70	–	–	11.42	–	11.42
Average X-Large LVLMs	20.00	40.83	38.64	17.50	37.38	39.62	18.70	40.80	40.16	16.86	36.76	37.09	23.16	35.89	42.24	12.74	31.30	35.24	18.16	37.16	38.83	31.38
Average Large LVLMs	17.81	40.59	36.02	15.63	37.45	35.55	17.14	40.24	36.98	18.42	33.75	33.85	19.70	34.74	38.03	12.67	31.03	30.72	16.90	36.30	35.20	29.46
Average Medium LVLMs	11.81	33.49	28.47	10.89	29.80	27.68	11.14	32.79	28.53	13.24	28.12	28.98	13.94	29.20	28.62	7.86	21.47	22.99	11.48	29.15	27.55	24.18
Average Small LVLMs	8.82	28.62	26.02	8.54	25.03	26.13	8.56	26.72	25.69	10.72	25.29	24.24	12.89	23.35	26.38	8.86	19.19	22.55	9.73	24.70	25.17	19.87
Average LVLMs	12.77	33.79	30.13	11.67	30.24	29.96	12.15	32.83	30.39	13.60	29.08	29.14	15.70	28.88	31.11	9.60	23.30	25.68	12.58	29.69	29.40	24.41
Average X-Large LLMs	–	47.55	–	–	45.17	–	–	50.62	–	–	42.70	–	–	44.47	–	–	37.95	–	–	44.74	–	44.74
Average Large LLMs	–	42.45	–	–	38.93	–	–	42.52	–	–	35.94	–	–	37.50	–	–	33.39	–	–	38.46	–	38.46
Average Medium LLMs	–	37.63	–	–	34.98	–	–	37.61	–	–	33.19	–	–	33.42	–	–	26.41	–	–	33.87	–	33.87
Average Small LLMs	–	26.94	–	–	23.79	–	–	27.06	–	–	22.82	–	–	22.59	–	–	18.48	–	–	23.62	–	23.62
Average LLMs	–	34.55	–	–	31.54	–	–	34.89	–	–	29.84	–	–	30.25	–	–	25.12	–	–	31.03	–	31.03
Average X-Large	20.00	43.07	38.64	17.50	39.98	39.62	18.70	44.07	40.16	16.86	38.74	37.09	23.16	38.75	42.24	12.74	33.52	35.24	18.16	39.68	38.83	35.83
Average Large	17.81	41.52	36.02	15.63	38.19	35.55	17.14	41.38	36.98	18.42	34.84	33.85	19.70	36.12	38.03	12.67	32.21	30.72	16.90	37.38	35.20	33.96
Average Medium	11.81	34.87	28.47	10.89	31.53	27.68	11.14	34.40	28.53	13.24	29.81	28.98	13.94	30.61	28.62	7.86	23.12	22.99	11.48	30.72	27.55	27.41
Average Small	8.82	27.78	26.02	8.54	24.41	26.13	8.56	26.89	25.69	10.72	24.05	24.24	12.89	22.97	26.38	8.86	18.84	22.55	9.73	24.16	25.17	21.74
Average Open	12.77	34.11	30.13	11.67	30.79	29.96	12.15	33.70	30.39	13.60	29.40	29.14	15.70	29.46	31.11	9.60	24.07	25.68	12.58	30.26	29.40	27.21
Average Proprietary	40.64	53.67	56.08	33.51	50.77	52.78	36.85	54.25	54.60	27.65	48.82	50.04	38.76	51.47	54.07	22.66	43.57	46.26	33.35	50.43	52.31	45.36
Average	20.11	37.27	36.96	17.42	34.01	35.96	18.65	37.02	36.76	17.30	32.53	34.64	21.77	33.01	37.15	13.04	27.22	31.10	18.05	33.51	35.43	30.14

Table 17: Average Judge Score for the GIMMICK Cultural Knowledge Question Answering (CKQA) – Describing. The columns **I**, **T**, and **I+T** stand for **image-only**, **text-only**, and **image+text** input to the model.

	West EU & North Am.	East EU	Asian & Pacific	Latin-America & Caribbean	Arab	Subsaharian Africa	Average
GPT-4o	37.79	32.57	37.68	30.15	38.03	28.42	34.11
Claude 3.5 Sonnet	40.27	33.63	39.29	25.71	38.16	24.25	33.55
GPT-4o Mini	34.46	28.73	33.08	23.67	34.87	25.89	30.12
Centurio Qwen	18.69	19.10	21.97	18.67	25.46	15.96	19.98
Gemini Pro	16.91	15.60	16.71	11.13	17.30	10.55	14.70
Gemini Flash	15.77	16.27	14.87	11.60	14.61	11.30	14.07
InternVL2.5 38B	14.06	12.60	16.24	10.71	21.12	8.36	13.85
Phi 3.5 Vision	15.17	13.67	13.54	12.45	14.28	10.75	13.31
InternVL2.5 78B	12.08	14.73	14.89	7.35	15.72	7.53	12.05
InternVL2.5 26B	11.51	10.50	13.16	7.65	14.34	7.74	10.82
InternVL2.5 1B	10.20	9.43	10.42	10.71	14.80	8.22	10.63
Qwen2 VL 72B	11.04	10.07	9.96	7.40	11.45	8.56	9.75
MiniCPM V 2.6	8.89	8.60	11.42	4.74	10.99	9.79	9.07
Centurio Aya	6.95	6.57	6.06	8.78	5.20	7.40	6.83
InternVL2.5 2B	6.31	6.80	6.17	7.14	8.49	3.08	6.33
InternVL2.5 4B	6.28	5.47	5.07	6.02	9.28	5.00	6.19
InternVL2.5 8B	6.51	5.30	4.54	6.48	9.28	3.77	5.98
Qwen2 VL 2B	5.40	4.27	7.35	3.62	5.53	3.63	4.97
Qwen2 VL 7B	5.27	5.63	4.78	4.03	6.32	3.70	4.96
Average X-Large LVLMs	11.56	12.40	12.42	7.38	13.58	8.04	10.90
Average Large LVLMs	12.78	11.55	14.70	9.18	17.73	8.05	12.34
Average Medium LVLMs	9.26	9.04	9.75	8.54	11.45	8.12	9.36
Average Small LVLMs	8.67	7.93	8.51	7.99	10.48	6.14	8.29
Average LVLMs	9.88	9.48	10.40	8.27	12.30	7.39	9.62
Average X-Large	11.56	12.40	12.42	7.38	13.58	8.04	10.90
Average Large	12.78	11.55	14.70	9.18	17.73	8.05	12.34
Average Medium	9.26	9.04	9.75	8.54	11.45	8.12	9.36
Average Small	8.67	7.93	8.51	7.99	10.48	6.14	8.29
Average Open	9.88	9.48	10.40	8.27	12.30	7.39	9.62
Average Proprietary	29.04	25.36	28.33	20.45	28.59	20.08	25.31
Average	14.92	13.66	15.12	11.47	16.59	10.73	13.75

Table 18: Average Judge Score for the GIMMICK Cultural Knowledge Question Answering (CKQA) – Naming.

VI

Centurio: On Drivers of Multilingual Ability of Large Vision-Language Model

Bibliographic Entry

Gregor Geigle*, **Florian Schneider***, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, and Goran Glavaš. 2025. Centurio: On Drivers of Multilingual Ability of Large Vision-Language Model. In *Proceedings of the Association for Computational Linguistics: ACL 2025,* in press. Vienna, Austria: Association for Computational Linguistics [1]

1. * The authors contributed equally to this work.

`Centurio`: On Drivers of Multilingual Ability of Large Vision-Language Model

Gregor Geigle[12*] **Florian Schneider**[3*] **Carolin Holtermann**[4]
Chris Biemann[3] **Radu Timofte**[2] **Anne Lauscher**[4] **Goran Glavaš**[1]
[1]WüNLP, [2]Computer Vision Lab, CAIDAS, University of Würzburg
[3]Language Technology Group, [4]Data Science Group, University of Hamburg
(gregor.geigle|florian.schneider-1)@uni-(wuerzburg|hamburg).de
gregor-ge.github.io/Centurio

Abstract

Most Large Vision-Language Models (LVLMs) to date are trained predominantly on English data, which makes them struggle to understand non-English input and fail to generate output in the desired target language. Existing efforts mitigate these issues by adding multilingual training data, but do so in a largely ad-hoc manner, lacking insight into how different training mixes tip the scale for different groups of languages. In this work, we present a comprehensive investigation into the training strategies for massively multilingual LVLMs. First, we conduct a series of multi-stage experiments spanning 13 downstream vision-language tasks and 43 languages, systematically examining: (1) the number of training languages that can be included without degrading English performance and (2) optimal language distributions of pre-training as well as (3) instruction-tuning data. Further, we (4) investigate how to improve multilingual text-in-image understanding, and introduce a new benchmark for the task. Surprisingly, our analysis reveals that one can (i) include as many as 100 training languages simultaneously (ii) with as little as 25-50% of non-English data, to greatly improve multilingual performance while retaining strong English performance. We further find that (iii) including non-English OCR data in pre-training and instruction-tuning is paramount for improving multilingual text-in-image understanding. Finally, we put all our findings together and train `Centurio`, a 100-language LVLM, offering highly competitive performance in an evaluation covering 14 tasks and 56 languages.

1 Introduction

Large Vision-Language Models (LVLMs) extend Large Language Models (LLMs) (Brown et al., 2020) to natively understand images as input (Li et al., 2023; Liu et al., 2023). This leverages the impressive language generation and reasoning abilities of recent LLMs (Llama Team, 2024; Yang et al., 2024) for vision-language tasks like image captioning or visual question answering.

While there exists a plethora of LVLMs (Zhang et al., 2024), most models are trained with just English data (Liu et al., 2024a; Dai et al., 2023; Liu et al., 2024b). This limits the access for speakers of other languages as the resulting models have several limitations even if the underlying LLMs exhibit multilingual capabilities: the models fail to understand non-English instructions (Schneider and Sitaram, 2024), struggle with non-English text content in images (Tang et al., 2024), and often fail to reply in the correct language, i.e., they have problems with *language fidelity* (Hinck et al., 2024). To ameliorate these issues, LVLMs need to be trained on a multilingual data composition. As the amount of data one can train on is always limited—by time, computing resources, financial costs, or other constraints—an effective distribution of the data across different languages is paramount. Existing multilingual LVLM work has, however, given minimal consideration to this central question of optimal training data composition (e.g., Geigle et al., 2023; Sun et al., 2024; Maaz et al., 2025).

In this work, we comprehensively investigate the space of language distributions of LVLM training mixes, focusing on the presumed trade-off between the number of included languages and performance across languages—grouped by the amount of data available for them—under a *fixed training budget*. We train several models with different data compositions obtained by machine-translating high-quality English data and benchmark them across 13 downstream tasks covering 43 diverse languages—from low-resource languages like Igbo to high-resource languages like German. We focus on four research questions, each building on the previous one, designed to identify an optimal multilingual training mix: **RQ1:** What is the optimal

[*] Equal contribution.

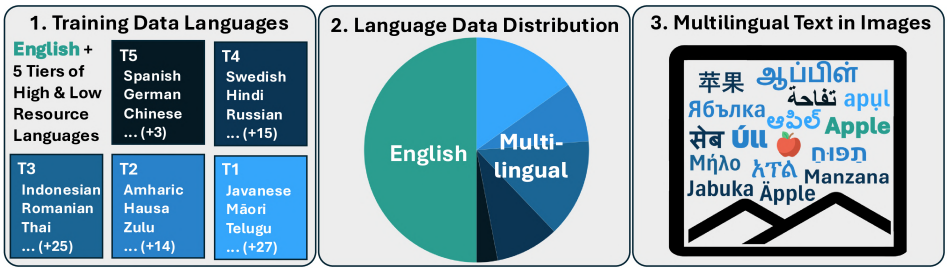


Figure 1: Exploring drivers of multilingual ability: (1) Languages in the training data; (2) Distribution of languages in the training data; (3) Incorporating multilingual OCR samples to understand non-English text in images.

number of training languages? **RQ2 & RQ3**: What is the optimal distribution of data across languages in (RQ3) pre-training data and (RQ2) instruction-tuning? **RQ4:** How to improve the understanding of multilingual text in images? To measure progress for RQ4, we introduce SMPQA (Synthetic Multilingual Plot Question Answering), a novel dataset for testing multilingual OCR capabilities, spanning 11 languages and 7 scripts.

Our findings are encouraging, albeit surprising. **1.** We do not observe the infamous "curse of multilinguality" (Conneau et al., 2020; Pfeiffer et al., 2022b) and find that gradually increasing the number of languages incurs only a negligible "performance tax": scaling from 7 to 100 languages greatly improves performance for languages newly added to the training data, especially with respect to language fidelity, while largely retaining performance levels for all previously added languages. **2.** We find that exposure to a language matters more than increasing the training portion of the language or, in particular, that the majority of the training data can still be in English, which lowers the cost of acquiring training data in other languages (e.g., via machine translation). Concretely, we find that turning 25 to 50% of training data multilingual yields strong performance, with more data sometimes even degrading performance; in *pre-training*, having a larger share of multilingual data is more beneficial, but it also saturates after 50%. **3.** We obtain mixed results for text-in-image problems: while incorporating (synthetic) OCR data with 5k samples per language rapidly boosts the performance for Latin-script languages, the same does not hold for languages with other scripts.

Finally, to demonstrate the practical impact of our findings, we train `Centurio`, a massively multilingual LVLM with 100 languages, following what we found to be an "optimal" data distribution across languages for both training stages. `Centurio` achieves state-of-the-art results over 14 tasks , matching the performance of popular multilingual open-weight LVLMs like Qwen2-VL (Wang et al., 2024b), InternVL 2.5 (Chen et al., 2024d) and Pangea (Yue et al., 2025) on English and other high-resource languages while outperforming them on low(er)-resource languages.

2 Drivers of Multilingual Ability

The design space for training (multilingual) LVLMs is extensive, ranging from the choice of the image encoder and the alignment module between the image encoder and LLM to the selection of training data. (Karamcheti et al., 2024; Laurençon et al., 2025; Tong et al., 2024). Exhaustively searching through the cross-product of all choices is not feasible. In this work, we focus on extensive evaluation of language distributions of training data in both pre-training and instruction-tuning. Intuitively, this should be a major factor affecting the multilingual ability of an LVLM. Figure 1 illustrates the scope of our analysis. We keep adding groups of languages—from highest- to lowest-resourced, following the "resourceness" tiers of (Joshi et al., 2020)—into the training mix while keeping the data size fixed. Besides the number of languages, our main focus is on the division of the training budget between English and all other languages. Finally, we posit that, besides understanding instructions and generating outputs in different languages, truly multilingual LVLMs must be able to "understand" multilingual text in images. We thus pay special attention to training adaptions for multilingual text-in-image problems.

2.1 Experimental Setup

Architecture. For our experiments, we adopt the popular LLaVA architecture (Liu et al., 2023, 2024a): An image encoder encodes images into

a sequence of visual tokens, which are projected with a 2-layer MLP into the LLM embedding space. These projected visual tokens are then concatenated with the text tokens and fed through the LLM to generate textual output.

For all our experiments, we chose `SigLIP SO400`[1] (Zhai et al., 2023) as the image encoder because it seems to be the best-performing "CLIP-like", i.e., language-supervised, embedding model, as shown by Tong et al. (2024). Further, we decided against other current image encoders for the following two reasons:1) The multilingual language understanding and text generation capabilities evaluated in RQ1, RQ2, and RQ3 are mainly attributed to the LLM, with the image encoder expected to play only a minor role. 2) When it comes to "text-in-image" understanding capabilities assessed in RQ4, where the image encoder is thought to play a more important role, all relevant image encoders are trained primarily with English image captions. Hence, we expect all of them to struggle with low-resource languages and especially with non-Latin scripts due to the lack of exposure to respective data during training.

We chose `Phi 3.5` (Abdin et al., 2024b) as our LLM because it exhibits strong multilingual performance while its small size (3.8B parameters) allows for more computationally efficient experimentation while still exhibiting strong multilingual performance. To show that our findings generalize to other LLMs, we repeat a subset of the analysis experiments with `Llama 3` (8B) (Llama Team, 2024) as the LLM backbone (see Appendix D.1).

Training Setup. Following previous work (Liu et al., 2024a; Tong et al., 2024), we split the training into two phases : *1) pre-training:* the model is trained only on image captioning, with dense image captions; *2) instruction tuning:* the model is trained on a mix of diverse vision-language tasks using several public datasets. While pre-training benefits downstream performance, it is not strictly necessary for the LVLM to perform well on the downstream tasks (Karamcheti et al., 2024). To reduce the computational cost of our analysis (i.e., to avoid coupling each language distribution over pre-training data with every language distribution of instruction-tuning data), we skip pre-training while searching for an optimal language distribution for instruction tuning. Then, with instruction-tuning data fixed, we search for an optimal language distribution for pre-training data. In both stages, we freeze the image encoder and only update the MLP and LLM (with LoRA (Hu et al., 2022)) weights. We provide further details in Appendix A.

Training Data. Our controlled experiments require comparable data over a wide range of languages. Existing multilingual datasets, available only for some tasks and in a handful of languages[2] thus do not meet our needs. Instead, we resort to machine translation (MT) and use the open NLLB model (Costa-jussà et al., 2022)[3] to translate readily available English datasets.[4] While MT results in lower data quality, especially for lower-resource languages, it is the only option to obtain multilingual vision-language training data at scale. Moreover, gains from "low-quality" MT data are guaranteed to be met or even surpassed with higher-quality translations (e.g., commercial MT or human translators). Our instruction-tuning data is adapted from LLaVA-Next (Liu et al., 2024b) and contains 0.77M samples. For pre-training, we use the 1.3M dense captions from ShareGPT4v (Chen et al., 2024b). We provide further details in Appendix B.

Evaluation. We curate an extensive test suite of 13 tasks covering 43 languages to assess the multilingual abilities of our models. Following Joshi et al. (2020), we cluster the tested languages into five *tiers*, with *T5* encompassing the high-resource languages (e.g., German, Chinese) and *T1* extremely low-resource languages (e.g., Maori, Telugu). The tasks contained in our test suite are twofold: (1) *discriminative* tasks with questions that require binary ("yes/no") or multiple-choice answers and (2) *open generation* tasks, where the models need to generate output in the target language (e.g., an image caption or a free-form answer). Generative tasks additionally evaluate a model's *language fidelity*, i.e., the ability to generate the answer in the language of the instruction. The full list of evaluation tasks and languages, along with further details, is in Appendix C. We report the results for language tiers (*T1–T5*), averaging the scores over all tasks and all tier languages.[5] We separately report English performance and **exclude** it from *T5*.

[1] `google/siglip-so400m-patch14-384`

[2] See, for example, datasets collected by Yue et al. (2025)

[3] `facebook/nllb-200-distilled-1.3B`

[4] We do **not** translate text-in-image datasets as that would result in mismatches between the instruction/output language and the English text in the image.

[5] While tasks use different measures, all are on the 0-100% scale, so no task skews the average.

2.2 RQ1: Number of Training Languages

We first investigate on *how many* languages to actually train with: does training on few high-resource languages and (zero-shot) cross-lingual transfer to unseen languages suffice, as suggested, e.g., by Shaham et al. (2024); Chen et al. (2024c); Kew et al. (2024), or do we need to explicitly include each targeted languages? Conversely, does training with more languages harm the per-language performance, with a smaller portion of the training data now allocated to each language?

Setup. We focus on the instruction-tuning step: 50% of the data is kept in English[6], while the other 50% split between N other languages equally, i.e., each language gets $\frac{50}{N}\%$ of the data budget. We gradually increase N, starting with the highest-resource tier (T5) and then including tiers of lower-resource languages (T4 to T1), one at a time. This results in the following setups: **T5** ($N = 6$), **T5-T4** ($N = 24$), **T5-T3** ($N = 52$), **T5-T2** ($N = 69$), and finally **L100** ($N = 99$). In **L100**, in addition to languages from **T5-T2**, we include T1 languages to cover XM3600 (Thapliyal et al., 2022) and otherwise randomly to reach 99 languages.

Results. Table 1 summarizes the results. Expectedly, we find that including a language (tier) in instruction-tuning improves their performance (Table 1a, top half). Nevertheless, the (negative) effect of adding new languages on performance of previously included languages is negligible, if at all present. This makes training massively multilingual LVLMs feasible with only minor performance drawbacks for any given language. In-language training leads to dramatic improvements in language fidelity (i.e., the model producing the output in the correct language), as shown in Table 1b. Interestingly, the more multilingual the training, the larger the fidelity gains also for languages not included in training; explicit in-language training, expectedly, then further improves fidelity for any given language (see Table 27 in the Appendix for detailed per language results). Even when excluding tasks where language fidelity plays a role (Table 1a bottom), we observe the same trends: steady improvements from in-language training, with negligible (if any) performance drops for other languages. A subset of experiments with Llama 3 (setups: English, T5, and L100) in Table 13 in the Appendix confirms these trends observed with Phi 3.5: in fact, we see even larger gains over all tasks when training with more languages.

[6]More specifically, 50% of the 80% of non-text-in-image data, which is excluded from translation.

Train Lang.	T1	T2	T3	T4	T5	en
All tasks						
English	14.4	30.4	24.4	23.6	28.5	53.6
T5	16.5	31.0	26.3	26.7	34.0	53.7
T5-4	17.4	30.6	27.9	**29.6**	33.5	51.5
T5-3	17.7	31.4	**32.1**	29.0	34.1	52.7
T5-2	17.0	**34.5**	30.0	28.2	33.4	**54.1**
L100	**19.3**	32.6	30.7	28.9	**34.4**	52.6
Tasks unaffected by language fidelity						
English	33.0	32.5	36.3	38.5	**42.9**	55.7
T5	35.3	33.2	36.4	38.7	42.4	56.0
T5-T4	35.8	32.6	37.8	**40.1**	42.2	55.7
T5-T3	35.9	33.6	**40.5**	39.7	42.6	**56.3**
T5-T2	35.2	**36.5**	38.5	39.5	42.8	55.5
L100	**36.1**	34.3	39.1	39.8	42.7	54.6

(a) Scores are averaged over results from all tasks grouped by language tier. The performance on the following tasks is affected by language fidelity: XM3600, MaXM, MTVQA.

Train Lang.	T1	T2	T3	T4	T5	en
English	0.2	0.2	0.1	2.4	6.2	100.0
T5	39.1	36.1	82.2	83.9	**99.1**	100.0
T5-T4	61.8	84.6	87.5	**99.2**	98.4	100.0
T5-T3	72.9	84.4	**98.2**	95.2	97.9	100.0
T5-T2	68.5	**99.0**	97.9	98.4	98.1	100.0
L100	**72.9**	98.2	95.4	97.8	98.2	100.0

(b) Average language fidelity on XM3600 in %.

Table 1: **RQ1** (§2.2) results for models trained with different sets of languages. We emphasize the **best** and second-best result in each column.

2.3 RQ2: Language Distribution in Instruction-Tuning

RQ1 experiments show that massively multilingual instruction-tuning data is beneficial across the board. We now analyze *how much* of the training data should be multilingual. On the one hand, intuitively, increasing the non-English portion of the training data budget could then lead to further gains. On the other hand, the gains from more multilingual training are, at some point, likely to be offset by the fact that we are adding noisy (MT-obtained) data at the expense of clean (English) data.

Setup. We opt for the full set of 100 languages (*L100*) in this experiment due to their robust multilingual performance. However, we adjust the language distribution by keeping $E\%$ of the data budget in English and splitting the remaining $100 - E\%$ equally across the other 99 languages[7]. We consider the following six setups: $E \in \{1, 10, 25, 50, 75, 90\}$.

Results. We present the results in Table 2. We ob-

[7]We observed no benefits from an unequal allocation that up-samples low(er)-resource languages; see §D.2)

English %	T1	T2	T3	T4	T5	en
1	19.1	30.3	28.8	27.1	31.7	48.9
10	18.1	32.4	29.4	27.4	32.5	50.1
25	**19.7**	**35.5**	29.9	27.9	33.0	50.3
50	19.3	32.6	30.7	**28.9**	34.4	52.6
75	18.5	31.5	**30.7**	28.4	**34.6**	54.1
90	15.9	31.2	27.6	26.9	34.1	**54.8**

Table 2: **RQ2** (§2.3) results for models trained with different ratios of English to multilingual data in the instruction-tuning phase. Scores are averaged over results from all tasks grouped by language tier.

serve peak performance for all language tiers when between 25% and 75% of the training data is in English. For some tasks (e.g., XM3600, MaXM, BIN-MC), we see weaker performance with more English data, while for others (e.g., MTVQA, xGQA, MaRVL) more multilingual data leads to slight performance drops (see per-task results in F.1). Overall, lower-resource languages benefit from more multilingual data and, conversely, higher-resource languages benefit from more English data. However, this is in part also a consequence of language coverage across tasks: XM3600 and BIN-MC profit from a more multilingual training mix; at the same time, they are the tasks that encompass the most low(er)-resource languages.

Results obtained with the Llama 3 backbone (see Table 14 in the Appendix) follow the same pattern: we observe the best performance in T1 and T2 with $E = 10$; and for T5 and English with $E = 90$; $E = 50$ yields the best results overall, considering all tiers. Our findings align with concurrent work by Yue et al. (2025), who found that anywhere between 20 and 80% of English data yields good global performance. Following these results, we choose $E = 50$ as a robust value for the training.

2.4 RQ3: Language Distribution in Pre-Training

As hinted by (Liu et al., 2023, 2024b) and explicitly demonstrated by Tong et al. (2024), pre-training on image-caption pairs improves the LVLM's performance. We therefore, after identifying an effective distribution of instruction-tuning data, next explore the effect of different distributions of pre-training data across languages. Specifically, we test if balancing out the English and multilingual portions delivers better performance than unbalanced distributions, that assign more training budget to English or the multilingual mix, respectively.

Setup. In these experiments, we fix the instruction-tuning mix to *L100* with $E_{IT} = 50\%$ of data in English, which we found in the previous section to produce overall most balanced results. For the pre-training data mix, we select the same 100 languages, varying the portion of English image-caption pairs, $E_{PT} \in \{100\%, 50\%, 1\%\}$; as in instruction-tuning, the non-English data budget is equally distributed across the other 99 languages.

English %	T1	T2	T3	T4	T5	en
No pre-training	19.3	32.6	30.7	28.9	34.4	52.6
100	19.3	33.3	32.1	29.4	34.5	**55.2**
50	**22.8**	**39.5**	**33.8**	30.8	**35.7**	54.9
1	22.7	38.9	33.7	**31.2**	35.4	55.1

Table 3: **RQ3** (§2.4) results with different English-to-multilingual ratios (*L100*) for pre-training. All variants are identically instruction-tuned (*L100, 50% En.*).

Results. Scores in Table 3 reveal that while English-only pre-training yields downstream benefits on English tasks, it has a largely negligible effect on other languages. The multilingual mixes substantially improve the performance for virtually all language tiers, with gains being the most prominent for lowest-resource languages from T2 and T1. In contrast to instruction-tuning, a very low proportion of clean English data does not result in tangible performance degradation, but it generally does not improve the multilingual performance either. We thus select $E_{PT} = 50\%$ as the "optimal" choice for subsequent experiments. Experiments with Llama 3, with 1% and 100% of English data (see Table 15 in the Appendix) support this finding that having a highly multilingual pre-training benefits multilingual downstream performance.

2.5 RQ4: Improving on Multilingual Text-in-Image Tasks

Finally, we focus on the models' multilingual understanding of text in images and how to improve it. Unlike tasks based on natural images, text-in-image tasks cannot be translated trivially from English: even if the prompt and output text are translated, the text in the image remains in English. Because of this, we test how *synthetic* multilingual OCR data, which can be generated at scale in any number of languages, can help improve performance.

Novel Evaluation Dataset. To this end, we introduce **SMPQA** (Synthetic Multilingual Plot QA) a new multilingual evaluation dataset, which focuses on two fundamental skills required in text-in-image tasks: 1) **reading** (and outputting) the text from an

Setup	SMPQA Ground			SMPQA Read		
	en	Latin	other	en	Latin	other
No pre-training	69.6	67.2	51.9	33.4	12.8	0.1
No OCR	76.1	73.0	55.3	41.8	23.1	0.2
100% Eng.	78.4	74.7	57.9	**55.8**	39.9	3.9
50% Eng.	**81.2**	76.7	60.0	53.8	41.8	7.1
50% (frozen)	76.1	70.8	56.3	47.2	34.1	3.5
1% Eng.	81.0	**78.3**	**64.1**	54.8	**43.5**	8.0
Latin down	78.9	74.2	59.5	54.6	41.0	**9.9**

Table 4: **RQ4** (§2.5) results of models trained with additional synthetic OCR data on SMPQA for English, Latin-script languages, and languages with other scripts. **No pre-training**: from Table 2; **No OCR**: from Table 3; **frozen**: image encoder frozen; **N% Eng.**: $N\%$ of OCR data is English, rest uniform distributed over L100 languages; **Latin down**: 2.5k samples for all Latin-script languages, 10k samples for others.

image and 2) **grounding** the input text (given as part of the prompt) to the corresponding text in the image (via balanced 'yes/no' questions, e.g., "Is the bar with label *$Label* the largest?"). We provide further details on the construction and examples in Appendix C.5.[8] We construct SMPQA to cover (i) 5 Latin-script languages, one from each tier, and (ii) 6 major languages with different non-Latin scripts.

Setup. We generate multilingual synthetic text-in-image data for training following the Synthdog approach (Kim et al., 2022) (see B.3 for details). We again adopt the training setup *L100* with 50% English data, both in pre-training and fine-tuning, now adding 500k Synthdog samples to pre-training and a subset of 50k instances to the instruction-tuning mix. As before, we select $E \in \{100, 50, 1\}\%$ English samples, distributing the rest of the budget equally over the other 99 languages. We test an additional *Latin-down* distribution: we double the budget allocated to 32 non-Latin-script languages and cut the training budget for Latin-script languages (other than English) in half. Importantly, in these experiments we *unfreeze* the image encoder and fine-tune its parameters as well.

Results. Table 4 summarizes the results. The models from prior experiments, *No pre-training* and *No OCR*, succeed for English and other Latin-script languages but utterly fail on non-Latin scripts with near-random performance. We note that the model with the pre-training step (without the additional OCR data) already performs better than the model trained just via instruction-tuning; this is likely due to the presence of images with text coupled with captions that explicitly mention this text. Training with synthetic data greatly improves the performance across all languages even if all of the OCR data is in English (100% Eng.). Nonetheless, using multilingual synthetic OCR data is very effective and, importantly, does not degrade English SMPQA performance even if English constitutes only 1% of the training data. We note that *unfreezing* and training the image encoder is critical for optimal performance in *all scripts*. Despite all this, we still observe a large performance gap between Latin- and non-Latin-script languages, even if we skew the training budget towards the non-Latin scripts (*Latin-down*). We hypothesize that orders of magnitude more text-in-image training data for other scripts are required for adequate performance.[9]

3 `Centurio`: Applying Lessons Learned

Our answers to **RQ1**–**RQ4** (see §2) point to the feasibility of training massively multilingual LVLMs supporting 100 languages with a "sweet spot" of roughly 50% of the English data being MT-translated to the languages covered. For improving multilingual OCR capabilities, training on large-scale synthetic data with an unfrozen image encoder has proven effective. Demonstrating the practicability of our findings, we now train state-of-the-art multilingual LVLMs applying our lessons learned, which we call `Centurio`. We briefly describe further design choices below.

3.1 Design Choices

Text Encoder. The choice of the LLM greatly impacts multilingual performance. We benchmark several LLMs (with 7-9B parameters) following the evaluation setup described in §2 for L100 languages and translations for 50% of the English instruct data to find candidates for `Centurio` (details in Appendix D.3). The best performances where obtained with Aya-Expanse (Dang et al., 2024) and Qwen 2.5 (Yang et al., 2024) as backbones.

Image Tiling and Projection. Image tiling methods (Lin et al., 2024; Liu et al., 2024b) increase the image resolution by concatenating encodings of n non-overlapping tiles of an input image together, which significantly helps with 'reading' small text

[8] While MTVQA and M3Exam also require OCR capabilities, they also require input image resolution that is far greater than what we use in our experiments (384px); SMPQA uses bigger letters, making performance effects from multilingual training on text-in-image understanding easier to measure.

[9] Concurrently, in a preliminary exploration of text-in-image capabilities, Yue et al. (2025) noted steady gains with 50k samples per language but also observed worse performance for non-Latin-script languages.

	AVG.	XM3600			MT-VQA	SMPQA G.		SMPQA N.		M3Exam		xMMMU		C-VQA
		en	mul	fid.		en	mul	en	mul	en	mul	en	mul	
Parrot	25.8	5.6	0.4	25.0	2.0	51.0	49.9	0.0	0.0	46.6	36.2	35.3	32.4	41.1
PALO 7B	28.7	65.9	13.5	72.0	5.8	55.5	52.8	22.4	2.7	41.0	29.1	31.8	30.9	37.1
PALO 13B	29.9	67.3	17.0	60.1	6.3	54.0	51.5	25.6	4.0	45.2	28.3	32.4	28.9	39.6
Llama-Vision 3.2 11B	*32.3	35.9	7.2	33.3	15.2	91.1	84.8	58.4	22.8	—	—	—	—	38.8
Maya	33.4	55.9	14.6	65.7	5.3	51.4	50.9	14.6	1.8	49.2	36.3	37.9	33.3	39.8
Pixtral 12B	38.1	26.5	22.1	**96.8**	14.1	91.1	71.0	85.0	35.9	49.4	33.7	30.3	26.2	33.5
Phi 3.5 Vision	39.5	32.3	6.3	40.8	11.1	**92.2**	79.4	84.8	35.9	56.3	40.7	41.7	37.4	40.9
Qwen2VL 2B	41.2	68.8	5.2	13.2	19.0	85.0	83.5	68.8	47.4	47.9	40.5	36.8	35.5	33.6
MiniCPM 2.6	41.7	**87.5**	14.2	92.3	16.1	89.0	74.3	80.8	39.3	55.0	48.2	39.1	36.5	34.1
InternVL 2.5 4B	45.3	38.9	17.5	91.0	**25.1**	87.0	78.3	77.8	47.5	63.2	50.3	49.2	42.7	48.1
InternVL 2.5 8B	47.4	38.3	15.7	91.1	25.0	91.0	79.2	80.6	48.2	**67.0**	**53.3**	**50.7**	**45.2**	48.6
Qwen2VL 7B	47.7	50.3	24.6	90.0	23.2	91.2	**90.9**	85.0	**64.9**	56.1	49.7	43.0	40.7	37.6
Pangea	48.2	70.1	34.6	87.9	19.3	87.2	72.2	72.0	23.8	58.0	45.5	43.1	42.0	**55.2**
`Centurio Aya`	48.5	78.4	**39.2**	95.7	11.1	83.1	74.2	60.0	30.1	53.0	41.2	37.6	37.2	49.4
`Centurio Qwen`	**51.6**	79.1	34.4	95.2	11.9	84.8	76.1	65.2	31.7	61.2	46.9	46.4	43.0	52.9

	MAXM		xGQA		BIN-MC		XVNLI		MaRVL		VGR		VLOD	
	en	mul	en	mul	en	mul	en	mul	en	mul	en	mul	en	mul
Parrot	28.2	3.6	37.7	21.2	30.5	25.7	28.7	31.4	63.5	55.1	59.2	52.9	0.0	0.0
PALO 7B	54.0	22.5	59.1	36.6	58.7	38.6	58.0	53.4	62.7	24.1	48.3	25.6	5.8	6.8
PALO 13B	51.7	33.1	58.0	27.8	61.4	41.1	56.6	53.6	63.8	33.1	63.3	26.2	2.5	4.9
Llama-Vision 3.2 11B	0.0	4.7	39.3	27.6	75.6	50.8	—	—	—	—	—	—	—	—
Maya	55.4	17.3	58.2	49.1	54.0	43.2	50.1	43.9	60.3	56.3	46.7	42.3	20.0	20.1
Pixtral 12B	59.4	43.4	59.9	3.8	71.0	54.2	60.9	52.7	67.7	60.7	55.8	47.7	9.2	12.4
Phi 3.5 Vision	43.6	17.9	**65.2**	38.0	63.1	36.8	58.9	53.3	73.4	46.4	81.7	50.3	45.8	**31.5**
Qwen2VL 2B	53.7	26.5	60.5	38.2	78.2	47.2	61.9	56.2	67.9	55.9	61.7	50.5	22.5	20.4
MiniCPM 2.6	53.4	22.3	57.9	45.7	72.6	47.4	71.9	65.4	70.2	57.9	52.5	49.1	9.2	14.6
InternVL 2.5 4B	46.0	42.5	63.6	28.0	68.4	45.4	69.0	58.7	74.9	59.0	72.5	49.7	24.2	21.0
InternVL 2.5 8B	45.6	38.2	63.4	32.0	70.3	44.2	73.5	66.4	83.0	63.3	87.5	51.6	**57.5**	29.0
Qwen2VL 7B	54.7	31.2	62.5	49.3	**80.7**	**57.5**	62.1	59.6	69.8	60.2	60.0	52.9	5.8	13.2
Pangea	**61.4**	**55.0**	64.6	**60.4**	70.3	52.1	69.0	65.2	75.8	70.5	69.2	58.9	0.0	6.7
`Centurio Aya`	55.7	49.3	59.1	53.2	69.7	54.7	65.0	62.4	85.0	77.9	82.5	66.8	12.5	20.7
`Centurio Qwen`	60.1	47.7	60.6	54.8	72.7	56.2	**75.4**	**70.2**	**89.6**	**81.7**	87.5	**73.1**	28.3	27.0

Table 5: Comparison of `Centurio` and 13 other LVLMs across 14 tasks. We highlight the **best** and second-best results. Scores are accuracy (CIDEr for XM3600). **en** & **mul** are the English and averaged multilingual results. **XM3600 fid.** is the language fidelity over all languages; **SMPQA G. & N** are **G**rounding and **N**aming. ***:** supports only single-image input. **AVG.**: average over all tasks. Details on the setup and models are provided in Appendix C.

Model	T1	T2	T3	T4	T5	en
`Centurio Aya`	35.1	46.4	47.0	46.7	48.3	60.6
`Centurio Qwen`	38.1	**51.0**	**48.3**	**47.0**	**50.9**	**66.6**
InternVL 2.5 8B	29.9	37.0	37.4	41.0	50.5	64.4
Qwen2VL 7B	30.6	36.8	40.5	46.2	48.0	56.8
Pangea	**38.5**	38.6	46.9	44.2	49.9	59.8
Without multi-image tasks (MaRVL, VGR, VLOD):						
`Centurio Aya`	35.1	44.5	45.7	46.2	47.7	60.7
`Centurio Qwen`	38.1	**49.5**	45.6	45.8	49.6	**66.0**
InternVL 2.5 8B	29.9	40.4	35.2	39.4	49.7	62.3
Qwen2VL 7B	30.6	38.7	40.8	**46.8**	48.3	61.7
Pangea	**38.5**	46.5	**47.7**	44.4	**49.9**	64.9

Table 6: Comparison between `Centurio` and the top-3 models of Table 5. Scores are averages over results from all 14 tasks grouped by language tier.

in images. However, they also greatly increase the input length: a 2×2 tiling would yield 3,645 tokens per image with our model.[10] Instead, we adopt the method by Shi et al. (2024), which concatenates the tokens of the whole image and the tiles along the *feature dimension* before projection by the MLP. This gives an efficient trade-off between computing cost—the number of tokens stays constant—and performance gains for fine-grained content.

[10]The whole image plus four tiles, each with 729 tokens.

Training Data. We increase the amount of the pre-training and instruct tuning data to further improve performance beyond our analysis setup. For pre-training, we add the 0.7M ALLaVA captions (Chen et al., 2024a) to the ShareGPT-4V captions and we use all synthetic OCR data generated in §2.5 (1.16M total: 500k English, 5k for Latin-script language, 10k for other scripts). For instruction-tuning, we incorporate additional datasets from the Cambrian collection (Tong et al., 2024) along with several text-only instruction-tuning datasets (see Appendix B.2 for a list). We translate the data to the *L100 50% En.* setup, excluding text-heavy datasets and others that are problematic for MT.

3.2 Results

We compare our `Centurio` models against 13 other multilingual LVLMs across the 13 tasks used in §2, and additionally evaluate them on CVQA[11], testing the models' capabilities across 56 languages. We provide details for all models in Appendix C.6.

On average, `Centurio` achieves the best results

[11]CVQA has a private test set and only allows limited submissions hence we left it out for our analysis experiments.

across all 14 tasks (cf. Table 5). Focusing on multilingual single-task performance, our model sometimes lags behind the best-performing models but still achieves best or second-best results in 9 of 14 tasks. While `Centurio` performs strongly on English, it is more often surpassed by other models. These results prove the effectiveness of our training composition: we are able to retain high English performances while maximizing the models' multilingual capabilities. When analyzing these results grouped by language tier (Table 6), we find that our models shine in the low-resource tiers T1[12] and T2, with competitive results for higher-resource languages—even when excluding multi-image tasks (VGR, MaRVL, VLOD), where our models greatly outperform most others.

For text-heavy tasks (primarily MTVQA and SMPQA), we find that `Centurio` falls behind. While we show the importance of multilingual OCR training—`Centurio` succeeds at the SMPQA reading task in more languages than, for example, Pangea—the limited input resolution and magnitudes less OCR data compared to Qwen2-VL and others result in comparably poor performance.

4 Related Work

Multilingual LVLMs. Building on the success of monolingual LVLMs like BLIP-2 (Li et al., 2023) and LLaVA (Liu et al., 2023, 2024a), researchers extended the English training protocols to include multilingual data for obtaining massively multilingual LVLMs (e.g., Maaz et al., 2025; Geigle et al., 2023). As such, Google's PaLI models (Chen et al., 2022, 2023) were the first closed-weight models trained on multilingual captions and VQA data with the recent open-weight PaliGemma (Beyer et al., 2024) following a similar training strategy. Geigle et al. (2023) presented with mBLIP the first open model, trained with image captions and a limited mix of instruct data translated to 98 languages. Subsequent models similarly followed an established procedure by directly translating parts of the English training data (Maaz et al., 2025; Hu et al., 2024; Alam et al., 2024). For the concurrent Pangea, Yue et al. (2025) optimized for multicultural aspects and used a mix of machine-translated data, existing multilingual data, and synthetically generated data. While they analyze the ratio between English and multilingual data, they do not vary the number of languages, fixing it at 39. Interestingly, most researchers either (*i*) did not properly motivate their multilingual data mix (e.g., Geigle et al., 2023; Alam et al., 2024; Beyer et al., 2024), or (*ii*) did not provide any details on the training data composition (e.g., Wang et al., 2024b; Yao et al., 2024; Chen et al., 2024d))

Multilingual OCR with LVLMs. While OCR recently gained popularity for English LVLMs (Lu et al., 2024; Tong et al., 2024), multilingual OCR has largely been neglected in prior work. As an exception, Qwen2-VL (Wang et al., 2024b) and InternVL 2.5 (Chen et al., 2024d) exhibit excellent multilingual OCR capabilities, but no training details are known. Towards open knowledge on improving multilingual OCR, Yue et al. (2025) performed preliminary experiments leveraging data in 10 languages. However, such efforts are still hindered by the lack of evaluation resources: MTVQA (Tang et al., 2024) and M3Exam (Zhang et al., 2023a) only cover up to 9 languages and conflate language understanding (in the text input) with understanding text on images. In this work, we push multilingual OCR research by presenting the novel SMPQA dataset dedicated to evaluation of multilingual OCR. We further explore how synthetic training data can improve models' capabilities.

Multilingual Instruction Tuning of LLMs.. While older LLMs struggled in multilingual tasks (Ahuja et al., 2024), more recent ones like Qwen 2.5 (Yang et al., 2024), Llama 3 (Llama Team, 2024), Gemma 2 (Team et al., 2024), or Aya (Aryabumi et al., 2024) have greatly improved in that respect, making them usable in many languages besides English. Still, current LLMs often fail to respond faithfully to the prompting language if that language is not English, especially for low-resource languages (Holtermann et al., 2024; Kew et al., 2024; Marchisio et al., 2024). To mitigate this issue, several works have analyzed the importance of multilingual instruction tuning. Weber et al. (2024) demonstrated that multilingual training is crucial for downstream performance even if the base models are pre-trained on multilingual data mixtures. Others showed that just a small set of languages is sufficient to improve cross-lingual transfer for multilingual downstream tasks significantly (Shaham et al., 2024; Chen et al., 2024c; Kew et al., 2024). However, they focus on a small set of primarily higher-resource languages, while we consider the problem in the vision-language

[12] Despite 4 of 7 T1 CVQA languages **not** in our training data.

context for a wider language selection.

In (Soykan and Sahin, 2024), the authors propose methods to select the optimal mix of languages for instruction tuning in a "linguistically-informed manner". However, they find no general best selection, and instead a task- and model-dependent selection is necessary. Therefore, in our work, we do not apply these techniques and instead choose languages based on the taxonomy introduced by Joshi et al. (2020).

5 Conclusion

In this study, we systematically investigated the optimal data composition for training a multilingual LVLM through four progressively refined analysis setups. Our findings reveal that massively multilingual training with 100 languages is highly effective, achieving comparable results to configurations with fewer languages. Moreover, only 25–50% of the training data needs to be non-English, keeping the cost of multilingual data production low. To enhance multilingual text understanding in images, we introduced a novel evaluation benchmark and demonstrated the importance and effectiveness of including multilingual synthetic OCR data in the training mix. Finally, we apply our findings to train `Centurio`, massively multilingual LVLMs trained with 100 languages, and achieve state-of-the-art results on our evaluation suite covering 14 tasks and 56 language tasks against 13 other LVLMs.

6 Limitations

Lack of Explicit Multicultural Training The focus of this work is on *language understanding* in a massively multilingual setup, that is, how to train the model to maximize its ability to understand and generate text in various languages. We do not consider the multicultural aspect, that is, training a model so that it is also more knowledgeable about concepts from the countries whose languages it can understand as measured by benchmarks like CVQA or CulturalVQA (Nayak et al., 2024). While the two aspects — multilingual and multicultural knowledge — can be intermingled in practice, they require distinct approaches in training: Multilingual data is necessary for multilingual language understanding, as we have shown. However, multicultural knowledge can be learned from multilingual resources as created by, for example, by Yue et al. (2025), but also from fully English resources like Wikipedia (Srinivasan et al., 2021).

Using Machine-Translated Training Data We train our model using machine-translated (MT) data derived from high-quality English datasets. This is advantageous because it allows us to create comparable setups for our analyses with full control over the languages and their proportions. While the data proves effective in increasing multilingual performance, MT data, especially for low-resource languages, can be of low quality and, even in higher-resource languages, might exhibit unwanted "translationese" artifacts. This can negatively impact the quality of generated text in a way that the metrics employed in our evaluation suite do not adequately measure. While native multilingual training data is available, it is not available for all tasks or languages equally, or, for most languages, not at all. Future work should consider evaluation setups to quantify the effect the MT data has on the final model, work on better MT pipelines, or create more data through native speakers.

Using Synthetically Generated OCR Data The text-heavy, "real-world" tasks in some datasets of our instruction tuning mix, which cover diverse image types such as plots, scans, application screenshots, or screenshots of webpages, are still entirely in English. Due to the issues that arise when translating such samples, we do not translate them. Hence, our methods to improve the understanding of multilingual texts in images are limited to only using synthetically generated images. While we have seen that our synthetic data positively impacts the performance of models on the respective tasks, future work should explore methods for collecting or generating more diverse data in different languages beyond our synthetic OCR data.

Another limitation regarding OCR capabilities is our relatively small image input resolution compared to models like Qwen2-VL or InternVL 2.5 — both of which support image inputs in native resolution at the cost of thousands of tokens per image —, which limits the performance of `Centurio` for images with small text.

Acknowledgments

Simulations were performed with computing resources granted by WestAI under project 9148.

Simulations were performed with computing resources from Julia 2. Julia 2 was funded as DFG project as "Forschungsgroßgerät nach Art 91b GG" under INST 93/1145-1 FUGG

The work of Gregor Geigle was in part supported by the Alexander von Humboldt Foundation.

The work of Carolin Holtermann and Anne Lauscher is funded under the Excellence Strategy of the German Federal Government and the States.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024a. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219*.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024b. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219*.

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. TallyQA: Answering Complex Counting Questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8076–8084.

Željko Agić and Natalie Schluter. 2018. Baselines and Test Data for Cross-Lingual Inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3890–3894, Miyazaki, Japan.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. 2024. Pixtral 12B. *arXiv preprint arXiv:2410.07073*.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2598–2637, Mexico City, Mexico.

AI@Meta. 2024. Llama 3 Model Card.

Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, SM Uddin, Shayekh Bin Islam, et al. 2024. Maya: An Instruction Finetuned Multilingual Multimodal Model. *arXiv preprint arXiv:2412.07112*.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan N. Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open Weight Releases to Further Multilingual Progress. *arXiv preprint arXiv:2405.15032*.

Jonas Belouadi, Anne Lauscher, and Steffen Eger. 2024. AutomaTikZ: Text-Guided Synthesis of Scientific Vector Graphics with TikZ. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, Vienna, Austria.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey A. Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bosnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier J. Hénaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. 2024. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*.

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. Scene Text Visual Question Answering. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, pages 4290–4300, Seoul, Korea (South).

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, Virtual.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulic. 2022. IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages. In *International Conference on Machine Learning*, pages 2370–2392, Baltimore, MD, USA. PMLR.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. MaXM: Towards Multilingual Visual Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2667–2682, Singapore.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. ALLaVA: Harnessing GPT4V-synthesized Data for A Lite Vision-Language Model. *arXiv preprint arXiv:2402.11684*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024b. ShareGPT4V: Improving Large Multi-modal Models with Better Captions. In *European Conference on Computer Vision*, pages 370–387, Milan, Italy. Springer.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024c. Monolingual or Multilingual Instruction Tuning: Which Makes a Better Alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, A. J. Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI-X: On Scaling up a Multilingual Vision and Language Model. *arXiv preprint arXiv:2305.18565*.

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2022. PaLI: A Jointly-Scaled Multilingual Language-Image Model. In *The Eleventh International Conference on Learning Representations*, Kigali, Rwanda.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024d. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024e. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv preprint arXiv:2412.05271*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R. Routledge, and William Yang Wang. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 3697–3711, Virtual Event / Punta Cana, Dominican Republic.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2022. HiTab: A Hierarchical Table Dataset for Question Answering and Natural Language Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 1094–1110, Dublin, Ireland.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 8440–8451, Online.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram

Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, volume abs/2305.06500, pages 49250–49267, New Orleans, LA, USA.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier. *arXiv preprint arXiv:2412.04261*.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL, USA.

Peter Devine. 2024. Tagengo: A Multilingual Chat Dataset. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 106–113, Miami, FL, USA.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. 2023. G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model. In *The Thirteenth International Conference on Learning Representations*, Singapore.

William Gaviria Rojas, Sudnya Diamos, Keertan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. 2022. The Dollar Street Dataset: Images Representing the Geographic and Socioeconomic Diversity of the World. In *Advances in Neural Information Processing Systems*, volume 35, pages 12979–12990.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavas. 2023. mBLIP: Efficient Bootstrapping of Multilingual Vision-LLMs. *arXiv preprint arXiv:2307.06930*.

Gregor Geigle, Radu Timofte, and Goran Glavas. 2024a. African or European Swallow? Benchmarking Large Vision-Language Models for Fine-Grained Object Classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 2653–2669, Miami, FL, USA.

Gregor Geigle, Radu Timofte, and Goran Glavaš. 2024b. Babel-ImageNet: Massively Multilingual Evaluation of Vision-and-Language Representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5064–5084, Bangkok, Thailand.

Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334, Honolulu, HI, USA.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions From Blind People. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pages 3608–3617, Salt Lake City, UT, USA.

Musashi Hinck, Carolin Holtermann, Matthew L. Olson, Florian Schneider, Sungduk Yu, Anahita Bhiwandiwalla, Anne Lauscher, Shao-Yen Tseng, and Vasudev Lal. 2024. Why do LLaVA Vision-Language Models Reply to Images in English? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13402–13421, Miami, FL, USA.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the Elementary Multilingual Capabilities of Large Language Models with MultiQ. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4476–4494, Bangkok, Thailand.

Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Srinivas Sunkara, Victor Carbune, Jason Lin, Maria Wang, Yun Zhu, and Jindong Chen. 2025. ScreenQA: Large-Scale Question-Answer Pairs Over Mobile App Screenshots. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9427–9452, Albuquerque, New Mexico, USA.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022*, Virtual.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Large Multilingual Models Pivot Zero-Shot Multimodal Learning across Languages. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, Vienna, Austria.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 6700–6709, Long Beach, CA, USA.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to Describe Differences Between Pairs of Similar Images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4024–4034, Brussels, Belgium.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 1988–1997, Honolulu, HI, USA.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 6282–6293, Online.

Kushal Kafle, Brian L. Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: Understanding Data Visualizations via Question Answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pages 5648–5656, Salt Lake City, UT, USA.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic VLMs: investigating the design space of visually-conditioned language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 23123–23144, Vienna, Austria.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. 2023. GeomVerse: A Systematic Evaluation of Large Models for Geometric Reasoning. In *AI for Math Workshop@ ICML 2024*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A Diagram is Worth a Dozen Images. In *Computer Vision – ECCV 2016*, pages 235–251, Amsterdam, The Netherlands.

Aniruddha Kembhavi, Min Joon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are You Smarter Than a Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 5376–5384, Honolulu, HI, USA.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2024. Turning English-centric LLMs Into Polyglots: How Much Multilinguality Is Needed? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13097–13124, Miami, Florida, USA.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-Free Document Understanding Transformer. In *Computer Vision - ECCV 2022 - 17th European Conference*, volume 13688, pages 498–517, Tel Aviv, Israel.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vision*, 123(1):32–73.

Hugo Laurençon, Léo Tronchon, and Victor Sanh. 2024. Unlocking the conversion of Web Screenshots into HTML Code with the WebSight Dataset. *arXiv preprint arXiv:2403.09029*.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2025. What matters when building vision-language models? In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1372–1381.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*, pages 19730–19742, Honolulu, HI, USA. PMLR.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal ArXiv: A Dataset for Improving Scientific Comprehension of Large Vision-Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pages 14369–14387, Bangkok, Thailand.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision*

- *ECCV 2014 - 13th European Conference*, volume 8693, pages 740–755, Zurich, Switzerland.

Ziyi Lin, Dongyang Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Yu Qiao, and Hongsheng Li. 2024. SPHINX: A Mixer of Weights, Visual Embeddings and Image Scales for Multi-modal Large Language Models. In *Computer Vision - ECCV 2024 - 18th European Conference*, volume 15120, pages 36–55, Milan, Italy.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. CLEVR-Math: A Dataset for Compositional Language, Visual and Mathematical Reasoning. In *Proceedings of the 16th International Workshop on Neural-Symbolic Learning and Reasoning as part of the 2nd International Joint Conference on Learning & Reasoning (IJCLR 2022)*, volume 3212 of *CEUR Workshop Proceedings*, pages 155–170, Windsor Great Park, UK.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually Grounded Reasoning across Languages and Cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 10467–10485, Virtual Event / Punta Cana, Dominican Republic.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 26296–26306, Seattle, WA, USA.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916.

Llama Team. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. DeepSeek-VL: Towards Real-World Vision-Language Understanding. *arXiv preprint arXiv:2403.05525*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021a. Inter-GPS: Interpretable Geometry Problem Solving with Formal Language and Symbolic Reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers)*, pages 6774–6786, Virtual Event.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *Advances in Neural Information Processing Systems 35, NeurIPS 2022*, New Orleans, LA, USA.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. Dynamic Prompt Learning via Policy Gradient for Semi-structured Mathematical Reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023*, Kigali, Rwanda.

Pan Lu, Liang Qiu, Jiaqi Chen, Tanglin Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021b. IconQA: A New Benchmark for Abstract Diagram Understanding and Visual Language Reasoning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*, Virtual.

Muhammad Maaz, Hanoona Rasheed, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M Anwer, Tim Baldwin, Michael Felsberg, and Fahad S Khan. 2025. Palo: A Polyglot Large Multimodal Model for 5B People. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1745–1754, Tucson, Arizona, USA.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pages 11–20, Las Vegas, NV, USA.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and Mitigating Language Confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 6653–6677, Miami, FL, USA.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3199, Long Beach, CA, USA.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the As-*

sociation for Computational Linguistics: ACL 2022, pages 2263–2279, Dublin, Ireland.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. InfographicVQA. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022*, pages 2582–2591, Waikoloa, HI, USA.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. DocVQA: A Dataset for VQA on Document Images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021*, pages 2199–2208, Waikoloa, HI, USA.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. OCR-VQA: Visual Question Answering by Reading Text in Images. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019*, pages 947–952, Sydney, Australia.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking Vision Language Models for Cultural Understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 5769–5790, Miami, FL, USA.

Jason Obeid and Enamul Hoque. 2020. Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020*, pages 138–147, Dublin, Ireland.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulic, and Iryna Gurevych. 2022a. xGQA: Cross-Lingual Visual Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022b. Lifting the Curse of Multilinguality by Pre-training Modular Transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, WA, USA.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D' Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Naome Etori, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 11479–11505, Vancouver, Canada.

Florian Schneider and Sunayana Sitaram. 2024. M5 – A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4309–4345, Miami, FL, USA.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 25278–25294, New Orleans, LA, USA.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge. In *Computer Vision - ECCV 2022 - 17th European Conference*, volume 13668, pages 146–162, Tel Aviv, Israel.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual Instruction Tuning With Just a Pinch of Multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand.

Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. 2024. When Do We Not Need Larger Vision Models? In *Computer Vision - ECCV 2024 - 18th European Conference*, volume 15066, pages 444–462, Milan, Italy.

Chenglei Si, Yanzhe Zhang, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. 2024. Design2Code: How Far

Are We From Automating Front-End Engineering? *arXiv preprint arXiv:2403.03163*.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand.

Gurkan Soykan and Gözde Gül Sahin. 2024. Linguistically-Informed Multilingual Instruction Tuning: Is There an Optimal Set of Languages to Tune? *arXiv preprint arXiv:2410.07809*.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 2443–2449, New York, NY, USA.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy.

Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, et al. 2024. Parrot: Multilingual Visual Instruction Tuning. *arXiv preprint arXiv:2406.02539*.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. VisualMRC: Machine Reading Comprehension on Document Images. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 13878–13888, Virtual.

Benny J. Tang, Angie W. Boggust, and Arvind Satyanarayan. 2023. VisText: A Benchmark for Semantically Rich Chart Captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 7268–7298, Toronto, Canada.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. 2024. MTVQA: Benchmarking Multilingual Text-Centric Visual Question Answering. *arXiv preprint arXiv:2405.11985*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118*.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pages 87310–87356, Vancouver, Canada.

Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. 2024. How Many Are in This Image A Safety Evaluation Benchmark for Vision LLMs. In *Computer Vision – ECCV 2024: 18th European Conference*, page 37–55, Milan, Italy.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pages 4566–4575, Boston, MA, USA.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023. To See is to Believe: Prompting GPT-4V for Better Visual Instruction Tuning. *arXiv preprint arXiv:2311.07574*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset. In *Advances in Neural Information Processing Systems*, volume 37, pages 95095–95169, Vancouver, Canada.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Alexander Arno Weber, Klaudia Thellmann, Jan Ebert, Nicolas Flores-Herr, Jens Lehmann, Michael Fromm, and Mehdi Ali. 2024. Investigating Multilingual Instruction-Tuning: Do Polyglot Models Demand for Multilingual Instructions? In *Proceedings of the*

2024 Conference on Empirical Methods in Natural Language Processing, pages 20829–20855, Miami, Florida, USA.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual Entailment: A Novel Task for Fine-grained Image Understanding. *arXiv preprint arXiv:1901.06706*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing. *CoRR*, abs/2406.08464.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 Technical Report. *CoRR*, abs/2407.10671.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *CoRR*, abs/2408.01800.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume abs/2311.16502, pages 9556–9567, Seattle, WA, USA.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2025. Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages. In *The Thirteenth International Conference on Learning Representations*, Singapore, Republic of Singapore.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. *CoRR*, abs/2303.15343.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. RAVEN: A Dataset for Relational and Analogical Visual REasoNing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 5317–5327, Long Beach, CA, USA.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. MM-LLMs: Recent Advances in MultiModal Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12401–12430, Bangkok, Thailand.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023a. M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pages 5484–5505, New Orleans, LA, USA.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023b. LLaVAR: Enhanced Visual Instruction Tuning for Text-Rich Image Understanding. *CoRR*, abs/2306.17107.

Yilun Zhao, Chen Zhao, Linyong Nan, Zhenting Qi, Wenlin Zhang, Xiangru Tang, Boyu Mi, and Dragomir Radev. 2023. RobuT: A Systematic Study of Table QA Robustness Against Human-Annotated Adversarial Perturbations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 6064–6081, Toronto, Canada.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers)*, pages 3277–3287, Virtual Event.

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded Question Answering in Images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pages 4995–5004, Las Vegas, NV, USA.

VI. Centurio:
On Drivers of Multilingual Ability of Large Vision-Language Model

Appendix Overview

Due to the number of experiments, the general density of our work, and our aim to be as transparent as possible in the sense of open science, the following appendix is extensive. Hence, we provide a brief outline of its content to ease navigation and to get an overview quickly.

A Training Setup
Hyperparameters and Languages
B Training Data
Training Datasets
C Evaluation Setup
Hyperparameters, Metrics, Datasets
D Additional Experiments
Ablation Studies on Architecture and Data
E Qualitative Examples
Showcasing `Centurio`
B Full Results
All Models, Datasets, and Languages

A Training Setup

All models are trained with the following hyperparameters: AdamW optimizer (Loshchilov and Hutter, 2019) with cosine learning rate schedule and 3% linear warmup; LORA (Hu et al., 2022) is used with rank 256 and α512 and applied to all matrices in the LLM – the LLM is otherwise frozen; the image encoder is frozen in the first three experiments and jointly trained with the model otherwise; weight decay is 0; batch size is 32 using gradient accumulation; learning rate is $1e-6$ for the image encoder, $1e-4$ for LORA and the MLP in general except when training `Centurio`, we use $5e-5$ during pretraining and $3e-5$ during instruct tuning. Models are always trained for one epoch on the entire data. The training loss is causal language modeling and we mask both image and input prompt tokens for calculating the loss.

For going from the pretraining to the instruct tuning phase, we found it best to continue training the same LORA adapter; merging the LORA weights after pretraining and initializing a new adapter gave worse results.

Hyperparameter (LORA rank & α, learning rates, weight decay) were tuned for Phi 3.5 and transferred to the other LLMs.

All models were trained with 4 H100 GPUs. Training `Centurio` took $\approx$ 6 days (half for pretraining, half for instruct tuning).

Training of one Phi 3.5 model for §2 takes 8-10h for instruct tuning, and for pre-training 12 to 20h (with synthetic OCR data and unfrozen image encoder).

A.1 Training Languages

We list the 100 languages used in training in Table 7.

B Training Data

B.1 For Analysis Experiments

The collections of datasets used in the instruct tuning phase for the analysis experiments (§2) is adapted from LLaVA-Next (Liu et al., 2024b). As multiple evaluation datasets contain multiple images in the input (MaRVL, VGR, VLOD, M3Exam, xMMMU), we include additional datasets to improve capabilities for this situation. See Table 8 for the full list.

B.2 For Training `Centurio`

For training `Centurio`, we combine the datasets from Table 8 with additional datasets listed in Table 9.

B.3 Synthetic OCR Data

We use the official Synthdog code[13] to generate the samples using the Google Noto font and with images from the ImageNet train split as background. Text is sampled from Wikipedias of the respective languages.

We consider the following 32 languages as not using the Latin script: *am, ar, as, azb, be, bg, bn, bo, el, fa, he, hi, ja, ka, kk, km, ko, lo, mr, my, pa, ru, sa, sd, sr, ta, te, th, ti, uk, ur, zh*.

C Evaluation Setup

This section describes the details of our evaluation setup.

C.1 Generation Parameters

In all experiments of our test suite, we use greedy decoding (`temperature=0.0; do_sample=False;`).

C.2 Metrics

Depending on the dataset and task, we employ either CIDEr (Vedantam et al., 2015), the exact match accuracy, or a relaxed match accuracy (see Table C.4).

[13] https://github.com/clovaai/donut/tree/master/synthdog

Name	Script	ISO-639	Flores-200	Tier
Arabic	Arabic	ar	arb_Arab	5
Chinese	Trad. Han	zh	zho_Hant	5
English	Latin	en	eng_Latn	5
French	Latin	fr	fra_Latn	5
German	Latin	de	deu_Latn	5
Japanese	Japanese	ja	jpn_Jpan	5
Spanish	Latin	es	spa_Latn	5
Basque	Latin	eu	eus_Latn	4
Catalan	Latin	ca	cat_Latn	4
Croatian	Latin	hr	hrv_Latn	4
Czech	Latin	cs	ces_Latn	4
Dutch	Latin	nl	nld_Latn	4
Finnish	Latin	fi	fin_Latn	4
Hindi	Devanagari	hi	hin_Deva	4
Hungarian	Latin	hu	hun_Latn	4
Italian	Latin	it	ita_Latn	4
Korean	Hangul	ko	kor_Hang	4
Persian	Arabic	fa	pes_Arab	4
Polish	Latin	pl	pol_Latn	4
Portuguese	Latin	pt	por_Latn	4
Russian	Cyrillic	ru	rus_Cyrl	4
Serbian	Cyrillic	sr	srp_Cyrl	4
Swedish	Latin	sv	swe_Latn	4
Turkish	Latin	tr	tur_Latn	4
Vietnamese	Latin	vi	vie_Latn	4
Afrikaans	Latin	af	afr_Latn	3
Bangla	Bengali	bn	ben_Beng	3
Belarusian	Cyrillic	be	bel_Cyrl	3
Bosnian	Latin	bs	bos_Latn	3
Bulgarian	Cyrillic	bg	bul_Cyrl	3
Cebuano	Latin	ceb	ceb_Latn	3
Danish	Latin	da	dan_Latn	3
Egyptian Arabic	Arabic	ar-eg	arz_Arab	3
Estonian	Latin	et	est_Latn	3
Galician	Latin	gl	glg_Latn	3
Georgian	Georgian	ka	kat_Geor	3
Greek	Greek	el	ell_Grek	3
Indonesian	Latin	id	ind_Latn	3
Kazakh	Cyrillic	kk	kaz_Cyrl	3
Latin	Latin	la	NO	3
Latvian	Latin	lv	lvs_Latn	3
Lithuanian	Latin	lt	lit_Latn	3
Malay	Latin	ms	zsm_Latn	3
Romanian	Latin	ro	ron_Latn	3
Slovak	Latin	sk	slk_Latn	3
Slovenian	Latin	sl	slv_Latn	3
Tagalog	Latin	tl	tgl_Latn	3
Tamil	Tamil	ta	tam_Taml	3
Thai	Thai	th	tha_Thai	3
Ukrainian	Cyrillic	uk	ukr_Cyrl	3

Name	Script	ISO-639	Flores-200	Tier
Urdu	Arabic	ur	urd_Arab	3
Uzbek	Latin	uz	uzn_Latn	3
Hebrew	Hebrew	iwhe	heb_Hebr	3
Amharic	Ethiopic	am	amh_Ethi	2
Haitian	Latin	ht	hat_Latn	2
Hausa	Latin	ha	hau_Latn	2
Icelandic	Latin	is	isl_Latn	2
Irish	Latin	ga	gle_Latn	2
Lao	Lao	lo	lao_Laoo	2
Maltese	Latin	mt	mlt_Latn	2
Marathi	Devanagari	mr	mar_Deva	2
Punjabi	Gurmukhi	pa	pan_Guru	2
Sanskrit	Devanagari	sa	san_Deva	2
Swahili	Latin	sw	swh_Latn	2
Tigrinya	Ethiopic	ti	tir_Ethi	2
Tswana	Latin	tn	tsn_Latn	2
Wolof	Latin	wo	wol_Latn	2
Xhosa	Latin	xh	xho_Latn	2
Yoruba	Latin	yo	yor_Latn	2
Zulu	Latin	zu	zul_Latn	2
Albanian	Latin	sq	als_Latn	1
Assamese	Bengali	as	asm_Beng	1
Azerbaijani	Arabic	azb	azb_Arab	1
Bambara	Latin	bm	bam_Latn	1
Burmese	Myanmar	my	mya_Mymr	1
Esperanto	Latin	eo	epo_Latn	1
Igbo	Latin	ig	ibo_Latn	1
Javanese	Latin	jv	jav_Latn	1
Khmer	Khmer	km	khm_Khmr	1
Kikuyu	Latin	ki	kik_Latn	1
Lingala	Latin	ln	lin_Latn	1
Luxembourgish	Latin	lb	ltz_Latn	1
Maori	Latin	mi	mri_Latn	1
Norwegian	Latin	no	nob_Latn	1
Occitan	Latin	oc	oci_Latn	1
Quechua	Latin	qu	quy_Latn	1
Samoan	Latin	sm	smo_Latn	1
Sango	Latin	sg	sag_Latn	1
Sardinian	Latin	sc	srd_Latn	1
Scottish Gaelic	Latin	gd	gla_Latn	1
Sindhi	Arabic	sd	snd_Arab	1
Somali	Latin	so	som_Latn	1
Swati	Latin	ss	ssw_Latn	1
Telugu	Telugu	te	tel_Telu	1
Tibetan	Tibetan	bo	bod_Tibt	1
Tok Pisin	Latin	tpi	tpi_Latn	1
Tsonga	Latin	ts	tso_Latn	1
Twi	Latin	tw	twi_Latn	1
Waray	Latin	war	war_Latn	1
Welsh	Latin	cy	cym_Latn	1

Table 7: The list of 100 languages used in our training experiments. The "Tier" column represents the tier in the taxonomy proposed by Joshi et al. (2020), where a higher tier indicates more available resources, i.e., data, in the respective language.

Dataset	Size (Images)	Translated?
Natural Image:		
LLaVA Instruct (Liu et al., 2023)	160k	yes
VQAv2 (Goyal et al., 2017)	83k	yes
GQA (Hudson and Manning, 2019)	72k	yes
OKVQA (Marino et al., 2019)	9k	yes
A-OKVQA (Schwenk et al., 2022)	30k	yes
RefCOCO (Kazemzadeh et al., 2014; Mao et al., 2016)	48k	yes
VG (Krishna et al., 2017)	86k	yes
MSCOCO (Lin et al., 2014)	50k (subset)	yes
Multiple Images:		
NLVR (Suhr et al., 2019)	86k	yes
Spot-the-difference (Jhamtani and Berg-Kirkpatrick, 2018)	8k	yes
OCR:		
OCRVQA (Mishra et al., 2019)	50k (subset)	no
DocVQA (Mathew et al., 2021)	10k	no
AI2D (Kembhavi et al., 2016)	3k	no
ChartQA (Masry et al., 2022)	18k	no
DVQA (Kafle et al., 2018)	50k (subset)	no
ScienceQA (Lu et al., 2022)	6k	no
Total	766k	

Table 8: List of datasets included in the *instruct tuning phase* in our analysis experiments. All sizes are based on unique images; examples about the same image are packed into one sequence.

For the relaxed match accuracy, we consider an answer correct if it starts with the correct choice letter. For example, answers like "A." are also counted as correct if the gold label is "A".

C.3 Prompts

We list the prompts for each dataset in our test suite used for all models in Figure 2.

C.4 Datasets

In the following, datasets included in our test suite are briefly introduced. An overview is provided in Table 10. Details about the languages covered by the datasets are listed in Table 11.

xGQA The xGQA dataset (Pfeiffer et al., 2022a) is a cross-lingual visual question-answering dataset. It extends the well-known English-only GQA dataset (Hudson and Manning, 2019) by manually translating the questions in the balanced *test-dev* set. Each of the 9666 questions is available in eight languages covering five scripts, while the answers are in English only. The dataset holds 300 unique images from Visual Genome (Krishna et al., 2017).

MaXM The MaXM dataset was introduced by Changpinyo et al. (2023) and is a VQA dataset comprising seven languages in five scripts. In MaXM, the questions and their respective answers are in the same language. The images are a subset of the XM3600 (Thapliyal et al., 2022) dataset and are chosen to match a region where the language of the question-answer pair is spoken. This ensures cultural diversity in the images in addition to the language diversity in the question-answer texts.

Dataset	Size (Images)	Translated?
Natural Image:		
ALLaVA Instruct[1] (Chen et al., 2024a)	760k	yes
LVIS Instruct4V (Wang et al., 2023)	223k	yes
Visual7W (Zhu et al., 2016)	14k	no
VizWiz QA (Gurari et al., 2018)	21k	no
TallyQA (Acharya et al., 2019)	133k	yes
SketchyVQA (Tu et al., 2024)	4k	yes
OODVQA (Tu et al., 2024)	3k	no
OCR:		
ScienceQA (Cambrian version)	6k	no
AI2D (Cambrian version)	4k	no
Rendered Text [2]	10k	no
ScreenQA (Hsiao et al., 2025)	33k	no
LLaVAR (Zhang et al., 2023b)	20k	no
ArxivQA (Li et al., 2024)	54k	no
Chart2Text (Obeid and Hoque, 2020)	25k	no
InfographicVQA (Mathew et al., 2022)	2k	no
VisText (Tang et al., 2023)	10k	no
TQA (Kembhavi et al., 2017)	1k	no
STVQA (Biten et al., 2019)	17k	no
TAT-QA (Zhu et al., 2021)	2k	no
TabMWP (Lu et al., 2023)	23k	no
HiTab (Cheng et al., 2022)	2k	no
IconQA (Lu et al., 2021b)	27k	no
VisualMRC (Tanaka et al., 2021)	3k	no
RobuT (Zhao et al., 2023)	113k	no
FinQA (Chen et al., 2021)	5k	no
Math & Code:		
WebSight (Laurençon et al., 2024)	10k	yes
Design2Code (Si et al., 2024)	0k	yes
DaTikz (Belouadi et al., 2024)	48k	no
CLEVR (Johnson et al., 2017)	70k	yes
CLEVR-Math (Lindström and Abraham, 2022)	70k	yes
Geo170k (Gao et al., 2023)	9k	no
GeomVerse (Kazemi et al., 2023)	9k	no
Inter-GPS (Lu et al., 2021a)	1k	no
MathVision (Wang et al., 2024a)	3k	no
Raven (Zhang et al., 2019)	42k	no
Text (no images):		
Aya Dataset (Singh et al., 2024)	202k	–
Tagengo-GPT4 (Devine, 2024)	70k	–
Magpie[2] (Xu et al., 2024)	400k	–
Total	2.47M	

Table 9: Datasets used on top of the datasets from Table 8 for the instruct tuning phase of `Centurio`. [1]: also contains web-scraped images from LAION (Schuhmann et al., 2022) which contain textual elements. 2:`https://huggingface.co/datasets/wendlerc/RenderedText`. [2]: Combining *magpie-ultra-v0.1* (50k), *Magpie-Qwen2-Pro-200K-English* (200k), *Magpie-Llama-3.1-Pro-MT-300K-Filtered* (150k subset).

XVNLI The XVNLI dataset (Bugliarello et al., 2022) introduces the task of Cross-lingual Visual Natural Language Inference where a model needs to predict whether a textual hypothesis *entails*, *contradicts*, or is *neutral* concerning a visual premise. XVNLI comprises five languages covering three scripts and 357 unique images from Visual Genome. It is based on a combination of the text-only SNLI (Bowman et al., 2015) dataset and its cross-lingual (Agić and Schluter, 2018) and cross-modal (Xie et al., 2019) equivalents.

MaRVL The MaRVL dataset (Liu et al., 2021) aims to benchmark models on Multicultural Reasoning over Vision and Language. A task sample

SMPQA

```
<IMG>{QUESTION}\nAnswer the question using a single word or phrase.
```

CVQA

```
<IMG>{QUESTION}\nThere are several options:\nA. {OPTION A}\nB. {OPTION B}\nC. {OPTION C}\nD.
{OPTION D}\nAnswer with the option's letter from the given choices directly.
```

xMMMU

```
{QUESTION}\nThere are several options:\nA. {OPTION A}\nB. {OPTION B}\nC. {OPTION C}\nD. {OPTION
D}\nAnswer with the option's letter from the given choices directly.
```

MTVQA

```
<IMG>{QUESTION}\nAnswer the question using a single word or phrase.\nAnswer in {LANGUAGE}.
```

M3Exam

```
{QUESTION}\nOptions:\nA. {OPTION A}\nB. {OPTION B}\nC. {OPTION C}\nD. {OPTION D}\n Answer with
the option's letter from the given choices directly.
```

BIN-MC

```
<IMG>Which of these choices (in English) is shown in the image?\n Choices:\nA. {CHOICE A}\nB.
{CHOICE B}\nC. {CHOICE C}\nD. {CHOICE D}\n Answer with the letter from the given choices directly.
```

xGQA

```
<IMG>{QUESTION}?\nAnswer the question using a single word or phrase.\nAnswer in English.
```

MaXM

```
<IMG>{QUESTION}?\nAnswer the question using a single word or phrase.\nAnswer in {LANGUAGE}.
```

MaRVL

```
<IMG>Given the two images <IMG><IMG>, is it correct to say "{HYPOTHESIS}"? Answer yes or no.'
```

XVNLI

```
<IMG>Is it guaranteed true that "{HYPOTHESIS}"? Yes, no, or maybe? Answer in English.
```

M5-VGR

```
Given the two images <IMG><IMG>, is it correct to say "{HYPOTHESIS}"? Answer yes or no.'
```

M5-VLOD

```
Based on the 5 images <IMG><IMG><IMG><IMG><IMG> ordered from top-left to bottom-right, which
image does not match the hypothesis "{HYPOTHESIS}"? Choose one from [A, B, C, D, E] and only
output a single letter:
```

XM3600

```
Briefly describe the image in {LANGUAGE} in one sentence.
```

Figure 2: Prompts used for the different datasets of our test suite. For M3Exam and xMMMU, the questions contain images at individual positions, and also the options can consist of images. In total, a sample of M3Exam can contain up to 8 images and 8 options, and a sample of xMMMU can contain up to 4 images and 4 options.

Dataset	Task	Visual Input	Textual Input	Target Output	Metric	#Lang.
MaXM	VQA	Single-Image	Question (TL)	WoP (TL)	E. Acc.	6
xGQA	VQA	Single-Image	Question (TL)	WoP (EN)	E. Acc.	8
XVNLI	VNLI	Single-Image	Hypothesis (TL)	'yes' / 'no' / 'maybe'	E. Acc.	5
M5B-VLOD	VLOD	Multi-Image	Hypothesis (TL)	LoC	R. Acc.	12
M5B-VGR	VGR	Multi-Image	Hypothesis (TL)	'yes' / 'no'	E. Acc.	12
MaRVL	VGR	Multi-Image	Hypothesis (TL)	'yes' / 'no'	E. Acc.	6
MTVQA	TH VQA	Single-Image	Question (TL)	WoP (TL)	E. Acc.	9
SMPQA - Name	TH VQA	Single-Image	Question (TL)	WoP (TL)	E. Acc.	11
SMPQA - Ground	TH VGR	Single-Image	Question (TL)	'yes' / 'no'	E. Acc.	11
M3Exam	TH MC VQA	Single or Multi-Image	Question (TL)	LoC	R. Acc.	7
MMMU	TH MC VQA	Single or Multi-Image	Question (EN)	LoC	R. Acc.	1
xMMMU	TH MC VQA	Single or Multi-Image	Question (TL)	LoC	R. Acc.	7
BabelImageNet-MC	MC VQA	Single-Image	Question (TL)	LoC	R. Acc.	20
CVQA	MC VQA	Single-Image	Question (TL)	LoC	R. Acc.	39
XM3600	Captioning	Single-Image	Prompt (EN)	Caption (TL)	CIDEr	36

Table 10: List of datasets contained in our test suite. In the Task column, "VQA" "VNLI", "VLOD", "VGR", "TH", and "MC" are acronyms for "Visual Question Answering", "Visual Natural Language Inference", "Visio-Linguistic Outlier Detection", "Visually Grounded Reasoning", "Text-Heavy", and "Multiple-Choice", respectively. In the "Textual Input" and "Target Output" columns, the acronyms "WoP", "LoC", "TL", and "EN" stand for "(Single) Word or Phrase", "Letter of the correct Choice", "Target Language", and "English", respectively. Further, "E. Acc." is "Exact Accuracy" and "R. Acc." is "Relaxed Accuracy". CVQA is not used in §2 due to its hidden test set with limited submissions.

comprises two images, a textual statement, and a binary true or false answer grounded in the images. MaRVL comprises five languages covering three scripts and 4914 culturally diverse images that match the respective languages. The images in a sample are chosen to match the culture of the annotator who has written the textual statement in his or her native language.

XM3600 The XM3600 dataset (Thapliyal et al., 2022) is a large multilingual image captioning dataset comprising 36 languages with 261375 captions covering 13 different scripts for 100 unique images per language. The images are selected to match the language's cultural background, ensuring cultural and linguistic diversity. The captions were not automatically translated but manually created by professional annotators who are native speakers of the respective language.

We only use a subset of 500/3600 images (selected randomly) per language when evaluating XM3600 due to its size.

Babel-ImageNet (multiple-choice) (BIN-MC) Babel-ImageNet (Geigle et al., 2024b) translates the labels of ImageNet (Deng et al., 2009) to nearly 300 languages, which allows us to test if models are capable of recognizing and linking the diverse objects of ImageNet to their correct label in the tested language. Testing all 300 languages would be too expensive, instead we use it to deepen our evaluation in languages appearing in only 1 or 2 other datasets, plus English and select few high-resource languages. Also, we only use 10 images per class instead of 50, again, to keep computational cost reasonable.

We formulate the task as a multiple-choice problem, following the approach by Geigle et al. (2024a) to mine hard negative options from the total label pool. This avoids problems of unclear or underspecified answers that appear in a traditional open-ended VQA formulation. We mine negatives with the English labels, filtering out all candidates not translated by Babel-ImageNet in the target language, that is, in the end, we select the three most similar negative labels that appear in the Babel-ImageNet labels of a given language.

SMPQA We propose **SMPQA** (Synthetic Multilingual Plot QA) as a novel test dataset for evaluating multilingual OCR capabilities in images – bar plots and pie charts to be specific – in 11 languages, covering different scripts and resource levels. See §C.5 for details.

M5B-VGR The M5B-VGR dataset is a Visually Grounded Reasoning dataset similar to MaRVL and was introduced by (Schneider and Sitaram, 2024). A sample comprises two images, a textual statement, and a binary true or false answer grounded in the images. It comprises 12 languages covering 7 scripts and culturally diverse photos taken in regions where the respective language is spoken. The images are sampled from the Dollar Street (Gaviria Rojas et al., 2022) dataset. For each language, there are 120 samples.

Name	Tier	ISO-639-3	ISO-639-1	Datasets
Afrikaans	3	afr	af	BabelImageNet-MC, M3Exam
Amharic	2	amh	am	BabelImageNet-MC, CVQA, M5B-VGR, M5B-VLOD
Arabic	5	ara	ar	MTVQA, SMPQA, XM3600, xMMMU, XVNLI
Bengali	3	ben	bn	CVQA, M5B-VGR, M5B-VLOD, xGQA, XM3600
Berber (macrolanguage)	0	ber	-	M5B-VGR, M5B-VLOD
Breton	1	bre	br	CVQA
Bulgarian	3	bul	bg	CVQA
Chinese	5	zho	zh	CVQA, M3Exam, MaRVL, MaXM, SMPQA, xGQA, XM3600
Croatian	4	hrv	hr	BabelImageNet-MC, XM3600
Cusco Quechua	1	quz	-	XM3600
Czech	4	ces	cs	BabelImageNet-MC, XM3600
Danish	3	dan	da	XM3600
Dutch	4	nld	nl	BabelImageNet-MC, XM3600
Egyptian Arabic	3	arz	-	CVQA
English	5	eng	en	BabelImageNet-MC, M3Exam, M5B-VGR, M5B-VLOD, MaRVL, MaXM, MME, MMMU, SMPQA, xGQA, XM3600, xMMMU, XVNLI
Filipino	3	fil	-	CVQA, M5B-VGR, M5B-VLOD, XM3600
Finnish	4	fin	fi	BabelImageNet-MC, XM3600
French	5	fra	fr	MaXM, MTVQA, XM3600, xMMMU, XVNLI
German	5	deu	de	M5B-VGR, M5B-VLOD, MTVQA, SMPQA, xGQA, XM3600
Hausa	2	hau	ha	BabelImageNet-MC, M5B-VGR, M5B-VLOD
Hebrew	3	heb	he	XM3600
Hindi	4	hin	hi	M5B-VGR, M5B-VLOD, MaXM, SMPQA, XM3600, xMMMU
Hungarian	4	hun	hu	BabelImageNet-MC, XM3600
Igbo	1	ibo	ig	CVQA
Indonesian	3	ind	id	CVQA, MaRVL, SMPQA, xGQA, XM3600, xMMMU
Irish	2	gle	ga	CVQA
Italian	4	ita	it	M3Exam, MTVQA, SMPQA, XM3600
Japanese	5	jpn	ja	BabelImageNet-MC, CVQA, MTVQA, XM3600, xMMMU
Javanese	1	jav	jv	CVQA
Kanuri	0	kau	kr	CVQA
Kinyarwanda	1	kin	rw	CVQA
Korean	4	kor	ko	CVQA, SMPQA, xGQA, XM3600
Malay (macrolanguage)	3	msa	ms	CVQA
Maori	1	mri	mi	BabelImageNet-MC, XM3600
Mi-gkabau	1	min	-	CVQA
Modern Greek	3	ell	el	BabelImageNet-MC, XM3600
Mongolian	1	mon	mn	CVQA
Norwegian	1	nor	no	BabelImageNet-MC, CVQA, XM3600
Oromo	1	orm	om	CVQA
Persian	4	fas	fa	BabelImageNet-MC, XM3600
Polish	4	pol	pl	BabelImageNet-MC, XM3600
Portuguese	4	por	pt	CVQA, M3Exam, xGQA, XM3600, xMMMU
Romanian	3	ron	ro	BabelImageNet-MC, CVQA, MaXM, XM3600
Russian	4	rus	ru	CVQA, M5B-VGR, M5B-VLOD, MTVQA, SMPQA, xGQA, XM3600, XVNLI
Sinhala	0	sin	si	CVQA
Spanish	5	spa	es	BabelImageNet-MC, CVQA, XM3600, XVNLI
Sundanese	1	sun	su	CVQA
Swahili (macrolanguage)	2	swa	sw	CVQA, M5B-VGR, M5B-VLOD, MaRVL, XM3600
Swedish	4	swe	sv	XM3600
Tamil	3	tam	ta	BabelImageNet-MC, CVQA, MaRVL
Telugu	1	tel	te	BabelImageNet-MC, CVQA, XM3600
Thai	3	tha	th	M3Exam, M5B-VGR, M5B-VLOD, MaXM, MTVQA, SMPQA, XM3600
Turkish	4	tur	tr	MaRVL, XM3600
Ukrainian	3	ukr	uk	XM3600
Urdu	3	urd	ur	CVQA
Vietnamese	4	vie	vi	M3Exam, MTVQA, XM3600
Zulu	2	zul	zu	BabelImageNet-MC, M5B-VGR, M5B-VLOD, SMPQA
Unique Languages	56 (43 without CVQA)			

Table 11: List of languages covered in the datasets of our test suite. The "Tier" column represents the tier in the taxonomy proposed by Joshi et al. (2020), where a higher tier indicates more available resources, i.e., data, in the respective language. CVQA is not used in §2 due to its hidden test set with limited submissions.

M5B-VLOD The M5B-VLOD (Visio-Linguistic Outlier Detection) dataset was introduced by (Schneider and Sitaram, 2024). A sample comprises five images and a textual statement that is true for all but one of the images. The task is to find the outlier image, that is, the image that does not match the statement. It comprises the same 12 languages as M5B-VGR and images sampled with a similar strategy from the same dataset. For each language, there are 120 samples.

MTVQA The MTVQA dataset was introduced by (Tang et al., 2024) and comprises text-heavy Visual Question Answering (VQA) tasks. It features human expert annotations across 9 diverse

languages, consisting of a total of 6778 question-answer pairs across 2116 images. The images primarily contain text in the respective language and the question (and answer) related to that text. The images are sampled from different publicly available datasets.

CVQA The CVQA dataset was introduced by (Romero et al., 2024) and is a multilingual, culturally nuanced VQA benchmark that includes a diverse set of languages, many of them underrepresented and understudied in NLP. It consists of 10000 questions across 30 countries, covering 31 languages, and in 39 distinct country-language pairs (e.g., the dataset includes 7 different splits for Spanish because it contains 7 countries where Spanish is spoken). The images in the dataset were manually gathered by human annotators to match and depict the culture of the respective country-language pair.

A sample consists of one image and a question related to the image in the respective language. The authors did not release the test set publicly but allowed up to 5 daily submissions to their leaderboard to obtain evaluation results.

M3Exam The M3Exam dataset was introduced by (Zhang et al., 2023a). It contains real-world exam questions in 9 languages, which are either text-only or multi-modal. In our test suite, we only consider samples that require at least one image. Further, due to the low number of resulting samples for Swhalili and Javanese, we only include the remaining 7 languages. The remaining samples consist of multiple-choice questions in the target language and up to 8 images that can appear both in the question and the answer options. Further, the number of options ranges from 4 to 8 depending on the individual sample.

xMMMU The xMMMU was introduced by (Yue et al., 2025) and consists of college-level multiple-choice VQA samples across seven languages. It was automatically translated using GPT4o from a subset of 300 randomly selected questions from the MMMU (Yue et al., 2024) validation split.

C.5 Details for SMPQA

We propose **SMPQA** (Synthetic Multilingual Plot QA) as a test dataset for evaluating multilingual OCR capabilities, that is capabilities to identify and read text in various languages in images, specifically bar plots and pie charts.

We test the capabilities in two directions: i) *grounding* requires the model to ground a given label in the user prompt to the corresponding part in the plot to answer a yes/no question ("Is the bar with label $X the biggest?"); ii) *reading* requires the model to output the label of a specified part of the plot ("What is the label of the biggest slice?"). The questions are simple by design, requiring minimal reasoning, math, or multi-hop capabilities, as to isolate solely the OCR capabilities in the tested language. We show example plots and questions in Figure 3.

We use exact match accuracy for both tasks. For reading, edit distance to the correct word would be a fine-grained alternative but since word lengths differ between languages – Chinese can be 1-2 characters while Indonesian can be >10 – we opt against it to more easily compare results between languages. To have a fair comparison between languages, we construct the dataset in a way that plots and questions about them are identical between languages (except for labels in the respective languages, obviously).

Construction: SMPQA is constructed with a deterministic pipeline yielding identical results for each language.
1. We define a list of diverse pie charts and bar plots by randomly sampling the number of bars/slices, the size of each, their colors, the plot size and aspect ratio, and vertical/horizontal orientation for bar plots, and exploding some slices in pie charts. For each plot type, we define 50 configurations, so we have 100 plots/images in total per language.
2. Using word lists of common words in the languages, we sample words for use as labels for the bars and pie slices to fill and ultimately render the pre-defined plots. This means the plots are identical between languages except for the labels and some size adjustments caused by different word lengths.
3. For each plot, we use templates to generate 5 questions for reading and 8 questions for grounding (with balanced 'yes' and 'no' as answers). The questions are always the same for a plot, so each language has the same questions, just with different labels.

Language Selection: We selected the languages as follows: For Latin-script languages, we chose English and one language from Tier 5 to 2 to have both high- and low-resource languages: German, Italian, Indonesian, and Zulu. For non-Latin scripts,

HuggingFace Model ID	Params
Qwen/Qwen2-VL-2B-Instruct (Wang et al., 2024b)	2B
Qwen/Qwen2-VL-7B-Instruct (Wang et al., 2024b)	7B
microsoft/Phi-3.5-vision-instruct (Abdin et al., 2024a)	4B
neulab/Pangea-7B-hf (Yue et al., 2025)	7B
openbmb/MiniCPM-V-2_6 (Yao et al., 2024)	8B
meta-llama/Llama-3.2-11B-Vision-Instruct (AI@Meta, 2024)	11B
mistralai/Pixtral-12B-2409 (Agrawal et al., 2024)	12B
AIDC-AI/Parrot-7B (Sun et al., 2024)	7B
MBZUAI/PALO-7B (Maaz et al., 2025)	7B
MBZUAI/PALO-13B (Maaz et al., 2025)	13B
OpenGVLab/InternVL2_5-4B (Chen et al., 2024e)	4B
OpenGVLab/InternVL2_5-8B (Chen et al., 2024e)	8B
maya-multimodal/maya (Alam et al., 2024)	8B

Table 12: List of models considered in our evaluation experiments.

we select 6 languages to represent scripts with high usage in the world: Russian (Cyrillic), Chinese, Korean (Hangul), Hindi (Devanagari), Arabic, and Thai.

We note that our dataset construction can easily be extended to other languages if needed (as long as word lists are available) to test, for example, more scripts (Telugu, Greek, Hebrew, ...) or languages using the Latin script with heavy use of diacritics (Vietnamese, Turkish, ...). This makes SMPQA an ideal starting point for probing OCR capabilities in diverse languages.

C.6 Baseline Models

We list the evaluated baseline models in Table 12. In all baseline evaluation experiments, we use greedy decoding (`temperature=0.0; do_sample=False;`). Further, we do not preprocess the images in any way and use the provided code for inference with the respective model.

We use relaxed match accuracy for all tasks even if `Centurio` uses exact match for a fairer comparison because some models struggled with replying just 'yes'/'no' and similar issues.

D Additional Experiments

D.1 Analysis Results with Llama 3

We report the results with Llama 3 when repeating the experiments of §2.2, 2.3, 2.4 in Table 13, 14, 15.

D.2 Non-Uniform Language Allocation

In our experiments in §2, we distribute the non-English portion of the data uniformly over all languages. We now consider two *stratified* distributions that upsample low-resource languages. A language with taxonomy i will get allocated the following portion of the non-English data:

$$p(i) = \frac{f(i)}{\sum_{j \in \text{TrainLanguages}} f(j)} \quad (1)$$

with $f(i) = \frac{1}{i}$ for **Stratified-1** and $f(i) = \frac{1}{\exp i}$ for **Stratified-2**. This effectively doubles the allocated data for T1 languages, and divides the data for T5 languages by a factor 3 or 20 (depending on Stratified-1 or -2).

Results are reported in Table 16. We do observe a small decrease for T5 and T4 languages, and also for T3 with Stratified-2, but results for T1 and T2 languages stay relatively constant despite more data. This suggests that higher-resource languages can be quite sample efficient even with what amounts to a few hundred samples (at least in the instruct tuning phase) but as the stratified distributions fail to improve lower-resource languages, there is little reason in practice to not use the uniform distribution which makes no assumptions about the resource-level of a language.

Train Lang.	**T1**	**T2**	**T3**	**T4**	**T5**	**en**
English	16.1	34.7	26.3	24.3	26.2	56.4
T5	19.1	32.5	29.3	27.2	35.5	54.3
L100	**31.1**	**43.0**	**39.4**	**35.9**	**36.4**	**56.6**
Without tasks affected by language fidelity:						
English	36.6	37.1	39.0	39.6	40.0	54.6
T5	38.8	34.8	40.1	40.2	40.4	53.5
L100	**46.3**	**44.0**	**45.0**	**42.8**	**42.9**	**55.3**

Table 13: Experimental setup of Table 1 repeated with Llama 3 and the setups: just English, T5 languages, and L100 languages.

English %	**T1**	**T2**	**T3**	**T4**	**T5**	**en**
10	**32.9**	**43.1**	38.7	35.4	35.4	54.2
50	31.1	43.0	**39.4**	**35.9**	**36.4**	56.6
90	26.9	38.7	36.9	34.2	35.8	**56.6**

Table 14: Experimental setup of Table 2 repeated with Llama 3 and the setups: 10, 50, and 90% English instruct tune data.

D.3 LLM Comparison

We train several recent 7-9B parameter LLMs on the instruct tuning data mix used in our analysis with L100 languages and 50% English. All models are trained with the same hyperparameters. We compare Llama 3 (Llama Team, 2024)[14], Gemma 2

[14]While Llama 3.1 has officially better multilingual capabilities, we found Llama 3 to work better.

English %	T1	T2	T3	T4	T5	en
No pretrain	31.1	43.0	39.4	35.9	36.4	56.6
100	33.9	44.7	43.3	39.9	39.9	60.8
1	**37.8**	**47.4**	**45.0**	**41.1**	**40.7**	**61.4**

Table 15: Results of Table 3 repeated with Llama 3 and the setups: 1 and 100% English pre-train data.

Distribution	T1	T2	T3	T4	T5	en
Uniform	18.9	32.6	30.7	**28.8**	**34.4**	52.6
Stratified-1	18.6	32.5	**30.7**	28.0	33.8	**53.0**
Stratified-2	**19.2**	**32.6**	29.5	27.4	33.9	52.0

Table 16: Comparison between our uniform allocation of data compared to two stratified allocations that upsample low-resource languages.

(Team et al., 2024), Aya-Expanse (Aryabumi et al., 2024; Dang et al., 2024), and Qwen 2.5 (Yang et al., 2024).

Table 17 shows that Qwen and Aya yield the overall best results with Aya slightly ahead in T3-T5 and Qwen (with Llama 3) better in T1 and T2. Qwen achievs by far the best English results. Qwen is also notably strong in the exam tasks (M3Exam, xMMMU) and is the only model with better-than-guessing results on average for VLOD.

D.4 Qualitative Analysis of the Effect of Synthetic OCR-Training Data on Non-Latin Scripts

We performed a qualitative analysis of the predictions in the SMPQA Naming subtask to identify potential issues with non-Latin scripts when synthetic OCR training data is used. However, the results were not conclusive. For example, without the OCR training data, some models answered with colors or values (of the chart elements) but only in non-Latin scripts, while they tried to give a proper answer in other languages. We also saw (for models trained with OCR data) that in some scripts (e.g., Thai, Hindi), answers were sometimes nearly correct with only 1 or 2 characters wrong, but e.g., in Chinese or Korean, predicted answers were often completely wrong, i.e., did not just differ by 1 or 2 incorrect radicals.

E Qualitative Examples

We provide some qualitative examples of our `Centurio` models. Figure 4, 5, and 6 show results with (non-English) text in images with English prompts. Figure 7 and 8 show examples for multilingual prompts (and responses).

LLM	T1	T2	T3	T4	T5	en
`Phi-3.5-mini-instruct`	18.9	32.6	30.7	28.8	34.4	52.6
`gemma-2-9b-it`	29.2	40.9	36.4	33.5	35.3	52.8
`Meta-Llama-3-8B-Instruct`	**31.1**	43.0	39.4	35.9	36.4	56.6
`Qwen2.5-7B-Instruct`	30.7	**43.7**	42.0	38.1	40.5	**62.7**
`aya-expanse-8b`	28.3	42.5	**43.0**	**39.8**	**40.9**	59.9

Table 17: Comparison between different LLM backbones all trained with the instruct tuning data with L100 languages and 50% English (as in §2.3).

F Full Results

We report the full results of all tasks with all language-specific results. **Avg.** refers to the average without English. Metric are the same as in §C.

F.1 Analysis Experiments

We report the full results for all models trained for §2 (and also the LLMs tested for §D.3).

The following holds for all Tables: Models of the form **'Phi 3.5 - T5 50'** are to be interpreted as using the LLM Phi 3.5 with the T5 languages and 50% English with analog interpretation for other rows.

'Phi 3.5 - PT 1' means the model was pretrained with 1% English and then instruct-tuned with the L100 50% English mix (see §3).

'Phi 3.5 - OCR 1' means the model was pretrained with 50% English for the captions and 1% English for the OCR data and then instruct-tuned with the L100 50% English mix (see §4).

BIN-MC Table 18

M3Exam Table 19

VGR Table 20

VLOD Table 21

MaRVL Table 22

MaXM Table 23

MTVQA Table 24

xGQA Table 25

XM3600 Table 26. Language fidelity for §2.2 in Table 27.

XVNLI Table 28

xMMMU Table 29

SMPQA - Ground Table 30

SMPQA - Name Table 31

	en	avg.	af	am	cs	el	es	fa	fi	ha	hr	hu	ja	mi	nl	no	pl	ro	ta	te	zu
Phi 3.5 - English	64.7	38.1	43.3	29.7	41.5	35.5	55.9	33.6	36.4	24.5	43.3	39.0	49.8	27.8	47.3	44.2	41.4	42.8	30.0	27.1	31.1
Phi 3.5 - T5 50	66.0	39.6	46.0	30.3	43.1	36.3	56.3	33.4	36.5	35.1	45.1	40.7	50.9	30.1	48.4	46.2	41.1	43.1	31.0	29.6	29.1
Phi 3.5 - T5-4 50	65.2	40.6	46.8	29.6	44.6	37.9	59.1	36.7	37.5	29.0	46.4	42.5	52.0	31.1	50.7	47.4	43.0	43.5	31.5	29.0	32.4
Phi 3.5 - T5-3 50	65.5	40.6	50.0	28.8	43.3	37.4	58.6	34.4	38.4	33.0	46.1	41.4	50.9	31.2	49.7	47.0	41.9	43.8	32.5	29.6	32.7
Phi 3.5 - T5-2 50	64.8	39.1	47.2	25.6	41.9	35.8	57.9	34.0	36.0	29.8	44.8	39.5	50.0	30.5	47.7	45.8	41.2	42.4	30.4	29.2	33.9
Phi 3.5 - L100 50	64.7	39.9	48.1	28.2	42.8	36.8	57.2	34.7	37.0	28.2	44.7	40.4	51.2	31.6	47.8	46.4	40.9	43.8	30.6	30.1	37.5
Llama 3 - English	65.4	40.9	44.0	28.2	46.9	42.2	53.0	42.4	38.7	31.1	47.6	46.3	48.6	30.1	48.2	47.4	44.0	44.9	31.6	32.5	29.3
Llama 3 - T5 50	63.9	43.7	50.6	28.7	49.2	46.4	54.6	46.6	41.7	35.4	50.7	50.8	51.9	30.0	51.2	50.9	47.0	48.4	31.1	35.6	30.1
Llama 3 - L100 50	66.2	48.8	55.3	35.1	54.2	51.2	56.2	47.6	46.2	37.2	56.1	54.1	53.3	33.7	54.6	54.3	50.8	51.9	43.6	50.9	40.8
Phi 3.5 - L100 1	63.1	39.7	47.4	26.8	42.9	36.7	56.2	34.3	35.9	33.5	46.8	40.5	49.0	32.7	48.4	46.7	41.3	43.1	29.0	29.9	34.2
Phi 3.5 - L100 10	62.7	39.4	47.1	27.1	43.1	36.8	56.5	34.4	36.5	29.3	43.8	40.9	49.8	29.8	47.2	48.2	41.4	43.6	30.2	28.0	34.4
Phi 3.5 - L100 24	63.3	40.4	48.0	29.0	43.3	37.7	56.5	35.2	36.7	32.4	46.9	40.7	50.7	33.0	49.3	47.2	41.8	44.4	31.9	31.2	31.1
Phi 3.5 - L100 50	64.7	39.9	48.1	28.2	42.8	36.8	57.2	34.7	37.0	28.2	44.7	40.4	51.2	31.6	47.8	46.4	40.9	43.8	30.6	30.1	37.5
Phi 3.5 - L100 75	65.4	39.8	47.1	26.0	42.0	37.1	57.4	34.7	36.9	32.2	44.2	40.4	51.3	31.5	49.4	46.8	41.6	42.9	31.1	28.5	34.2
Phi 3.5 - L100 90	64.7	37.5	43.8	24.1	40.3	35.8	57.1	31.8	35.7	25.0	43.1	39.2	49.1	24.9	47.9	44.4	39.3	42.8	28.7	27.7	31.9
Llama 3 - L100 10	65.9	49.8	58.4	38.1	55.0	50.9	58.5	49.3	45.7	40.7	59.4	56.3	54.1	34.9	53.7	56.8	51.8	51.3	42.6	51.9	36.0
Llama 3 - L100 50	66.2	48.8	55.3	35.1	54.2	51.2	56.2	47.6	46.2	37.2	56.1	54.1	53.3	33.7	54.6	54.3	50.8	51.9	43.6	50.9	40.8
Llama 3 - L100 90	64.4	45.3	52.5	26.8	51.0	47.2	54.8	45.9	44.0	29.5	54.1	50.0	51.2	31.3	52.1	51.8	48.6	49.8	36.5	48.3	34.7
Phi 3.5 - L100 50	64.7	39.9	48.1	28.2	42.8	36.8	57.2	34.7	37.0	28.2	44.7	40.4	51.2	31.6	47.8	46.4	40.9	43.8	30.6	30.1	37.5
Phi 3.5 - PT 100	66.3	38.9	48.4	25.0	42.7	36.0	57.3	33.2	36.5	22.3	44.8	39.8	49.9	31.3	48.6	46.4	41.3	43.2	30.5	30.8	30.6
Phi 3.5 - PT 50	65.7	42.2	50.0	37.8	44.2	40.0	57.8	36.0	36.5	33.0	45.2	41.8	49.3	35.0	49.0	48.1	42.0	44.1	33.7	37.7	40.6
Phi 3.5 - PT 1	65.8	42.8	50.1	35.1	44.8	38.9	56.9	37.9	37.5	41.2	49.1	42.1	49.6	33.4	49.6	48.2	43.6	45.9	34.9	36.1	38.5
Llama 3 - L100 50	66.2	48.8	55.3	35.1	54.2	51.2	56.2	47.6	46.2	37.2	56.1	54.1	53.3	33.7	54.6	54.3	50.8	51.9	43.6	50.9	40.8
Llama 3 - PT 1	69.6	55.5	62.4	44.0	60.4	60.4	62.9	55.3	51.7	40.4	63.0	62.1	59.9	36.6	59.4	62.1	58.0	58.6	50.6	60.6	45.7
Llama 3 - PT 100	68.7	53.6	63.4	36.8	59.6	58.1	62.5	54.1	50.8	37.5	63.1	61.6	60.7	36.9	59.9	61.0	58.0	58.0	46.5	54.0	34.9
Gemma 2 - L100 50	60.5	44.8	49.1	42.5	47.5	45.3	52.0	44.8	41.6	30.9	50.7	47.6	51.4	32.8	49.8	51.1	47.2	47.5	41.8	45.1	32.1
Llama 3 - L100 50	66.2	48.8	55.3	35.1	54.2	51.2	56.2	47.6	46.2	37.2	56.1	54.1	53.3	33.7	54.6	54.3	50.8	51.9	43.6	50.9	40.8
Qwen 2.5 - L100 50	68.2	50.6	62.4	37.1	57.9	50.8	63.4	49.6	42.6	28.7	61.0	48.3	63.1	33.5	58.8	58.2	57.2	55.4	36.8	55.6	40.6
Aya-Expanse - L100 50	67.6	52.0	62.2	31.0	65.3	65.5	63.2	58.9	39.8	33.2	60.8	46.3	65.1	33.1	61.3	55.5	60.2	61.9	43.5	43.2	37.2
`Centurio Aya`	69.7	54.7	63.6	29.4	66.2	67.8	65.1	60.0	43.3	37.5	63.6	49.8	66.7	37.0	62.4	59.1	62.6	64.0	46.9	50.9	42.6
`Centurio Qwen`	72.7	56.2	65.3	47.4	62.2	56.7	67.0	53.6	48.8	36.7	65.4	54.1	67.6	39.1	63.7	63.6	60.4	58.5	45.2	63.4	49.5

Table 18: BIN-MC

F.2 Comparison with `Centurio`

BIN-MC Table 32

M3Exam Table 33

VGR Table 34

VLOD Table 35

MaRVL Table 36

MaXM Table 37

MTVQA Table 38

xGQA Table 39

XM3600 Table 40. Language fidelity Table 41.

XVNLI Table 42

xMMMU Table 43

CVQA Table 44

SMPQA - Ground Table 45

SMPQA - Name Table 46

	en	avg.	af	zh	it	pt	th	vi
Phi 3.5 - English	52.9	32.7	32.5	37.0	49.6	39.7	25.4	12.2
Phi 3.5 - T5 50	51.2	35.3	39.9	35.9	46.4	39.7	28.2	21.7
Phi 3.5 - T5-4 50	52.2	34.2	40.5	32.4	49.1	38.6	25.2	19.1
Phi 3.5 - T5-3 50	51.3	35.3	43.6	34.0	47.4	37.3	27.9	21.7
Phi 3.5 - T5-2 50	49.2	33.7	39.3	32.9	45.1	38.4	22.2	24.3
Phi 3.5 - L100 50	50.8	36.0	39.3	36.1	50.9	40.1	26.2	23.5
Llama 3 - English	46.1	32.5	38.6	32.6	41.6	35.0	25.9	20.9
Llama 3 - T5 50	45.0	33.8	40.5	34.3	41.9	34.1	25.7	26.1
Llama 3 - L100 50	46.6	34.2	44.2	31.0	42.4	34.6	27.2	26.1
Phi 3.5 - L100 1	50.3	35.1	39.9	35.4	46.6	39.2	23.9	25.2
Phi 3.5 - L100 10	48.8	33.9	35.0	33.6	48.1	36.1	24.7	26.1
Phi 3.5 - L100 24	50.8	36.5	41.7	37.0	51.6	35.9	27.7	25.2
Phi 3.5 - L100 50	50.8	36.0	39.3	36.1	50.9	40.1	26.2	23.5
Phi 3.5 - L100 75	48.0	36.1	44.2	35.9	47.1	38.4	26.7	24.3
Phi 3.5 - L100 90	51.7	35.1	36.8	38.0	48.1	36.8	26.4	24.3
Llama 3 - L100 10	43.7	33.6	41.7	29.4	44.9	35.3	23.7	27.0
Llama 3 - L100 50	46.6	34.2	44.2	31.0	42.4	34.6	27.2	26.1
Llama 3 - L100 90	43.3	34.6	37.4	32.2	44.9	35.3	30.2	27.8
Phi 3.5 - L100 50	50.8	36.0	39.3	36.1	50.9	40.1	26.2	23.5
Phi 3.5 - PT 100	50.3	35.8	41.7	37.5	49.4	36.6	24.2	25.2
Phi 3.5 - PT 50	49.7	33.1	41.1	36.1	44.4	35.0	21.7	20.0
Phi 3.5 - PT 1	48.4	33.8	41.7	35.9	46.4	34.8	23.2	20.9
Llama 3 - L100 50	46.6	34.2	44.2	31.0	42.4	34.6	27.2	26.1
Llama 3 - PT 1	50.2	37.9	44.8	34.7	48.1	40.6	31.4	27.8
Llama 3 - PT 100	52.9	37.1	50.3	33.8	46.6	37.5	30.2	24.3
Gemma 2 - L100 50	42.5	33.4	43.6	33.6	41.6	30.4	27.7	23.5
Llama 3 - L100 50	46.6	34.2	44.2	31.0	42.4	34.6	27.2	26.1
Qwen 2.5 - L100 50	53.6	39.6	46.0	44.7	50.6	42.4	29.7	24.3
Aya-Expanse - L100 50	49.3	36.5	46.6	36.8	51.9	39.0	26.2	18.3
`Centurio` Aya	53.0	41.2	52.8	40.3	51.4	47.7	27.4	27.8
`Centurio` Qwen	61.2	46.9	50.9	64.1	55.6	49.0	31.9	29.6

Table 19: M3Exam

	en	avg.	am	ber	bn	de	fil	ha	hi	ru	sw	th	zu
Phi 3.5 - English	80.8	54.1	45.0	50.8	41.5	71.7	55.8	41.7	62.7	85.0	35.8	68.3	36.2
Phi 3.5 - T5 50	75.8	50.9	49.2	49.2	40.7	72.5	55.0	42.5	54.2	60.8	37.5	60.8	37.9
Phi 3.5 - T5-4 50	83.3	55.1	51.7	43.3	49.2	70.8	65.8	42.5	61.9	70.8	38.3	75.0	36.2
Phi 3.5 - T5-3 50	83.3	56.6	43.3	50.8	50.8	74.2	69.2	42.5	57.6	76.7	43.3	71.7	42.2
Phi 3.5 - T5-2 50	81.7	57.5	45.8	52.5	44.1	73.3	64.2	39.2	59.3	73.3	60.0	60.8	59.5
Phi 3.5 - L100 50	76.7	56.4	46.7	46.7	54.2	71.7	60.0	45.0	57.6	70.8	57.5	65.8	44.0
Llama 3 - English	82.5	56.3	66.7	30.8	49.2	77.5	50.8	48.3	63.6	75.8	46.7	70.0	39.7
Llama 3 - T5 50	77.5	55.9	47.5	49.2	49.2	71.7	63.3	42.5	62.7	73.3	45.8	70.8	38.8
Llama 3 - L100 50	80.0	64.8	58.3	47.5	64.4	75.8	61.7	67.5	64.4	73.3	59.2	67.5	73.3
Phi 3.5 - L100 1	65.0	47.5	42.5	50.0	38.1	65.0	58.3	40.0	45.8	58.3	39.2	42.5	42.2
Phi 3.5 - L100 10	73.3	54.5	43.3	50.0	51.7	67.5	60.0	45.0	51.7	63.3	53.3	63.3	50.0
Phi 3.5 - L100 24	73.3	60.3	54.2	47.5	58.5	72.5	55.0	58.3	60.2	72.5	64.2	59.2	61.2
Phi 3.5 - L100 50	76.7	56.4	46.7	46.7	54.2	71.7	60.0	45.0	57.6	70.8	57.5	65.8	44.0
Phi 3.5 - L100 75	80.0	56.7	51.7	53.3	55.1	70.8	67.5	41.7	63.6	75.8	38.3	69.2	36.2
Phi 3.5 - L100 90	79.2	54.6	43.3	50.0	44.9	80.8	60.0	42.5	55.9	77.5	45.0	55.8	44.8
Llama 3 - L100 10	77.5	65.4	65.0	45.0	63.6	76.7	58.3	70.8	64.4	74.2	63.3	69.2	69.0
Llama 3 - L100 50	80.0	64.8	58.3	47.5	64.4	75.8	61.7	67.5	64.4	73.3	59.2	67.5	73.3
Llama 3 - L100 90	82.5	63.0	45.8	39.2	66.1	80.8	58.3	68.3	61.9	75.0	63.3	75.0	59.5
Phi 3.5 - L100 50	76.7	56.4	46.7	46.7	54.2	71.7	60.0	45.0	57.6	70.8	57.5	65.8	44.0
Phi 3.5 - PT 100	80.8	58.6	44.2	49.2	56.8	78.3	56.7	47.5	65.3	75.0	47.5	73.3	50.9
Phi 3.5 - PT 50	80.0	63.2	58.3	50.0	55.1	78.3	63.3	60.0	61.9	76.7	55.0	75.0	61.2
Phi 3.5 - PT 1	80.0	62.0	55.8	50.0	51.7	81.7	62.5	60.0	66.1	75.0	50.0	66.7	62.1
Llama 3 - L100 50	80.0	64.8	58.3	47.5	64.4	75.8	61.7	67.5	64.4	73.3	59.2	67.5	73.3
Llama 3 - PT 1	87.5	71.2	70.0	50.8	65.3	79.2	63.3	83.3	68.6	82.5	66.7	85.8	68.1
Llama 3 - PT 100	85.0	68.8	65.8	49.2	67.8	80.8	61.7	70.0	66.9	85.0	70.0	74.2	65.5
Gemma 2 - L100 50	77.5	61.8	64.2	52.5	48.3	70.8	51.7	64.2	58.5	71.7	54.2	70.8	73.3
Llama 3 - L100 50	80.0	64.8	58.3	47.5	64.4	75.8	61.7	67.5	64.4	73.3	59.2	67.5	73.3
Qwen 2.5 - L100 50	91.7	71.2	76.7	50.0	69.5	81.7	77.5	57.5	72.9	83.3	71.7	80.8	62.1
Aya-Expanse - L100 50	92.5	69.9	52.5	54.2	55.9	80.8	85.0	72.5	79.7	83.3	63.3	78.3	63.8
`Centurio` Aya	82.5	66.8	71.7	54.2	59.3	73.3	59.2	65.0	71.2	75.8	67.5	72.5	65.5
`Centurio` Qwen	87.5	73.1	77.5	49.2	62.7	80.8	78.3	76.7	72.9	85.0	70.0	81.7	69.0

Table 20: VGR

	en	avg.	am	ber	bn	de	fil	ha	hi	ru	sw	th	zu
Phi 3.5 - English	16.7	21.3	20.8	20.8	19.2	16.7	25.8	28.3	17.0	12.5	25.0	26.7	22.0
Phi 3.5 - T5 50	23.3	20.0	15.0	18.3	20.8	21.7	16.7	20.0	23.2	27.5	22.3	15.8	18.6
Phi 3.5 - T5-4 50	17.5	18.2	19.2	20.8	13.3	20.8	17.5	16.7	21.4	26.7	16.1	10.0	17.8
Phi 3.5 - T5-3 50	25.8	19.8	16.7	17.5	21.7	21.7	20.0	21.7	23.2	20.8	18.8	16.7	18.6
Phi 3.5 - T5-2 50	21.7	20.5	21.7	18.3	16.7	22.5	27.5	27.5	17.9	21.7	17.0	13.3	21.2
Phi 3.5 - L100 50	18.3	19.5	16.7	20.8	19.2	25.8	20.0	16.7	25.0	20.8	13.4	17.5	18.6
Llama 3 - English	12.5	20.8	18.3	21.7	20.0	10.8	24.2	29.2	15.2	12.5	28.6	29.2	19.5
Llama 3 - T5 50	20.8	20.1	18.3	19.2	17.5	16.7	25.0	21.7	24.1	15.0	19.6	23.3	20.3
Llama 3 - L100 50	12.5	20.6	19.2	20.8	20.0	10.8	24.2	30.0	15.2	10.8	28.6	27.5	19.5
Phi 3.5 - L100 1	24.2	19.3	15.0	21.7	17.5	20.0	29.2	22.5	17.9	14.2	16.1	22.5	16.1
Phi 3.5 - L100 10	23.3	19.2	23.3	15.0	16.7	21.7	20.8	20.8	20.5	24.2	10.7	15.8	22.0
Phi 3.5 - L100 24	25.0	18.3	20.8	18.3	16.7	20.8	16.7	20.8	17.9	21.7	14.3	16.7	16.9
Phi 3.5 - L100 50	18.3	19.5	16.7	20.8	19.2	25.8	20.0	16.7	25.0	20.8	13.4	17.5	18.6
Phi 3.5 - L100 75	16.7	18.0	15.0	20.0	19.2	19.2	16.7	23.3	17.0	13.3	17.9	15.8	20.3
Phi 3.5 - L100 90	22.5	19.0	20.0	16.7	15.8	20.0	16.7	23.3	21.4	23.3	16.1	15.8	19.5
Llama 3 - L100 10	13.3	20.4	18.3	21.7	19.2	10.8	23.3	26.7	17.9	10.0	28.6	28.3	19.5
Llama 3 - L100 50	12.5	20.6	19.2	20.8	20.0	10.8	24.2	30.0	15.2	10.8	28.6	27.5	19.5
Llama 3 - L100 90	12.5	19.9	18.3	21.7	15.0	10.8	22.5	28.3	15.2	10.8	28.6	28.3	19.5
Phi 3.5 - L100 50	18.3	19.5	16.7	20.8	19.2	25.8	20.0	16.7	25.0	20.8	13.4	17.5	18.6
Phi 3.5 - PT 100	23.3	20.0	16.7	16.7	24.2	20.0	25.0	21.7	19.6	15.0	20.5	20.0	20.3
Phi 3.5 - PT 50	20.0	18.6	18.3	17.5	15.0	15.8	14.2	21.7	17.9	23.3	20.5	20.8	19.5
Phi 3.5 - PT 1	25.0	19.4	21.7	22.5	19.2	22.5	16.7	15.8	20.5	21.7	16.1	15.0	22.0
Llama 3 - L100 50	12.5	20.6	19.2	20.8	20.0	10.8	24.2	30.0	15.2	10.8	28.6	27.5	19.5
Llama 3 - PT 1	19.2	20.5	15.8	19.2	22.5	15.0	23.3	23.3	17.9	13.3	25.9	28.3	21.2
Llama 3 - PT 100	13.3	20.8	18.3	21.7	20.0	12.5	23.3	29.2	17.0	10.8	28.6	28.3	19.5
Gemma 2 - L100 50	14.2	21.1	18.3	22.5	20.8	10.8	25.0	28.3	16.1	11.7	27.7	30.0	20.3
Llama 3 - L100 50	12.5	20.6	19.2	20.8	20.0	10.8	24.2	30.0	15.2	10.8	28.6	27.5	19.5
Qwen 2.5 - L100 50	26.7	27.3	25.0	21.7	26.7	27.5	27.5	25.0	29.5	25.0	29.5	40.0	22.9
Aya-Expanse - L100 50	12.5	20.7	18.3	21.7	20.0	10.8	24.2	29.2	15.2	10.8	28.6	29.2	19.5
`Centurio` Aya	12.5	20.7	18.3	21.7	20.0	11.7	24.2	29.2	15.2	10.8	28.6	29.2	19.5
`Centurio` Qwen	28.3	27.0	18.3	20.0	33.3	32.5	29.2	22.5	25.0	22.5	30.4	30.0	33.1

Table 21: VLOD

	en	avg.	id	sw	ta	tr	zh
Phi 3.5 - English	82.1	61.4	65.6	50.8	53.3	63.8	73.2
Phi 3.5 - T5 50	81.5	61.8	66.4	53.4	53.7	61.6	73.8
Phi 3.5 - T5-4 50	81.2	64.3	68.7	52.3	54.3	70.2	76.2
Phi 3.5 - T5-3 50	81.5	65.9	70.8	56.4	56.7	68.9	76.7
Phi 3.5 - T5-2 50	79.7	66.4	70.2	62.2	57.5	66.7	75.4
Phi 3.5 - L100 50	79.6	64.4	69.0	59.0	53.6	67.5	73.0
Llama 3 - English	85.2	65.0	68.8	52.5	54.3	69.7	79.8
Llama 3 - T5 50	84.5	67.1	73.8	55.7	53.6	72.7	79.6
Llama 3 - L100 50	83.7	74.2	75.3	71.4	68.4	79.8	76.0
Phi 3.5 - L100 1	71.9	61.4	65.1	56.1	54.3	65.2	66.1
Phi 3.5 - L100 10	74.1	63.4	66.8	58.1	57.2	65.1	70.0
Phi 3.5 - L100 24	76.0	61.6	63.4	57.6	56.9	64.0	66.3
Phi 3.5 - L100 50	79.6	64.4	69.0	59.0	53.6	67.5	73.0
Phi 3.5 - L100 75	81.7	64.7	71.3	54.4	56.1	64.8	77.0
Phi 3.5 - L100 90	83.1	64.3	70.7	56.3	53.8	62.8	77.8
Llama 3 - L100 10	80.0	72.9	71.9	70.8	71.7	75.7	74.2
Llama 3 - L100 50	83.7	74.2	75.3	71.4	68.4	79.8	76.0
Llama 3 - L100 90	85.1	71.1	73.4	63.7	65.1	75.7	77.6
Phi 3.5 - L100 50	79.6	64.4	69.0	59.0	53.6	67.5	73.0
Phi 3.5 - PT 100	82.0	65.6	68.6	59.4	57.9	67.6	74.5
Phi 3.5 - PT 50	82.5	69.9	75.2	64.0	64.1	71.1	74.9
Phi 3.5 - PT 1	81.9	67.9	74.0	64.0	60.2	68.0	73.4
Llama 3 - L100 50	83.7	74.2	75.3	71.4	68.4	79.8	76.0
Llama 3 - PT 1	87.5	80.4	82.5	75.5	77.1	84.5	82.3
Llama 3 - PT 100	86.5	78.9	81.3	73.0	75.1	83.4	81.5
Gemma 2 - L100 50	82.5	73.0	72.6	71.4	68.3	76.4	76.2
Llama 3 - L100 50	83.7	74.2	75.3	71.4	68.4	79.8	76.0
Qwen 2.5 - L100 50	89.6	79.4	84.8	73.9	65.2	86.6	86.6
Aya-Expanse - L100 50	87.0	80.2	83.9	75.6	71.7	86.9	83.0
`Centurio` Aya	85.0	77.9	79.5	70.9	73.4	83.4	82.4
`Centurio` Qwen	89.6	81.7	85.0	76.8	76.0	84.2	86.7

Table 22: MaRVL

	en	avg.	fr	hi	he	ro	th	zh
Phi 3.5 - English	53.0	9.2	14.3	11.9	7.9	7.2	7.0	7.2
Phi 3.5 - T5 50	51.3	25.6	41.0	30.6	17.5	15.6	27.5	21.5
Phi 3.5 - T5-4 50	51.0	33.1	45.4	50.7	27.0	23.7	32.5	19.5
Phi 3.5 - T5-3 50	53.7	36.7	41.0	45.9	33.0	36.6	40.4	23.5
Phi 3.5 - T5-2 50	53.4	35.9	42.3	48.0	33.3	35.1	32.8	23.8
Phi 3.5 - L100 50	54.4	36.6	43.0	48.0	30.8	35.1	39.1	23.5
Llama 3 - English	55.4	7.7	9.2	10.9	6.7	4.5	8.3	6.8
Llama 3 - T5 50	41.3	20.2	45.1	12.6	2.9	24.3	14.6	21.8
Llama 3 - L100 50	52.7	42.3	42.3	54.4	40.6	40.5	52.6	23.1
Phi 3.5 - L100 1	48.0	33.8	39.9	45.2	32.4	32.4	32.8	19.9
Phi 3.5 - L100 10	52.0	35.4	44.7	45.6	34.6	36.0	29.5	22.1
Phi 3.5 - L100 24	50.7	35.1	44.0	44.6	29.8	33.0	38.1	21.2
Phi 3.5 - L100 50	54.4	36.6	43.0	48.0	30.8	35.1	39.1	23.5
Phi 3.5 - L100 75	51.0	32.5	42.0	36.4	29.8	33.3	31.8	21.8
Phi 3.5 - L100 90	54.7	29.7	41.6	28.2	27.3	28.5	30.5	21.8
Llama 3 - L100 10	49.0	41.9	37.9	53.4	45.7	41.4	51.0	21.8
Llama 3 - L100 50	52.7	42.3	42.3	54.4	40.6	40.5	52.6	23.1
Llama 3 - L100 90	52.7	40.6	43.3	52.7	36.2	40.2	49.0	22.1
Phi 3.5 - L100 50	54.4	36.6	43.0	48.0	30.8	35.1	39.1	23.5
Phi 3.5 - PT 100	54.0	36.2	44.0	48.6	32.4	33.9	36.8	21.5
Phi 3.5 - PT 50	53.4	39.0	45.7	49.3	39.4	36.6	40.7	22.1
Phi 3.5 - PT 1	55.7	39.7	44.7	52.0	41.0	40.8	40.1	19.9
Llama 3 - L100 50	52.7	42.3	42.3	54.4	40.6	40.5	52.6	23.1
Llama 3 - PT 1	55.0	48.5	47.4	57.1	56.2	47.4	57.3	25.7
Llama 3 - PT 100	58.1	47.4	44.7	54.8	54.0	47.1	57.3	26.4
Gemma 2 - L100 50	51.7	41.5	39.6	52.4	44.1	39.3	48.7	24.8
Llama 3 - L100 50	52.7	42.3	42.3	54.4	40.6	40.5	52.6	23.1
Qwen 2.5 - L100 50	58.7	45.8	46.4	51.4	50.2	41.7	57.9	27.0
Aya-Expanse - L100 50	53.4	47.2	46.4	58.8	59.4	49.9	41.4	27.4
`Centurio` Aya	55.7	49.3	45.1	62.9	58.7	51.1	46.7	31.6
`Centurio` Qwen	60.1	47.7	47.1	56.8	45.1	47.7	57.0	32.2

Table 23: MaXM

	avg.	ar	de	fr	it	ja	ko	ru	th	vi
Phi 3.5 - English	3.2	0.9	6.5	9.3	8.1	0.8	0.7	1.6	0.0	1.1
Phi 3.5 - T5 50	5.7	1.7	12.0	15.9	10.1	2.4	3.8	2.6	0.9	1.8
Phi 3.5 - T5-4 50	5.9	2.7	14.0	15.1	9.6	3.5	3.8	1.9	0.9	1.6
Phi 3.5 - T5-3 50	5.8	2.0	13.5	14.6	9.4	3.9	3.8	2.4	0.9	2.0
Phi 3.5 - T5-2 50	6.6	5.3	15.9	15.1	9.4	4.1	3.8	2.5	0.4	2.7
Phi 3.5 - L100 50	6.3	2.8	15.8	16.8	8.9	3.9	2.7	2.8	0.4	2.9
Llama 3 - English	3.2	0.3	6.9	8.0	8.7	0.7	0.5	0.7	0.4	2.7
Llama 3 - T5 50	5.6	2.0	14.2	15.0	9.1	1.9	1.4	2.6	1.3	2.8
Llama 3 - L100 50	6.0	2.1	11.9	15.8	7.2	2.1	3.2	2.4	4.8	4.1
Phi 3.5 - L100 1	4.7	2.0	12.0	9.4	7.5	3.4	3.4	1.9	0.9	2.3
Phi 3.5 - L100 10	5.7	3.0	12.1	14.2	8.6	4.6	4.1	2.1	0.9	1.5
Phi 3.5 - L100 24	6.2	3.6	14.0	15.8	8.7	3.1	3.8	3.3	0.9	2.5
Phi 3.5 - L100 50	6.3	2.8	15.8	16.8	8.9	3.9	2.7	2.8	0.4	2.9
Phi 3.5 - L100 75	6.3	2.6	13.8	18.3	8.7	4.3	2.9	2.8	0.9	2.8
Phi 3.5 - L100 90	7.0	2.6	14.7	19.3	10.4	3.6	4.1	3.2	3.5	1.5
Llama 3 - L100 10	5.3	1.6	11.3	13.8	7.5	2.9	3.4	2.6	0.9	3.5
Llama 3 - L100 50	6.0	2.1	11.9	15.8	7.2	2.1	3.2	2.4	4.8	4.1
Llama 3 - L100 90	6.5	2.1	14.0	17.8	9.7	2.5	3.8	2.8	2.2	3.5
Phi 3.5 - L100 50	6.3	2.8	15.8	16.8	8.9	3.9	2.7	2.8	0.4	2.9
Phi 3.5 - PT 100	6.9	3.7	16.0	15.9	11.3	3.4	3.2	2.9	2.2	3.5
Phi 3.5 - PT 50	6.1	1.8	14.8	15.8	10.5	3.5	2.9	2.6	0.9	2.1
Phi 3.5 - PT 1	6.2	1.6	14.9	15.9	11.1	3.7	3.0	1.7	0.9	2.7
Llama 3 - L100 50	6.0	2.1	11.9	15.8	7.2	2.1	3.2	2.4	4.8	4.1
Llama 3 - PT 1	6.9	2.4	17.1	16.6	9.1	3.4	4.5	2.5	1.7	5.2
Llama 3 - PT 100	8.3	2.6	18.7	19.6	11.4	4.0	4.3	4.0	4.8	5.3
Gemma 2 - L100 50	4.3	1.7	11.1	8.1	7.1	3.0	2.3	2.1	1.7	1.7
Llama 3 - L100 50	6.0	2.1	11.9	15.8	7.2	2.1	3.2	2.4	4.8	4.1
Qwen 2.5 - L100 50	6.4	5.5	12.0	13.0	10.3	3.0	3.2	2.9	2.2	5.2
Aya-Expanse - L100 50	6.2	3.7	13.2	13.9	9.5	3.0	3.4	3.4	1.7	3.6
`Centurio` Aya	11.1	6.7	19.9	22.5	16.7	5.0	9.0	5.2	5.2	9.7
`Centurio` Qwen	11.9	4.6	22.7	26.5	18.6	5.9	9.9	5.0	5.2	8.9

Table 24: MTVQA

	en	avg.	bn	de	id	ko	pt	ru	zh
Phi 3.5 - English	59.7	37.2	4.9	47.8	33.2	38.2	47.1	42.1	47.2
Phi 3.5 - T5 50	54.1	34.1	2.6	44.6	34.3	36.3	43.8	36.4	41.0
Phi 3.5 - T5-4 50	52.0	37.4	5.7	45.6	38.7	40.4	45.2	43.4	42.7
Phi 3.5 - T5-3 50	54.8	40.6	22.7	46.5	42.1	39.8	46.0	43.6	43.6
Phi 3.5 - T5-2 50	57.8	45.3	27.4	50.3	46.0	46.4	48.6	49.5	48.9
Phi 3.5 - L100 50	56.6	45.1	27.0	51.4	44.8	44.9	50.8	48.2	48.7
Llama 3 - English	61.9	39.2	13.2	49.0	35.6	39.1	44.9	44.1	48.4
Llama 3 - T5 50	49.3	33.8	5.9	43.8	38.0	32.4	41.7	37.3	37.4
Llama 3 - L100 50	60.6	51.0	46.7	54.1	51.2	49.4	53.4	51.2	51.3
Phi 3.5 - L100 1	48.4	40.3	28.2	43.9	41.1	40.6	43.0	42.8	42.8
Phi 3.5 - L100 10	51.8	42.2	27.6	46.3	43.0	42.2	45.7	44.8	45.6
Phi 3.5 - L100 24	53.8	42.9	29.1	47.6	43.4	42.2	46.6	45.9	45.4
Phi 3.5 - L100 50	56.6	45.1	27.0	51.4	44.8	44.9	50.8	48.2	48.7
Phi 3.5 - L100 75	58.6	45.8	26.4	52.4	44.4	45.4	51.9	49.9	50.0
Phi 3.5 - L100 90	58.5	42.1	14.2	53.0	39.8	43.3	51.4	45.7	47.7
Llama 3 - L100 10	54.9	45.0	40.5	46.4	45.7	42.5	46.5	46.2	47.4
Llama 3 - L100 50	60.6	51.0	46.7	54.1	51.2	49.4	53.4	51.2	51.3
Llama 3 - L100 90	61.9	51.4	42.5	56.2	51.7	50.1	54.6	52.5	52.1
Phi 3.5 - L100 50	56.6	45.1	27.0	51.4	44.8	44.9	50.8	48.2	48.7
Phi 3.5 - PT 100	58.0	46.1	29.5	52.8	46.1	44.5	51.7	49.5	48.3
Phi 3.5 - PT 50	58.3	47.6	35.4	52.8	48.7	45.5	52.5	49.6	48.6
Phi 3.5 - PT 1	58.3	47.0	37.6	52.6	46.8	44.1	51.5	48.1	48.1
Llama 3 - L100 50	60.6	51.0	46.7	54.1	51.2	49.4	53.4	51.2	51.3
Llama 3 - PT 1	61.1	55.1	52.8	56.6	56.0	53.9	56.0	55.4	55.0
Llama 3 - PT 100	61.6	53.0	50.4	54.9	53.6	52.4	53.0	53.1	53.4
Gemma 2 - L100 50	56.5	47.5	43.9	51.6	47.6	44.2	50.1	47.5	47.5
Llama 3 - L100 50	60.6	51.0	46.7	54.1	51.2	49.4	53.4	51.2	51.3
Qwen 2.5 - L100 50	60.3	51.9	44.2	54.8	53.1	51.3	54.3	53.2	52.8
Aya-Expanse - L100 50	60.5	52.5	45.2	54.6	53.8	51.7	54.7	53.9	53.4
`Centurio` Aya	59.1	53.2	43.4	56.9	54.4	53.6	56.2	54.0	54.3
`Centurio` Qwen	60.6	54.8	49.9	57.0	54.9	53.5	57.2	55.8	55.6

Table 25: xGQA

	en	avg.	ar	bn	cs	da	de	el	es	fa	fi	fil	fr	he	hi	hr	hu	id
Phi 3.5 - English	33.6	1.2	0.0	0.0	0.7	1.1	1.7	0.0	10.5	0.0	0.4	1.6	4.4	0.0	0.0	0.5	0.6	1.5
Phi 3.5 - T5 50	33.0	9.5	7.8	0.6	3.9	8.2	24.7	0.6	34.4	0.4	1.8	1.9	39.1	3.4	3.5	2.6	4.0	7.9
Phi 3.5 - T5-4 50	25.2	11.8	6.9	1.0	13.8	10.8	24.4	1.4	27.3	8.6	7.0	3.6	31.3	3.5	9.7	9.2	10.5	8.4
Phi 3.5 - T5-3 50	32.7	13.6	6.5	5.9	13.8	18.2	25.4	7.0	31.0	7.1	5.7	11.9	30.7	7.6	7.2	9.4	8.2	22.9
Phi 3.5 - T5-2 50	29.9	11.1	4.5	5.5	10.7	12.4	20.4	6.4	20.4	5.7	5.5	10.1	24.8	7.4	6.7	8.0	7.1	17.5
Phi 3.5 - L100 50	31.0	13.2	5.6	3.3	10.9	18.0	26.4	4.5	30.9	4.1	4.4	11.1	38.5	6.7	6.3	7.3	7.8	22.4
Llama 3 - English	75.6	1.1	0.1	0.0	1.3	2.3	1.6	0.1	2.1	0.1	0.8	2.9	3.4	0.0	0.0	0.7	1.3	2.1
Llama 3 - T5 50	76.1	12.6	27.9	0.5	1.6	20.7	29.2	0.6	61.6	0.4	1.1	2.9	58.2	0.0	0.4	0.9	1.6	19.6
Llama 3 - L100 50	72.6	28.5	25.6	14.0	30.5	38.1	27.6	23.3	56.0	24.3	13.8	29.0	50.9	15.5	22.5	19.7	18.0	39.9
Phi 3.5 - L100 1	43.3	13.3	5.2	4.3	11.1	15.3	24.5	5.3	25.4	5.4	6.1	13.3	37.1	7.0	7.2	7.9	6.7	22.6
Phi 3.5 - L100 10	38.9	12.7	4.7	4.1	11.4	12.5	23.6	5.9	28.1	4.8	5.3	14.2	33.3	8.5	8.6	8.2	6.3	22.6
Phi 3.5 - L100 24	31.5	13.2	5.2	5.0	11.6	15.1	22.6	6.9	30.9	4.9	6.1	12.0	39.2	8.2	7.7	8.2	5.1	22.4
Phi 3.5 - L100 50	31.0	13.2	5.6	3.3	10.9	18.0	26.4	4.5	30.9	4.1	4.4	11.1	38.5	6.7	6.3	7.3	7.8	22.4
Phi 3.5 - L100 75	36.5	12.0	4.4	2.5	9.1	13.6	25.0	3.0	25.7	3.4	3.8	7.1	33.6	6.3	5.2	5.9	7.0	20.2
Phi 3.5 - L100 90	34.2	9.4	4.0	1.9	6.7	9.2	21.8	2.8	23.4	2.0	3.8	4.1	26.2	4.4	3.7	4.7	4.9	12.5
Llama 3 - L100 10	74.8	28.9	23.0	11.9	25.8	43.6	26.0	24.6	53.7	24.9	16.0	30.2	52.6	17.1	20.1	20.5	18.5	43.3
Llama 3 - L100 50	72.6	28.5	25.6	14.0	30.5	38.1	27.6	23.3	56.0	24.3	13.8	29.0	50.9	15.5	22.5	19.7	18.0	39.9
Llama 3 - L100 90	73.6	23.0	18.2	7.8	24.1	36.7	23.0	19.5	54.2	17.6	10.5	24.0	51.9	9.7	20.2	15.4	15.3	33.0
Phi 3.5 - L100 50	31.0	13.2	5.6	3.3	10.9	18.0	26.4	4.5	30.9	4.1	4.4	11.1	38.5	6.7	6.3	7.3	7.8	22.4
Phi 3.5 - PT 100	35.9	13.5	5.3	5.0	13.9	15.7	26.5	5.9	29.6	5.4	4.1	9.1	33.5	8.3	6.9	8.8	7.1	22.3
Phi 3.5 - PT 50	37.1	17.3	7.7	9.0	16.5	21.2	27.8	9.3	38.0	8.2	7.0	15.2	42.4	10.9	10.4	11.8	9.7	28.5
Phi 3.5 - PT 1	33.1	17.4	6.3	9.3	17.2	22.1	26.9	8.2	37.5	9.1	7.2	13.9	40.6	12.2	9.1	11.5	11.1	28.9
Llama 3 - L100 50	72.6	28.5	25.6	14.0	30.5	38.1	27.6	23.3	56.0	24.3	13.8	29.0	50.9	15.5	22.5	19.7	18.0	39.9
Llama 3 - PT 1	80.8	35.3	30.6	15.4	35.5	51.3	34.0	28.2	65.4	32.3	17.9	36.3	62.5	24.6	27.4	26.9	24.7	49.2
Llama 3 - PT 100	77.9	31.8	26.1	14.4	35.4	43.5	33.4	27.0	60.7	23.0	14.6	31.7	58.9	18.2	24.6	22.2	22.6	45.2
Gemma 2 - L100 50	66.6	27.5	24.5	17.9	28.6	35.1	26.0	18.2	54.7	29.4	13.7	26.8	54.3	22.8	21.6	17.7	20.1	43.8
Qwen 2.5 - L100 50	74.8	27.8	28.6	13.9	26.1	35.4	29.6	11.6	58.7	17.1	10.2	26.0	55.6	22.2	16.3	19.8	11.7	40.3
Aya-Expanse - L100 50	75.6	33.4	40.4	12.2	39.4	37.7	32.7	31.9	69.2	41.6	8.4	26.2	67.6	42.5	24.5	19.1	12.5	50.3
`Centurio` Aya	78.4	39.2	40.4	18.5	33.9	40.0	38.6	35.3	69.7	55.8	11.0	34.0	71.3	47.1	26.3	24.9	19.6	58.3
`Centurio` Qwen	79.1	34.4	36.6	17.1	29.7	43.1	32.0	19.2	69.2	31.2	12.0	33.6	67.6	27.6	20.3	22.0	18.7	50.4

	it	ja	ko	mi	nl	no	pl	pt	quz	ro	ru	sv	sw	te	th	tr	uk	vi	zh
Phi 3.5 - English	9.2	0.1	0.0	0.0	1.6	1.5	0.7	1.9	0.1	0.7	0.4	1.1	0.6	0.0	0.2	0.3	0.1	0.6	0.0
Phi 3.5 - T5 50	28.5	27.6	1.6	0.1	20.4	8.8	4.8	30.2	0.7	5.4	17.5	11.8	1.3	0.0	4.0	3.2	5.7	2.6	12.8
Phi 3.5 - T5-4 50	24.9	27.9	3.1	2.1	21.7	12.0	14.1	24.0	0.5	6.0	22.5	24.1	2.1	0.0	5.8	9.4	4.6	18.8	10.6
Phi 3.5 - T5-3 50	26.0	27.3	3.0	1.8	28.5	14.1	13.6	22.1	0.4	11.0	16.7	23.8	1.5	0.0	12.9	7.6	10.0	15.9	20.2
Phi 3.5 - T5-2 50	18.5	27.0	2.5	2.0	18.9	11.2	10.0	18.1	0.4	7.6	21.1	17.8	7.0	0.0	12.6	7.9	10.3	12.7	9.5
Phi 3.5 - L100 50	30.6	23.7	2.5	2.6	24.2	18.8	10.8	29.0	1.1	8.1	18.6	17.8	8.1	1.5	10.2	7.2	8.3	14.7	15.6
Llama 3 - English	1.5	0.2	0.0	0.2	3.0	1.8	1.1	2.6	0.8	1.2	0.7	2.0	0.8	0.0	0.4	0.5	0.3	0.5	0.3
Llama 3 - T5 50	26.8	35.8	0.1	0.2	34.9	15.7	11.3	12.5	1.3	5.4	12.8	18.3	0.6	0.0	4.4	6.1	0.2	10.1	17.2
Llama 3 - L100 50	44.1	33.8	13.4	24.9	50.1	41.5	26.9	45.1	1.3	22.6	23.5	42.7	28.9	11.0	27.7	21.8	21.9	49.7	16.9
Phi 3.5 - L100 1	28.1	25.4	2.1	4.0	22.3	17.7	12.0	24.9	1.4	10.7	21.4	17.8	11.1	1.4	12.6	7.6	8.2	17.2	16.9
Phi 3.5 - L100 10	23.3	25.2	2.4	3.0	23.0	15.6	12.0	25.8	0.7	8.2	20.0	15.9	9.9	1.8	11.8	6.7	11.9	14.7	11.8
Phi 3.5 - L100 24	24.7	25.8	3.3	4.8	24.8	18.5	13.6	17.7	0.7	9.1	17.8	17.5	9.3	2.3	13.3	6.5	9.6	15.4	14.4
Phi 3.5 - L100 50	30.6	23.7	2.5	2.6	24.2	18.8	10.8	29.0	1.1	8.1	18.6	17.8	8.1	1.5	10.2	7.2	8.3	14.7	15.6
Phi 3.5 - L100 75	29.7	24.8	2.0	3.3	23.0	17.1	9.8	27.8	0.8	6.2	19.6	15.2	5.0	1.4	10.9	5.7	8.9	13.0	18.5
Phi 3.5 - L100 90	21.3	21.3	2.0	1.2	12.7	11.8	7.3	22.5	0.8	5.9	16.3	16.1	5.6	0.6	7.5	3.9	7.8	8.8	20.3
Llama 3 - L100 10	40.3	35.0	13.9	29.4	53.4	41.9	25.6	44.8	1.6	19.8	25.3	44.0	30.3	13.8	28.8	22.1	21.1	47.4	20.2
Llama 3 - L100 50	44.1	33.8	13.4	24.9	50.1	41.5	26.9	45.1	1.3	22.6	23.5	42.7	28.9	11.0	27.7	21.8	21.9	49.7	16.9
Llama 3 - L100 90	38.0	25.6	10.4	17.5	46.1	33.1	19.8	41.2	0.2	17.1	20.6	38.1	14.6	5.8	23.2	16.3	0.3	43.9	13.3
Phi 3.5 - L100 50	30.6	23.7	2.5	2.6	24.2	18.8	10.8	29.0	1.1	8.1	18.6	17.8	8.1	1.5	10.2	7.2	8.3	14.7	15.6
Phi 3.5 - PT 100	30.3	25.7	2.7	3.9	21.6	20.1	12.0	21.8	0.9	9.5	19.5	18.9	8.5	1.4	13.6	7.5	8.5	14.9	23.9
Phi 3.5 - PT 50	33.9	26.2	3.2	7.2	30.0	24.7	16.1	29.1	2.5	14.7	21.3	24.1	15.3	4.3	18.6	8.0	9.8	19.4	22.3
Phi 3.5 - PT 1	34.8	30.5	2.9	7.9	27.7	26.4	14.9	31.2	2.3	14.4	22.4	23.8	14.7	4.4	18.4	10.8	10.5	18.8	21.1
Llama 3 - L100 50	44.1	33.8	13.4	24.9	50.1	41.5	26.9	45.1	1.3	22.6	23.5	42.7	28.9	11.0	27.7	21.8	21.9	49.7	16.9
Llama 3 - PT 1	51.6	38.7	15.2	35.9	59.1	49.2	32.5	51.1	2.9	30.7	32.3	51.8	38.0	17.4	36.2	29.3	26.4	58.1	16.5
Llama 3 - PT 100	49.2	35.9	14.0	26.5	55.3	45.8	32.3	51.6	0.9	25.2	30.7	47.7	29.3	12.4	32.0	27.0	25.2	56.0	15.2
Gemma 2 - L100 50	39.7	36.3	11.5	21.5	46.2	38.7	25.1	45.3	1.8	21.7	24.0	37.9	27.2	13.1	28.5	19.0	0.2	50.1	18.5
Qwen 2.5 - L100 50	43.8	39.0	13.4	21.8	48.9	36.6	25.0	52.3	0.9	16.9	33.6	36.7	18.9	8.0	38.1	17.4	18.1	59.5	19.9
Aya-Expanse - L100 50	53.6	39.5	18.4	20.1	59.6	36.3	35.5	50.6	0.7	31.3	31.9	34.9	21.3	9.2	19.0	29.8	27.7	68.2	26.1
`Centurio` Aya	60.4	49.1	21.3	33.7	61.7	42.5	37.9	59.3	1.7	34.6	38.0	45.9	29.9	15.1	26.0	30.6	30.6	72.7	56.9
`Centurio` Qwen	53.7	43.5	13.4	34.9	56.2	41.4	30.0	59.9	2.1	23.4	39.2	42.7	30.2	13.5	42.3	23.3	20.3	69.4	33.8

Table 26: XM3600

	ar	bn	cs	da	de	el	en	es	fa	fi	fil	fr	he	hi	hr	hu	id
Phi 3.5 - L100	93.8	99.0	99.0	91.8	100.0	88.3	100.0	100.0	98.8	100.0	96.7	100.0	99.6	97.7	76.2	100.0	95.3
Phi 3.5 - T5-2	94.7	100.0	99.4	98.4	100.0	99.8	100.0	100.0	100.0	100.0	99.6	100.0	100.0	98.8	79.1	100.0	88.5
Phi 3.5 - T5-3	92.2	99.8	99.2	97.9	100.0	99.8	100.0	100.0	100.0	100.0	99.6	100.0	99.8	99.0	77.9	100.0	91.8
Phi 3.5 - T5-4	96.5	96.1	99.2	78.3	100.0	98.4	100.0	100.0	100.0	100.0	99.6	100.0	99.4	98.6	89.8	100.0	98.0
Phi 3.5 - T5	98.2	97.3	78.7	87.5	100.0	50.2	100.0	99.8	2.3	76.4	49.4	100.0	96.7	99.2	73.6	99.2	95.1
Phi 3.5 - English	0.0	0.0	0.0	0.2	0.2	0.0	100.0	31.4	0.0	0.8	0.0	5.3	0.0	0.0	0.2	0.0	0.6

	it	ja	ko	mi	nl	no	pl	pt	quz	ro	ru	sv	sw	te	th	tr	uk	vi	zh
Phi 3.5 - L100	100.0	99.2	99.8	100.0	100.0	98.4	100.0	97.7	0.2	100.0	99.8	98.6	98.2	93.2	99.8	100.0	92.6	100.0	96.5
Phi 3.5 - T5-2	100.0	99.8	100.0	97.3	100.0	85.7	100.0	99.2	27.5	100.0	99.8	99.6	99.0	63.5	100.0	100.0	96.9	100.0	93.9
Phi 3.5 - T5-3	100.0	100.0	95.9	100.0	75.2	100.0	52.1	37.5	100.0	100.0	99.2	84.4	82.8	100.0	100.0	96.5	100.0	95.3	
Phi 3.5 - T5-4	100.0	100.0	100.0	95.7	100.0	71.5	100.0	100.0	12.7	100.0	100.0	100.0	84.6	67.2	99.0	100.0	30.9	100.0	94.1
Phi 3.5 - T5	100.0	100.0	99.8	1.6	99.8	92.0	96.5	100.0	0.2	96.3	100.0	98.6	36.1	62.5	90.2	94.9	91.6	39.8	96.7
Phi 3.5 - English	31.2	0.0	0.0	0.0	0.2	1.0	0.0	2.5	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.6	0.0	0.0	0.0

Table 27: XM3600 language fidelity (§1b)

	en	avg.	ar	es	fr	ru
Phi 3.5 - English	59.6	55.0	52.3	54.9	57.6	55.2
Phi 3.5 - T5 50	59.9	51.8	49.7	51.3	55.2	51.0
Phi 3.5 - T5-4 50	58.9	48.3	47.7	47.1	51.2	47.3
Phi 3.5 - T5-3 50	58.6	50.5	46.6	51.3	52.6	51.3
Phi 3.5 - T5-2 50	58.5	53.6	50.7	54.7	55.1	54.0
Phi 3.5 - L100 50	59.6	53.3	49.9	53.7	56.8	52.6
Llama 3 - English	46.1	36.3	33.4	37.0	36.5	38.2
Llama 3 - T5 50	45.4	37.5	36.6	36.1	38.7	38.7
Llama 3 - L100 50	59.7	54.8	53.0	54.7	56.3	55.4
Phi 3.5 - L100 1	55.2	48.2	42.2	50.6	51.1	48.8
Phi 3.5 - L100 10	58.3	53.4	50.5	54.3	55.7	53.1
Phi 3.5 - L100 24	58.2	48.4	43.5	50.3	52.7	47.3
Phi 3.5 - L100 50	59.6	53.3	49.9	53.7	56.8	52.6
Phi 3.5 - L100 75	61.9	54.5	50.3	56.2	57.5	54.2
Phi 3.5 - L100 90	60.0	50.5	43.6	53.8	55.1	49.5
Llama 3 - L100 10	60.8	55.3	56.3	52.6	55.0	57.1
Llama 3 - L100 50	59.7	54.8	53.0	54.7	56.3	55.4
Llama 3 - L100 90	57.8	51.1	48.9	52.3	52.1	51.0
Phi 3.5 - L100 50	59.6	53.3	49.9	53.7	56.8	52.6
Phi 3.5 - PT 100	54.3	45.4	40.1	47.8	49.4	44.4
Phi 3.5 - PT 50	58.9	52.5	49.0	53.8	54.6	52.5
Phi 3.5 - PT 1	56.8	49.7	46.8	49.6	53.9	48.6
Llama 3 - L100 50	59.7	54.8	53.0	54.7	56.3	55.4
Llama 3 - PT 1	61.7	59.4	58.8	59.0	60.0	59.7
Llama 3 - PT 100	60.3	57.3	56.5	56.5	58.3	57.8
Gemma 2 - L100 50	59.9	55.0	53.1	54.6	57.1	55.1
Llama 3 - L100 50	59.7	54.8	53.0	54.7	56.3	55.4
Qwen 2.5 - L100 50	57.8	52.6	55.7	47.5	52.5	54.8
Aya-Expanse - L100 50	58.2	54.7	54.7	54.0	56.4	53.5
`Centurio` Aya	65.0	62.4	61.7	61.0	64.3	62.7
`Centurio` Qwen	75.4	70.2	68.8	70.9	70.5	70.8

Table 28: XVNLI

	en	**avg.**	**ar**	**fr**	**hi**	**id**	**ja**	**pt**
Phi 3.5 - English	38.4	36.2	36.2	41.9	29.9	35.4	34.2	39.7
Phi 3.5 - T5 50	36.7	36.2	31.5	38.9	31.6	37.0	34.9	43.4
Phi 3.5 - T5-4 50	37.0	33.9	33.2	39.9	29.2	32.3	31.2	37.7
Phi 3.5 - T5-3 50	37.3	35.8	32.5	39.3	32.3	37.0	36.8	37.0
Phi 3.5 - T5-2 50	37.6	35.1	32.2	40.3	32.6	34.7	32.3	38.7
Phi 3.5 - L100 50	36.6	32.0	28.5	35.9	27.8	32.0	31.2	36.7
Llama 3 - English	33.2	32.4	30.9	34.2	30.6	32.7	30.5	35.7
Llama 3 - T5 50	33.4	32.4	34.9	36.6	28.9	31.3	30.9	31.6
Llama 3 - L100 50	33.0	31.7	31.5	34.6	34.0	31.6	27.9	30.6
Phi 3.5 - L100 1	37.3	34.1	32.5	40.3	30.9	31.3	31.6	37.7
Phi 3.5 - L100 10	36.1	30.9	27.5	33.9	28.2	28.6	32.7	34.7
Phi 3.5 - L100 24	34.4	31.9	28.5	35.9	29.2	30.3	33.5	34.0
Phi 3.5 - L100 50	36.6	32.0	28.5	35.9	27.8	32.0	31.2	36.7
Phi 3.5 - L100 75	36.2	33.2	31.9	38.9	29.2	32.7	29.0	37.4
Phi 3.5 - L100 90	37.1	31.9	30.5	35.6	25.8	31.0	33.8	34.7
Llama 3 - L100 10	32.6	30.0	26.8	31.5	26.8	31.6	32.0	31.3
Llama 3 - L100 50	33.0	31.7	31.5	34.6	34.0	31.6	27.9	30.6
Llama 3 - L100 90	32.7	33.5	30.5	35.9	30.9	35.4	31.2	37.0
Phi 3.5 - L100 50	36.6	32.0	28.5	35.9	27.8	32.0	31.2	36.7
Phi 3.5 - PT 100	33.4	30.2	28.5	32.9	28.9	30.0	27.5	33.3
Phi 3.5 - PT 50	35.0	33.4	30.9	39.3	33.7	31.0	30.5	35.0
Phi 3.5 - PT 1	36.0	31.3	26.5	35.9	29.2	32.0	28.3	36.0
Llama 3 - L100 50	33.0	31.7	31.5	34.6	34.0	31.6	27.9	30.6
Llama 3 - PT 1	38.6	35.2	33.9	34.2	34.0	35.0	36.1	38.0
Llama 3 - PT 100	36.9	36.1	34.6	36.2	36.8	36.7	36.1	36.0
Gemma 2 - L100 50	32.8	32.0	32.5	30.9	33.0	30.6	32.7	32.0
Llama 3 - L100 50	33.0	31.7	31.5	34.6	34.0	31.6	27.9	30.6
Qwen 2.5 - L100 50	39.8	39.7	38.6	40.3	34.4	40.7	38.7	45.5
Aya-Expanse - L100 50	36.8	35.4	34.9	35.2	37.5	36.4	34.6	33.7
`Centurio` Aya	37.6	37.2	36.2	38.9	38.8	39.7	34.2	35.4
`Centurio` Qwen	46.4	43.0	39.6	45.0	41.6	44.1	43.5	44.1

Table 29: xMMMU

	en	avg.	avg. Latin	avg. other	ar	de	hi	id	it	ko	ru	th	zh	zu
Phi 3.5 - English	65.8	55.8	62.3	51.5	50.2	63.5	58.5	61.4	64.0	49.0	52.1	49.1	49.8	60.2
Phi 3.5 - T5 50	75.2	60.2	70.9	53.1	50.2	70.8	65.4	71.8	71.6	49.8	54.1	51.0	48.0	69.4
Phi 3.5 - T5-4 50	74.2	60.8	71.4	53.7	52.2	71.5	65.5	72.8	73.1	51.1	53.9	49.6	49.6	68.4
Phi 3.5 - T5-3 50	70.4	58.7	67.7	52.8	51.6	66.9	61.0	69.6	67.2	50.0	53.6	48.9	51.4	67.0
Phi 3.5 - T5-2 50	68.4	56.2	64.2	50.8	49.5	64.5	58.4	65.4	64.9	50.0	50.5	48.8	47.9	62.0
Phi 3.5 - L100 50	69.6	58.0	67.2	51.9	49.9	68.0	62.4	69.0	67.9	48.6	52.5	49.6	48.4	64.1
Llama 3 - English	72.0	60.5	69.6	54.4	53.5	69.9	67.2	71.1	70.9	48.9	57.5	50.1	49.4	66.5
Llama 3 - T5 50	73.4	62.2	72.5	55.4	54.5	72.2	67.1	74.4	71.5	50.5	56.6	51.6	51.9	72.0
Llama 3 - L100 50	72.0	58.4	67.9	52.1	51.6	69.6	62.0	65.9	70.4	49.9	52.0	48.8	48.4	65.6
Phi 3.5 - L100 1	58.4	52.6	55.7	50.5	50.2	55.2	53.5	57.5	56.4	49.6	50.9	48.2	50.5	53.5
Phi 3.5 - L100 10	56.9	51.6	54.9	49.4	48.5	54.8	49.6	55.1	56.8	50.5	48.2	49.6	50.0	52.9
Phi 3.5 - L100 24	60.4	54.0	58.8	50.8	51.8	58.9	54.5	58.1	60.0	50.0	51.1	48.4	49.2	58.0
Phi 3.5 - L100 50	69.6	58.0	67.2	51.9	49.9	68.0	62.4	69.0	67.9	48.6	52.5	49.6	48.4	64.1
Phi 3.5 - L100 75	74.5	61.2	71.6	54.2	53.2	71.2	63.8	74.0	70.5	50.5	54.2	51.9	51.8	70.6
Phi 3.5 - L100 90	71.6	59.4	69.2	52.9	51.0	70.2	60.5	69.4	71.2	49.6	54.1	50.4	51.8	66.1
Llama 3 - L100 10	65.9	56.6	62.6	52.6	51.5	62.1	59.5	62.6	65.8	50.8	54.5	50.4	48.8	59.9
Llama 3 - L100 50	72.0	58.4	67.9	52.1	51.6	69.6	62.0	65.9	70.4	49.9	52.0	48.8	48.4	65.6
Llama 3 - L100 90	73.1	59.4	68.4	53.3	51.0	67.4	65.8	71.0	69.0	50.6	52.6	49.9	50.1	66.2
Phi 3.5 - L100 50	69.6	58.0	67.2	51.9	49.9	68.0	62.4	69.0	67.9	48.6	52.5	49.6	48.4	64.1
Phi 3.5 - PT 100	79.5	63.3	74.8	55.6	52.8	75.8	68.5	76.2	76.5	50.8	59.6	50.9	51.0	70.8
Phi 3.5 - PT 50	76.1	62.4	73.0	55.3	52.4	72.2	69.6	73.6	73.8	49.2	59.9	50.0	50.6	72.2
Phi 3.5 - PT 1	78.1	64.5	74.5	57.7	57.0	74.0	72.8	76.8	75.0	52.8	62.2	51.4	50.2	72.4
Llama 3 - L100 50	72.0	58.4	67.9	52.1	51.6	69.6	62.0	65.9	70.4	49.9	52.0	48.8	48.4	65.6
Llama 3 - PT 1	76.9	65.1	74.4	58.9	55.0	74.8	73.0	75.5	74.4	53.4	65.9	52.5	53.8	72.9
Llama 3 - PT 100	79.9	65.2	77.4	57.0	52.6	77.6	73.4	78.1	78.2	51.0	64.0	49.1	51.8	75.8
Phi 3.5 - OCR English	78.4	64.6	74.7	57.9	59.1	77.1	70.9	73.6	74.5	50.6	66.5	51.1	49.0	73.6
Phi 3.5 - OCR 50	81.2	66.7	76.7	60.0	61.4	78.6	72.1	76.0	77.1	51.5	71.5	52.1	51.6	75.0
Phi 3.5 - OCR 1	81.0	69.8	78.3	64.1	66.8	78.0	76.8	78.5	79.1	56.9	73.2	58.6	52.4	77.6
Phi 3.5 - OCR Latin-down	78.9	65.4	74.2	59.5	57.8	75.5	67.6	75.0	75.0	56.4	67.8	55.0	52.5	71.1
Phi 3.5 - OCR 50 (frozen)	76.1	62.1	70.8	56.3	59.2	73.2	63.2	66.2	76.1	50.0	68.0	47.8	49.8	67.8
Gemma 2 - L100 50	59.9	53.5	57.1	51.1	49.6	59.1	56.5	56.8	58.9	49.9	51.0	50.6	49.2	53.6
Llama 3 - L100 50	72.0	58.4	67.9	52.1	51.6	69.6	62.0	65.9	70.4	49.9	52.0	48.8	48.4	65.6
Qwen 2.5 - L100 50	82.8	62.5	75.1	54.0	51.5	76.4	66.5	76.5	76.5	50.1	55.2	51.0	49.8	71.1
Aya-Expanse - L100 50	79.1	63.5	75.2	55.7	53.9	77.2	71.4	75.6	75.0	50.6	56.0	51.1	51.0	73.1
`Centurio` Aya	83.1	74.2	80.9	69.7	75.9	82.1	80.1	81.4	80.6	68.8	73.5	66.5	53.4	79.5
`Centurio` Qwen	84.8	76.1	82.7	71.8	76.9	83.5	82.4	83.8	83.1	72.4	75.6	64.4	58.9	80.2

Table 30: SMPQA Ground

	en	avg.	avg. Latin	avg. other	ar	de	hi	id	it	ko	ru	th	zh	zu
Phi 3.5 - English	36.2	5.0	12.4	0.0	0.0	17.4	0.0	12.6	15.2	0.0	0.0	0.0	0.0	4.4
Phi 3.5 - T5 50	36.4	5.4	13.6	0.0	0.0	21.2	0.0	13.2	16.0	0.0	0.0	0.0	0.0	3.8
Phi 3.5 - T5-4 50	35.0	5.8	14.4	0.0	0.0	20.0	0.0	14.6	16.6	0.0	0.0	0.0	0.0	6.4
Phi 3.5 - T5-3 50	34.6	5.8	14.4	0.0	0.0	16.0	0.0	16.6	20.4	0.0	0.0	0.0	0.0	4.8
Phi 3.5 - T5-2 50	35.8	5.8	14.5	0.0	0.0	18.0	0.0	14.8	19.6	0.0	0.0	0.0	0.0	5.6
Phi 3.5 - L100 50	33.4	5.2	12.8	0.1	0.0	17.4	0.0	14.0	14.6	0.0	0.2	0.2	0.0	5.2
Llama 3 - English	41.0	8.5	21.1	0.0	0.0	24.4	0.0	21.6	23.8	0.0	0.0	0.2	0.0	14.8
Llama 3 - T5 50	41.4	8.2	20.4	0.0	0.0	25.2	0.0	21.8	23.4	0.0	0.0	0.2	0.0	11.2
Llama 3 - L100 50	39.2	7.3	18.2	0.0	0.0	21.6	0.0	18.8	21.6	0.0	0.0	0.2	0.0	10.8
Phi 3.5 - L100 1	22.0	4.0	10.1	0.0	0.0	12.0	0.0	9.0	14.0	0.0	0.0	0.0	0.0	5.2
Phi 3.5 - L100 10	24.6	4.1	10.3	0.0	0.0	11.6	0.0	10.0	14.2	0.0	0.0	0.0	0.0	5.4
Phi 3.5 - L100 24	26.0	3.8	9.5	0.1	0.0	12.2	0.0	8.4	12.6	0.0	0.0	0.4	0.0	4.8
Phi 3.5 - L100 50	33.4	5.2	12.8	0.1	0.0	17.4	0.0	14.0	14.6	0.0	0.2	0.2	0.0	5.2
Phi 3.5 - L100 75	38.4	6.0	15.1	0.0	0.0	21.0	0.0	14.8	18.6	0.0	0.2	0.0	0.0	5.8
Phi 3.5 - L100 90	39.8	6.5	16.1	0.0	0.0	21.0	0.0	17.0	21.8	0.0	0.0	0.0	0.0	4.8
Llama 3 - L100 10	32.0	6.3	15.6	0.1	0.0	17.8	0.0	15.8	19.2	0.0	0.0	0.4	0.0	9.6
Llama 3 - L100 50	39.2	7.3	18.2	0.0	0.0	21.6	0.0	18.8	21.6	0.0	0.0	0.2	0.0	10.8
Llama 3 - L100 90	40.0	7.5	18.8	0.0	0.0	21.2	0.0	21.0	20.4	0.0	0.0	0.2	0.0	12.6
Phi 3.5 - L100 50	33.4	5.2	12.8	0.1	0.0	17.4	0.0	14.0	14.6	0.0	0.2	0.2	0.0	5.2
Phi 3.5 - PT 100	44.0	9.9	24.5	0.2	0.0	31.4	0.0	25.6	26.8	0.0	1.2	0.2	0.0	14.0
Phi 3.5 - PT 50	41.8	9.4	23.1	0.2	0.0	27.8	0.0	24.4	25.0	0.0	1.2	0.2	0.0	15.0
Phi 3.5 - PT 1	42.2	9.5	23.7	0.1	0.0	27.2	0.0	24.4	29.0	0.0	0.4	0.0	0.0	14.0
Llama 3 - L100 50	39.2	7.3	18.2	0.0	0.0	21.6	0.0	18.8	21.6	0.0	0.0	0.2	0.0	10.8
Llama 3 - PT 1	48.4	11.4	27.9	0.4	0.0	29.6	0.2	30.6	30.6	0.0	1.6	0.4	0.0	20.6
Llama 3 - PT 100	48.8	10.5	25.0	0.8	0.0	28.8	2.6	26.2	28.4	0.2	1.8	0.4	0.0	16.6
Phi 3.5 - OCR English	55.8	18.3	39.9	3.9	5.2	38.6	2.4	43.2	41.6	0.0	15.2	0.4	0.0	36.4
Phi 3.5 - OCR 50	53.8	21.0	41.8	7.1	14.4	42.2	6.4	45.8	42.6	0.2	21.2	0.6	0.0	36.4
Phi 3.5 - OCR 1	54.8	22.2	43.5	8.0	17.2	43.8	6.2	46.4	42.8	1.2	21.4	1.8	0.0	40.8
Phi 3.5 - OCR Latin-down	54.6	22.4	41.0	9.9	20.2	41.6	7.0	42.6	43.0	2.8	25.6	3.4	0.6	36.8
Phi 3.5 - OCR 50 (frozen)	47.2	15.7	34.1	3.5	5.2	36.4	3.8	37.2	33.0	0.0	11.8	0.2	0.0	29.6
Gemma 2 - L100 50	28.6	3.8	9.4	0.1	0.0	13.8	0.0	10.4	8.4	0.0	0.0	0.4	0.0	5.0
Llama 3 - L100 50	39.2	7.3	18.2	0.0	0.0	21.6	0.0	18.8	21.6	0.0	0.0	0.2	0.0	10.8
Qwen 2.5 - L100 50	48.8	10.1	25.1	0.1	0.0	32.0	0.0	23.8	29.0	0.0	0.2	0.2	0.0	15.6
Aya-Expanse - L100 50	46.6	10.2	25.4	0.1	0.0	27.4	0.0	28.8	27.4	0.0	0.0	0.4	0.0	18.0
`Centurio` Aya	60.0	30.1	49.8	17.0	29.2	50.2	17.6	52.6	51.2	11.2	38.2	4.8	0.8	45.2
`Centurio` Qwen	65.2	31.7	54.3	16.6	21.4	53.2	21.4	55.4	56.6	16.2	34.8	5.2	0.6	52.2

Table 31: SMPQA Name

	en	avg.	af	am	cs	el	es	fa	fi	ha	hr	hu	ja	mi	nl	no	pl	ro	ta	te	zu
Centurio Aya	69.7	54.7	63.6	29.4	66.2	**67.8**	65.1	**60.0**	43.3	**37.5**	63.6	49.8	66.7	37.0	62.4	59.1	62.6	64.0	46.9	50.9	42.6
Centurio Qwen	72.7	56.2	65.3	**47.4**	62.2	56.7	67.0	53.6	48.8	36.7	65.4	54.1	67.6	**39.1**	63.7	63.6	60.4	58.5	45.2	63.4	**49.5**
Parrot	30.5	25.7	26.0	22.8	26.1	25.5	27.3	25.9	26.4	23.7	25.3	25.6	26.7	25.4	28.0	26.6	26.5	26.8	25.5	23.9	24.0
PALO 13B	61.4	41.1	48.4	25.9	47.9	35.8	53.2	37.5	42.7	26.1	52.3	47.9	49.1	31.0	48.9	51.2	46.1	46.5	28.9	32.2	28.3
PALO 7B	58.7	38.6	44.2	28.4	43.6	33.5	49.9	36.9	39.1	24.5	49.6	45.4	48.8	27.8	45.1	45.8	42.0	44.0	26.7	30.1	28.3
InternVL 2.5 4B	68.4	45.4	53.2	31.3	53.2	42.3	60.8	45.4	38.3	26.3	55.2	42.1	60.5	29.5	56.6	53.7	53.1	49.7	35.3	50.1	26.5
InternVL 2.5 8B	70.3	44.2	54.4	29.1	52.8	43.3	57.8	40.5	41.3	25.8	55.6	44.9	57.3	30.0	51.8	54.8	50.3	48.9	33.2	41.2	27.3
Qwen2-VL 2B	78.2	47.2	56.6	30.3	56.7	47.2	64.0	48.7	41.7	26.1	57.1	48.0	62.2	30.0	59.2	57.8	54.6	54.5	31.9	43.4	27.6
Qwen2-VL 7B	**80.7**	**57.5**	**68.9**	37.2	**68.5**	62.2	**72.6**	59.8	**55.1**	27.1	**72.2**	61.8	**71.8**	29.5	**69.5**	**69.6**	**67.5**	**65.6**	42.7	62.3	29.3
Maya	54.0	43.2	50.6	27.1	53.3	53.6	52.7	48.7	35.3	23.7	50.5	39.3	55.2	28.6	51.4	46.4	50.0	51.3	31.9	36.9	33.4
Llama-Vision	75.6	50.8	65.1	30.6	61.3	42.9	65.1	49.9	51.5	31.1	60.9	**65.0**	46.3	32.8	61.5	61.8	55.7	57.3	42.0	51.6	31.9
Phi 3.5 Vision	63.1	36.8	40.9	28.7	41.0	34.7	52.7	33.5	34.9	27.1	40.5	36.8	45.9	28.2	43.6	44.4	38.5	39.8	30.9	28.1	28.1
Pixtral 12B	71.0	54.2	62.3	34.3	61.6	58.3	66.1	57.3	52.0	27.7	67.1	60.4	64.8	31.9	58.6	62.1	59.8	59.0	**56.7**	**64.5**	25.0
Pangea	70.3	52.1	61.4	34.3	59.6	54.2	64.4	54.9	45.4	27.9	63.0	49.8	65.5	29.6	61.0	64.1	59.5	60.6	42.4	62.7	29.3
MiniCPM 2.6	72.6	47.4	56.0	29.9	55.1	46.6	62.1	48.5	41.8	22.9	59.5	44.9	62.9	29.0	57.8	55.2	54.7	52.7	34.5	53.9	33.4

Table 32: BIN-MC

	en	avg.	af	zh	it	pt	th	vi
Centurio Aya	53.0	41.2	52.8	51.4	47.7	27.4	27.8	40.3
Centurio Qwen	61.2	46.9	50.9	55.6	49.0	31.9	29.6	64.1
Parrot	46.6	36.2	38.0	37.8	36.8	25.9	23.5	55.1
PALO 13B	45.2	28.3	33.1	31.3	36.5	19.3	20.2	29.2
PALO 7B	41.0	29.1	34.4	31.5	32.7	21.8	21.1	33.4
InternVL 2.5 4B	63.2	50.3	46.0	60.9	50.3	34.9	39.1	70.4
InternVL 2.5 8B	**67.0**	**53.3**	**57.7**	**61.7**	**53.2**	33.0	39.1	**75.2**
Qwen2-VL 2B	47.9	40.5	38.0	51.6	36.4	36.2	26.1	54.9
Qwen2-VL 7B	56.1	49.7	50.9	58.6	46.8	34.7	38.3	69.0
Maya	49.2	36.3	48.5	46.4	36.6	25.9	20.0	40.3
Phi 3.5 Vision	56.3	40.7	51.5	54.4	44.1	25.2	24.3	44.4
Pixtral 12B	49.4	33.7	39.9	53.6	34.4	19.5	7.0	47.7
Pangea	58.0	45.5	50.3	58.6	49.0	32.2	27.8	55.3
MiniCPM 2.6	55.0	48.2	44.2	54.6	44.3	**36.9**	38.3	70.8

Table 33: M3Exam

	en	avg.	am	ber	bn	de	fil	ha	hi	ru	sw	th	zu
Centurio Aya	82.5	66.8	71.7	54.2	59.3	73.3	59.2	65.0	71.2	75.8	67.5	72.5	65.5
Centurio Qwen	87.5	**73.1**	**77.5**	49.2	**62.7**	**80.8**	**78.3**	**76.7**	**72.9**	85.0	**70.0**	**81.7**	**69.0**
Parrot	59.2	52.9	45.0	64.2	53.4	63.3	49.2	41.7	62.7	62.5	35.8	67.5	36.2
PALO 13B	63.3	26.2	25.0	55.0	0.8	44.2	47.5	40.0	0.0	5.8	32.5	0.0	37.1
PALO 7B	48.3	25.6	40.8	75.0	0.0	0.0	49.2	40.0	0.0	0.0	39.2	0.0	37.9
InternVL 2.5 4B	72.5	49.7	43.3	50.0	40.7	62.5	56.7	41.7	42.4	63.3	35.8	74.2	36.2
InternVL 2.5 8B	87.5	51.6	43.3	50.0	41.5	64.2	49.2	41.7	59.3	75.8	36.7	68.3	37.1
Qwen2-VL 2B	61.7	50.5	44.2	50.0	43.2	65.0	53.3	41.7	61.0	52.5	38.3	67.5	38.8
Qwen2-VL 7B	60.0	52.9	48.3	50.0	46.6	60.0	50.0	46.7	48.3	63.3	58.3	60.8	49.1
Maya	46.7	42.3	43.3	48.3	33.9	50.8	51.7	40.8	42.4	45.8	34.2	38.3	35.3
Phi 3.5 Vision	81.7	50.3	45.8	49.2	56.8	73.3	54.2	41.7	56.8	**85.8**	38.3	15.0	36.2
Pixtral 12B	55.8	47.7	51.7	32.5	47.5	63.3	51.7	44.2	16.1	54.2	65.8	53.3	44.0
Pangea	69.2	58.9	45.8	**90.0**	53.4	61.7	55.0	41.7	60.2	74.2	54.2	75.8	36.2
MiniCPM 2.6	52.5	49.1	45.0	55.8	49.2	45.8	48.3	40.8	44.1	59.2	48.3	65.8	37.9

Table 34: VGR

	en	avg.	am	ber	bn	de	fil	ha	hi	ru	sw	th	zu
Centurio Aya	12.5	20.7	18.3	21.7	20.0	11.7	24.2	29.2	15.2	10.8	28.6	29.2	19.5
Centurio Qwen	28.3	27.0	18.3	20.0	**33.3**	32.5	29.2	22.5	25.0	22.5	30.4	30.0	**33.1**
Parrot	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 13B	2.5	4.9	6.7	5.0	6.7	5.0	5.8	2.5	3.6	4.2	5.4	5.0	4.2
PALO 7B	5.8	6.8	8.3	9.2	10.0	5.8	6.7	4.2	9.8	5.0	4.5	5.8	5.1
InternVL 2.5 4B	24.2	21.0	18.3	26.7	17.5	20.8	20.0	23.3	22.3	20.0	23.2	20.0	18.6
InternVL 2.5 8B	**57.5**	29.0	25.0	22.5	25.8	**38.3**	**36.7**	25.8	**41.1**	**35.8**	15.2	30.0	22.9
Qwen2-VL 2B	22.5	20.4	17.5	20.0	13.3	26.7	25.0	24.2	20.5	16.7	21.4	15.8	23.7
Qwen2-VL 7B	5.8	13.2	14.2	15.8	13.3	11.7	10.0	15.0	12.5	12.5	13.4	13.3	13.6
Maya	20.0	20.1	20.0	25.8	19.2	20.8	15.0	25.8	17.9	23.3	21.4	15.8	16.1
Phi 3.5 Vision	45.8	**31.5**	**27.5**	**29.2**	23.3	36.7	30.0	**31.7**	33.9	29.2	**37.5**	**35.8**	31.4
Pixtral 12B	9.2	12.4	17.5	13.3	10.0	16.7	10.0	16.7	3.6	14.2	8.9	12.5	13.6
Pangea	0.0	6.7	0.0	0.8	0.0	20.8	24.2	15.8	6.2	0.8	3.6	0.8	0.8
MiniCPM 2.6	9.2	14.6	11.7	19.2	12.5	10.8	10.0	22.5	10.7	12.5	19.6	11.7	19.5

Table 35: VLOD

	en	avg.	id	sw	ta	tr	zh
Centurio Aya	85.0	77.9	79.5	70.9	73.4	83.4	82.4
Centurio Qwen	**89.6**	**81.7**	**85.0**	**76.8**	**76.0**	**84.2**	**86.7**
Parrot	63.5	55.1	56.6	51.2	50.7	58.6	58.2
PALO 13B	63.8	33.1	58.7	50.9	2.6	53.1	0.2
PALO 7B	62.7	24.1	33.6	47.8	0.4	38.5	0.0
InternVL 2.5 4B	74.9	59.0	65.7	50.7	50.9	64.2	63.5
InternVL 2.5 8B	83.0	63.3	63.2	51.4	54.6	67.2	79.9
Qwen2-VL 2B	67.9	55.9	60.9	51.8	52.2	59.0	55.8
Qwen2-VL 7B	69.8	60.2	61.1	53.1	60.9	65.3	60.7
Maya	60.3	56.3	60.3	50.7	50.6	58.9	61.2
Phi 3.5 Vision	73.4	46.4	56.4	51.3	50.8	58.0	15.7
Pixtral 12B	67.7	60.7	62.5	54.4	61.8	65.5	59.1
Pangea	75.8	70.5	74.3	70.9	66.6	71.1	69.6
MiniCPM 2.6	70.2	57.9	57.8	54.2	57.2	63.3	57.2

Table 36: MaRVL

	en	avg.	fr	hi	he	ro	th	zh
Centurio Aya	55.7	49.3	45.1	58.7	**62.9**	51.1	46.7	31.6
Centurio Qwen	60.1	47.7	47.1	45.1	56.8	47.7	57.0	32.2
Parrot	28.2	3.6	2.7	2.9	1.4	1.2	3.0	10.7
PALO 13B	51.7	33.1	42.0	17.5	53.4	34.2	20.9	30.6
PALO 7B	54.0	22.5	39.9	9.2	30.6	16.8	12.3	26.4
InternVL 2.5 4B	46.0	42.5	45.7	37.1	38.8	31.5	51.0	50.8
InternVL 2.5 8B	45.6	38.2	**51.2**	27.9	24.5	35.7	36.4	**53.4**
Qwen2-VL 2B	53.7	26.5	40.3	10.8	9.5	15.6	38.1	44.6
Qwen2-VL 7B	54.7	31.2	38.6	18.7	13.9	37.2	42.1	36.8
Maya	55.4	17.3	19.1	13.0	21.1	18.0	11.6	20.8
Llama-Vision	0.0	4.7	0.0	0.6	2.4	0.3	24.8	0.0
Phi 3.5 Vision	43.6	17.9	23.5	12.1	16.3	7.8	20.9	27.0
Pixtral 12B	59.4	43.4	46.8	31.7	54.4	44.1	44.4	39.1
Pangea	**61.4**	**55.0**	47.4	**61.0**	53.7	**52.9**	**67.2**	47.9
MiniCPM 2.6	53.4	22.3	14.3	12.1	5.1	19.5	53.6	29.3

Table 37: MaXM

	avg.	ar	de	fr	it	ja	ko	ru	th	vi
Centurio Aya	11.1	6.7	19.9	22.5	16.7	5.0	9.0	5.2	5.2	9.7
Centurio Qwen	11.9	4.6	22.7	26.5	18.6	5.9	9.9	5.0	5.2	8.9
Parrot	2.0	1.4	1.9	0.9	1.6	1.6	2.7	2.0	5.2	0.9
PALO 13B	6.3	2.6	15.6	12.1	10.4	4.0	4.3	4.0	0.0	4.2
PALO 7B	5.8	1.8	14.3	13.3	8.3	3.4	3.2	3.6	0.4	4.1
InternVL 2.5 4B	**25.1**	11.2	**34.4**	**38.4**	33.5	18.4	29.0	9.8	16.5	**34.6**
InternVL 2.5 8B	25.0	11.5	33.8	37.4	**35.3**	**19.7**	**30.3**	10.4	16.5	30.4
Qwen2-VL 2B	19.0	6.1	26.8	30.9	30.7	13.5	21.1	9.3	10.0	22.4
Qwen2-VL 7B	23.2	**16.9**	27.3	31.7	35.2	16.1	24.6	**10.8**	15.6	30.7
Maya	5.3	2.8	13.1	12.2	6.6	2.8	4.8	2.9	0.4	2.3
Llama-Vision	15.2	7.4	24.0	18.7	25.3	9.4	14.5	6.1	15.2	15.8
Phi 3.5 Vision	11.1	3.3	18.2	20.2	25.2	5.6	8.8	5.4	3.0	10.5
Pixtral 12B	14.1	4.3	25.7	27.3	25.2	5.9	9.1	7.5	5.2	16.6
Pangea	19.3	8.3	29.5	35.2	29.2	9.3	14.5	7.4	10.8	29.2
MiniCPM 2.6	16.1	2.3	23.9	27.5	32.7	11.7	12.7	7.3	10.0	16.5

Table 38: MTVQA

	en	avg.	bn	de	id	ko	pt	ru	zh
Parrot	37.7	21.2	20.2	23.2	19.8	22.8	21.7	19.7	21.2
PALO 13B	58.0	27.8	26.3	14.7	29.6	30.9	17.8	30.9	44.1
PALO 7B	59.1	36.6	42.8	34.5	30.0	40.8	27.7	32.2	47.9
InternVL 2.5 4B	63.6	28.0	28.1	29.2	15.4	38.3	27.2	31.5	25.9
InternVL 2.5 8B	63.4	32.0	17.4	23.8	25.0	38.2	27.6	36.4	55.2
Qwen2-VL 2B	60.5	38.2	18.6	43.2	32.6	39.0	39.9	44.1	50.3
Qwen2-VL 7B	62.5	49.3	37.4	51.1	48.4	50.3	51.8	52.1	54.1
Maya	58.2	49.1	40.1	53.2	49.7	47.2	52.5	50.6	50.1
Llama-Vision	39.3	27.6	26.0	29.2	26.8	24.9	27.9	30.7	27.9
Phi 3.5 Vision	**65.2**	38.0	5.0	51.9	37.3	35.6	50.6	45.9	39.5
Pixtral 12B	59.9	3.8	0.7	5.4	14.0	0.3	3.6	0.4	1.9
Pangea	64.6	**60.4**	**59.1**	**61.6**	**60.7**	**58.8**	**62.1**	**60.7**	**59.6**
MiniCPM 2.6	57.9	45.7	33.9	49.0	46.3	42.1	51.0	48.7	48.6

Table 39: xGQA

	en	avg.	ar	bn	cs	da	de	el	es	fa	fi	fil	fr	he	hi	hr	hu	id
Centurio Aya	78.4	**39.2**	**40.4**	18.5	**33.9**	40.0	38.6	**35.3**	**69.7**	**55.8**	11.0	**34.0**	**71.3**	**47.1**	**26.3**	**24.9**	**19.6**	58.3
Centurio Qwen	79.1	34.4	36.6	17.1	29.7	**43.1**	32.0	19.2	69.2	31.2	**12.0**	33.6	67.6	27.6	20.3	22.0	18.7	50.4
Parrot	5.6	0.4	0.6	0.0	0.2	0.0	0.0	0.0	3.3	2.3	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0
PALO 13B	67.3	17.0	23.5	22.9	7.9	30.3	32.4	0.2	57.0	1.5	6.6	8.4	66.2	0.6	25.0	9.9	2.7	22.7
PALO 7B	65.9	13.5	17.3	18.8	5.8	18.5	23.3	0.1	48.3	1.5	4.0	2.7	59.1	0.2	21.2	2.8	6.3	20.2
InternVL 2.5 4B	38.9	17.5	12.1	3.7	9.4	13.6	28.0	2.0	39.7	10.1	2.6	6.2	49.2	8.5	5.4	5.6	3.8	39.9
InternVL 2.5 8B	38.3	15.7	7.9	4.0	10.7	19.2	27.8	2.9	35.0	8.7	5.0	10.9	47.0	8.2	8.3	7.9	5.3	24.7
Qwen2-VL 2B	68.8	5.2	0.8	0.0	1.7	7.2	7.0	0.2	5.1	9.0	1.2	2.9	9.4	0.4	0.0	1.4	2.1	8.9
Qwen2-VL 7B	50.3	24.6	17.9	11.5	23.8	32.3	36.1	13.5	38.9	23.6	8.0	8.3	50.6	13.7	6.7	11.6	15.5	45.4
Maya	55.9	14.6	20.6	18.4	11.4	10.6	23.6	10.7	38.2	1.5	0.5	2.1	47.3	18.9	15.0	2.0	0.9	19.4
Llama-Vision	35.9	7.2	0.0	0.0	0.9	15.5	22.4	0.5	14.7	0.0	4.0	13.1	32.1	0.0	0.0	2.9	13.2	2.2
Phi 3.5 Vision	32.3	6.3	2.8	0.0	0.6	10.5	21.3	0.1	21.9	0.1	0.9	2.5	32.5	1.0	0.1	1.5	2.6	4.2
Pixtral 12B	26.5	22.1	18.6	9.6	16.8	24.4	33.2	8.9	36.5	20.5	10.4	15.3	47.8	18.0	6.3	18.7	15.6	44.6
Pangea	70.1	34.6	33.3	**30.8**	19.4	25.2	**39.4**	13.0	61.4	25.4	4.2	6.7	69.7	42.7	21.5	9.5	3.6	**70.9**
MiniCPM 2.6	**87.5**	14.2	6.7	3.3	8.5	8.7	27.5	1.7	44.0	5.8	3.2	5.0	52.1	1.5	3.0	6.1	5.8	24.6

	it	ja	ko	mi	nl	no	pl	pt	quz	ro	ru	sv	sw	te	th	tr	uk	vi	zh
Centurio Aya	**60.4**	49.1	**21.3**	33.7	**61.7**	42.5	**37.9**	59.3	1.7	34.6	38.0	**45.9**	29.9	15.1	26.0	**30.6**	**30.6**	72.7	**56.9**
Centurio Qwen	53.7	43.5	13.4	**34.9**	56.2	41.4	30.0	59.9	**2.1**	23.4	39.2	42.7	30.2	13.5	42.3	23.3	20.3	69.4	33.8
Parrot	0.2	0.0	0.0	0.0	1.6	0.0	0.0	4.0	0.2	0.4	0.4	0.0	0.0	0.8	0.0	1.0	0.0	0.0	0.0
PALO 13B	40.4	19.7	0.2	0.3	36.5	31.0	9.1	13.8	0.8	14.5	21.3	33.9	0.8	0.0	0.5	0.6	2.6	15.6	37.0
PALO 7B	31.0	29.8	2.4	0.3	29.8	16.5	8.4	8.9	0.5	2.6	19.7	23.3	0.0	0.0	0.0	0.5	0.1	17.4	29.7
InternVL 2.5 4B	33.1	33.1	8.9	0.8	29.3	14.2	12.9	39.0	0.2	9.9	23.4	17.1	0.6	1.1	27.9	7.8	7.1	61.3	44.1
InternVL 2.5 8B	27.5	22.0	6.7	0.9	26.8	16.6	12.0	35.0	0.8	12.0	22.6	20.5	1.0	2.6	7.2	9.3	4.7	46.2	40.1
Qwen2-VL 2B	5.9	8.4	1.0	0.3	2.9	5.2	1.5	21.0	1.0	3.7	1.1	13.0	1.1	0.0	1.3	0.9	0.6	7.9	49.5
Qwen2-VL 7B	38.7	32.0	9.1	0.9	39.1	35.7	30.1	48.8	0.9	19.0	37.9	43.1	2.4	3.9	31.2	15.8	16.6	55.6	41.8
Maya	34.4	26.3	8.9	0.3	28.8	9.4	15.8	16.4	0.6	22.0	19.9	11.4	0.5	0.0	0.2	13.5	1.5	31.8	26.9
Llama-Vision	33.5	0.2	0.1	0.8	30.1	2.8	2.4	15.7	0.2	23.4	0.3	11.2	6.8	0.0	1.2	0.6	0.1	0.8	0.0
Phi 3.5 Vision	23.6	8.0	0.3	0.2	19.8	10.7	1.7	25.8	0.4	3.0	0.5	10.2	0.5	0.0	1.0	1.7	0.1	2.6	8.1
Pixtral 12B	32.8	21.8	12.0	5.9	29.7	26.0	19.6	42.4	1.0	20.2	33.8	30.0	10.4	6.2	23.8	14.9	18.4	51.7	28.1
Pangea	53.5	**63.3**	20.3	0.3	44.9	**48.5**	24.1	**64.6**	1.7	**38.7**	**47.3**	20.1	**40.7**	**21.8**	**61.4**	30.2	20.7	**81.3**	50.7
MiniCPM 2.6	24.6	18.6	4.4	2.2	27.8	12.0	12.0	36.0	0.2	10.0	20.0	17.0	1.5	0.5	20.9	8.0	7.5	25.8	39.4

Table 40: XM3600

	en	avg.	ar	bn	cs	da	de	el	es	fa	fi	fil	fr	he	hi	hr	hu	id
Centurio Aya	100.0	95.7	93.6	100.0	97.7	**96.7**	100.0	100.0	99.8	100.0	99.8	100.0	100.0	99.8	99.6	84.6	99.8	99.2
Centurio Qwen	99.8	95.2	95.1	100.0	98.6	93.9	100.0	100.0	99.4	100.0	**100.0**	100.0	100.0	98.8	99.0	80.9	100.0	96.7
Parrot	100.0	25.0	**100.0**	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PALO 13B	100.0	60.1	98.6	93.9	47.1	87.5	100.0	60.7	99.6	0.0	74.0	71.5	99.8	35.4	98.2	70.1	9.0	66.4
PALO 7B	100.0	72.0	99.6	98.8	47.5	93.4	100.0	58.2	99.8	0.0	91.8	52.7	100.0	30.7	98.8	27.0	90.8	96.9
InternVL 2.5 4B	100.0	91.0	96.7	93.9	97.1	82.8	100.0	99.0	99.8	98.8	96.1	95.3	100.0	96.7	88.9	91.4	96.1	96.9
InternVL 2.5 8B	100.0	91.1	99.4	95.3	97.7	82.8	100.0	100.0	99.4	97.9	98.2	96.3	100.0	98.4	95.1	83.2	98.2	96.7
Qwen2-VL 2B	100.0	13.2	8.2	0.0	0.0	9.6	12.9	0.2	5.9	58.4	0.2	10.9	10.0	4.5	0.0	3.1	0.2	12.7
Qwen2-VL 7B	100.0	90.0	96.5	98.2	93.9	86.1	99.8	99.4	99.2	99.2	95.7	96.3	98.2	98.2	60.2	79.1	75.4	86.9
Maya	100.0	65.7	99.0	96.1	67.6	85.5	98.6	92.0	99.8	0.2	12.1	1.0	100.0	77.0	98.4	20.7	60.7	40.6
Llama-Vision	100.0	33.3	0.0	0.0	4.9	68.8	95.5	7.0	52.7	0.0	35.0	80.7	88.3	0.0	0.0	17.0	91.0	1.6
Phi 3.5 Vision	100.0	40.8	58.4	0.6	1.4	85.4	99.2	16.2	99.4	0.0	15.2	30.1	99.8	14.8	4.7	25.2	56.4	31.6
Pixtral 12B	100.0	**96.8**	99.8	99.6	**98.8**	95.9	100.0	99.4	100.0	100.0	99.8	99.8	100.0	99.8	**100.0**	**93.4**	100.0	**99.6**
Pangea	99.8	87.9	98.8	99.0	97.9	19.1	99.6	99.8	99.2	98.4	91.6	68.9	100.0	**100.0**	98.2	67.8	93.6	97.9
MiniCPM 2.6	99.8	92.3	94.7	96.5	95.5	96.3	100.0	99.8	99.8	99.0	98.4	97.9	100.0	62.9	92.6	77.3	94.5	93.6

	it	ja	ko	mi	nl	no	pl	pt	quz	ro	ru	sv	sw	te	th	tr	uk	vi	zh
Centurio Aya	99.6	98.8	100.0	98.8	100.0	**97.3**	99.8	100.0	1.8	100.0	99.6	98.8	90.6	100.0	99.6	100.0	99.8	100.0	89.6
Centurio Qwen	99.8	98.8	100.0	**100.0**	100.0	95.7	99.6	99.4	3.7	100.0	99.4	98.2	86.5	100.0	99.8	99.6	99.2	100.0	86.5
Parrot	0.0	0.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	100.0	0.0	0.0	0.0
PALO 13B	99.2	61.5	9.8	0.0	99.8	92.2	41.2	27.5	0.0	95.5	68.2	94.9	1.6	26.4	68.0	9.4	11.3	54.7	88.9
PALO 7B	99.4	99.2	91.6	0.0	99.4	95.1	95.5	27.0	0.0	69.1	100.0	96.9	0.0	56.8	91.6	84.4	0.0	99.6	99.8
InternVL 2.5 4B	99.6	100.0	99.2	48.2	99.6	83.0	99.6	100.0	7.0	98.8	99.2	97.7	34.6	90.8	98.2	97.7	95.7	96.1	99.8
InternVL 2.5 8B	100.0	99.8	99.2	66.8	99.2	86.5	99.6	99.8	1.2	99.8	99.8	98.2	54.7	99.4	99.2	98.4	40.6	100.0	99.8
Qwen2-VL 2B	3.9	19.3	18.0	0.2	0.0	5.3	0.0	34.0	0.0	15.4	0.0	23.6	4.5	0.0	1.0	1.0	0.8	12.7	98.8
Qwen2-VL 7B	99.0	98.8	99.0	64.5	99.2	94.1	95.7	95.3	0.2	98.2	99.4	97.9	72.1	95.1	98.8	89.8	83.2	98.2	99.0
Maya	99.6	99.8	91.4	0.0	99.8	80.1	92.0	43.9	0.0	95.7	100.0	96.7	1.6	0.0	47.9	96.3	7.2	62.5	99.8
Llama-Vision	94.7	0.0	0.0	93.0	99.6	9.0	9.2	48.2	0.8	92.6	0.2	33.8	73.6	0.0	2.3	0.2	0.0	0.2	0.0
Phi 3.5 Vision	99.0	58.6	9.0	1.0	93.6	89.8	27.3	63.9	0.0	56.8	0.0	85.0	2.9	13.7	3.5	38.1	0.0	51.8	37.9
Pixtral 12B	100.0	100.0	100.0	99.6	100.0	95.5	**100.0**	100.0	9.4	99.6	100.0	**100.0**	**95.9**	99.4	100.0	99.8	**100.0**	100.0	99.8
Pangea	100.0	99.4	100.0	0.8	99.6	95.7	99.6	100.0	0.0	100.0	99.6	67.0	82.4	99.8	100.0	99.8	91.0	99.8	99.0
MiniCPM 2.6	100.0	98.4	99.2	99.2	99.6	95.5	96.5	99.8	**10.0**	98.2	97.3	99.4	66.4	85.5	96.3	95.1	99.2	90.6	97.1

Table 41: XM3600 Language Fidelity

	en	avg.	ar	es	fr	ru
Centurio Aya	65.0	62.4	61.7	61.0	64.3	62.7
Centurio Qwen	**75.4**	**70.2**	**68.8**	**70.9**	**70.5**	**70.8**
Parrot	28.7	31.4	34.0	24.3	30.0	37.4
PALO 13B	56.6	53.6	51.8	52.7	54.9	55.0
PALO 7B	58.0	53.4	52.5	52.3	53.7	55.1
InternVL 2.5 4B	69.0	58.7	55.7	58.8	61.4	59.0
InternVL 2.5 8B	73.5	66.4	61.8	68.0	68.4	67.3
Qwen2-VL 2B	61.9	56.2	52.9	55.3	58.6	57.9
Qwen2-VL 7B	62.1	59.6	59.2	58.9	60.0	60.3
Maya	50.1	43.9	45.3	42.7	45.8	41.8
Phi 3.5 Vision	58.9	53.3	49.7	52.7	56.4	54.3
Pixtral 12B	60.9	52.7	36.0	57.9	59.0	58.1
Pangea	69.0	65.2	64.5	64.3	66.3	65.7
MiniCPM 2.6	71.9	65.4	61.1	67.5	67.0	66.1

Table 42: XVNLI

	en	avg.	ar	fr	hi	id	ja	pt
Centurio Aya	37.6	37.2	36.2	38.9	38.8	39.7	34.2	35.4
Centurio Qwen	46.4	43.0	39.6	45.0	**41.6**	44.1	43.5	44.1
Parrot	35.3	32.4	31.9	34.9	26.1	31.3	34.9	35.4
PALO 13B	32.4	28.9	24.2	34.9	24.2	31.6	26.4	32.3
PALO 7B	31.8	30.9	28.2	33.6	27.3	30.6	32.3	33.3
InternVL 2.5 4B	49.2	42.7	**41.6**	45.6	33.7	43.4	44.2	47.8
InternVL 2.5 8B	**50.7**	**45.2**	40.3	**48.7**	41.2	43.1	**47.6**	**50.2**
Qwen2-VL 2B	36.8	35.5	31.5	41.3	30.2	36.7	36.1	37.0
Qwen2-VL 7B	43.0	40.7	36.9	42.6	38.5	41.1	41.3	43.8
Maya	37.9	33.3	32.6	36.6	31.3	31.6	32.0	36.0
Phi 3.5 Vision	41.7	37.4	34.9	44.3	29.2	37.7	35.7	42.4
Pixtral 12B	30.3	26.2	19.1	28.5	19.2	27.3	28.6	34.7
Pangea	43.1	42.0	37.6	43.0	38.5	**46.8**	41.6	44.8
MiniCPM 2.6	39.1	36.5	30.5	38.9	33.7	37.7	37.2	40.7

Table 43: xMMMU

	avg.	amh-ethiopia	arz-egypt	ben-india	bre-france	bul-bulgaria	fil-philippines	gle-ireland	hin-india	ibo-nigeria
Centurio Aya	49.4	32.1	52.7	45.8	30.4	50.1	48.8	41.7	**67.2**	31.0
Centurio Qwen	52.9	**38.0**	52.7	54.9	30.6	49.6	51.7	48.5	65.7	30.5
Parrot	41.1	31.6	35.5	33.9	31.1	38.8	45.3	45.1	41.8	35.5
PALO 13B	39.6	26.1	31.5	33.9	31.6	35.3	45.3	41.1	40.3	32.5
PALO 7B	37.1	20.5	25.1	30.1	27.7	32.6	43.3	37.1	43.3	31.0
InternVL 2.5 4B	48.1	36.3	38.9	42.7	33.1	42.0	47.8	45.7	51.7	29.0
InternVL 2.5 8B	48.6	29.5	41.4	42.3	29.4	47.4	46.8	47.5	50.2	33.5
Qwen2-VL 2B	33.6	27.4	31.0	33.6	25.9	32.9	32.0	31.3	34.8	35.5
Qwen2-VL 7B	37.6	31.2	35.5	31.5	31.1	35.6	40.9	37.1	39.3	31.0
Maya	39.8	30.3	41.9	38.8	30.6	36.7	35.0	34.4	46.8	31.0
Llama-Vision	38.8	32.1	5.4	**60.1**	13.6	22.4	43.8	35.9	46.3	28.5
Phi 3.5 Vision	40.9	28.6	38.4	28.7	28.9	33.7	45.3	40.2	42.8	38.0
Pixtral 12B	33.5	22.6	27.1	21.7	24.9	30.5	35.5	38.7	41.3	26.5
Pangea	**55.2**	35.5	49.3	53.5	33.1	**52.3**	**56.7**	**53.1**	66.2	**40.0**
MiniCPM 2.6	34.1	26.9	31.5	27.6	25.9	32.6	31.5	37.1	32.3	36.0

	ind-indonesia	jav-indonesia	jpn-japan	kin-rwanda	kor-south korea	mar-india	min-indonesia	mon-mongolia	msa-malaysia	nor-norway
Centurio Aya	53.6	41.1	44.8	32.8	61.7	56.9	42.6	29.2	52.7	55.5
Centurio Qwen	54.9	46.1	44.8	**42.1**	66.2	55.9	41.8	33.3	55.9	58.2
Parrot	43.4	37.7	34.5	32.8	47.6	32.2	36.3	34.3	42.9	49.5
PALO 13B	41.0	35.4	33.0	28.9	42.8	37.6	37.8	26.3	44.1	54.5
PALO 7B	37.9	33.7	29.1	29.4	42.4	33.7	31.1	25.6	36.8	47.5
InternVL 2.5 4B	54.6	44.4	39.4	34.9	65.9	48.0	41.0	27.6	53.0	52.8
InternVL 2.5 8B	54.9	44.8	41.4	32.8	56.9	43.6	44.6	33.0	54.3	55.2
Qwen2-VL 2B	36.9	31.6	25.1	31.5	37.6	24.3	27.9	31.1	32.1	40.5
Qwen2-VL 7B	40.8	32.3	36.0	30.2	43.4	31.2	33.5	34.0	43.5	42.8
Maya	36.2	34.7	29.1	31.5	50.0	42.6	33.9	31.1	44.4	47.5
Llama-Vision	42.0	34.0	25.1	18.3	45.9	38.1	34.3	27.9	40.6	48.5
Phi 3.5 Vision	40.8	35.4	36.9	36.2	44.1	33.7	39.0	35.3	41.3	48.2
Pixtral 12B	36.9	32.3	27.6	25.1	39.7	24.3	28.3	24.7	36.8	32.1
Pangea	**60.0**	**50.5**	42.9	33.6	**68.3**	**57.9**	**48.2**	**40.7**	**60.3**	58.2
MiniCPM 2.6	36.2	31.6	33.5	30.6	31.0	32.2	29.9	29.2	34.6	40.5

	orm-ethiopia	por-brazil	ron-romania	rus-russia	sin-sri lanka	spa-argentina	spa-chile	spa-colombia	spa-ecuador	spa-mexico
Centurio Aya	36.4	65.1	61.6	65.5	28.9	60.0	58.5	56.0	55.2	52.6
Centurio Qwen	35.0	**70.8**	57.0	67.5	42.7	63.8	66.2	59.8	58.8	**61.9**
Parrot	34.6	60.9	44.0	45.0	28.0	47.9	52.1	48.5	43.9	44.3
PALO 13B	29.4	53.9	52.6	44.0	24.9	47.5	50.0	47.3	47.8	44.9
PALO 7B	51.1	49.3	45.5	24.0	47.5	49.1	45.2	45.9	42.4	57.5
InternVL 2.5 4B	34.6	66.5	49.7	65.5	34.7	60.4	59.8	54.4	56.6	53.9
InternVL 2.5 8B	32.2	64.4	60.3	62.5	29.8	60.4	65.4	56.4	59.7	57.0
Qwen2-VL 2B	33.2	39.8	32.5	33.0	24.9	40.0	40.2	40.7	39.0	35.0
Qwen2-VL 7B	**37.4**	47.9	34.8	47.5	30.7	44.5	47.9	40.7	42.3	40.2
Maya	29.9	53.5	51.3	42.0	30.2	44.9	47.4	45.2	45.6	39.3
Llama-Vision	23.4	38.7	47.7	52.0	**48.4**	47.9	55.1	51.0	48.1	48.3
Phi 3.5 Vision	33.6	62.0	43.7	45.0	29.3	55.1	59.0	51.9	52.8	48.0
Pixtral 12B	21.5	41.2	28.8	40.5	23.6	49.1	54.3	43.6	48.3	40.6
Pangea	36.4	69.7	**62.9**	**73.5**	36.0	63.8	**67.1**	**61.4**	**62.4**	60.7
MiniCPM 2.6	36.0	45.1	33.4	37.0	26.7	37.7	44.4	40.7	38.7	35.9

	spa-spain	spa-uruguay	sun-indonesia	swa-kenya	tam-india	tel-india	urd-india	urd-pakistan	zho-china	zho-singapore
Centurio Aya	68.2	40.3	42.0	50.2	49.5	37.0	47.3	50.5	64.6	65.6
Centurio Qwen	70.4	42.5	41.5	56.4	43.0	50.5	52.7	56.9	71.1	**73.1**
Parrot	65.7	36.2	31.5	40.7	36.9	29.5	31.8	32.9	64.3	55.2
PALO 13B	63.2	39.4	37.5	39.6	31.3	29.5	35.9	40.3	46.6	39.6
PALO 7B	57.5	36.2	31.5	31.1	29.0	28.5	34.1	40.3	46.3	42.0
InternVL 2.5 4B	68.2	44.8	44.0	45.8	35.5	41.0	47.7	41.7	74.6	66.5
InternVL 2.5 8B	72.3	44.8	41.5	50.9	35.5	39.5	44.1	39.8	**78.5**	72.6
Qwen2-VL 2B	42.1	34.6	33.5	37.7	25.2	27.0	30.5	31.9	44.1	41.5
Qwen2-VL 7B	47.8	41.6	31.5	34.1	26.6	26.5	37.3	31.0	51.4	43.4
Maya	55.7	34.3	33.5	38.8	32.2	29.0	43.2	48.1	50.5	50.9
Llama-Vision	70.1	37.1	29.5	52.0	**60.7**	**62.5**	24.5	15.3	36.3	23.1
Phi 3.5 Vision	64.2	**45.1**	32.0	46.5	29.4	32.5	29.5	26.9	51.1	40.6
Pixtral 12B	48.4	40.3	27.0	42.1	22.9	19.0	27.7	23.6	47.3	39.6
Pangea	**73.0**	44.8	**49.0**	**65.6**	46.7	55.0	**57.7**	**65.7**	71.7	68.4
MiniCPM 2.6	42.5	36.2	29.0	34.1	24.3	28.5	32.7	22.7	48.2	46.2

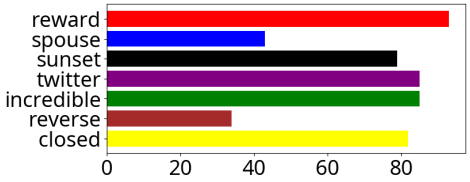
Table 44: CVQA

	en	avg.	avg. Latin	avg. other	ar	de	hi	id	it	ko	ru	th	zh	zu
Centurio Aya	83.1	74.2	80.9	69.7	75.9	82.1	80.1	81.4	80.6	68.8	73.5	66.5	53.4	79.5
Centurio Qwen	84.8	76.1	82.7	71.8	76.9	83.5	82.4	83.8	83.1	72.4	75.6	64.4	58.9	80.2
Parrot	51.0	49.9	50.5	49.5	50.4	51.6	49.6	51.0	49.8	50.4	50.5	48.2	47.8	49.5
PALO 13B	54.0	51.5	52.7	50.7	50.9	53.2	51.2	52.5	52.8	51.0	49.5	51.0	50.7	52.1
PALO 7B	55.5	52.8	55.4	51.0	50.4	56.9	51.0	55.0	54.1	51.6	51.1	51.4	50.2	55.8
InternVL 2.5 4B	87.0	78.3	86.9	72.6	54.9	87.6	59.8	87.0	88.2	89.4	86.4	55.1	90.4	84.8
InternVL 2.5 8B	91.0	79.2	88.7	72.8	55.8	89.8	54.9	89.1	89.1	92.5	86.9	53.1	93.6	86.9
Qwen2-VL 2B	85.0	83.5	83.4	83.5	70.6	84.4	86.5	84.1	83.5	88.1	78.8	86.4	90.4	81.8
Qwen2-VL 7B	91.2	**90.9**	90.1	**91.4**	**83.4**	90.5	**94.8**	91.0	90.8	**93.8**	87.5	**94.1**	**94.9**	88.2
Maya	51.4	50.9	51.6	50.4	50.4	53.4	50.1	51.5	50.0	49.9	49.5	51.1	51.6	51.6
Llama-Vision	91.1	84.8	89.9	81.5	63.2	90.1	91.1	89.5	**91.9**	87.4	83.0	84.8	79.5	88.0
Phi 3.5 Vision	**92.2**	79.4	90.2	72.2	53.1	**91.9**	83.8	89.2	90.9	77.9	86.6	55.5	76.5	**88.8**
Pixtral 12B	91.1	71.0	**90.5**	58.0	50.4	91.5	53.6	**91.1**	90.9	49.5	**88.2**	52.9	53.4	88.4
Pangea	87.2	72.2	85.7	63.1	51.5	86.6	69.4	86.2	87.1	71.4	79.2	54.4	52.9	82.9
MiniCPM 2.6	89.0	74.3	88.0	65.2	52.0	89.0	53.1	87.9	89.0	54.8	84.0	53.1	94.5	86.0

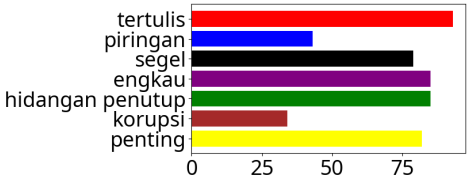
Table 45: SMPQA Ground

	en	avg.	avg. Latin	avg. other	ar	de	hi	id	it	ko	ru	th	zh	zu
Centurio Aya	60.0	30.1	49.8	17.0	**29.2**	50.2	17.6	52.6	51.2	11.2	38.2	4.8	0.8	45.2
Centurio Qwen	65.2	31.7	54.3	16.6	21.4	53.2	21.4	55.4	56.6	16.2	34.8	5.2	0.6	52.2
Parrot	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0
PALO 13B	25.6	4.0	9.9	0.1	0.0	12.0	0.0	10.2	12.4	0.4	0.0	0.0	0.0	5.0
PALO 7B	22.4	2.7	6.7	0.1	0.0	8.4	0.0	7.0	7.0	0.4	0.0	0.0	0.0	4.4
InternVL 2.5 4B	77.8	47.5	67.7	34.0	0.0	71.0	0.0	69.8	69.6	69.0	54.4	0.2	80.2	60.4
InternVL 2.5 8B	80.6	48.2	68.1	34.9	0.0	69.2	0.0	70.4	70.8	67.2	61.2	0.2	80.8	62.2
Qwen2-VL 2B	68.8	47.4	60.0	39.0	0.2	61.2	24.8	59.4	61.2	66.0	46.8	24.0	72.0	58.2
Qwen2-VL 7B	85.0	**64.9**	**76.2**	**57.4**	1.8	**80.6**	**58.6**	**75.8**	79.2	**77.6**	**70.6**	**43.8**	**92.0**	**69.2**
Maya	14.6	1.8	4.3	0.1	0.0	8.2	0.0	3.6	4.6	0.4	0.0	0.0	0.0	0.8
Llama-Vision	58.4	22.8	46.6	6.9	0.0	55.4	2.4	38.4	37.2	8.4	13.0	6.0	11.8	55.4
Phi 3.5 Vision	84.8	35.9	69.4	13.5	0.2	70.8	12.0	69.4	76.6	15.4	40.4	0.2	12.8	61.0
Pixtral 12B	85.0	35.9	73.3	10.9	0.0	71.8	0.0	75.4	**81.6**	0.4	64.6	0.4	0.0	64.6
Pangea	72.0	23.8	54.4	3.4	0.0	58.6	0.2	57.2	64.4	0.4	19.2	0.4	0.0	37.4
MiniCPM 2.6	80.8	39.3	67.5	20.6	0.0	67.2	0.0	69.8	71.4	1.0	38.4	0.4	83.6	61.6

Table 46: SMPQA Name



(a) Example of a bar plot in SMPQA for English.
Questions for Grounding: "Is the bar with label 'reward' the biggest?", "Is the bar with label 'incredible' the biggest?", "Is the bar with label 'reverse' the smallest?", "Is the bar with label 'sunset' the smallest?", "Is the bar with label 'closed' colored in yellow?", "Is the bar with label 'closed' colored in purple?", "Is the bar with label 'twitter' colored in purple?", "Is the bar with label 'twitter' colored in red?"
Questions for Reading: "What is the label of the biggest bar?", "What is the label of the smallest bar?", "What is the label of the yellow bar?", "What is the label of the red bar?", "What is the label of the purple bar?"



(b) The same plot in Indonesian. Note that all questions refer to the same parts of the plot as the English version just with different words for labels.
Questions for Grounding: "Is the bar with label 'tertulis' the biggest?", "Is the bar with label 'hidangan penutup' the biggest?", "Is the bar with label 'korupsi' the smallest?", "Is the bar with label 'segel' the smallest?", "Is the bar with label 'penting' colored in yellow?", "Is the bar with label 'penting' colored in purple?", "Is the bar with label 'engkau' colored in purple?", "Is the bar with label 'engkau' colored in red?"
Questions for Reading: "What is the label of the biggest bar?", "What is the label of the smallest bar?", "What is the label of the yellow bar?", "What is the label of the red bar?", "What is the label of the purple bar?"

Figure 3: Examples of one plot configuration in SMPQA for English and Indonesian.

Can you explain the joke of this German meme?

`Centurio Aya`: The humor in this image comes from juxtaposing a biblical figure, Jesus, with an everyday scene that is both ordinary and unexpected. The man's appearance as Jesus, complete with long hair and a beard, contrasts with the mundane setting of a caravan park populated by caravans (trailers). This creates a humorous effect because it takes something familiar (Jesus) and places it into a context that is not only unfamiliar but also somewhat banal or trivial. It plays on the idea of finding divinity in the most unexpected places, which can be seen as a commentary on how we perceive holiness or significance. Additionally, the caption "Jesus und seine Anhänger" further emphasizes the absurdity by referencing followers, implying that even in such an unremarkable location, one could find disciples or supporters, thus adding another layer to the comedic effect.

`Centurio Qwen`: The joke in this German meme plays on a pun. The phrase "Jesus und seine Anhänger" translates to "Jesus and his followers." However, the word "Anhänger," which means "followers," is also slang for caravans or RVs (recreational vehicles) in some contexts. Therefore, the image humorously suggests that Jesus has modern-day followers who travel together in caravans instead of following him spiritually as he did with his original disciples. This juxtaposition creates a humorous effect by combining religious iconography with contemporary culture.

Figure 4: `Centurio` explaining a German meme. Both models correctly read the caption but only `Qwen` 'understands' the pun and gets the explanation (mostly) correct.

VI. Centurio:
On Drivers of Multilingual Ability of Large Vision-Language Model

	AVG.	XM3600			MT-VQA	SMPQA G.		SMPQA N.		M3Exam		xMMMU		C-VQA
		en	mul	fid.		en	mul	en	mul	en	mul	en	mul	
Parrot	25.8	5.6	0.4	25.0	2.0	51.0	49.9	0.0	0.0	46.6	36.2	35.3	32.4	41.1
PALO 7B	28.7	65.9	13.5	72.0	5.8	55.5	52.8	22.4	2.7	41.0	29.1	31.8	30.9	37.1
PALO 13B	29.9	67.3	17.0	60.1	6.3	54.0	51.5	25.6	4.0	45.2	28.3	32.4	28.9	39.6
Llama-Vision 3.2 11B	*32.3	35.9	7.2	33.3	15.2	91.1	84.8	58.4	22.8	—	—	—	—	38.8
Maya	33.4	55.9	14.6	65.7	5.3	51.4	50.9	14.6	1.8	49.2	36.3	37.9	33.3	39.8
Pixtral 12B	38.1	26.5	22.1	**96.8**	14.1	91.1	71.0	85.0	35.9	49.4	33.7	30.3	26.2	33.5
Phi 3.5 Vision	39.5	32.3	6.3	40.8	11.1	**92.2**	79.4	84.8	35.9	56.3	40.7	41.7	37.4	40.9
Qwen2VL 2B	41.2	68.8	5.2	13.2	19.0	85.0	83.5	68.8	47.4	47.9	40.5	36.8	35.5	33.6
MiniCPM 2.6	41.7	**87.5**	14.2	92.3	16.1	89.0	74.3	80.8	39.3	55.0	48.2	39.1	36.5	34.1
InternVL 2.5 4B	45.3	38.9	17.5	91.0	**25.1**	87.0	78.3	77.8	47.5	63.2	50.3	49.2	42.7	48.1
InternVL 2.5 8B	47.4	38.3	15.7	91.1	25.0	91.0	79.2	80.6	48.2	**67.0**	**53.3**	**50.7**	**45.2**	48.6
Qwen2VL 7B	47.7	50.3	24.6	90.0	23.2	91.2	**90.9**	85.0	**64.9**	56.1	49.7	43.0	40.7	37.6
Pangea	48.2	70.1	34.6	87.9	19.3	87.2	72.2	72.0	23.8	58.0	45.5	43.1	42.0	**55.2**
Centurio Aya	48.5	78.4	**39.2**	95.7	11.1	83.1	74.2	60.0	30.1	53.0	41.2	37.6	37.2	49.4
Centurio Qwen	**51.6**	79.1	34.4	95.2	11.9	84.8	76.1	65.2	31.7	61.2	46.9	46.4	43.0	52.9

	MAXM		xGQA		BIN-MC		XVNLI		MaRVL		VGR		VLOD	
	en	mul	en	mul	en	mul	en	mul	en	mul	en	mul	en	mul
Parrot	28.2	3.6	37.7	21.2	30.5	25.7	28.7	31.4	63.5	55.1	59.2	52.9	0.0	0.0
PALO 7B	54.0	22.5	59.1	36.6	58.7	38.6	58.0	53.4	62.7	24.1	48.3	25.6	5.8	6.8
PALO 13B	51.7	33.1	58.0	27.8	61.4	41.1	56.6	53.6	63.8	33.1	63.3	26.2	2.5	4.9
Llama-Vision 3.2 11B	0.0	4.7	39.3	27.6	75.6	50.8	—	—	—	—	—	—	—	—
Maya	55.4	17.3	58.2	49.1	54.0	43.2	50.1	43.9	60.3	56.3	46.7	42.3	20.0	20.1
Pixtral 12B	59.4	43.4	59.9	3.8	71.0	54.2	60.9	52.7	67.7	60.7	55.8	47.7	9.2	12.4
Phi 3.5 Vision	43.6	17.9	**65.2**	38.0	63.1	36.8	58.9	53.3	73.4	46.4	81.7	50.3	45.8	**31.5**
Qwen2VL 2B	53.7	26.5	60.5	38.2	78.2	47.2	61.9	56.2	67.9	55.9	61.7	50.5	22.5	20.4
MiniCPM 2.6	53.4	22.3	57.9	45.7	72.6	47.4	71.9	65.4	70.2	57.9	52.5	49.1	9.2	14.6
InternVL 2.5 4B	46.0	42.5	63.6	28.0	68.4	45.4	69.0	58.7	74.9	59.0	72.5	49.7	24.2	21.0
InternVL 2.5 8B	45.6	38.2	63.4	32.0	70.3	44.2	73.5	66.4	83.0	63.3	87.5	51.6	**57.5**	29.0
Qwen2VL 7B	54.7	31.2	62.5	49.3	**80.7**	**57.5**	62.1	59.6	69.8	60.2	60.0	52.9	5.8	13.2
Pangea	**61.4**	**55.0**	64.6	**60.4**	70.3	52.1	69.0	65.2	75.8	70.5	69.2	58.9	0.0	6.7
Centurio Aya	55.7	49.3	59.1	53.2	69.7	54.7	65.0	62.4	85.0	77.9	82.5	66.8	12.5	20.7
Centurio Qwen	60.1	47.7	60.6	54.8	72.7	56.2	**75.4**	**70.2**	**89.6**	**81.7**	87.5	**73.1**	28.3	27.0

Table 6: Comparison of Centurio and 13 other LVLMs over 14 tasks. We mark **best** and second-best results. Scores are accuracy (CIDEr for XM3600). **en** & **mul** are the English and averaged multilingual results. **XM3600 fid.** is the language fidelity over all languages; **SMPQA G. & N** are Grounding and Naming. ***:** only single-image input was supported. **AVG.** is the average over all task. Details on the evaluation setup and models in Appendix B.

Summarize the table for me.

`Centurio Aya`: This is a detailed summary of the table:
The table compares Centurio and 13 other LLM models on 14 different tasks, using CIDEr as an evaluation metric. It presents accuracy scores (AVG., MTXMUL) and their standard deviations (SD.) across various datasets like SQuAD G & N, Image captioning, etc. Models are listed in descending order based on AVG score, with Paratrol being at the top followed by others like PALO-TB, Lama Vision, etc. There's also information about model names, task descriptions, and additional metrics like BIN_MC, XNLI, MARVL, VGRU, VLOD, and C.Q.A. A note explains that only one image input was used for evaluating language fidelity over all languages, and it mentions "CENTURIO QEN" which might be a specific version or variant of the Centurio model. The bottom part of the table provides further details about the evaluation setup and models' performance on these tasks. Overall, this table seems to evaluate the Centurio model against several others in terms of its ability to perform well on multiple NLP tasks.

`Centurio Qwen`: Here a summary:
The best model is Centurio Aya.
The worst model is Palo 7B.
The average score of all models is around 50%.
The most accurate model (Centurio Aya) has an accuracy over 63% in all tasks, while the least accurate one (Palo 7B) has an accuracy under 28%.
In multilingual results, the best model is also the most accurate one and vice versa.
The most accurate model has an overall language fidelity over 94%, whereas the least accurate one has an overall language fidelity below 13%.
All models have an input supported AVG. above 40%.
The highest number of errors made by any model was 128, which occurred on the Inten/VL 2.5 SB task.
The lowest number of errors made by any model was 1, which occurred on the MiniCPM 2.6 task.

What model is best? Explain why.

`Centurio Aya`: Centurio! QEN, with an AVG., MAMM/U 32.0 and C.QAL 41.8. It has the highest average over all tasks compared to other models listed in table 6, indicating its superior performance across multiple linguistic tests.

`Centurio Qwen`: Centurio Aya Centurio'Queen. It has the highest accuracy score (65.7) and lowest error rate (20.3).

Figure 5: `Centurio` struggle with images containing small text.

What is written on this bottle?

`Centurio Aya`: The text on the bottle, in German, reads "Sasxia Natürliches Mineralwasser" which translates to "Sasxia Natural Mineral Water." Below that, it says "Wort am Rhein," likely referring to a specific source or brand name associated with the Rhine River. The phrase "ohne Kohlensäure" indicates that the water contains no carbonation. There's also some additional information such as size ("0,5L") and possibly an expiration date or batch number at the bottom of the label, but these are not clearly visible in the image provided.

`Centurio Qwen`: The text on the label of the bottle reads "Saskia," followed by a description in German that translates to "Natural Mineral Water" and "Mineral water from Württemberg am Rhein." Below, it states "still," indicating the type of mineral water. The bottom part mentions "ohne Kohlenstoffe," which means "without carbonation," confirming that the water is still (non-carbonated). Additionally, there's a logo or emblem above the brand name "Saskia," but due to the angle and focus of the image, specific details about the logo cannot be discerned. There are also some numbers at the very top right corner of the label, likely related to product information such as batch number or size, but they are not clearly visible. The overall design suggests a natural, possibly organic origin for the mineral water. However, without clearer visibility of all elements, more detailed information would require further examination or context. If you're looking for precise nutritional facts or chemical composition, those specifics might not be available from the visual content alone. For accurate information regarding the contents or any health-related aspects of the beverage, consulting the official product packaging or manufacturer's website would be advisable.
Please note: This analysis focuses solely on what can be seen within the provided image and does not include speculative or unverifiable claims beyond what is directly observable. Any additional interpretation goes beyond the scope of the given visual data.
If you need assistance with something else or have another question, feel free to ask!
In summary:
Brand Name: Saskia Type: Natural Mineral Water Origin: Württemberg am Rhein Carbonation Status: Still (Non-Carbonated) Additional Information: Without Carbon Dioxide (Kohlensäure) Label Design Suggests Organic/Healthy Origin Specific Nutritional Facts or Chemical Composition Not Clearly Visible in Image
For comprehensive details, refer to the actual product packaging or manufacturer's resources.

Figure 6: `Centurio` summarizing the label of a water bottle. While mostly correct, there are some OCR mistakes, e.g., with the city name ('Wörth')

Opišite sliku što detaljnije. (*Describe the image in as much detail as possible.*)

`Centurio Aya`: Slika prikazuje mirno i živopisno prizoru vodopada koji se nalazi u bujnim zelenim krajoliku. Vodopad, s više padina, stvara dinamičnu scenu sa čistim plavom vodi koja kontrastira sa okolnom zelenilom. Okolište je gusto naseljen drvećem, a nekoliko oblaka može se vidjeti na nebu, dodajući dubinu sceni. (*The image depicts a peaceful and picturesque scene of a waterfall set in a lush green landscape. The waterfall, with multiple slopes, creates a dynamic scene with its clear blue water contrasting with the surrounding greenery. The surroundings are densely populated with trees, and a few clouds can be seen in the sky, adding depth to the scene.*)

`Centurio Qwen`: Slika prikazuje mirnu vodopadsku scenu, karakterizirana svojim više padavina koji se spuštaju u čisto plavo-zeleno jezero ispod njih. Voda izgleda glatko i neometano, sugerirajući dubinu i čistost. Okolišna vegetacija je bogata i živahna, ukazujući na dobro održavan ekosistem sa obiljem svježe vode. Nebo je jasno i plavo, s minimalnim oblaci, što sugeriše sunčan dan bez vidljive atmosferske promene ili udaljenog događaja. Sama vodopada nema ljudskih elemenata kao što su mostovi, klupi ili barijeri, što bi moglo značiti da je lokacija priloživa i možda nepovremena za putnike. (*The image depicts a tranquil waterfall scene, characterized by its multiple falls cascading into a clear blue-green lake below. The water appears smooth and undisturbed, suggesting depth and purity. The surrounding vegetation is rich and vibrant, indicating a well-maintained ecosystem with an abundance of fresh water. The sky is clear and blue, with minimal clouds, suggesting a sunny day with no visible atmospheric change or distant event. The waterfall itself has no human elements such as bridges, benches, or barriers, which could indicate that the location is approachable and perhaps uninhabited by travelers.*)

Figure 7: `Centurio` describing an image in Croatian. Translations by Google Translate are provided in brackets for the reader.

Cine este prezentat în acest portret? (*Who is shown in this portrait?*)

`Centurio Aya`: Portretul prezintă un om care pare a fi o figură istorică sau fictivă, probabil de rang înalt datorită îmbrăcămintei sale regale și a bijuteriilor. Stilul picturii sugerează că ar putea fi din perioada Renașterii sau barocă, având în vedere detaliile meticuloase ale costumului și tehnica picturilor cu ulei. Fără context suplimentar, nu este posibil să se identifice persoana exactă reprezentată, dar hainele și accesoriile sunt caracteristice celor care au deținut poziții semnificative în istoria europeană. (*The portrait depicts a man who appears to be a historical or fictional figure, likely of high rank due to his regal attire and jewelry. The style of the painting suggests that it may be from the Renaissance or Baroque period, given the meticulous detail of the costume and the oil painting technique. Without further context, it is not possible to identify the exact person depicted, but the clothing and accessories are characteristic of those who held significant positions in European history.*)

`Centurio Qwen`: Persoana din portret pare a fi un om de natare, probabil o figură istorică sau nobilă, având în vedere îmbrăcămintea și simbolismul capului. În mod specific, el poartă o coroană cu modele complicate, care ar putea indica putere sau statut. Stila îmbrăcămintei sugerează că ar putea fi din secolul al XVIII-lea, ceea ce indică faptul că figura poate fi asociată cu această epocă. Fără context suplimentar, nu este posibil să se determine identitatea exactă, dar caracteristicile sale distincte sugerează că ar putea fi un lider politic, un regiune sau un alt individ important dintr-o perioadă istorică. Portretul servește pentru a transmite autoritatea și statusul sa prin utilizarea elementelor visuale ale îmbrăcămintei și accesoriilor. Este important de menționat că, deși detaliul exact al persoanei este neclar, natura portretului sugerează că ea are un rol semnificativ în istoria respectivă. Portretul servește pentru a transmite autoritatea și statusul sa prin utilizarea elementelor vizuale ale îmbrăcămintei și accesoriilor. (*The person in the portrait appears to be a man of character, perhaps a historical or noble figure, given the clothing and symbolism of the head. Specifically, he is wearing an intricately patterned crown, which could indicate power or status. The style of the clothing suggests that it may be from the 18th century, indicating that the figure can be associated with this era. Without additional context, it is not possible to determine the exact identity, but his distinctive features suggest that he may be a political leader, a regional leader, or another important individual from a historical period. The portrait serves to convey his authority and status through the use of visual elements of his clothing and accessories. It is important to note that while the exact detail of the person is unclear, the nature of the portrait suggests that he has a significant role in the history in question. The portrait serves to convey his authority and status through the use of visual elements of his clothing and accessories.*)

Figure 8: `Centurio` answering a question in Romanian at length. Still, neither model correctly identifies the famous portrait of Vlad III Dracula and both models are wrong with their guess of creation time (16th century). Translations by Google Translate are provided in brackets for the reader.