Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Active Clarification in Human-Robot Interaction via Object State Perception and Reasoning

Dissertation

submitted to the University of Hamburg
with the aim of achieving a doctoral degree at
the Faculty of Mathematics, Informatics and Natural Sciences,
Department of Informatics

Xiaowen Sun
Hamburg, 2026

Submission of the dissertation:
16.02.2026

Day of oral defense:
18.06.2026

Dissertation Committee:

Prof. Dr. Stefan Wermter (reviewer, advisor)
Dept. of Computer Science
University of Hamburg, Germany

Prof. Dr. Jianwei Zhang (reviewer, deputy chair)
Dept. of Computer Science
University of Hamburg, Germany

Prof. Dr. Chris Biemann (chair)
Dept. of Computer Science
University of Hamburg, Germany

Abstract (English)

Embodied agents have benefited from advances in Artificial Intelligence (AI), enabling them to operate in human-centered environments such as domestic or restaurant settings. In such environments, Human-Robot Interaction (HRI) is characterized by Uncertainty, Vagueness, and Ambiguity, particularly when interaction refers to fine-grained object states (e.g., clean vs. dirty) rather than object categories alone. This increased granularity introduces significant challenges for embodied agents in integrating language, perception, and action.

This thesis addresses these challenges by investigating how embodied agents can acquire and process object states for robust communication and interaction. It begins with a comprehensive review of existing approaches from an architectural perspective, synthesizing recent trends that have advanced the development of communication models, conversational agents, and transformer-based language and vision-language models. Building on this foundation, the thesis first clarifies the concepts of uncertainty, vagueness, and ambiguity within the interdisciplinary domains of linguistics and HRI, and then discusses object states. This analysis motivates four methodological contributions, each addressing different combinations of the identified conceptual, representational, and methodological problems.

First, a novel HYbrid Neural Architecture (HYNA) is introduced to advance visually grounded human–robot dialog by resolving ambiguities between visual context and natural language. Next, the Object State-Sensitive Agent (OSSA) is proposed to evaluate and improve the performance of large language models (LLMs) and vision-language models (VLMs) in robotic task planning. OSSA incorporates object state representations and commonsense knowledge to support long-horizon instruction execution in dynamic environments, including user-adaptive and multimodal interaction scenarios. Additionally, the Dialog Bridge mechanism is formulated to assess LLMs' capabilities in user age-sensitive adaptation and multimodal processing in HRI. Finally, the State-sensitive Vision-Language Model (StateVLM) is presented to enhance the numerical reasoning capabilities of VLMs. StateVLM is improved through an efficient fine-tuning strategy with an auxiliary regression loss, enabling better object localization and affordance reasoning at the object state-level with limited annotated data and computational resources.

Overall, this thesis advances embodied intelligence by integrating knowledge-driven reasoning with visual perception and language models, thereby improving communication and visual understanding for embodied agents operating in dynamic real-world HRI scenarios.

Zusammenfassung (Deutsch)

Verkörperte Agenten profitieren erheblich von Fortschritten in der Künstlichen Intelligenz (KI), die es ihnen ermöglichen, in Umgebungen mit Menschen, wie Haushalten oder Restaurants, zu agieren. In solchen Kontexten ist die Mensch-Roboter-Interaktion (Human-Robot Interaction, HRI) durch Unsicherheit, Unschärfe und Mehrdeutigkeit geprägt, insbesondere wenn die Interaktion von fein-granularen Objektzuständen (z.B. sauber vs. schmutzig) abhängig ist, und nicht lediglich von groben Objektkategorien. Diese feinere Granularität stellt verkörperte Agenten vor erhebliche Herausforderungen bei der Integration von Sprache, Wahrnehmung und Handlung.

Die vorliegende Arbeit beschäftigt sich mit diesen Herausforderungen, indem sie untersucht, wie verkörperte Agenten Objektzustände wahrnehmen und verarbeiten können, um eine robuste Kommunikation und Interaktion zu gewährleisten. Sie beginnt mit einer umfassenden Übersicht über bestehende Ansätze aus Software-Architektur-Perspektive und fasst aktuelle Entwicklungen zusammen, die durch weiterentwickelte Kommunikationsmodelle, konversationale Agenten sowie transformerbasierte language- und vision-language modelle ermöglicht werden. Darauf aufbauend werden zunächst die Konzepte von Unsicherheit, Unschärfe und Mehrdeutigkeit im interdisziplinären Kontext von Linguistik und HRI präzisiert, bevor auf feingranulare Objektzustände eingegangen wird. Diese Analyse bildet die Grundlage für vier methodische Beiträge, die jeweils unterschiedliche Kombinationen der identifizierten konzeptuellen, repräsentationalen und methodischen Herausforderungen adressieren.

Zunächst wird die neuartige HYbrid Neural Architecture (HYNA) vorgestellt, die den visuell gestützten Mensch-Roboter-Dialog verbessert, indem sie Mehrdeutigkeiten zwischen visuellem Kontext und natürlicher Sprache auflöst. Darauf aufbauend wird der Object State-Sensitive Agent (OSSA) vorgestellt, um Large Language Models (LLMs) und Vision-Language Models (VLMs) für die Aufgabenplanung eines Roboters zu evaluieren und deren Performanz zu erhöhen. OSSA integriert die Repräsentation von fein-granularen Objektzuständen mit Allgemeinwissen, um längerdauernde Aufgaben in dynamischen Umgebungen zu unterstützen - einschließlich nutzeradaptiver und multimodaler Interaktionsszenarien. Zusätzlich wird der Dialog Bridge Mechanismus entwickelt, um die Fähigkeiten von LLMs zu evaluieren, sich an den Benutzer altersgemäß anzupassen und multimodale Verarbeitung im Rahmen von HRI zu leisten. Schließlich wird das State-sensitive Vision-Language Model (StateVLM) vorgestellt, das die numerischen Fähigkeiten von VLMs verbessert. StateVLM profitiert von effizientem Finetuning mit einer zusätzlichen Regressions-Zielfunktion, wodurch eine präzisere Objektlokalisierung und Affordanzen-Erkennung unter Beachtung der feingranularen Objektzustände ermöglicht wird, wobei wenige annotierte Daten und Rechenressourcen nötig sind.

Insgesamt leistet diese Arbeit einen Beitrag zur Weiterentwicklung verkörperter Intelligenz, indem wissensbasiertes Argumentieren mit visueller Wahrnehmung und

Sprachmodellen integriert wird und somit die Kommunikation sowie das visuelle Verständnis verkörperter Agenten in dynamischen, realen HRI-Szenarien verbessert werden.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Daily life typically encompasses multiple users, such as grandparents, parents, and children. Developing an intelligent robot to assist in their everyday activities raises several important challenges that must be carefully considered. To further explore these challenges, we use a simple, representative common everyday scenario to analyze and discuss it. As shown in Figure 1.1, a humanoid intelligent robot stands in the living room with a mother, a grandfather sitting on the sofa, and a child sitting on the ground.

The child's linguistic and cognitive abilities are still developing, and their goal-oriented requests are often unspecified or lack clear intent. For instance, in this scenario, the child may only be able to refer to some objects by color rather than by their specific names. In such cases, the robot should be able to recognize and resolve referential ambiguity.

In contrast, the mother and grandfather possess well-developed linguistic and cognitive abilities. Their requests are generally goal-oriented and imply an expectation that the robot can autonomously execute the required tasks. In daily life, humans interact with objects at the state level. When clearing the table, the mother would store the intact apple in the cabinet and usually discard leftovers, such as a bitten apple, in the trash bin. However, the grandfather would also store the intact apple in the cabinet and may instead keep leftovers, such as a bitten apple, in the refrigerator. Therefore, the robot should be able to comprehensively understand long-horizon instructions that involve preferences and interact with objects at the same level as humans.

Furthermore, the grandfather may prefer the robot to announce its actions, such as its movements, in order to avoid potential collisions during long-horizon task execution. The mother is not interested in the process, as she possesses sufficient situational awareness and reaction capability to avoid collisions with the robot.

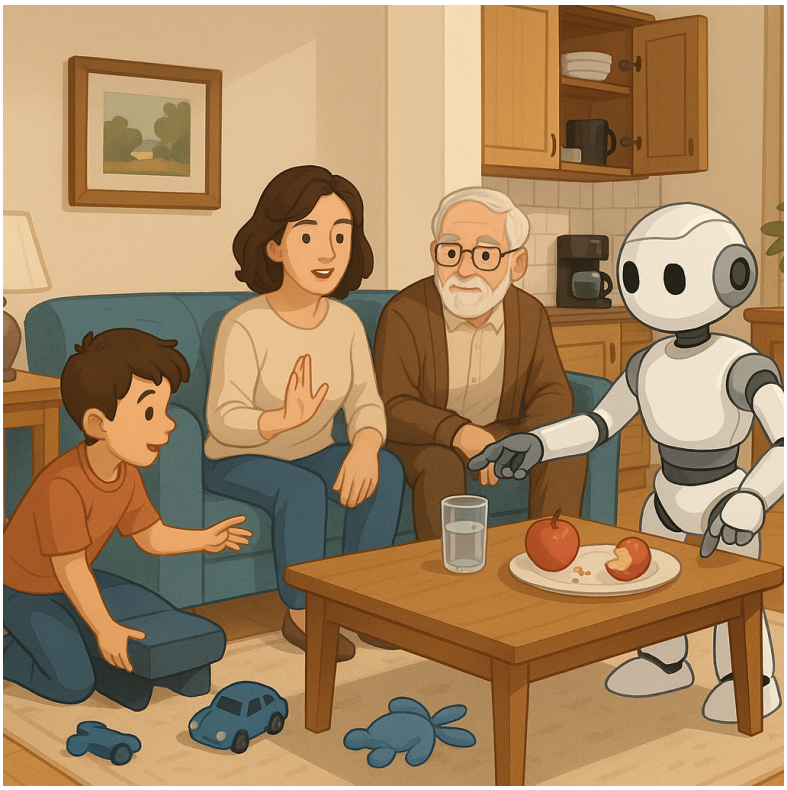


Figure 1.1: A common everyday life scenario involving a humanoid intelligent robot: a humanoid intelligent robot is involved in a simple, representative everyday scenario. It stands in the living room with a mother, a grandfather sitting on the sofa, and a child sitting on the floor. Source: AI-generated via ChatGPT.

Overall, to develop an intelligent robot that is capable of living in harmony with and assisting humans, it is essential to enable the robots to comprehend human linguistic requests accurately, and perceive their surroundings precisely, then act in ways that reflect awareness and adaptability of the dynamic environment, and give feedback to humans. Therefore, communication, visual perception, decision-making, and execution are the fundamental abilities required to accomplish the mission outlined above.

In this dissertation, we investigate methodologies for enhancing communication and visual perception capabilities in household robots. We focus on identifying communication challenges arising from Uncertainty, Vagueness[1], and Ambiguity in common everyday scenarios. By analyzing the underlying sources of these challenges, we develop frameworks designed to detect, clarify, and resolve them as they arise. Furthermore, we aim to facilitate the development of human-like, object state-level interactions that endow robots with advanced visual perception capabilities, thereby supporting effective HRI.

[1]We clarify here that our focus is on the linguistic and human-robot interaction (HRI) aspects of vagueness, as "Fuzziness" and "Vagueness" are often confused in the academic literature.

1.2 Research Objectives

Viewing a household robot as coexisting harmoniously with and assisting humans, the human user assumes the role of specifying goals and intentions through language. The robot, in turn, is responsible for interpreting these instructions and autonomously executing the corresponding tasks. To fulfill the user's intentions, the robot must follow human instructions while also interacting autonomously with its environment to determine and execute appropriate actions. Communication [28, 239], visual perception [255, 228], interaction [259, 155, 313], and execution [53, 156] are therefore essential cognitive capabilities for a robot to achieve this overarching goal.

Visual perception serves as a bridge between the robot's physical embodiment and its cognitive processes, enabling it to understand and interact with its surroundings. Humans typically perceive and interact with the environment at the object state-level, such as distinguishing between a clean plate and a dirty plate. To align the robot's internal representations with the user's mental model, the robot must likewise perceive and represent the environment at the same level.

This state-level alignment is equally critical in communication, which serves as a bridge between the user and the robot. Uncertainty, Vagueness, and Ambiguity (UVA-phenomena) frequently arise in communication due to factors such as the user's linguistic limitations, the robot's perceptual or reasoning capabilities, and the complexity of the environment. These challenges are particularly pronounced when communication refers to fine-grained object states (e.g., clean vs. dirty) rather than object categories alone. In this thesis, UVA-phenomena is used as a concise term to denote these distinct but partially overlapping concepts.

A more fundamental challenge is that the concepts of UVA-phenomena remain poorly defined, characterized by mixed terminology, uneven emphasis, and the absence of a unified conceptual framework. The literature exhibits substantial inconsistency in how key terms are defined and used; in some studies, ambiguity and uncertainty are treated interchangeably, resulting in conceptual confusion and obscuring the precise nature of the problem. For instance, Senanayake [250] explicitly states that "uncertainty, also known as ambiguity, stems from the stochasticity of the world."

Several studies regard the lack of clear or complete information about possible outcomes as an ambiguity problem [44, 174, 202, 130, 75, 57, 225]. In contrast, other studies frame the issue primarily in terms of uncertainty [279, 108, 144, 248]. Vagueness, however, is often overlooked entirely or implicitly folded into ambiguity or uncertainty. To the best of our knowledge, the collaborative behavior-based approach of Wang et al. [290] is the only study that explicitly considers UVA-phenomena simultaneously in the context of service robots. This study underscores the importance of representing human behaviors, robot behaviors, and object states in ways that support reasoning under all three forms of indeterminacy. However, this study focuses solely on natural language sentence parsing, thereby overlooking the

dynamics and complexity of the environment. **Accordingly, further discussion of uncertainty, vagueness, and ambiguity is required to achieve conceptual clarity, develop a rigorous academic formulation, and address the problem.**

On the other hand, transformer-based methods have recently revolutionized artificial intelligence (AI), particularly in the fields of natural language processing (NLP) and computer vision (CV). A key shift in contemporary methodology has been the move from rule-based systems and single-task models toward general-purpose models. Instead of designing task-specific, hand-crafted models or training independent deep learning (DL) models for individual tasks, researchers now fine-tune pretrained general models, such as large language models (LLMs) and vision-language models (VLMs), for specialized applications. These advancements provide remarkable capabilities, including linguistic processing, visual perception, and commonsense knowledge, which enable us to tackle more complex and practical problems in embodied agents. Such problems include commonsense reasoning, long- and short-horizon task planning for object manipulation and navigation, as well as decision-making.

This thesis investigates methods to enhance the communication and visual perception capabilities of embodied agents for domestic tasks by training a deep neural network (DNN) or leveraging pretrained foundation models. It specifically focuses on object state-level perception, which often gives rise to increased Uncertainty, Vagueness, and Ambiguity (UVA-phenomena) in HRI.

With this overarching goal in mind, we further targeted the following concrete objectives while considering the different factors. Considering that children's linguistic and cognitive abilities are still developing, their requests within goal-oriented requests are often less defined or explicit. This leads us to formulate our first objective.

Objective 1: Develop a hybrid neural architecture that integrates visual content into a conversational agent, enabling a domestic robot to resolve linguistic and visual ambiguities in situated HRI.

In contrast, adults, with their well-developed linguistic and cognitive abilities, typically make highly goal-oriented requests. These imply an expectation for the robot to autonomously execute tasks. Adults may provide complex, long-horizon instructions and hold varied personal preferences. Therefore, we propose our second objective.

Objective 2: Design and evaluate an effective approach to leverage foundation models for enabling a robot to recognize objects, interpret user instructions, and generate contextually appropriate object manipulation plans at the object state-level, while aligning with commonsense knowledge and user preferences. This approach is expected to enable the domestic robot to autonomously complete both short- and long-horizon tasks, while remaining robust in the face of ambiguous situations.

Regarding younger and older groups, users may expect different behaviors from the robot, even when performing the same task. We therefore formulate our third objective.

Objective 3: Develop and formalize a mechanism leveraging foundation models for age-sensitive and multimodal processes in HRI. This mechanism is expected to connect users and the robot, enabling the robot to behave differently by default for different user groups while allowing users to modify their preferences.

Beyond adapting to users' linguistic and cognitive abilities, embodied agents must also contend with the limitations of their own foundational models. Such models, often trained primarily for language tasks, lack inherent capabilities for numerical prediction, which is essential for the object manipulation of embodied agents. To address this core adaptation challenge, we define our fourth objective.

Objective 4: Establish an efficient strategy for modifying general models, large VLMs, to perform object localization and, further, affordance reasoning at the object state-level, all within constrained computational resources and using open-source or small-scale general models.

1.3 Research Questions

Since this thesis commenced before the methodological turning point that shifted the field from training independent DL models for individual tasks toward leveraging or fine-tuning pretrained general models for specialized applications, our research covers both methodological paradigms. Before this methodological turning point, the community primarily focused on end-to-end neural dialog for conversational agents [285, 180, 343, 341, 170]. However, these approaches were limited by their lack of visual content. Conversely, transformer-based vision-language models such as VilBERT [182], UNITER [40], and VL-BERT [265] were developed for visual understanding and reasoning but lack conversational capabilities. Although datasets like GuessWhat?! [59], Visual Dialog [56], and CLEVR-Dialog [140] address dialog challenges, they remain either overly restricted (e.g., binary yes/no) or synthetic and unnatural. In this context, a hybrid framework becomes attrac-

tive for task-oriented dialog modeling, as it can integrate the strengths of multiple approaches [87, 306]. Inspired by [306, 139], we pose the **Research Question 1** to address the **Objective 1**.

> **Question 1**: To what extent can a hybrid neural architecture enable the conversational agent to recognize and resolve ambiguity from the reciprocity of visual content and natural language instructions?

Recently, general-purpose foundation models, particularly LLMs and VLMs, have come to dominate the domains of conversational agents and visually grounded dialog, owing to their exceptional performance and, in some cases, superhuman capabilities. Consequently, our focus has shifted from merely discussing objects to leveraging these models for embodied agents, such as those engaged in object manipulation planning.

In everyday scenarios, human interaction with objects is fundamentally governed by object states. These states are determined by object properties such as color, shape, size, weight, texture, cleanliness, and rigidity. Accurately understanding and managing object states is essential for robust manipulation, as misinterpreting an object's state can lead to unintended or hazardous outcomes. As a result, object states exert a substantial influence on manipulation strategies. Despite their importance, however, incorporating state-sensitive knowledge into robotic systems remains a significant challenge.

Prior research has primarily addressed object state recognition [90, 89, 136] and the anticipation of state changes, as well as object category-level manipulation and task planning [309, 109, 342, 345, 257]. Although few studies have explored object state manipulation for deformable objects, such as clothing, within the fashion domain [158, 263], the use of general-purpose foundation models, including LLMs and VLMs, for object manipulation and task planning at the object state-level remains largely unexplored. To address this gap, we introduce **Research Question 2**, which directly supports **Objective 2**.

> **Question 2**: What architectural paradigms, a modular method with a perception and reasoning module, or a monolithic method, best enable agents to perceive fine-grained object states and leverage this understanding to generate contextually appropriate manipulation plans that align with user preferences in complex environments?

Previous research has not considered the relationship between the user's age and the robot's state. For example, Zhang et al. [336] investigated the potential of LLMs as zero-shot human models in HRI. Similarly, GPT-3 has been integrated with the Aldebaran Pepper and NAO robots, enabling participants to chat with them [19]. Their goal is to examine the possibilities and limitations of LLM in an alive HRI system. However, these studies overlook the fact that diverse user groups may

expect different robot behaviors, even when performing the same task. To further explore this problem, we propose an adaptive ROS-based framework [221] with an open-source code base for HRI involving diverse user groups, enabling the robot to interact with users via natural language speech. Users can interrupt the robot at minor and major levels. For the core aspect of multimodal processes, we formulate **Research Question 3** to investigate **Objective 3**.

Question 3: How can we leverage the language-processing power of LLMs to interpret multimodal inputs from the robot's symbolic planner and the user?

Object localization is essential for object manipulation. However, numerical prediction remains challenging for monolithic VLMs. The existing VLMs tokenize object locations into discrete text tokens, often wrapped with special start and end markers. These approaches [166, 38, 37, 12, 338, 226, 325, 339, 293] treat continuous numerical values as discrete symbols, which introduces an abstraction that departs from visual reality. To mitigate this mismatch, NExT-Chat [335] incorporates a dedicated box decoder to produce continuous location outputs. Nevertheless, its performance remains slightly below that of Shikra-7B [38], despite relying on a similar detection-oriented training recipe. The authors attribute this gap to two factors: (1) the difficulty of balancing token-classification and numerical-regression losses, and (2) the fact that the core LLM within the VLM is not pretrained for regression tasks. Building on these insights, we articulate **Research Question 4** to guide **Objective 4**:

Question 4: To what extent can VLMs be fine-tuned for numerical prediction tasks, incorporating an auxiliary regression loss to improve performance, while continuing to infer within the original VLMs framework?

1.4 Overview of Novel Contributions

This thesis makes two directional contributions to the field of multimodal interaction in embodied AI: **Synthesis and Foundations** and **Methodological Advancements**.

Synthesis and Foundations: We establish comprehensive theoretical foundations for conversation, communication, and perception in embodied agents. We highlight two aspects of contributions: (1) reviewing foundations and (2) providing conceptual clarification.

(1) We review dialog systems from an architectural perspective (Section 2.2) as well as transformer-based language (Section 2.3) and vision-language (Section 3.3) models from the perspective of training objectives.

(2) We provide conceptual clarification on UVA-phenomena as they pertain to embodied agents (Section 3.5.1), alongside a discussion of the object state for HRI (Section 3.5.2).

Methodological Advancements: We propose four novel methodologies designed to enhance the capabilities of embodied agents performing household tasks. These methods are informed by and address the core challenges formalized in our review, specifically the modeling of UVA-phenomena and reasoning about object states in the context of HRI.

Chapter 4 (Learning Visually Grounded Human-Robot Dialog in a Hybrid Neural Architecture): We propose a novel architecture, HYbrid Neural Architecture (HYNA), for visually grounded human-robot dialog learning. HYNA comprises an object detection (OD), a multimodal dialog state tracker (M-DST), and a Human-robot **i**nteraction Policy (HiPolicy). We introduce a neural dialog model for the M-DST that processes user input based on visual inputs and dialog history. The model uses bag-of-words (BoW) for its simplicity, utterance embedding (UE) due to their successful application in other natural language tasks, and a combination of both (BoW+UE) as multimodal[2] representations. We train the M-DST on a new visual dialog dataset, evaluate different forms of input representations, and validate the model on unseen examples. We further assess its performance in handling HRI scenarios. We demonstrate that HYNA can handle a wide range of unseen visual inputs and verbal instructions. Finally, we deploy the proposed architecture on the Neuro-Inspired COmpanion (NICO) robot. HYNA satisfies **Objective 1** by integrating visual perception into a conversational agent. The results show that using BoW to represent multimodal inputs for the dialog state tracking model is an effective approach for recognizing ambiguous instructions and scenarios, thereby addressing **Research Question 1**.

Chapter 5 (Object State-Sensitive Neurorobotic Task Planning): We introduce a task-planning agent, Object State-Sensitive Agent (OSSA), which is empowered by pretrained foundation models. We propose two methods for OSSA: (i) a modular model consisting of a pretrained vision processing module (dense captioning model, DCM) and a natural language processing model (LLM), and (ii) a monolithic model consisting only of a VLM. To quantitatively evaluate the performance of the two methods, we use tabletop scenarios where the task is to *clear the table*. We contribute a multimodal benchmark dataset that takes object states into consideration. OSSA aims to enable the domestic robot to autonomously complete long-horizon tasks and remain robust in the face of ambiguous, vague, and uncertain situations, which meets **Objective 2**. Our results demonstrate that both methods can be used for object state-sensitive tasks, but the monolithic approach outperforms the modular approach, which answers **Research Question 2**.

[2]The spelling "multimodal" (unhyphenated) is used consistently throughout this thesis. Contemporary literature in machine learning (ML) favors this form, though the hyphenated "multi-modal" appears in older or parallel research. The terms are functionally equivalent.

Chapter 6 (Multimodal Language Processing for Cognitive Systems): We propose a dialog bridge that connects the user and the robot within an adaptive ROS-based framework [221]. A prompted LLM serves as this dialog bridge, processing information and generating responses between the two. To quantitatively evaluate the dialog bridge, we constructed a Command and Object Properties Recognition (Cooper) task. The dialog bridge's individual evaluation shows that the pretrained foundation model GPT-3.5 can extract the command and target properties for the robot. The trial evaluation of the ROS-based framework demonstrates that the LLM was able to distinguish the user's age and generate differentiated responses at the beginning of each interaction. However, this behavior diminished after several iterations due to the LLM's memory limitations. As new conversation turns were appended to the history, the LLM tended to forget the initial prompt. Therefore, we answer **Research Question 3**.

Chapter 7 (State-Sensitive Vision-Language Model for Affordance Reasoning): We propose StateVLM, a novel VLM designed to perceive fine-grained scene information, specifically objects and their corresponding states. To enhance representational balance, StateVLM incorporates an auxiliary regression loss that emphasizes spatial location prediction during fine-tuning. We systematically evaluate the influence of this auxiliary regression loss by conducting extensive experiments on referring expression comprehension (REC), a classic region-level image understanding task. Our evaluation spans three established REC benchmarks: RefCOCO, RefCOCO+, and RefCOCOg. Additionally, we contribute an open-source benchmark dataset for object state affordance reasoning, generated via a diffusion model to support further research in fine-grained visual reasoning. StateVLM achieves strong performance across both REC and object state affordance reasoning tasks. Our results demonstrate that the proposed auxiliary regression loss significantly improves model adaptation during fine-tuning, confirming its utility in enhancing spatial and state-sensitive visual understanding. Therefore, we answer the **Research Question 4**.

1.5 Structure of this Thesis

This dissertation is organized into seven chapters, with their interconnections illustrated in Figure 1.2. The scientific "we" convention is adopted throughout[3].

In Chapter 2, we evaluate conversational agents from an architectural perspective and examine the dominant transformer-based approach to language modeling. In Chapter 3, we review key dataset challenges and models in vision-language research and then detail two critical concepts for embodied agents: UVA-phenomena and the object-state representations.

In Chapters 4 to 7, we present our novel methodologies proposed for each objec-

[3]For transparency, AI tools, including ChatGPT and DeepSeek, were used only for grammatical review; all interpretations and contributions are the authors' own.

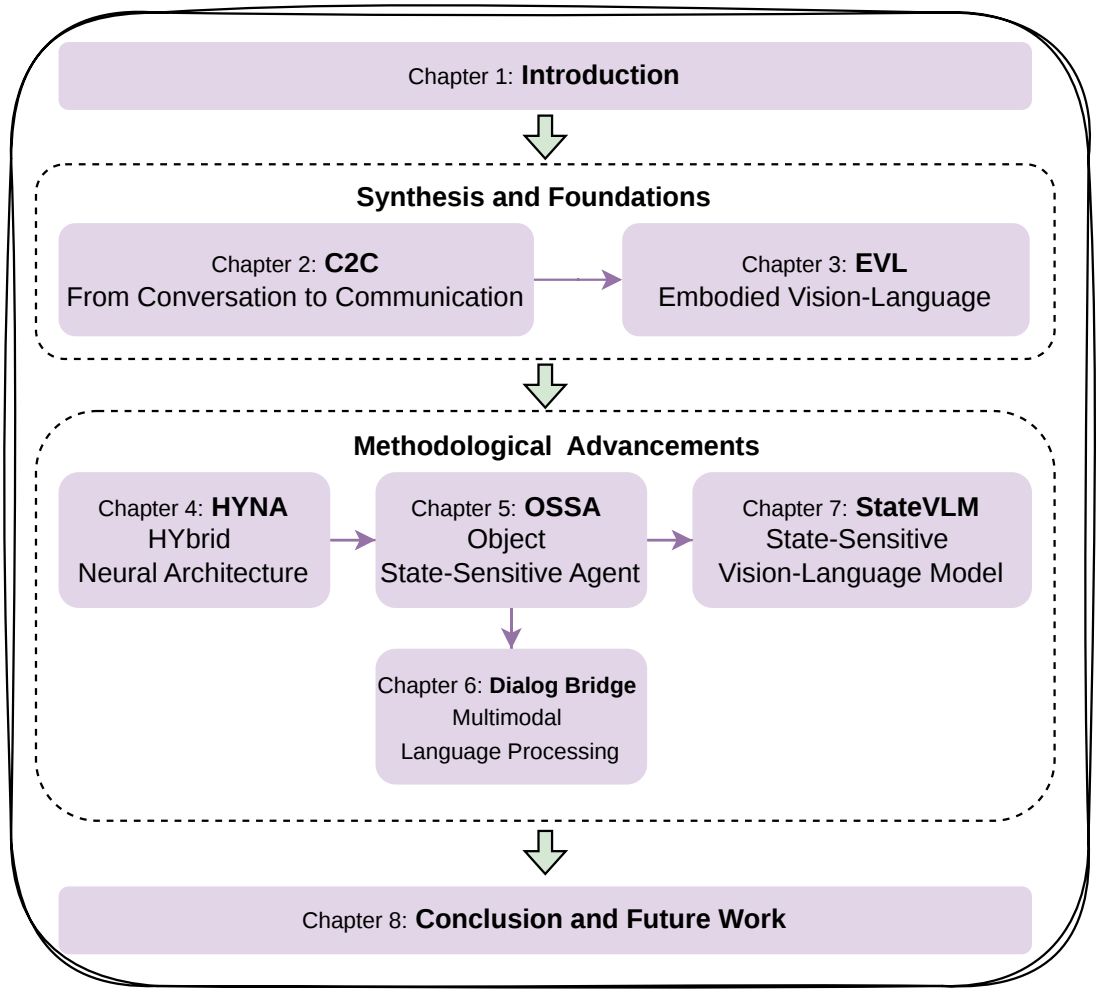


Figure 1.2: The structure of this thesis.

tive, along with the datasets generated for each goal, and analyze our experimental results in detail. Finally, Chapter 8 summarizes the key contributions of this dissertation, critically evaluates its strengths and limitations, and outlines directions for future research.

Chapter 2

From Conversation to Communication

Communication is a broad term describing the sharing of thoughts, ideas, information, and feelings between two or more individuals. Humans use primarily spoken words, written text, gestures, body language, or digital means to communicate [200]. A more specific form of communication is conversation, which is inherently interactive and reciprocal, typically carried out through language. Building on this concept, algorithmic computer systems are designed to simulate conversational interactions. These systems, which are able to interact with human users via text-based, verbal, or multimodal communication, are referred to as *dialog systems* or *conversational agents* [192]. In recent years, these systems have gained popularity and have been employed in a variety of domains, ranging from virtual to physical assistants.

In this chapter, we survey conversational agents in Section 2.1 from past to present, delve into the foundational principles underlying dialog system architectures in Section 2.2, highlight the current dominant language models, transformer-based models, in Section 2.3. Finally, we summarize the language-centric dialog system and generalize the concept of conversation within the broader context of communication, multimodal dialog systems in Section 2.4.

2.1 Conversational Agents: Past and Present

The earliest mainstream *virtual personal assistant*, Siri[1], which interacts with humans by natural language speech, was launched in 2011 by Apple. It can assist the user with multiple tasks: controlling the main functions of the phone, such as making phone calls, sending and reading messages, retrieving information from the internet, such as getting the updated weather and reporting stock prices, managing productivity tasks, such as making calendar appointments, writing and sending

[1] `https://www.apple.com/siri/`

emails, and supporting entertainment, such as playing music and chit-chat. Subsequently, many virtual personal assistants were launched to the market by internet commercial giants, such as Google Assistant[2], Microsoft's Cortana[3], Samsung's Bixby[4], and Amazon Alexa[5], and others.

Physical personal assistant smart speakers, such as Amazon Echo[6] and Google Nest[7], have become part of everyday life and are particularly specialized in home interactions, for example, controlling smart devices such as lights, thermostats, locks, cameras, and cookers. The Google Nest 2 and Amazon Echo Show even feature a screen that can play videos. Apart from engaging in non-task-oriented, free-form conversations with humans, these assistants also perform specific, task-oriented functions. Task-oriented dialog systems help humans achieve specific goals by allowing them to issue language commands to the assistant. RASA[8] and DeepPavlov[9] are software platforms that are being built for developers to develop task-oriented dialog systems. Overall, it is a fast-moving area that has attracted researchers in academia and industry.

However, solely language-based conversational agents lack the multimodal nature of human communication. Humans possess eyes for observing their environment, hands for tactile interaction and object manipulation, and legs for locomotion through space. Therefore, multimodal conversational agents have started to be investigated [23, 20, 49, 32]. Recently, there has been a growing body of research focused on integrating dialog systems into robots, for example, social robots (e.g., Pepper[10]) or cognitive robots (e.g., MiPA[11] and iCub[12]).

2.2 Architectures of Dialog Systems

Dialog systems, a subfield of natural language processing (NLP), are designed to simulate conversational interactions with human users across textual, verbal, or multimodal formats. The evolution of dialog systems can be traced through several major turning points, beginning with the cornerstone system ELIZA, as illustrated in Figure 2.1. ELIZA [300] is a *rule-based dialog system* that pioneered human-machine natural language communication. It represents an early prototype of the *traditional modular dialog system* [74, 288, 327], which was gradually formalized in later studies.

[2] https://assistant.google.com/
[3] https://www.microsoft.com/en-us/cortana
[4] https://www.samsung.com/global/galaxy/what-is/bixby/
[5] https://alexa.com
[6] https://www.amazon.com/smart-home-devices/b?ie=UTF8&node=9818047011
[7] https://store.google.com/magazine/compare_nest_speakers_displays
[8] https://rasa.com/
[9] https://deeppavlov.ai/
[10] https://www.softbankrobotics.com/
[11] https://neura-robotics.com/products/mipa
[12] https://icub.iit.it/

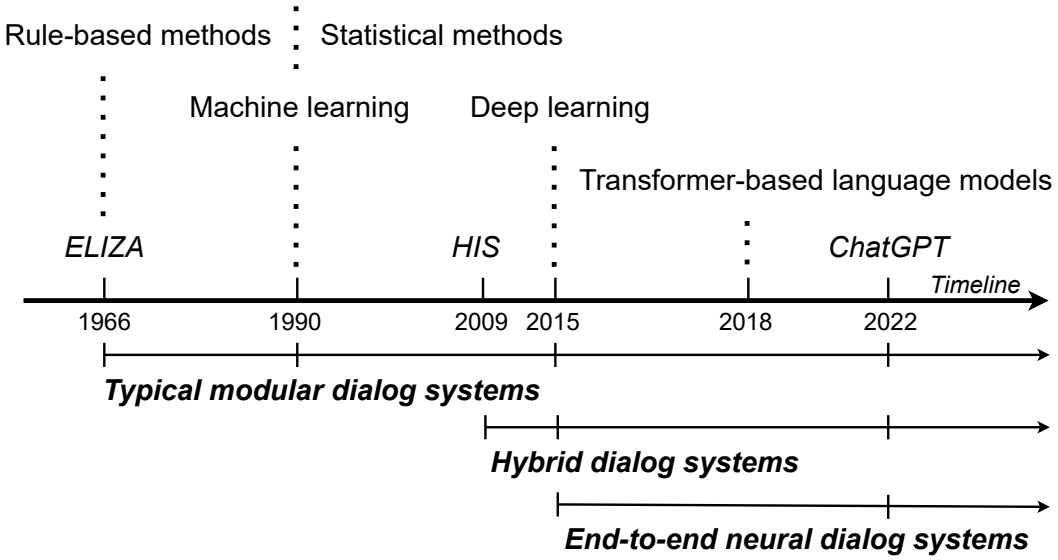


Figure 2.1: The timeline of dialog systems. ELIZA [300] is the earliest system in dialog research; Hidden Information State (HIS) [326] is a pioneering hybrid dialog architecture; ChatGPT [3] marks a major advancement in conversational agent technology.

The first major turning point was the introduction of probability theory into dialog systems, which transformed them from rule-based approaches to statistical ones. The early adoption of statistical methods relied on machine learning models such as support vector machines (SVMs) [249], partially observable Markov decision processes (POMDPs), dynamic decision networks (DDNs), and artificial neural networks (ANNs). These models primarily enhanced individual components of modular dialog systems. Therefore, the overall system architecture remained *modular*.

The second major turning point was the architectural innovation of *hybrid dialog systems*, exemplified by frameworks such as the Berkeley Restaurant Project (BeRP) [128], which combined neural acoustic-phonetic modeling with symbolic and probabilistic language components in a tightly interconnected pipeline. The Symbolic Connectionist Robust Enterprise for Natural Language (SCREEN) [303] integrated connectionist learning with symbolic rules for robust spoken language understanding. The Hidden Information State (HIS) [326] model fused structured representations with probabilistic reasoning via POMDPs to enable flexible belief-state tracking and decision-making. More recently, a significant milestone occurred with the release of ChatGPT[13] by OpenAI[14].

Subsequently, the emergence of big data and deep learning (DL) transformed dialog system architectures. Deep neural networks (DNNs) enabled sequence-to-sequence

[13] `https://www.chatgpt.com/`
[14] `https://openai.com/`

(Seq2Seq) learning, giving rise to what are known as *end-to-end neural dialog systems* [285]. In particular, transformer-based language models have come to dominate research on natural language processing and conversational agents.

Accordingly, this section presents a detailed discussion of dialog systems from an architectural perspective: *traditional modular dialog systems*, *end-to-end neural dialog systems*, and *hybrid dialog systems*, along with recent state-of-the-art (SOTA) studies.

2.2.1 Typical Modular Dialog Architecture

ELIZA [300] is a simple rule-based approach with pattern matching and substitution, which was launched in 1966. Since then, the era of rule-based dialog systems had started. As more research was carried out, it was gradually formalized as a typical structural system: modular dialog system architecture, which consists of: automatic speech recognition (ASR), natural language understanding (NLU), dialog manager (DM), knowledge sources (KSs), natural language generation (NLG), and text-to-speech (TTS) synthesis, as Figure 2.2 depicts. Among these components, the DM is the central component of the dialog system, like the brain for humans. NLU and NLG are necessary components for the system. ASR and TTS are not always essential, as their necessity depends on whether the dialog system involves spoken or text-based interaction. Requirements determine whether KSs are needed.

There have been two main phases in the history of modular dialog systems. In the first phase, dialog systems in academic and industrial research laboratories were developed using rules that defined each component. These components were then combined to determine the system's overall behavior. With the emergence of machine learning (ML) techniques, researchers have started to use statistical data-driven methods to optimize the components of the architecture, particularly the NLU, DM, and NLG components. We provide a detailed development of each component in the following sections and discuss its impact on the current study of conversational agents.

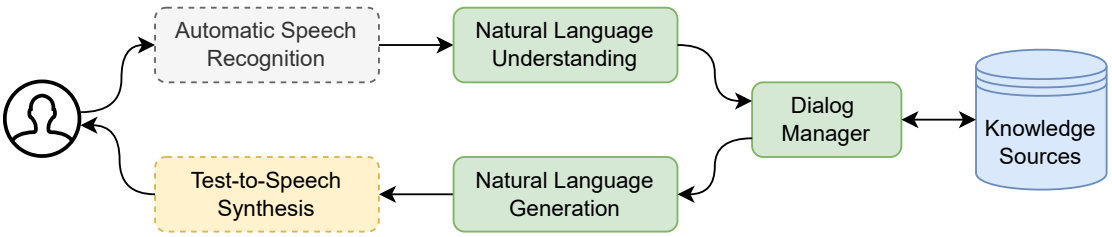


Figure 2.2: The typical modular dialog system architecture.

Natural Language Understanding (NLU)

The objective of the NLU component in a dialog system is to analyze the user's input utterance and extract a structured semantic representation of its meaning. This representation may include various semantic elements, such as *intent*, *entities*, *slots*, *dialog acts*, and *domain*. Specifically, the *intent* denotes the user's goal or action they wish to perform. *Entities* refer to the relevant parameters or pieces of information required to fulfill an intent. *Slots* are predefined templates for specific pieces of information that help complete an intent. The *domain* represents a specific topic area or task space, each associated with a distinct set of intents and slots. *Dialog acts* describe the communicative function of the utterance, such as a request, confirmation, or question.

In the early development of dialog systems, researchers primarily relied on *handcrafted grammars* and then built a parser to extract these semantic elements from user inputs. These *handcrafted grammars* encompassed a variety of approaches, including semantic grammars, context-free grammars, finite-state models, slot-filling frameworks, and domain-specific grammars. The selection of a particular method typically depends on the requirements and constraints of the specific application domain.

A *handcrafted grammar* [186] is supposed to contain all the rules required to cover all anticipated inputs, and then build a parser to apply the rules to the inputs. In practical application, however, it is impossible to guarantee that all possible inputs have been covered by the rules of the designed parser. If an input sentence does not match the rules exactly, it cannot be parsed, so small variations in the input require additional rules in order to be accepted by the designed parser. Furthermore, natural human language itself has an ambiguous characteristic; for example, one sentence can be understood in different ways because it can be structured differently grammatically, and each structure leads to a different meaning.

Later, probabilistic theory was introduced to *handcrafted grammars*, resulting in *probabilistic grammars* [127]. This approach enabled parsers to handle uncertainty and ambiguity. Probabilistic context-free grammars (PCFGs) [118] are a widely used type of *probabilistic grammar* in which a probability is associated with each production rule. These probabilities are typically learned from a corpus of parsed sentences. To enhance the accuracy of PCFGs, lexical heads can be incorporated, resulting in lexicalized PCFGs (L-PCFGs). When annotated data is limited, Bayesian priors can be employed in unsupervised grammar induction to learn grammatical structures, as in Bayesian PCFGs. Alternatively, when semantics take precedence, probabilistic extensions are applied to semantic grammars to improve intent recognition or slot filling, which are called probabilistic semantic grammars.

Reconceptualization defines the extraction of semantic elements, such as *domain*, *intent*, *entities*, *slots*, and *dialog acts*, from user inputs as classification problems. Consequently, statistical data-driven methods have been widely adopted in NLU. Compared to handcrafted grammar rules, ML approaches offer greater robustness

and scalability, which has spurred significant research in this domain. Traditional ML methods, such as SVMs [193] for intent recognition and conditional random fields (CRFs) for entity recognition [220], were among the earliest used. However, DL techniques have become dominant due to their superior performance. For instance, deep belief networks (DBNs) [247] have been employed for utterance classification and intent determination [247]. To improve accuracy, efficiency, and scalability, DBNs have been enhanced through integration with deep convex networks (DCNs) [61].

Convolutional neural networks (CNNs) [245] have also been used for utterance classification at first, but were replaced by recurrent neural networks (RNNs) in sequence-based tasks later. Variants of RNNs, including bi-directional RNNs [72], have shown strong performance in slot filling and are capable of jointly classifying intents and extracting entities [237].

Dialog Manager (DM)

As the central component of a dialog system, the DM accepts the interpreted inputs from the ASR or NLU components, interacts with external KSs, produces messages to respond to the user, and controls the dialog flow. The objective of DM is to decide the response action to the given user's input and realize the current state of the dialog.

In the era of rule-based dialog systems, the DM has been formalized into the dialog context and decision models. The dialog context model keeps track of information relevant to the dialog, including *dialog history*, *task record*, *domain model*, *model of conversational competence*, *user preference model*, in order to support the process of dialog management. The dialog decision model determines the system's reply action given the user's input utterance and the dialog-relevant information in the dialog context model. It is impossible to account for all possible interactions in a hand-crafted model due to the *uncertainties* present at every level of dialog.

Later on, statistical methods were introduced into dialog systems. Therefore, various probabilistic and neural models have been explored to model the dialog context and decision-making processes, which are termed as dialog state tracker (DST) and dialog policy model (DPM), respectively. The DST replaces the traditional rule-based context model, while a learned DPM takes the role of the rule-based decision model.

DST aims to represent the current dialog state (e.g., user goals, intents, and filled slots) typically using a belief state, which is a probability distribution over possible dialog states. For example, Hidden Markov Models (HMMs) capture the user's intent over time based on observed inputs [84]. Bayesian Networks perform probabilistic reasoning over user intent and slot values [280]. Conditional random fields predict sequences of state labels conditioned on observations [275]. RNNs [169] and transformers [34] are used to model the dialog state directly from the conversation history. Based on these states, the DPM determines the next system action (e.g.,

asking a question, confirming a slot, or making an API call to access the KSs). For example, ML methods such as SVMs [6] and ANNs [274] were trained as classifiers to map dialog states to system actions.

Another line of research in DM focuses on employing the Markov decision process (MDP) to model the dialog system [52]. A key limitation of this approach is the assumption within the MDP framework that the system state is fully observable. To more effectively account for uncertainties and ambiguities in users' goals and intentions, a POMDP is employed for dialog modeling. The POMDP framework can be combined with reinforcement learning (RL) to derive optimal policies that maximize an objective function defined over dialog trajectories [101, 199, 51]. More recently, neural policy networks have been used to learn response policies directly from dialog history [204, 205].

Natural Language Generation (NLG)

The objective of the NLG component is to produce natural language texts according to the output of the DM, enabling appropriate replies to user inputs, as shown in Figure 2.2. Similar to the NLU, NLG has evolved alongside advances in artificial intelligence (AI) technologies.

In the early stages, NLG received less attention, as it was considered a fairly trivial task, which uses handcrafted methods, involving either canned text or inserting retrieved items into pre-defined response templates. A more advanced approach frames generation as a pipeline process comprising content planning, determination, and realization [238]. Over time, NLG gained more attention as the quality of generated responses became essential to the user experience of the overall system. With the emergence of statistical methods, data-driven models were developed, enabling the generation of more natural utterances across large, real-world domains. However, many of these studies focused primarily on the realization level of generation [187, 214]. To better handle *uncertainty*, information presentation [241, 148] treats NLG as a planning problem under uncertainty, employing RL to optimize the dialog policy of the NLG component.

With the advancement of deep learning, its study in NLG has been extensively explored [235]. Text generation has expanded to a wide range of tasks, including summarization, text expansion for more detailed information, and revision for producing alternative expressions [64], rather than being limited to dialog systems and user response generation. However, this thesis focuses exclusively on generating responses to users. A comprehensive evaluation of template-based, rule-based, data-driven, and Seq2Seq methods found that Seq2Seq approaches perform best [71]. The Seq2Seq architecture is detailed in Section 2.2.3.

2.2.2 Hybrid Dialog Architecture

The widely deployed conversational agent ChatGPT is a highly sophisticated hybrid system that integrates multiple advanced components, including large lan-

guage models (LLMs) (such as GPT-4-turbo), multimodal language models (such as GPT-4o), diffusion models (for example, DALL-E 2), a toolformer architecture, function-calling capabilities, and output parsers, among others. Other contemporary conversational platforms can similarly be characterized as hybrid systems, including DeepSeek[15], Claude[16], LLaMA[17], and Gemini[18].

The notion of hybrid systems can be traced back to early studies [302] such as the Berkeley Restaurant Project (BeRP) [128] and Symbolic Connectionist Robust Enterprise for Natural Language (SCREEN) [303]. BeRP represents a classical hybrid spoken dialog system, combining neural acoustic-phonetic modeling with probabilistic language, semantic, and dialog components in a largely pipeline-based but tightly interconnected architecture. SCREEN combines data-driven learning, which allows syntactic and semantic structure to emerge from real spoken data, with connectionist networks, ensuring that analysis degrades gracefully rather than collapsing when speech is noisy, ungrammatical, or partially incorrect.

A landmark example of a hybrid dialog framework is the Hidden Information State (HIS) [326] model, which effectively integrates probabilistic modeling with structured, slot-based representations. By leveraging partially observable Markov decision processes (POMDPs), the HIS model enables approximate belief-state tracking and more flexible, data-driven decision-making.

Hybrid approaches have also emerged within neural network architectures. A representative example is the Mixture-of-Experts (MoE) layer [254], which combines multiple specialized expert networks through a learned gating mechanism. Initially applied to feed-forward networks for tasks such as language modeling and machine translation, MoE architectures were later scaled and incorporated into transformer models [150]. This evolution led to more efficient MoE variants within transformer feed-forward layers [76]. More recently, MoE techniques have been extended to universal transformers (e.g., MoEUT [50]) and large language models (e.g., DeepSeekMoE [54]).

Today, the term hybrid dialog architecture has broadened considerably. It generally refers to a system that integrates modules implemented using different technical methods, such as transformer-based large-scale models, rule-based components, traditional machine learning, and deep learning, to achieve robust, flexible, and scalable dialog behavior.

2.2.3 End-to-End Neural Dialog Architecture

In the early stages of DL emergence, researchers used DNNs to replace the component of the traditional modular dialog system [196], as we discussed in Section 2.2.1. Since its remarkable superior performance, it has become the domi-

[15] `https://www.deepseek.com/`
[16] `https://claude.ai`
[17] `https://www.llama.com/`
[18] `https://gemini.google.com/`

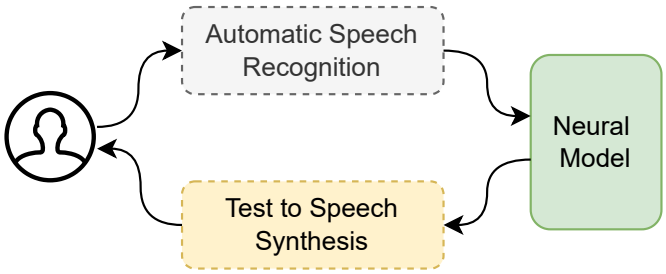


Figure 2.3: The end-to-end neural dialog system architecture.

nant technique in the field of dialog systems. Inspired by previous approaches that used RNNs to model the dialog systems for response generation in short conversations [252, 348, 260], end-to-end dialog systems [285, 180, 343] were proposed for open-domain conversations (or chit-chat) at first. More recently, end-to-end dialog systems have also been applied to task-oriented dialog [341, 170]. However, it presents additional challenges [229], for example, data scarcity, state tracking, external KSs integration, and ensuring goal completion across domains.

A spoken end-to-end neural dialog system includes ASR, TTS, and a neural model, as shown in Figure 2.3. The ASR module produces a textual utterance that is fed into a neural model, and the neural model generates the textual response, which is then passed to the TTS component to produce a spoken utterance. Therefore, the ASR and TTS components are separate from the neural model.

The neural model follows a Seq2Seq architecture, which consists of several key components: tokenization, detokenization, an encoder, a decoder, and input/output embeddings, as depicted in Figure 2.4. Tokenization converts a textual utterance into a sequence of discrete tokens, whereas detokenization reconstructs human-readable text from the generated tokens. The actual representation of language, however, lies in the embeddings. The input embeddings $(x_1, ..., x_n)$ are dense and continuous vectors that encode both semantic and syntactic information about the textual utterance. After processing the input embeddings, the encoder produces the final hidden states, known as the context vectors $(h_1, ..., h_n)$, which summarize the entire input and are used to initialize the decoder. Then the decoder generates the corresponding output embeddings $(y_1, ..., y_n)$. The decoder and encoder are usually deep neural networks, for example, RNNs [126, 73], long short-term memorys (LSTMs) [104], gated recurrent units (GRUs) [45], and Transformers [284].

Therefore, in the following sections, we first present approaches for natural language representation. We then review the architecture of DNNs, trace the core innovation of the attention mechanism, and examine its evolution.

Natural Language Representation

First, natural languages (e.g., English, Chinese, German) need to be converted into a structured format that computers can process, understand, and analyze.

However, natural languages are typically in an unstructured format with multiple granularities and domains. There are multiple levels of language entries, including but not limited to characters, words, phrases, sentences, paragraphs, and documents. The world's languages represent a diverse set of domains. According to the latest edition of Ethnologue[19], there are a total of 7,159 living languages, though such counts are always approximate due to challenges in defining and distinguishing languages. The number of written languages is significantly lower than this, and around 100 are widely used online (e.g., websites, Wikipedia pages).

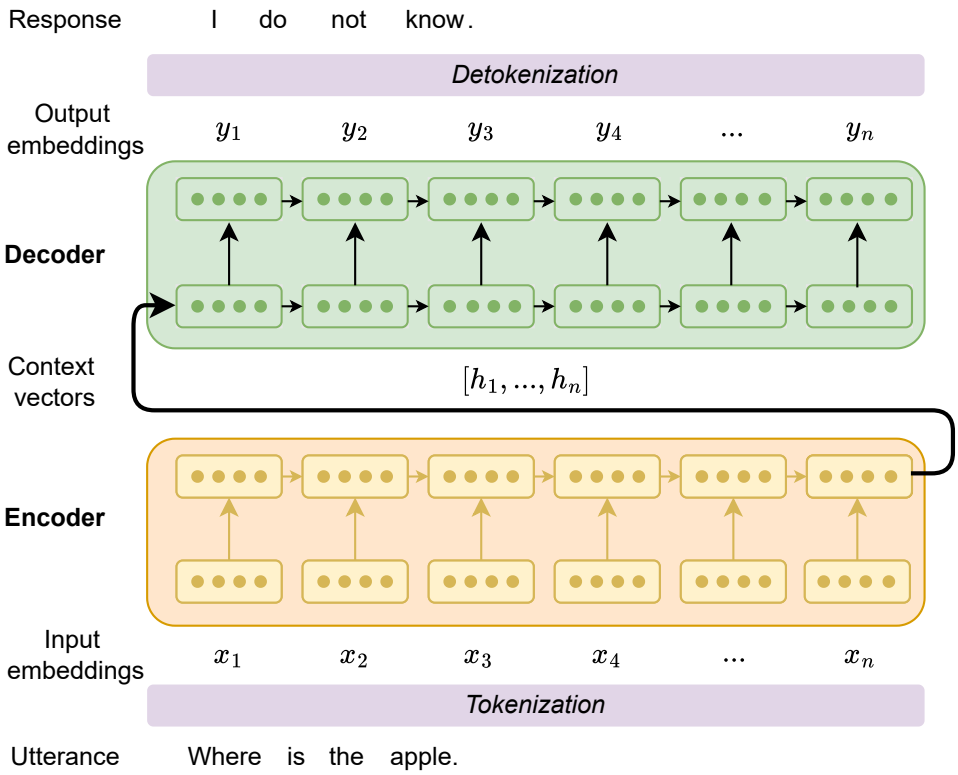


Figure 2.4: Overview of sequence-to-sequence models. The encoder and decoder are trainable deep neural networks (DNNs), whereas tokenization and detokenization are fixed natural language processing operations.

Conventional methods of natural language representation were typically rule-based. One-hot encoding is a classic rule-based approach to representing words, which assigns a unique index (1) for each word in the vector and sets it to the same index (0) for the rest of the vector. The length of the vectors is the vocabulary size. In one-hot encoding, each word is represented as a separate dimension. However, in theory, such high dimensionality is unnecessary, as many words are semantically similar or related. Another limitation of one-hot encoding is that the semantic information about words is missing.

[19] https://www.ethnologue.com/insights/how-many-languages/

Therefore, the distributional concept was introduced to the representation of words. The distributional hypothesis is that the linguistic items that have similar meanings are in similar distributions. N-gram models [253] are one of the earliest word representation methods that are coherent with the distributional hypothesis. Document representation models based on bag-of-words (BoW) models [100] rely on the distributional hypothesis, and the term frequency-inverse document frequency (TF-IDF) weighting scheme builds on this representation [261]. Distributed representation reduces the dimensionality of the representation, which can help prevent sparsity issues, and has been proven to be more efficient. With the great success of DL, distributed representations have become the most common approach for NLP.

The curse of dimensionality is a fundamental problem for language models that rely on distributed representations. However, the introduction of neural networks has largely mitigated this issue. The neural probabilistic language model [17] is a pioneering practice of distributed representation with neural networks. It assigns each word in the vocabulary a distributed word feature vector and models the joint probability function of word sequences based on these vectors. The distributed word feature vectors and the parameters of the associated probability function are learn simultaneously. Due to the projection and the hidden layer being densely connected, this requires complex computations.

To reduce the complexity of computations, word to vector (Word2Vec) represents each word as a low-dimensional vector to capture the similarities between words. For example, continuous bag-of-words (CBOW) models remove the non-linear hidden layer and use a hierarchical softmax, which employs a Huffman binary tree to represent the vocabulary. Therefore, all words share the projection layer. Another architecture, continuous skip-gram, has a similar network structure to CBOW, but it uses each current word as input to a log-linear classifier to predict words within a certain range before and after the target word. The results show that the increased range improves the quality of the resulting word vectors; however, it still suffers from computational complexity [194]. Instead of using the hierarchical softmax, a Huffman binary tree [116] is used to represent the vocabulary. Thus, this model's training speed got a significant increase, and learned more regular word representations [195]. Pennington et al. propose Global Vectors for Word Representation (GloVe) [222], a specific weighted least squares model that trains on global word-word co-occurrence counts, which produces a word vector space with meaningful substructure.

Representing syntax, semantics, and polysemy is another challenge. Peters et al. trained a bidirectional LSTM [104] with a coupled language model objective on a large text corpus, which they named as embeddings from language models (ELMo) representations [223]. Existing language models often overlook morphology by assigning a distinct vector to each word, thereby ignoring their internal structure, which can be problematic for morphologically rich languages (e.g., Turkish or Finnish). FastText [22] is the method that aims to solve this limitation. It is based on the skip-gram architecture and uses a bag of character n-grams to represent

each word. For the rare words representation, Macherey et al. divide words into a limited set of common sub-word units ("wordpieces") [310] for both input and output. This provides a good balance between the flexibility of character-delimited models and the efficiency of word-delimited models, naturally handles translation of rare words, and ultimately improves the overall accuracy of the system.

In contrast to the methods discussed above, which operate at the word level, transformer-based models [63, 231] utilize a global tokenizer to perform both tokenization and detokenization at the subword level. Transformer-based models, trained on massive amounts of unlabeled data, have demonstrated promising performance, representing a revolution in the field of NLP. The most common types of tokenizers include WordPiece [310], used by Bidirectional Encoder Representations from Transformers (BERT) [63]; byte-pair encoding (BPE) [251], used by the Generative Pre-trained Transformer (GPT) [231] family and RoBERTa [351]; and SentencePiece [143], used by Text-to-Text Transfer Transformer (T5) [234].

Deep Neural Networks

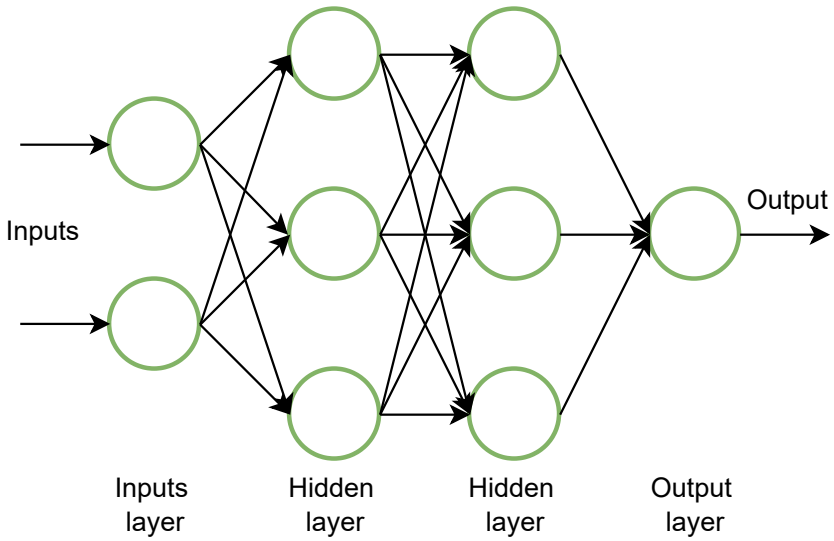


Figure 2.5: A multilayer feed-forward neural network.

The canonical artificial neural network (ANN) is a multilayer feed-forward neural network, which includes an input layer, hidden layers, and an output layer, as shown in Figure 2.5. The input must be a fixed-size vector processed in parallel, and the output is also a fixed-size vector. Many diverse neural networks have been developed that extend feed-forward networks with novel ideas. Convolutional neural networks (CNNs) maintain the feed-forward network structure but replace the fully connected layers with convolutional layers. Each artificial neuron processes only a local region of the input, making CNNs well-suited for spatial data tasks (e.g., image processing). However, due to their lack of temporal memory, CNNs are not well-suited for sequential data tasks (e.g., text or dialog).

To address temporal dependencies, feedback connections, also known as recurrence units, were introduced into feed-forward networks, resulting in recurrent neural networks (RNNs). RNNs are effective for handling sequential data (e.g., text or time series). However, RNNs suffer from two major limitations: their sequential processing of tokens impedes parallelization and requires multiple steps to model long-range dependencies, and they are prone to the vanishing gradient problem, which complicates learning from large datasets.

Various attention mechanisms have been introduced to assist recurrence units in addressing the above limitations, which manage long-range dependencies more effectively by dynamically focusing on relevant encoder states during decoding. However, this does not fundamentally solve the problems of sequential processing and the vanishing gradient. Furthermore, self-attention and cross-attention mechanisms were introduced to replace recurrence [284], allowing each token to attend to all other tokens in a sequence. This shift led to the development of transformers [63], which form the foundation of LLMs.

Overall, RNNs and transformers are able to process sequential inputs of variable length, which have been widely used in speech and language studies and applications. Therefore, we outline the preliminaries of RNNs and transformers, along with their unique mechanisms.

Recurrent Neural Networks (RNNs)

RNNs were proposed many years ago. Some of their variants continue to perform effectively on dialog-related tasks, achieving SOTA results. Generally, modern RNNs can be classified into Jordan-type [126] and Elman-type [73] RNNs. They are shown in Figure 2.6, where x_t, h_t, and y_t are the inputs, hidden state, and output of time step t, respectively. W_h, W_y, and U_h are weight matrices. The biggest difference between them is that each hidden state is decided by the current input and either the output of the last time step or the hidden state of the last time step. Thus, at the t^{th} time step, the hidden state, h_t, and output y_t, of Jordan-type RNNs can be calculated as:

$$h_t = \sigma_h(W_h x_t + U_h y_{t-1} + b_h) \tag{2.1}$$

$$y_t = \sigma_y(W_y h_t + b_y) \tag{2.2}$$

The hidden state, h_t, and output, y_t, of Elman-type RNNs can be calculated as:

$$h_t = \sigma_h(W_h x_t + U_h h_{t-1} + b_h) \tag{2.3}$$

$$y_t = \sigma_y(W_y h_t + b_y) \tag{2.4}$$

Where b_h and b_y are biases. σ_h and σ_y are activation functions.

In practical training, long-range dependencies are difficult to learn because the backpropagation errors accumulate and increase over many time steps, which is known as gradient vanishing and gradient explosion [93].

Long Short-Term Memory (LSTM)

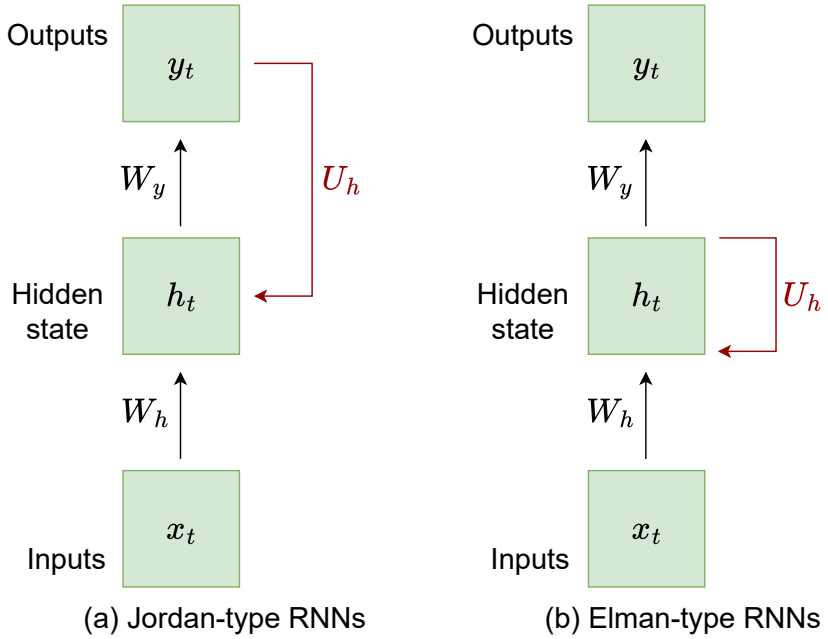


Figure 2.6: Graphical models of two basic types of Recurrent Neural Networks. The primary difference between these two types of methods lies in the fact that each hidden state is determined by the current input and either the output of the previous time step (**Jordan-type**) or the hidden state of the previous time step (**Elman-type**). Source: Adapted from [211]

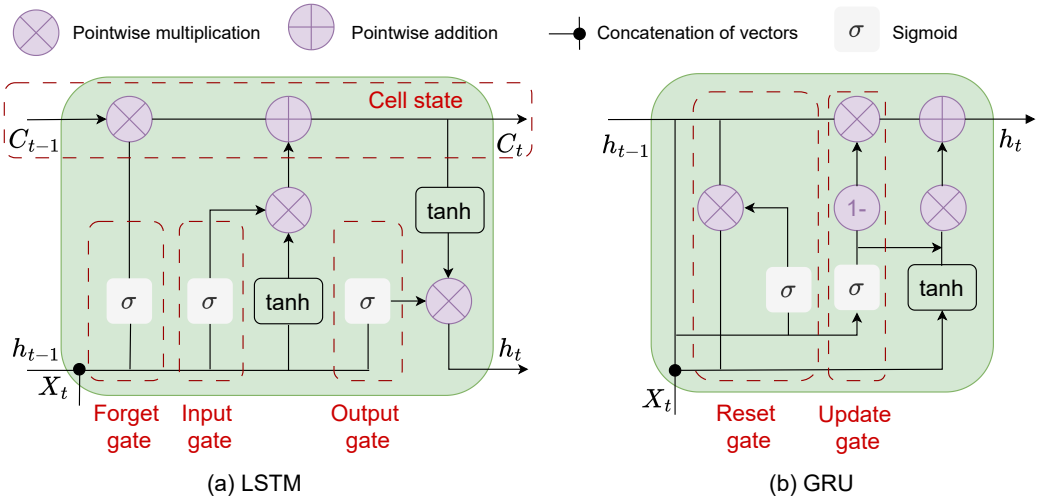


Figure 2.7: Overview of the LSTM and GRU architecture. Source: Adapted from [224].

To address the gradient vanishing problem, the gate mechanisms were introduced in LSTMs [104], as depicted in Figure 2.7 (a). Three gates, the forget gate, f, input gate, i, and output gate, O, were introduced to decide what information can be removed, what information is to be stored, and what information to output. They can be calculated as:

$$f_t = \sigma(W_f[h_{t-1}, X_t] + b_f) \tag{2.5}$$

$$i_t = \sigma(W_i[h_{t-1}, X_t] + b_i) \tag{2.6}$$

$$C_t^* = \tanh(W_c[h_{t-1}, X_t] + b_c) \tag{2.7}$$

$$C_t = f_t * C_{t-1} + i_t * C_t^* \tag{2.8}$$

$$O_t = \sigma(W_o[h_{t-1}, X_t] + b_o) \tag{2.9}$$

$$h_t = O_t * \tanh(C_t) \tag{2.10}$$

Where t represents the time step, X denotes the input, C^* denotes short-term memory, C (cell state) denotes long-term memory, b denotes bias, $[h_{t-1}, X_t]$ denotes the concatenation of h_{t-1} and X_t, $*$ denotes pointwise multiplication, and h denotes the final hidden state output.

Gated Recurrent Unit (GRU)

Inspired by the gate mechanism, GRUs [45] were proposed, see Figure 2.7 (b), which have a single update gate Z that combines the functions of the LSTM's forget and input gates. The cell state is removed. Hidden states h carry all information.

The update gate Z can be calculated as:

$$Z_t = \sigma(W_z[h_{t-1}, X_t] + b_z) \tag{2.11}$$

Where t represents the time step, X denotes the input, b denotes bias, $[h_{t-1}, X_t]$ denotes the concatenation of h_{t-1} and X_t, and h denotes the output, same as in the following equations.

The reset gate r controls how much past information is used when computing the candidate hidden state. It can be calculated as:

$$r_t = \sigma(W_r[h_{t-1}, X_t] + b_r) \tag{2.12}$$

Candidate hidden state h^* can be calculated as:

$$h_t^* = \tanh(W_h[r_t * h_{t-1}, X_t] + b_h) \tag{2.13}$$

Where $*$ denotes pointwise multiplication.

The final hidden state output h can be calculated as:

$$h_t = (1 - z_t) * h_{t-1} + Z_t * h_t^* \tag{2.14}$$

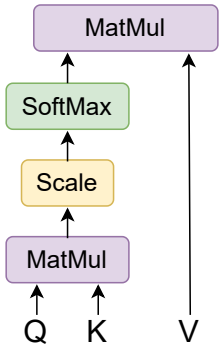


Figure 2.8: Overview of scaled dot-product attention. Q, K, and V represent query, key, and value, respectively. Source: Adapted from [284].

There are many variants that have been proven to perform better than standard LSTMs or GRUs; we will not detail them. The attention mechanism and the transformer have reformed the field of NLP and are now the dominant technologies. We will detail them in the next subsection.

The Evolution of Attention Mechanisms

The attention mechanism was first proposed to improve RNN models for machine translation and is known as additive attention. Additive attention enables the decoder to dynamically focus on different parts of the source sentence at each decoding step [10]. Subsequent study by Luong et al. [184] proposed two types of attention mechanisms: global attention, which considers all encoder hidden states, and local attention, which focuses on a small window around a predicted position. Memory-based models [267] further extend the use of attention by integrating it with an external memory.

Overall, early attention mechanisms, including additive, global, and local attention, served primarily as auxiliary components that augment recurrent models, such as LSTMs and GRUs. Their role was to help these networks manage long-range dependencies more effectively by dynamically focusing on relevant encoder states during decoding. However, the limitations of recurrent models have not been fundamentally resolved.

Scaled dot-product attention [284] models relationships in a sequence by weighting values according to query-key similarity. Unlike the previous attention to assist the recurrent models, scaled dot-product attention serves as a primary component. As shown in Figure 2.8, it can be formalized as Equation (2.15), which computes attention weights by taking the dot product of queries and keys and applying the softmax function to obtain weights for values.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.15}$$

Where Q, K, and V represent query, key, and value, respectively. d_k is the dimension of queries or key.

Transformer

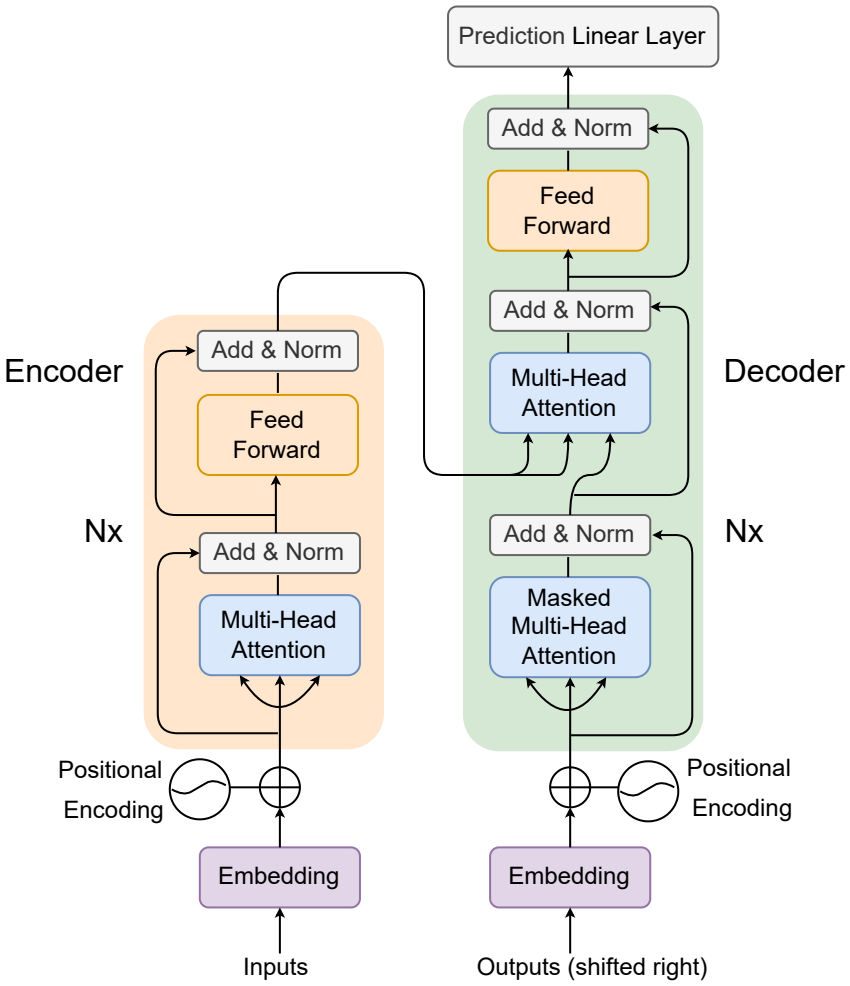


Figure 2.9: Overview of the transformer architecture. Source: Adapted from [284].

The Transformer's multi-head attention mechanism [284] is an extension of the fundamental scaled dot-product attention, running multiple heads in parallel to capture different types of relationships. It leverages two primary forms: **self-attention**, where Q, K, and V are derived from the same sequence, and **cross-attention**, where Q is derived from one sequence and K, V from another (e.g., connecting the decoder to the encoder). This architecture marked a paradigm shift in neural sequence modeling, resulting in more efficient training and superior performance across a wide range of tasks.

Similar to the vanilla Seq2Seq learning (as shown in Figure 2.4), the architecture of the canonical transformer consists of an encoder and a decoder, as depicted in Figure 2.9. Both the encoder and decoder are composed of a stack of N =

6 identical layers, as described in the original study. The encoder tokenizes the input sentence into tokens vectors $x = (x_1, ..., x_n)$ and maps it into continuous hidden states $h = (h_1, ..., h_n)$. The decoder then generates the output sequence $y = (y_1, ..., y_n)$ based on the encoder's hidden states. Each encoder layer consists of two sub-layers: a multi-head self-attention mechanism and a simple fully connected feed-forward network, both of which apply residual connections. The structure of a decoder layer is almost the same as that of an encoder layer, except for an additional encoder-decoder attention layer, which computes the attention between decoder hidden states of the current time step and the encoder output vectors. The input to the decoder is partially masked to ensure that each prediction is based on the previous tokens, thereby avoiding predictions in the presence of future information. Both the encoder and decoder inputs use a positional encoding mechanism [284, 211].

2.3 Transformer-based Language Models

Previous dominant language methods, such as RNNs, rely on a limited local context (a few preceding words), restricting the model to capture broader patterns in the data. Attention mechanisms address this limitation by internalizing linguistic structures and enabling models to process unstructured textual data more effectively. Therefore, transformer models develop a deeper statistical understanding of language, leading to better comprehension. In addition to encoder-decoder language models, there are two other major variants: encoder-only and decoder-only language models (see Figure 2.10). These three architectures serve as the foundational building blocks of modern transformer-based language models. Each of these models is trained on different objectives and so learns to solve different tasks. Figure 2.11 illustrates the relationships among SOTA LLMs.

2.3.1 Encoder-only Language Models

Encoder-only language models are trained for understanding tasks, such as sentence classification or embeddings. The dominant training objective is masked language modeling (MLM), which pre-trains the model to predict masked tokens in order to reconstruct the original sentence. Next sentence prediction (NSP) is another pre-training objective, designed to help the model learn whether two sentences are logically related. For example, BERT [63] (Bidirectional Encoder Representations from Transformers) is pretrained using both MLM and NSP to improve sentence reconstruction and understanding of inter-sentence relationships. Derivatives of BERT, such as RoBERTa [351], are trained on large-scale data to learn deep contextual word representations, while SimCSE [85] is trained using a contrastive learning approach to produce sentence-level embeddings.

Other objectives have been introduced to address different challenges. Sentence order prediction (SOP) was proposed to overcome the limitations of NSP by enhancing discourse-level understanding, as in ALBERT [145]. To improve the effi-

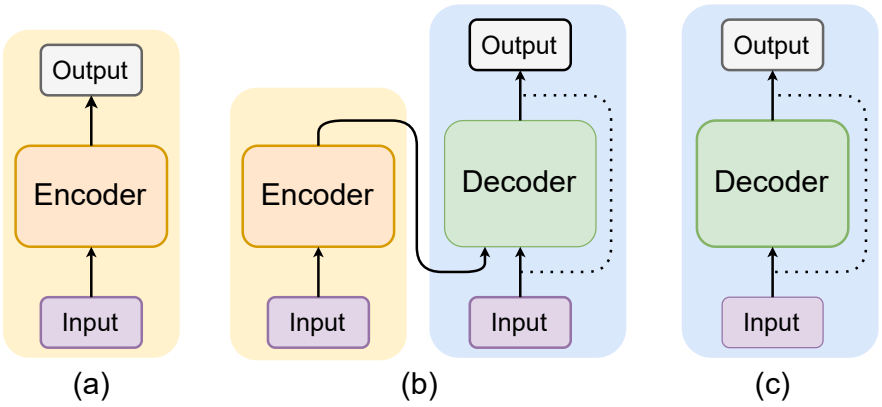


Figure 2.10: Transformer variants. **(a)** Encoder-only models, also known as autoencoding models (e.g., BERT [63], RoBERTa [351], SimCSE [85], ALBERT [145] et al.); **(b)** Encoder-decoder models, termed sequence-to-sequence models (e.g., BART [151], mBART [177], T5 [209], mT5 [316], UL2 [270], PALM [18], et. al.); **(c)** Decoder-only models, referred to as autoregressive models (e.g., GPT family [231, 232], LLaMA [278], Vicuna [43], Mixtral 7B [122], DeepSeek family [168, 95, 167], Claude [305], PaLM [46], et. al.).

ciency of MLM, replaced token detection (RTD) enables the model to learn from all tokens rather than only the masked ones, as in Electric [48].

2.3.2 Encoder-Decoder Language Models

Encoder-decoder language models are designed for input-output tasks such as summarization, question answering, and machine translation. Bidirectional and Auto-Regressive Transformers (BART) [151] combines a bidirectional encoder (similar to BERT) with a left-to-right autoregressive decoder (similar to GPT). It is pre-trained using a denoising objective, enabling the model to reconstruct the original text from the corrupted input. This training strategy allows BART to develop strong capabilities in both understanding and generating language. mBART [177] is a multilingual extension of BART, designed to handle a variety of languages in the same framework.

T5 [209] is another powerful encoder-decoder model that also uses a denoising pre-training objective. Unlike BART, T5 focuses solely on reconstructing the masked portions of the input text rather than the entire sentence. mT5 [316] extends T5 to multilingual settings, enabling the model to perform a wide range of tasks across different languages. Pre-training an Autoencoding & Autoregressive Language Model (PALM) [18] is an encoder-decoder model in which the encoder and decoder are jointly pretrained in two stages. In the first stage, the encoder is trained as a bidirectional autoencoder to reconstruct the original text from corrupted context. In the second stage, the encoder and decoder are jointly trained to generate text output autoregressively from the encoder's context representations.

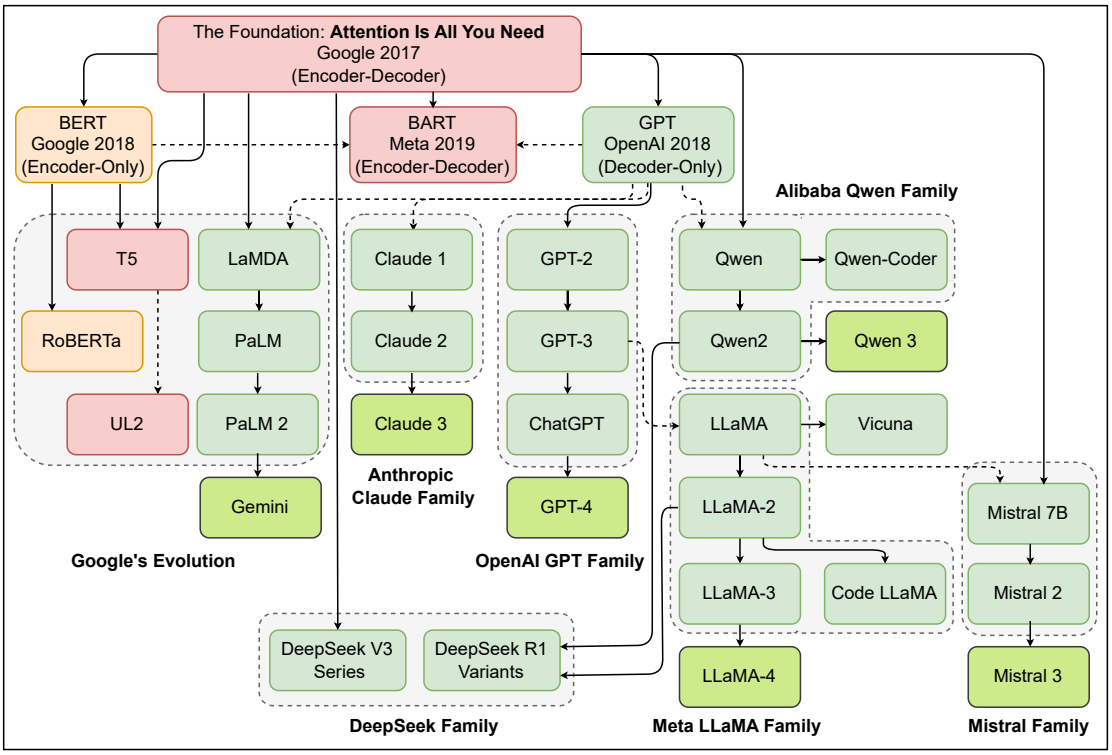


Figure 2.11: Evolutionary relationships of large language models. *Note:* Recently, these models have increasingly supported multimodal inputs. Gemini and Claude 3 are explicitly multimodal, whereas GPT-4, Qwen 3, Mistral 3, and LLaMA-4 are primarily language models, with vision-language capabilities introduced through specific multimodal variants. Their subsequent iterations increasingly adopt fully multimodal architectures, including Gemini 2, Gemini 3, Claude 4, and GPT-5.

2.3.3 Decoder-only Language Models

Decoder-only language models are experts at generative tasks such as story writing, conversational agents, and code generation. These models, also known as autoregressive models, are pretrained using causal language modeling (CLM), where the objective is to predict the next token in a sequence based solely on the preceding tokens. Notable examples include GPT [231], LaMDA [277], Mixtral 7B [122], Vicuna [43], LLaMA [278], Qwen [11], PaLM [46] and Claude [305], et. al.

CLM can be extended to multi-task pretraining, allowing the model to develop expertise across a wide range of domains. One such extension is prefix language modeling (PrefixLM), a hybrid of CLM and MLM. In PrefixLM, the model is conditioned on a given prefix and generates a span of text, rather than predicting only the next token. This approach is used in models like UL2 [270]. Multi-token prediction (MTP) extends the conventional CLM by training the model to predict multiple future tokens from each position, providing a denser training signal. Prior research and implementations, such as DeepSeek-V3 [168], demonstrate that this auxiliary objective can improve downstream performance and support more

complex reasoning and planning behaviors than standard next-token prediction.

Alignment and Fine-Tuning

In addition to conventional supervised and unsupervised strategies for fine-tuning LLMs, particularly decoder-only architectures. The SOTA systems such as ChatGPT, Google Gemini, Anthropic Claude, LLaMA-2-Chat, and DeepSeek-Chat employ reinforcement learning from human feedback (RLHF) [47, 216, 14] to improve output quality and better align model behavior with human goals and values. The recent success of DeepSeek-R1 [95] further illustrates that RL-based alignment methods can enhance a model's usefulness, honesty, and safety. Nevertheless, RLHF remains computationally demanding, difficult to stabilize, and labor-intensive due to its reliance on human preference data. RLHF [233] has emerged as a more efficient and scalable alternative to RLHF, and it is adopted in models such as Mistral-Instruct and Zephyr. Gemma-IT employs a distinct RL-based method, rejection sampling fine-tuning (RSFT) [272], which optimizes model behavior through filtered sampling rather than an explicit reward model.

Beyond RLHF and RSFT, LLMs can be improved through RL variants, including reinforcement learning from AI feedback (RLAIF) [146] and constitutional AI (CAI) [15]. RL-free methods include supervised instruction tuning [246, 298] and preference-based optimization methods such as rank responses to align human feedback (RRHF) [331] and direct preference optimization (DPO).

Additional advancements come from self-training and synthetic-data generation techniques [297, 350]. Models can also be adapted using parameter-efficient fine-tuning (PEFT) methods such as low-rank adaptation (LoRA) [107] and quantized low-rank adaptation (QLoRA) [62] as well as Knowledge Distillation [103], Retrieval-Augmented Training [77], and Helpfulness, Honesty, and Harmlessness (HHH)-oriented fine-tuning [9]. Together, these approaches offer a flexible and extensible toolbox for aligning, specializing, and scaling the capabilities of LLMs.

2.4 Chapter Summary

In this chapter, we reviewed conversational agents in Section 2.1. These personal assistants have become increasingly integrated into our daily life. We examined them from an architectural perspective in Section 2.2, including modular, neural, and hybrid dialog systems. The motivation for adopting an architectural perspective is that contemporary surveys of dialog systems typically categorize them as rule-based, statistical data-driven, end-to-end neural, or hybrid. However, this perspective presents two fundamental issues.

The first issue is that, in early research, although rule-based and data-driven dialog systems were built on fundamentally different technologies, both were based on modular architectures. Nowadays, statistical methods such as RNNs and transformers can be applied either as individual components within a modular dialog

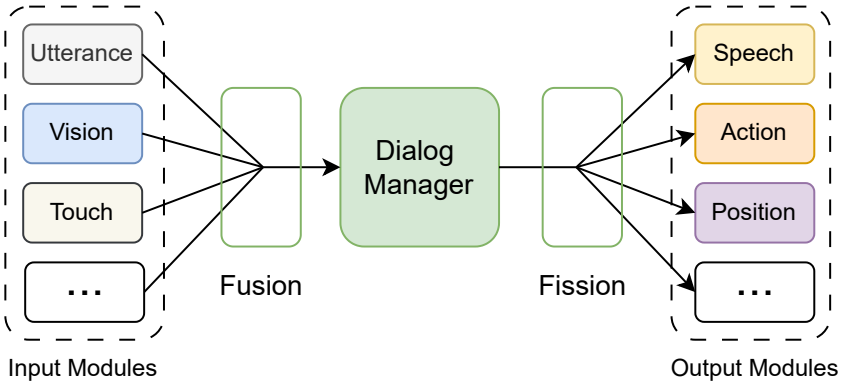


Figure 2.12: Overview of the high-level multimodal dialog architecture for an embodied agent. Source: adapted from [79].

system or as end-to-end neural dialog systems. Therefore, we reviewed the broader notion of modular dialog systems in Section 2.2.1 and end-to-end neural dialog systems in Section 2.2.3. The second issue is that, as technologies evolve, researchers have developed a deeper understanding of the strengths and limitations of these models. A more principled approach is to select the appropriate technical foundation based on specific requirements. In particular, the emergence of large-scale transformer-based methods, especially LLMs, has demonstrated remarkable capabilities. Consequently, the concept of hybrid dialog systems has evolved compared with early studies, such as BeRP, SCREEN, and HIS. We provide a review and summary of this new concept in Section 2.2.2.

Furthermore, we analyzed the prevailing paradigm of transformer-based language models in Section 2.3, which can be categorized into encoder-only, encoder-decoder, or decoder-only models, each with different training objectives, as illustrated in Figure 2.10. The detailed evolutionary relationships among these models are shown in Figure 2.11. One notable trend is that multimodal systems are gradually being valued and explored. It also aligns with the requirements of the research on embodied agents.

Conversation is arguably the most common form of human communication, typically involving spoken or written exchanges between two or more participants. In situated communication, humans rely on multiple senses: we listen with our ears (language), see with our eyes (vision), speak with our mouths (speech), and use our hands both to perceive touch and to manipulate objects. Multimodal dialog systems for embodied agents should move beyond a purely language-centric approach, incorporating a range of sensory inputs and outputs, as illustrated in Figure 2.12. Inputs include linguistic, environmental, haptic, and other sensory information, while outputs comprise speech responses, physical actions, and predictions of object positions in the environment. For an intelligent robot to assist in everyday human activities, environmental perception must be considered equally central to language processing.

Therefore, in the next chapter, we investigate how to extend conversational agents to incorporate sensory and contextual information beyond language, with a particular focus on integrating visual information for embodied agents.

Chapter 3

Embodied Vision-Language for Situated Communication

In the previous Chapter 2, we reviewed the remarkable success of conversational agents capable of communicating with humans in natural language. However, since human-human communication is usually situated and interactive, language-only, unimodal dialog systems are insufficient for achieving truly natural and intuitive interactions. As a result, researchers have developed situated, interactive dialog systems that engage in communication by processing and generating responses through both visual and linguistic input and output modalities [191, 42]. Situated and interactive communication refers to conversations that are based on visual input, such as images, videos, or sensory data from the real or virtual environments. These systems, known as multimodal dialog systems, integrate multiple modalities for both understanding and response.

Early multimodal dialog systems often focused on screen-based interactions, combining speech with a complementary input modality to control dynamic graphical outputs. For instance: Put-That-There [23] integrates speech and 3D hand-tracking data, outputting dynamic graphics where objects are moved, created, or deleted based on user commands. COMIC [33] combines speech with touchscreen gestures. The system responds with synthesized speech and updates a shared graphical design interface. QuickSet [49] uses speech and pen gestures as input to create and manipulate entities within a map-based simulation environment, with the results displayed graphically. MATCH [125] accepts input from speech and pen on a mobile device, providing spoken feedback and displaying relevant information and maps on its graphical interface. Consequently, the field has focused on input fusion [21] to combine modalities from the user, and output fission [286] to select appropriate responses. The statistical methods significantly advanced these fusion and fission techniques [21, 217], enabling systems to more effectively interpret multimodal inputs and generate appropriate outputs. Therefore, in this chapter, our main research question is *how to integrate the visual modality into a conversational agent?*

This chapter highlights the importance and challenges of integrating the visual modality into a language model for interactive and situated communication. We begin by outlining the key challenges of vision-language datasets (VLDs) in a specific task in Section 3.1 and then review general-purpose vision-language models (VLMs) in Section 3.3. We analyze the transition from specific task challenges to general-purpose VLMs in Section 3.2 and summarize the transfer from general-purpose VLMs to embodied agents, specifically addressing the robotic challenges in Section 3.4. Furthermore, we discuss the distinction between Uncertainty, Vagueness, and Ambiguity (UVA-phenomena) in human-robot interaction (HRI), and clarify the definitions and potential solutions in Section 3.5.1. Subsequently, we examine how the object states impact HRI in the following Section 3.5.2. Finally, we close this chapter with unresolved challenges in Section 3.6.

3.1 Vision-Language Datasets

Vision-language datasets (VLDs) are collections of image-text pairs that incorporate both natural language and visual data [185]. These datasets are used to train models, so their design logic directly impacts model performance. From a human perspective, the overarching goal is to empower an agent to communicate effectively with humans. The agent must be able to perceive environmental information in order to respond to human utterances or perform specific actions to fulfill human requests. Based on their objectives and interaction paradigms, we classify these datasets into three broad categories: *observation-only, retrieval-based, and interactive VLDs*. Accordingly, we examine the datasets from these three perspectives.

3.1.1 Observation-only VLDs

The objective of this type of dataset is to provide training data for agents that generate a correct natural language answer based on a given image and a corresponding natural language question. However, neither the human user nor the dialog agent directly manipulates or interacts with the environment, as illustrated in Figure 3.1(a). Instead, the conversation focuses on describing, understanding, or reasoning about the observed scene.

For example, visual question answering (VQA) [8] addresses the task of generating a correct natural language answer given an image and a related natural language question. Furthermore, Visual Dialog [56] aims to enable an artificial agent to conduct meaningful conversations with humans using natural, conversational language about visual content. CLEVR [123] and CLEVR-Ref+ [175] focus on building artificial agents capable of reasoning and answering questions about visual data. CLEVR-Dialog [141] is a larger diagnostic dataset designed for studying multi-round reasoning in visual dialog. GuessWhat?! [59] presents a multi-turn language-based task in which a questioner interacts with an oracle to identify an unknown target object in a complex image scene; the questioner must ask a sequence of questions and make a final guess, which the oracle then verifies.

Referring expression comprehension (REC) is the task of localizing in an image the object referred to by a given natural language expression. The classic datasets Referring Expressions in COCO (RefCOCO) and its variant of Referring Expressions in COCO Plus (RefCOCO+) [131] were collected via a two-player game on Common Objects in Context (COCO) image set, with the key difference that RefCOCO+ excludes explicit location words to force reliance on fine-grained visual details. In contrast, the Google version of Referring Expressions in COCO (RefCOCOg) [190] was collected via one-player annotation, yielding longer, more detailed expressions that provide richer contextual information.

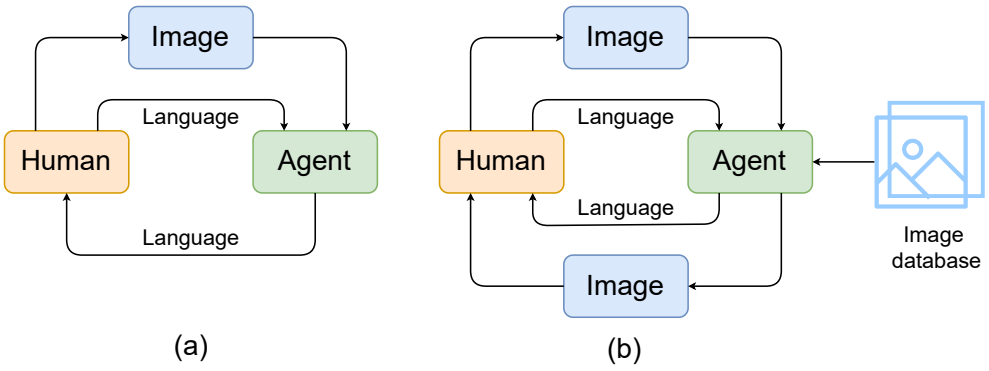


Figure 3.1: **(a)** Paradigm in observation-only vision-language datasets.
(b) Paradigm in retrieval-based vision-language datasets.

Situated and Interactive Multimodal Conversations (SIMMC) [201] aims to train the virtual assistants for two shopping scenarios: fashion (grounded in an evolving set of images) and furniture (grounded in a shared virtual environment). The virtual assistant is designed to incorporate multimodal inputs (e.g., visual context, memories of previous interactions, and users' utterances) and perform multimodal actions (e.g., displaying a route while generating a system utterance). However, to effectively help users in real-world multimodal environments, an assistant must understand conversational contexts within the user's perceived surroundings. The second generation, SIMMC 2.0 [139], was introduced, grounded in immersive and photo-realistic scenes. SIMMC 2.1 [138] extends SIMMC 2.0 with additional annotations for fine-grained referent candidates and new user utterances, which support the study and modeling of visual disambiguation.

3.1.2 Retrieval-based VLDs

The objective of this type of dataset is to overcome the challenge of bridging the semantic gap between textual and visual modalities. An agent assists users in finding images that match certain criteria expressed in natural language. Its role is to interpret the user's input-often in a conversational context-and return the most relevant results from the image database, as shown in Figure 3.1 (b).

Relative Captioning [98] introduces a dataset that captures comparative visual

differences, which are difficult to describe using only a predefined set of attributes. The goal is to enable the model to function as a dialog-based interactive retriever that can retrieve the user's desired images. Multimodal Dialog (MMD) [244] is a domain-aware conversational dataset that captures interactions between shoppers and salespersons in the retail domain, embodying the necessary capabilities for autonomous agents. MMDialog [78] focuses on multimodal open-domain conversation, encompassing both response generation and retrieval tasks.

3.1.3 Interactive VLDs

In addition to observing the environment for language reasoning, a key advancement of interactive vision-language study is that the agent can interact with the environment, either virtually or physically. Such conversations often involve task-oriented goals, where the agent must understand the current state of the environment and predict appropriate actions to achieve a specific outcome. The agent receives linguistic instructions from the user and interacts with the environment through visual perception. It may or may not provide feedback to the user. Similarly, the human may or may not physically interact with the environment. These interaction forms can be represented in four possible variants, as illustrated in the Figure 3.2.

Embodied Agents

In embodied agent research, robots primarily operate by navigating physical environments and manipulating objects. TOUCHDOWN [36] examines joint reasoning over language and vision in navigation and spatial tasks, requiring agents to generate actions that are consistent with both natural language instructions and object properties. ALFRED [256] introduces a benchmark within AI2-THOR[1] comprising five navigation and seven interaction actions, and emphasizes task execution without language-based responses. That interaction form is illustrated in Figure 3.2(a).

To incorporate linguistic responses into embodied interaction, Talk The Walk [58] presents the first large-scale dialog dataset grounded in action and perception. The dataset involves a guide and a tourist who collaborate through natural language, thereby integrating perception, navigation, and interactive dialog. That interaction form is shown in Figure3.2(c). Vision-and-Dialog Navigation [276] extends this paradigm by enabling agents to solicit assistance when encountering difficulties and to interpret human feedback. TEACh [219] contributes an interactive dataset involving human commanders and robot followers, where followers are permitted to query for clarification when instructions are ambiguous. Similarly, Imitating Interactive Intelligence [2] collects setter-solver interactions, in which setters issue instructions and solvers both execute them and produce language responses.

[1] `https://ai2thor.allenai.org/`

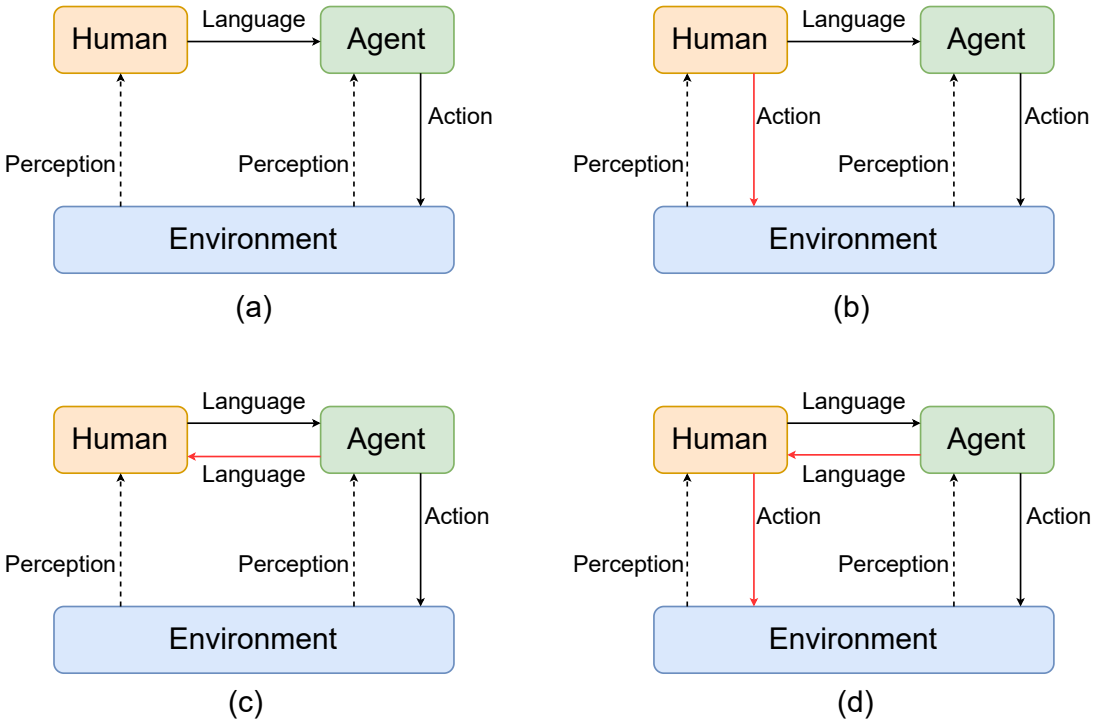


Figure 3.2: Interactive forms of vision-language datasets in four variants. **(a)** No agent feedback; no human interaction. For example, ALFRED [256] and TOUCHDOWN [36]; **(b)** No agent feedback; human interacts with the environment. Such as CEREALBAR [266]; **(c)** Agent provides feedback; no human interaction. Like, Minecraft Collaborative Building [207], TEACh [219], Talk The Walk [58], Imitating Interactive Intelligence [2], and Vision-and-Dialog Navigation [276]; **(d)** Agent provides feedback; human interacts with the environment (reflecting HRI in daily life).

Virtual Game Environments

In virtual game studies, game-playing agents execute actions conditioned on visual and linguistic cues. To support the development of fully interactive agents capable of collaborating effectively with humans, the Minecraft Dialog Corpus [207] was introduced, whose interaction form is illustrated in Figure 3.2(c). This corpus features a two-player game in which an Architect and a Builder play. The given target structure instructs the Builder via a text-based chat interface to construct a replica of the target within a designated build region. The dialog between the Architect and Builder serves to resolve ambiguities and correct construction errors.

CEREALBAR [266] introduces a dataset based on a collaborative game with two agents, a Leader and a Follower. It is designed to address the challenge of grounding the Leader's instructions in the Follower's actions. Beyond instruction-following, agents must reason about a dynamically evolving environment and interpret directions involving multiple, interdependent goals. That interaction form is shown

in Figure 3.2(b).

3.2 Transition: from Datasets Challenges to Models

In the era of data-driven statistical methods, datasets are often designed to address a specific challenge and are subsequently used to train corresponding models. Consequently, the nature and structure of these datasets largely determine the functionality and performance of the resulting models. However, the architecture of a model itself also plays a crucial role in shaping its effectiveness and capabilities.

In contrast to earlier approaches that focused on developing models tailored to a single, well-defined task, contemporary methodology has shifted toward the construction of general-purpose models [230, 333, 160]. Such models are trained on large-scale and diverse datasets to acquire the ability to perform a wide range of tasks and adapt to various scenarios. They can subsequently be fine-tuned or prompted to address specific downstream tasks with minimal additional supervision.

In the previous Chapter 2, we reviewed how general-purpose language models, grounded in transformer-based architectures, have revolutionized natural language processing (NLP) and conversational agents. Extending transformer-based language models by incorporating visual information gives rise to transformer-based VLMs. The fundamental principles of these models can be categorized into three paradigms: *discriminative, generative, and hybrid*. We will examine these paradigms in detail in the following section.

3.3 Transformer-based Vision-Language Models

3.3.1 Discriminative VLMs

Discriminative VLMs, also known as conditional models, are typically trained with *masked prediction objectives* (e.g., masked language modeling (MLM) conditioned on images, masked region/feature modeling, masked image modeling with feature regression), *matching objectives* (e.g., image-text matching, next-sentence mismatch detection), *contrastive objectives* (e.g., image-text contrastive), or *alignment objectives* (e.g., region-word or patch-token alignment). An encoder-based transformer model, like Bidirectional Encoder Representations from Transformers (BERT) [63], can effectively analyze and process textual data. Inspired by the profound impact of BERT on NLP, many BERT-style innovations have been introduced to advance the fusion of vision and language for understanding, reasoning, and alignment.

Understanding and Reasoning

ViLBERT [182] discretizes the space of visual inputs by clustering visual features into tokens within a BERT-based architecture, which are treated analogously to text tokens to form a vision stream. The text and vision streams are then processed in parallel within novel cross-modal co-attentional transformer layers, which enable learning of visual grounding from paired vision-language data, as shown in Figure 3.3(a). VisDial-BERT [206] adapts ViLBERT for multi-turn visually-grounded conversations.

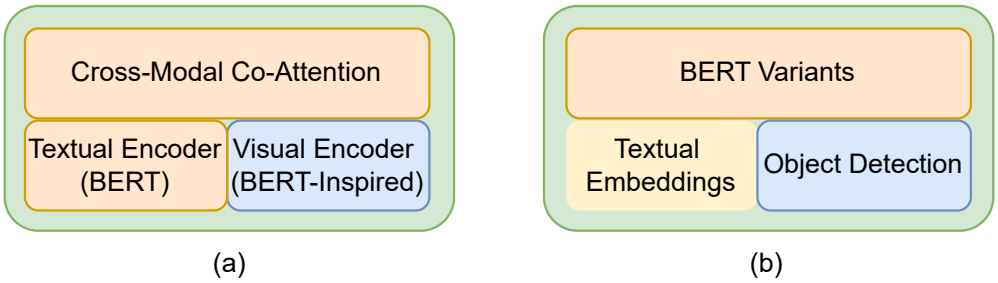


Figure 3.3: Schematic illustration of dual-stream and single-stream vision-language architectures. **(a)** Dual parallel streams approaches: ViLBERT [182] and VisDial-BERT [206]; **(b)** Single stream approaches: UNITER [40], VisualBERT [154], and VL-BERT [265].

Instead of processing vision and language as two parallel streams, VisualBERT [154], UNITER [40], and VL-BERT [265] use object detection models to extract image features and concatenate them with textual embeddings from BERT to form a single stream. The self-attention mechanism in BERT and its variants then implicitly aligns elements of the input text with regions in the input image, as depicted in Figure 3.3(b). The results from it show that a single-stream architecture performs better than two parallel streams.

Oscar [159] introduced object tags detected in images as anchor points to facilitate learning alignments between vision and language (see Figure 3.4(a)). Building on this, VinVL [340] developed a more powerful object detection model that produces richer visual features than the classical approach [5], which is fed into Oscar and substantially uplifts the state-of-the-art (SOTA) results on all major VL tasks across multiple public benchmarks. The achievements of Oscar and VinVL demonstrate that object tags as anchors enhance vision-language alignment. Notably, it was the first to apply contrastive loss on object tags, encouraging alignment between object labels and corresponding object regions. This idea was later extended more broadly to image-text alignment, which is termed contrastive learning.

Contrastive Learning

To facilitate a comprehensive understanding of vision, language, and their similarities, contrastive learning methods have been developed to align the represen-

tations of visual and textual modalities. As shown in Figure 3.4 (b), a simple dual-encoder architecture learns to align visual and language representations from image-text pairs. The textual encoder, like BERT [63], converts textual inputs into textual embeddings, while the visual encoder, such as convolutional neural networks (CNNs) [142] or vision transformers (ViTs) [68], extracts features from images to produce visual embeddings. Contrastive learning aligns related image-text pairs by minimizing the distance between their embeddings in a shared space, while maximizing the distance between embeddings of unrelated pairs. Pioneering models such as CLIP [230], ALIGN [121], and Florence [312] are pretrained on large-scale image-text datasets to learn transferable representations for downstream tasks.

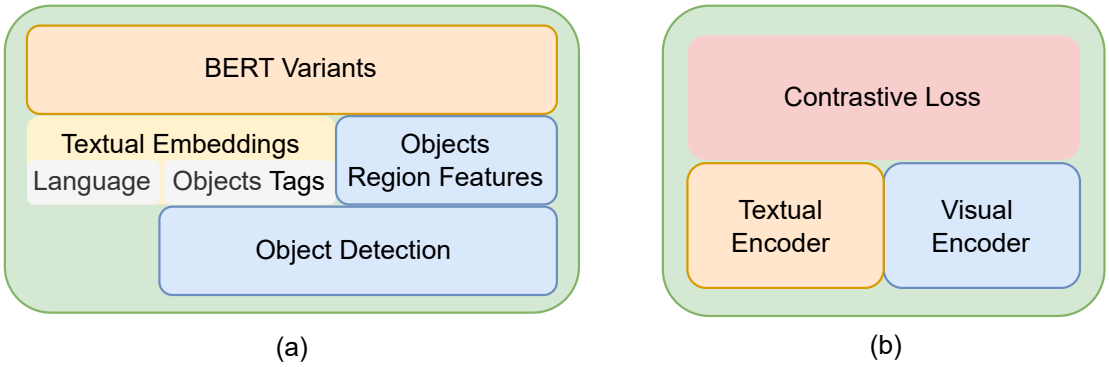


Figure 3.4: Pre-contrastive vs contrastive vision-language models. **(a)** Simplified diagram for pre-contrastive vision-language models: Oscar [159] and VinVL [340]; **(b)** Simplified diagram for contrastive vision-language models: CLIP [230], ALIGN [121], and Florence [312].

3.3.2 Text-Generating VLMs

Text-only generative VLMs are trained with objectives such as next-token prediction (e.g., standard causal or prefix language modeling (PrefixLM)), masked reconstruction, denoising, sequence-sequence, and permutation-based modeling. To enhance the model's communication capabilities, multimodal instruction tuning is employed during the fine-tuning stage. The most widely used large language models (LLMs), serving as backbones adapted for VLMs, include LaMDA [277], Mixtral 7B [122], LLaMA [278], Vicuna [43] (built upon LLaMA), and Qwen [11], all of which excel at text generation[2].

A straightforward approach to incorporating vision into an LLM is to employ a pretrained visual encoder to extract visual features, followed by a multilayer perceptron (MLP) or linear adapter to project these features into the textual embedding space, as shown in Figure 3.5(a). For example, LLaVA [173] employs a

[2]We primarily focus on open-source or open-weights models, since official details of closed-source models (e.g., GPT-4, Gemini, and Claude) remain undisclosed.

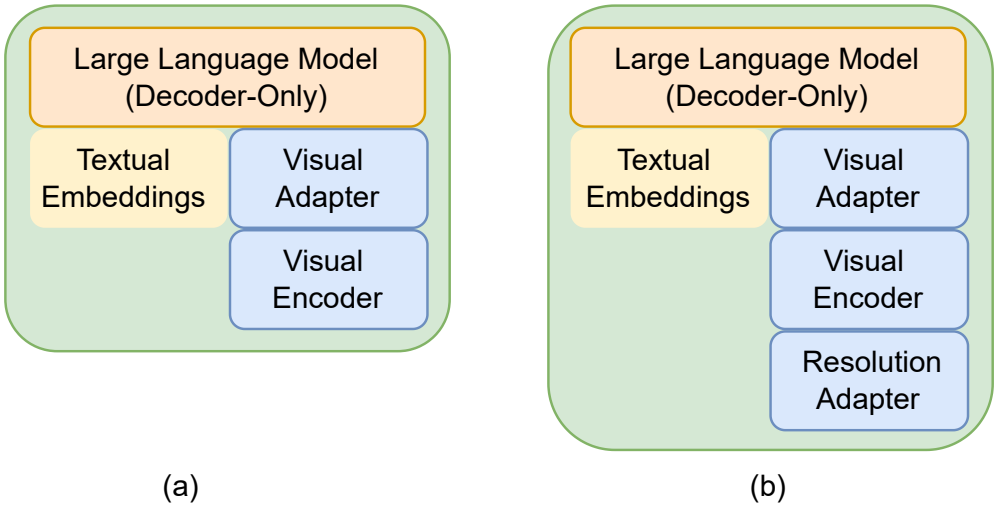


Figure 3.5: Simplified diagram for generative models. **(a)** The straightforward approaches: LLaVA [173], MiniGPT-4 [349], CogVLM [293], and CogVLM2 family [293]; **(b)** Image resolution adaptive approaches: Qwen-VL [12], Qwen2-VL [13], MiniCPM-V-2.6 [324], and DeepSeek-VL2 [311].

linear projection layer to combine a pretrained vision encoder, CLIP [230], with a pretrained language decoder, Vicuna [43], which updates all module weights across different training stages. MiniGPT-4 [349] uses a single linear projection layer to connect a visual encoder, BLIP-2 [152], into a language decoder, Vicuna [43], which only trains the linear projection layer to align the visual features with the language decoder. Although a single linear layer is simple and effective, it suffers from computational inefficiency when processing extended image sequences, as well as from limited expressive capacity. To address these limitations, a 2×2 convolutional layer is introduced into the visual adapter in CogVLM [293] and the CogVLM2 family [105].

Image resolution is another primary factor that impacts a model's performance. A naive solution is using a resolution adapter to resize all input images to a specific resolution, such as Qwen-VL [12], as shown in Figure 3.5(b). Furthermore, Qwen2-VL [13] introduces naive dynamic resolution, which dynamically converts images into a variable number of visual tokens. To handle high-resolution images with varying aspect ratios, GLM-4.5V [106] introduces two adaptations. First, 2D-RoPE [264] is integrated into the self-attention layers of the ViT. Second, the model's original learnable absolute positional embeddings are retained. MiniCPM-V-2.6 [324] proposes a divide-and-stride strategy that partitions the image into optimally scored slices based on resolution matching and a calculated ideal slice number, incorporating adjacent sets if needed, while capping the resolution for practical efficiency.

Designed for versatility, SPHINX [166] integrates a mixture of multi-LLMs, tuning tasks, visual embeddings, and high-resolution sub-images, as shown in Fig-

ure 3.6(a). Although it achieves impressive performance, its integration of numerous components renders it redundant and computationally inefficient. To mitigate this, the SPHINX-X family [171] removes the redundant visual encoder and simplifies the training stages.

To better process both low-resolution and high-resolution images, DeepSeek-VL introduces a hybrid visual encoder [181], as illustrated in Figure 3.6(b). This hybrid visual encoder combines a SAM-B encoder [137], which processes high-resolution (1024×1024) inputs to extract semantic features, with a SigLIP-L encoder [334], which handles low-resolution (384×384) inputs. However, this approach is still constrained by a fixed resolution. To address larger resolutions and extreme aspect ratios, DeepSeek-VL2 [311] employs a dynamic tiling strategy.

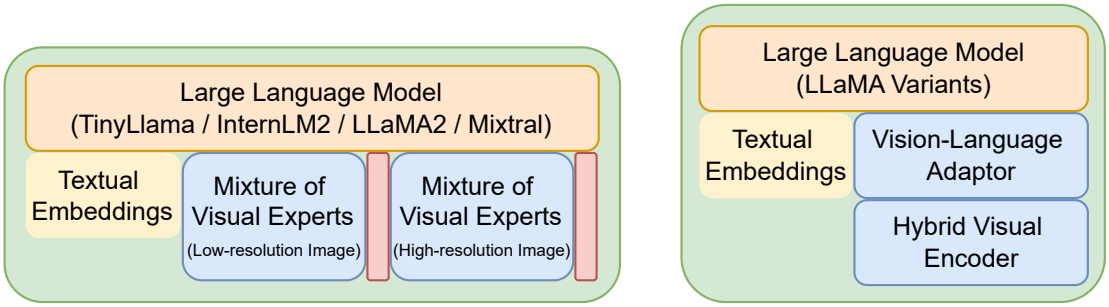


Figure 3.6: Simplified diagram for versatile generative models: **(a)** SPHINX-X family [171]; **(b)** DeepSeek-VL [181].

The above methods employ different strategies, such as resizing, padding, or partitioning, to adapt images with varying aspect ratios, thereby integrating a visual encoder into an LLM. However, these operations introduce layout distortions, disrupt spatial continuity between image regions, and impose additional computational overhead. Consequently, researchers have begun to explore methods that integrate perception and reasoning capabilities within a single, unified architecture, which is termed as visual encoder-free VLMs.

EVE [65] is a representative visual encoder-free VLM (see Figure 3.7) that processes images and text through a lightweight patch embedding layer, followed by a hierarchical patch-aligning layer that maps features into the pairwise space of an auxiliary vision encoder [230]. Its performance, however, is constrained by cross-modal interference in the pretrained LLM. EVEv2 [66] addresses this limitation by decoupling modalities within the unified decoder using modality-specific components, including attention, feed-forward, and normalization layers, along with modality-wise sparsity and routing. SOLO [39] instead applies a linear projection directly to raw image pixels within a unified transformer. Mono-InternVL [183] embeds visual experts into a pretrained LLM, thereby unifying visual encoding and language decoding within a single transformer architecture.

In a different vein, Chameleon [271], Emu3 [296], Show-o [314], and Transfusion [346] employ image tokenizers to encode and decode images for generative purposes. As this topic lies outside the scope of our discussion, we do not elaborate further.

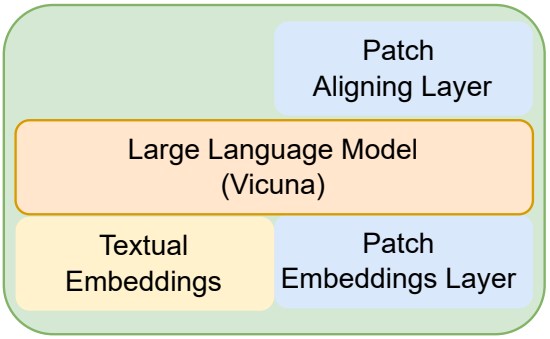


Figure 3.7: Visual encoder-free VLM: EVE [65].

3.3.3 Hybrid VLMs

Hybrid models are trained with mixed discriminative and generative training objectives, which leverage the strengths of both training to enable the models to become general-purpose multimodal assistants. For example, LXMERT [268] introduces a cross-modality encoder to align language semantics with visual concepts for vision-language reasoning (Figure 3.8(a)). The model is pretrained on a masked cross-modality language modeling task, a masked object prediction task, and cross-modality tasks including matching and question answering. CoCa [328] first pretrains a visual encoder on image classification, then a dual encoder with contrastive loss to align vision and language. A multimodal text decoder is subsequently trained with generative objectives for image captioning (Figure 3.8 (b)).

BLIP [153] trains a dual encoder with image-text contrastive loss, followed by an image-grounded text encoder that incorporates cross-attention layers trained with image-text matching loss. An image-grounded text decoder is then trained with a language modeling loss for caption generation (Figure 3.8 (c)). BLIP-2 [152] introduces Q-Former, which connects a frozen image encoder and text encoder. It is trained with image-text contrastive and matching losses for alignment, and with an image-grounded text generation loss for text synthesis.

3.4 Transfer: from General Models to Embodied Agents

The availability of massive online datasets and ever-increasing computational power has enabled the scientific community to make substantial progress in developing general-purpose LLMs and VLMs. These models are capable of performing tasks

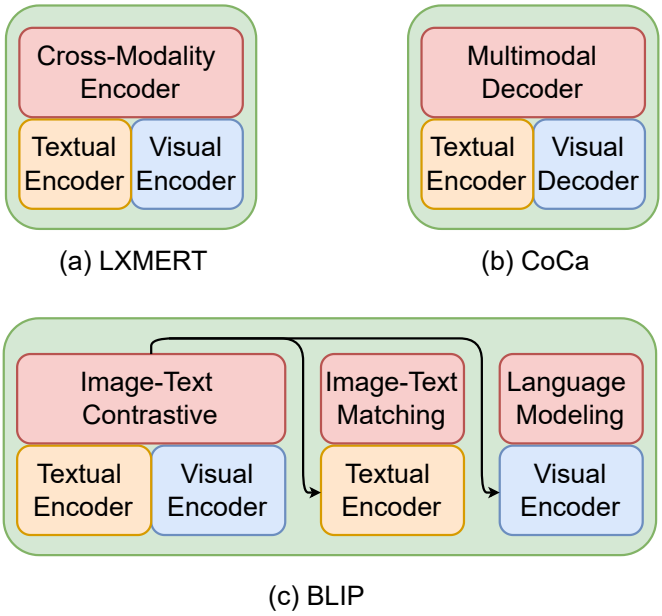


Figure 3.8: Simplified diagram for hybrid models: **(a)** LXMERT [268], **(b)** CoCa [328], and **(c)** BLIP [153].

relevant to robotics, such as long- or short-horizon task planning, commonsense reasoning, and multimodal perception. However, because their primary training data are collected from the web, they lack exposure to the physical world, which limits their abilities in tasks such as object affordances, handling uncertainty and ambiguity, spatial reasoning, and the dynamics of physical interactions.

Vision-language-action (VLA) models extend the capabilities of VLMs by training on datasets derived from physical robot interactions, thereby grounding the models in real-world dynamics. Notable examples include RT-2 [352] and PaLM-E [69]. A significant challenge for this paradigm is its reliance on massive, costly datasets and prohibitive computational resources. Consequently, advancing this direction may require a shift away from scaling transformer-based architectures and toward fundamental algorithmic innovations.

There are two other notable dominant strategies for adapting a general-purpose LLM or VLM to embodied agents. One is direct utilization: a general model can be applied directly through in-context learning or chain-of-thought (CoT) prompting. For example, PaLM-SayCan [28] leverages the LLM, Pathways Language Model (PaLM) [46], to provide high-level semantic knowledge about procedures for performing complex, temporally extended instructions, while value functions associated with low-level robotic skills provide the grounding necessary to connect this knowledge to a specific physical environment. KnowNo [239], as a follow-up to this research, directly utilizes the LLM, PaLM-2L [7], to model uncertainty in complex, multi-step planning tasks. TidyBot [309] leverages the summarization capabilities of the LLM, GPT-3, to infer users' preferences for sorting objects.

ViLa [109] utilizes the VLM, GPT-4V (vision), to perform long-horizon task planning for robotic operations. MOKA [172] leverages visual prompting on the VLM, GPT-4V, for open-vocabulary robot manipulation, annotating marks on images, and converting the prediction of keypoints and waypoints into a series of visual question-answering problems that the VLM can feasibly solve.

The second strategy is task-specific fine-tuning, where a general model is adapted using robotics datasets. This process tailors the model's embedded commonsense knowledge for specific embodied tasks. The standard methodology involves curating a domain-specific dataset and fine-tuning the model to specialize its capabilities. For instance, to teach VLMs about physical concepts, Gao et al. [83] created PhysObjects, an object-centric dataset with extensive annotations of physical concepts. Their study showed that fine-tuning InstructBLIP [55] on this dataset significantly improved its understanding of physical object concepts by capturing human priors from visual appearance. Similarly, KALIE [269] developed an automated pipeline for synthesizing high-quality, affordance-aware data. The authors demonstrated that after fine-tuning on just 50 data points, CogVLM [293] was able to robustly solve new manipulation tasks involving unseen objects. To address a broader lack of robotics knowledge, including affordances and physical concepts, ManipVQA [111] compiled a diverse dataset of interactive objects. This dataset presents challenges in tool detection, affordance prediction, and physical concept prediction. Furthermore, RT-Grasp [315] explored the use of VLMs for numerical predictions in robotics. Their results with LLaVA [173] confirm that VLMs can be successfully adapted for such precise, regression-based tasks.

The volume of research utilizing LLMs or VLMs in robotics is vast [161, 112], making exhaustive coverage impossible. However, much of this study operates at the object category-level, whereas in real life, we interact with objects based on their states. Although prior studies have explored uncertainty [115, 239] and ambiguity [162, 130, 57], they often fail to distinguish between these two concepts and overlook the essential concept of vagueness. A further limitation is that LLMs and VLMs themselves are not well-suited for spatial reasoning and object localization. The following chapters will analyze these challenges and detail our contributions to addressing them. First, the next section discusses the key challenges faced by embodied agents and formally defines the problem settings tackled in this thesis.

3.5 Challenges and Problem Formulation in Embodied Agents

Embodied agents are artificial intelligence (AI) robotic systems that integrate multimodal environmental perception, situated interaction, natural language understanding (NLU), and adaptive learning mechanisms [81]. The principal challenges of embodied agents, which we focus on in this thesis, concern visual perception and communication. As reviewed in the preceding sections, research on VLMs and

LLMs has demonstrated near-human-level capabilities in NLU, visual perception, commonsense reasoning, and content generation. These advances motivate further investigation into household robots that are capable of coexisting harmoniously with humans.

The constraints faced by embodied agents that interact with humans give rise to several *conceptual questions* that must be addressed before tackling the research objectives and research questions proposed at the beginning of this thesis:

- Why clarity about Uncertainty, Vagueness, and Ambiguity (UVA-phenomena) is needed (or not needed) in HRI? Why distinguish them, or should we simply merge them?
- Why is the object state important in HRI?
- How can an embodied agent formalize and handle object-state representations and UVA-phenomena?

3.5.1 Why Clarity about Uncertainty, Vagueness, and Ambiguity (UVA-phenomena) is Needed (or not needed) in HRI? Why Distinguish them, or Should we Simply Merge them?

Before discussing the clarity of the concepts of UVA-phenomena that arise in HRI, it is necessary to first review the definitions of these three words as provided by the official dictionaries, Oxford[3] and Cambridge[4]. *Uncertainty* refers to the state of being uncertain, or a situation that causes one to feel unsure or doubtful. *Vagueness* is the noun form of vague, which describes something that is unclear in a person's mind, lacking sufficient detail or information, indicating a lack of clear thought or attention, or lacking a well-defined shape or form. *Ambiguity* is defined as the state of having more than one possible meaning, or a word or statement that can be interpreted in multiple ways. It can also refer to something that is difficult to understand or explain due to its complexity or the involvement of many different aspects.

Those concepts manifest in different communicative and decision-making processes in HRI. These phenomena can occur independently or in combination [133, 281] in the interdisciplinary research areas of linguistics and cognitive robotics. Their tight relationships are depicted in Figure 3.9 by large circles, and their further categorization is shown by smaller circles. *Uncertainty* includes both epistemic and aleatoric uncertainty, each of which influences the agent's decision-making capabilities [1]. *Epistemic uncertainty* refers to whether the agent possesses sufficient cognitive capabilities and has been provided with adequate knowledge. *Aleatoric uncertainty* arises from the inherent stochasticity of a dynamic environment or

[3] `https://www.oxfordlearnersdictionaries.com/`

[4] `https://dictionary.cambridge.org/dictionary/english/`

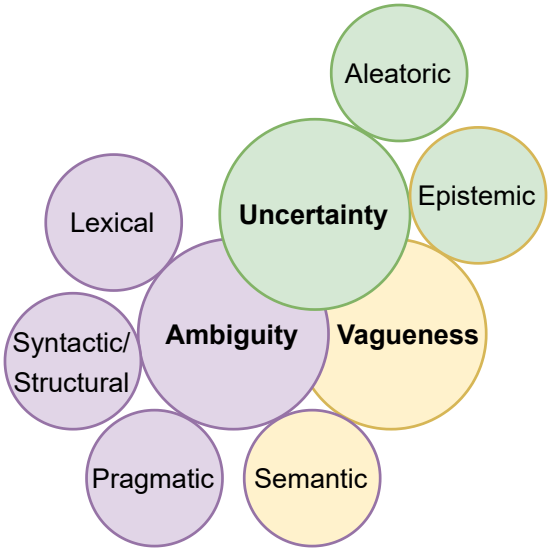


Figure 3.9: The principal problems of Uncertainty, Vagueness, and Ambiguity in the interdisciplinary research of linguistics and HRI. Uncertainty refers to states of incomplete knowledge that can be modeled as a probability problem. Vagueness pertains to concepts with indeterminate boundaries, where membership is a matter of degree rather than binary classification. Ambiguity arises when a linguistic or symbolic expression maps two or more distinct interpretations.

event. *Vagueness* indicates that conceptual boundaries are not clearly defined by the users, which complicates the agent's ability to make decisions [133, 283]. *Ambiguity* pertains to natural language instructions and involves multiple possible interpretations of words, phrases, or sentence structures. Although such instructions may be grammatically correct, they can still yield multiple interpretations in relation to the environment.

At its core, the primary source of *uncertainty* derives from the agent's lack of information, whether due to limitations of internal cognition or the absence of external data. The underlying cause of *vagueness* stems from the imprecise boundaries of concepts, which may result from the user's subjective cognition or from objective characteristics. In contrast, the root cause of *ambiguity* lies in the language instructions provided by the user, particularly when these instructions refer to the environment, leading to multiple possible interpretations or understandings [188]. Because these problems can arise from either different or the same causes, clarifying and distinguishing between them is essential in the context of HRI.

3.5.2 Why is the Object State Important in HRI?

Humans intuitively rely on their understanding of object states and commonsense knowledge to reason about how to interact with everyday objects. However, what is the clear definition of an object state? The concept of an object state from object-oriented programming [24] is: "The state of an object encompasses all of the

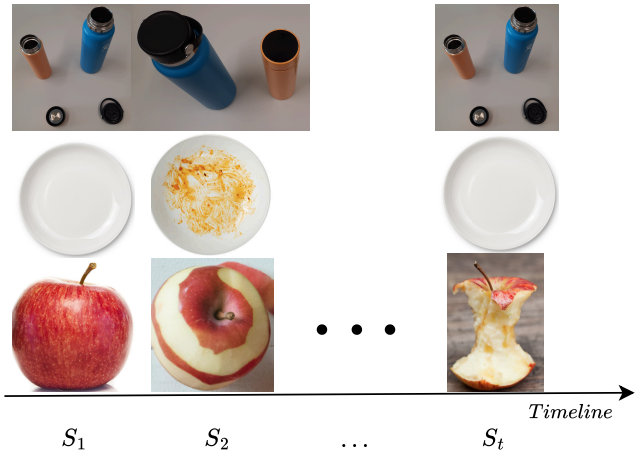


Figure 3.10: Illustration of dynamic object state changes. S_t represents object states at time t.

(usually static) properties of the object plus the current (usually dynamic) values of each of these properties." Therefore, in our context, the properties are object properties (usually static), such as color, shape, size, weight, texture, cleanliness, and rigidity. The values of these properties are dynamic, meaning they can change over time or remain constant, with different properties often changing at varying rates. As shown in Figure 3.10, the bottles and their lids are in a separated state at time S_1; they can be assembled into a combined state at time S_2. The plate can be in a clean state at time S_1 and in a dirty state at time S_2. The apple can be in an intact state at time S_1, a peeled state at time S_2, and a bitten state at time S_t. Among these examples, the bottle exhibits reversible state changes, as it can be separated into two parts again at time S_t. Similarly, a dirty plate can be washed to return to a clean state at time S_t. However, a bitten or cut apple cannot revert to its intact state.

Closely related to the concept of object state is that of affordances. In HRI, affordances [178, 86] refer to the potential actions an object enables for a perceiving agent, which are determined by the object's state and functionality. For instance, consider a bottle. Its affordances change with its state: a bottle with its lid on affords carrying and storing, while the same bottle with its lid removed affords the action of pouring water into it.

We formalize the relationship between the object state (S), property (P), and affordance (A). The observation model ($f_{\text{observation}}$) predicts the object state S from environment V according to property (P).

$$S = f_{\text{observation}}(V \mid P) \tag{3.1}$$

The affordance inference model ($f_{\text{affordance}}$) that infers affordances A from S using commonsense knowledge CK.

$$A = f_{\text{affordance}}(S \mid \text{CK}) \tag{3.2}$$

From Equation 3.2, we observe that object state and commonsense knowledge are crucial for determining affordances. Overlooking the object state can lead to unforeseen consequences [317]. Therefore, we conclude that object state is a fundamental component for effective HRI.

3.5.3 How Can an Embodied Agent Formalize and Handle Object State Representations and UVA-phenomena?

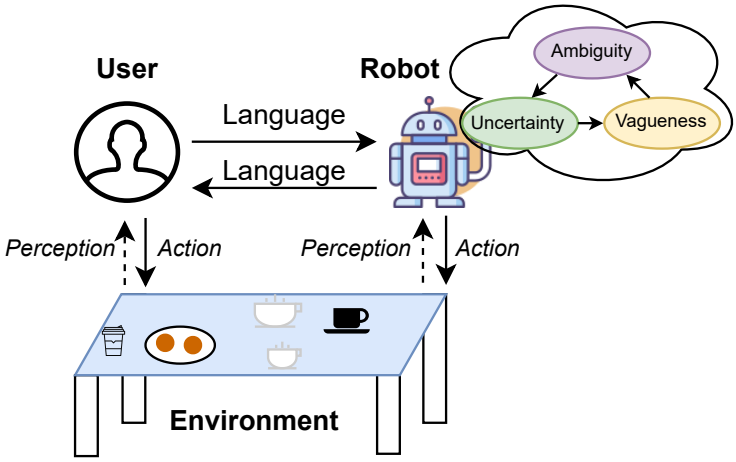


Figure 3.11: Formalization of a human-robot interaction scenario, as seen in Figure 1.1. The scenario comprises three main components: the user, the robot, and the environment. From the human-centered perspective, the user assumes a dominant role by interacting with the environment or issuing commands to the robot in natural language to achieve specific goals. From the agent-centered perspective, the robot perceives the environment and interprets the user's commands to execute actions and provide appropriate responses.

In exploring these problems within a common daily life scenario, we focus on embodied agents that assist humans in the context of HRI. We model the interaction diagram between the user (Language), the robot (Action), and the environment (Vision), as depicted in Figure 3.11. The robot perceives the environment and interprets the user's commands to execute actions and provide appropriate responses.

We extend the observation model (Equation 3.1) into a perception model ($f_{\text{perception}}$). This model fuses language (L) and vision (V) modalities to jointly infer the object state (S) as well as Uncertainty, Vagueness, and Ambiguity (UVA-phenomena).

$$S, \text{UVA-phenomena} = f_{\text{perception}}(L, V) \tag{3.3}$$

Similarly, we extend the affordance inference model (Equation 3.2) into a decision-making model (f_{decision}). This model uses commonsense knowledge (CK) to interpret the perceived object state (S) and UVA-phenomena, generating both corresponding actions (A) and natural language responses (L).

$$\text{A, L} = f_{\text{decision}}(S, \text{UVA-phenomena} \mid \text{CK}) \tag{3.4}$$

Specifically, limitations in users' expressive abilities lead to *linguistic ambiguity*, as users may express goals imprecisely or refer to undefined concepts. This, in turn, results in intentions that lack clear operational boundaries for the agent, referred to as *vague intentions*. These factors, together with robot cognition, further contribute to overall *uncertain decisions* [67, 147].

In traditional dialog systems, dialog managers (DMs) are commonly used to mitigate and interpret UVA-phenomena. More recently, LLMs have been increasingly leveraged to detect and resolve linguistic ambiguity. For instance, KnowNo [239] employs an LLM to generate multiple candidate actions, which aids in identifying linguistic ambiguities. Furthermore, this framework utilizes conformal prediction techniques for uncertainty estimation. The environment is modeled as a Partially Observable Markov Decision Process (POMDP) to account for incomplete information. More advanced VLMs integrate both language and vision for tasks involving understanding, reasoning, and generation. This integration enhances the agent's ability to achieve epistemic certainty. Additionally, fusing multiple modalities-such as speech, gesture, and gaze-has been shown to reduce uncertainty [279], although it introduces additional complexity. The reviews presented in Chapter 2 and earlier in this chapter show that LLMs and VLMs are beginning to bridge the gap between perception (recognizing objects and their states), communication (interacting with users), common sense (reasoning for decision-making), and action (predicting affordances). **This thesis explores the role of DMs, LLMs, and VLMs in understanding object state and UVA-phenomena.**

3.6 Chapter Summary

This chapter began by examining the key challenges inherent to vision-language datasets. Based on the nature of model inputs and outputs, we categorized existing datasets into three overarching paradigms: observation-only, retrieval-based, and interactive vision-language datasets. Within this framework, we positioned embodied agents and virtual game environments under the interactive paradigm, which constitutes the central focus of this thesis. Prior to the emergence of large-scale transformer-based general-purpose models, dataset construction and model training were conducted in a more isolated and task-specific manner. Contemporary transformer-based models, however, are trained on a comprehensive amalgamation of all aforementioned dataset types. Understanding these distinctions remains important for appreciating the developmental trajectory of the field.

Then, we reviewed the influence of architectural choices and training objectives on model behavior and performance, with particular attention to open-source or open-weights general-purpose VLMs. Such knowledge is essential for us, both for selecting models aligned with the objectives and for addressing inherent limitations.

Subsequently, we analyzed the transferability between general-purpose models and embodied agents. Owing to constraints imposed by the scarcity of large-scale physical-world datasets and the substantial computational cost of training, we argue that a purely transformer-based approach is currently unsuitable for advancing research on VLAs. More fundamentally, progress in this area will require methodological innovation rather than continued scaling alone. A more pragmatic strategy involves prompting or lightly fine-tuning general-purpose models with small, domain-specific datasets to better leverage their capabilities for robotic study.

The findings from our literature review, spanning Chapter 2 to Chapter 3.4, can be summarized in three points:

(1) The conceptual problem of UVA-phenomena remains poorly defined, characterized by inconsistent terminology, uneven emphasis, and a lack of a unified conceptual framework.

(2) The representational problem lies in the fact that current embodied agent studies operate primarily at the object category level during interaction, whereas humans interact with everyday objects at the state level.

(3) The methodological problem of existing general VLMs is that they are trained mainly for token classification tasks, whereas object localization, one of the core requirements for embodied agents, is a regression task.

We clarify the conceptual foundations of UVA-phenomena in Section 3.5.1, state the importance of object states in HRI in Section 3.5.2, and formalize an embodied agent that perceives at the object state level and handles UVA-phenomena in Section 3.5.3.

Building on the clarified concept of UVA-phenomena, we investigate solutions to linguistic and visual ambiguities in visually grounded human-robot dialog in Chapter 4. At the object state level, we examine architectural paradigms enabled by foundation models for robotic task planning in Chapter 5. A collaborative study addresses multimodal language processing, grounded in the foundations presented in Chapter 6. To fundamentally overcome the inherent limitations of LLMs, we explore and evaluate a fine-tuning strategy for VLMs applied to numerical prediction tasks in Chapter 7. Overall, each contribution is dedicated to addressing one or more of these challenges.

Chapter 4

Learning Visually Grounded Human-Robot Dialog

4.1 Introduction

Visually grounded human-robot interaction (HRI) is usually situated in an environment that can be perceived visually and with other sensors, and in which a robot can interact or execute commands. The interaction usually involves a multi-turn dialog. Especially in complex scenarios, ambiguities often arise when one person instructs another to perform a task. Multiple turns of conversation are needed to unambiguously determine the goal. Conducting a dialog that requires disambiguation remains a scientific challenge in visually grounded HRI.

The development of visually grounded dialog requires suitable multimodal datasets. Related areas like visual question answering (VQA) [91], visual dialog (VisDial) [56], image-grounded conversations [203], and CLEVR-dialog [140] are mainly focused on understanding the image information, not the interaction within an environment. A seller-buyer interaction in a virtual shop is covered in the Situated and Interactive Multimodal Conversations (SIMMC) dataset [201], where the focus is, however, on fashion and furniture, but not HRI.

To study the ambiguity problem in HRI, we designed the conversational scenario shown in Figure 4.1, proposed a new challenge, and collected a dataset in the area of robotic manipulation research. The robot, Neuro-Inspired COmpanion (NICO) [135], sits at a table on which there are some common household objects. The user uses natural language to instruct NICO to pick or point to one of those objects. We assume there are always three different objects on the table, and the robot cannot point to two objects at once. Sometimes, the user's command may be ambiguous. For example, the user instruction could contain two targets (e.g., `Show me the lemon and apple`). NICO can understand that this is an ambiguous instruction and give feedback. Another situation is that the user gives an unambiguous instruction (e.g., `Point to the red object`), but multiple objects have

the same color. In such situations, the robot needs to use visual information to understand the ambiguity of the command and request the user for additional input. Once the user and NICO reach a consensus, NICO will execute the appropriate action.

To achieve this goal, the agent must recognize objects in the environment and associate visual information with their corresponding names, colors, and spatial positions as conveyed through language.

(1) We explore the challenge of intermediate complexity for visually grounded human-robot conversations, focusing on the ambiguity between human instructions and the surrounding environment. To support this exploration, we introduce a multimodal dataset comprising images and dialog instances grounded in those images.

(2) We propose a novel hybrid neural architecture that tracks dialog state in visually grounded human-robot conversations.

(3) We demonstrate that the proposed hybrid architecture, when trained on the introduced multimodal dataset, can effectively address this intermediate-complexity challenge, particularly in handling linguistic ambiguity and visual perception.

The remainder of this chapter is structured as follows. Section 4.2 focuses on bridging gaps in visually grounded human-robot dialog. Section 4.3 outlines the task definition and provides a summary of the dataset. Our hybrid neural architecture is detailed in Section 4.4. Section 4.5 presents and analyzes the experimental results. Section 4.6 discusses these findings. Finally, Section 4.7 concludes the chapter and suggests directions for future work.

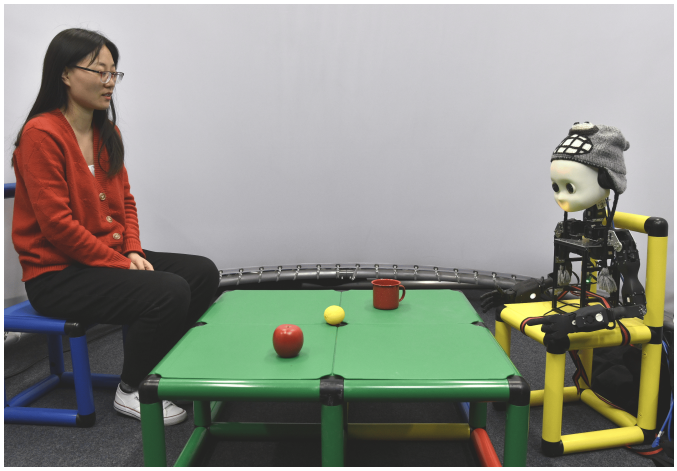


Figure 4.1: Visually grounded human-robot conversational scenario. The user is talking to the NICO robot about objects on the table.

4.2 Bridging Gaps in Visually Grounded Human-Robot Dialog

4.2.1 Multimodal Datasets

SIMMC incorporates multimodal inputs (e.g., vision, memories of previous interactions, and users' utterances) for multimodal actions (e.g., representing the search results while generating the agent's next utterance) [201]. The next generation of SIMMC 2.0 is still focused on a shopping scenario [139], which consists of four main benchmark tasks: Multimodal Disambiguation, Multimodal Coreference Resolution, Multimodal Dialog State Tracking, and Response Generation [139]. The primary task of SIMMC is dialog state tracking. To solve this task, many studies focus on leveraging transformer architecture, such as Bidirectional Encoder Representations from Transformers (BERT) [63, 114], Bidirectional and Auto-Regressive Transformers (BART) [151], and Generative Pre-trained Transformer (GPT) [231, 120, 227].

VQA [91] and VisDial [56] are used for commonsense learning of visual-language representations, which are both based on Microsoft Common Objects in Context (COCO) [165]. In contrast, GuessWhat? [59] is a two-player guessing game, which aims to find an unknown object in a rich image scene by question-answering strategies based on reinforcement learning (RL). The CLEVR-dialog [140] dataset focuses on multi-round reasoning learning in visual dialog, which constructs a dialog grammar that is grounded in the scene graphs of the images from the CLEVR dataset. Lu et al. [182] present Vision-and-Language BERT (VilBERT), which extends BERT [63] to process visual and linguistic input. Murahari et al. [206] pretrained the VilBERT on the VQA [91] dataset and fine-tuned it on the VisDial dataset, then created the VisDial-BERT for multi-turn visually grounded conversations. The augmented extended train robots dataset [134], which expands the extended train robots dataset [4], offers tasks for a robotic agent to reach for objects in three-dimensional space based on augmented reality and a simulation environment. However, in all the above approaches, no dialog studies focus on the visually grounded HRI domain.

4.2.2 Neural and Hybrid Approaches

A traditional dialog pipeline usually contains natural language understanding (NLU), dialog manager (DM), and natural language generation (NLG). The core part of a dialog system is the DM, including a dialog state tracker and dialog policy [210, 26]. Recurrent neural networks (RNNs) can be trained in an end-to-end manner to match an observable dialog history to output sentences [102]. A hybrid approach is also attractive for task-oriented dialog modelling, since it can combine multiple approaches, such as rule-based and data-driven [87]. Hybrid-code networks combine an RNN with domain-specific knowledge, which perform well on the bAbI dataset [25], and are applied to a real customer support domain [306].

Due to the lack of robotic, visually grounded datasets, we generated an artificial multimodal dataset for our HRI domain. This dataset primarily focuses on human language use, including naturally occurring ambiguities, to instruct a humanoid robot to point to an object in the environment. Drawing on the principles of Hybrid Code Networks [25], we train an RNN for multimodal dialog state tracker (M-DST) and define a Human-robot **i**nteraction Policy (HiPolicy) for our scenario.

4.3 Multimodal Dataset for Human-Robot Dialog Tasks

We propose a dataset consisting of two modalities, visual scenes and conversations in text form. The visual scenes were generated with Blender[1] and CoppeliaSim[2]. The conversations consist of the user and NICO discussing the characteristics of objects in the scene, such as their position, color, and name.

4.3.1 Task Definition

The task is set in the context of robot interaction with human instruction, which requires understanding user utterances and recognizing objects. The subtasks are:

Subtask 1: Opening greetings. The greeting is the start of the conversation.

Subtask 2: Receiving user requests. We assume that the user request is always related to the objects on the table. Nevertheless, there are still two types of ambiguities: ambiguous instructions from user utterances and ambiguous scenes that contain multiple objects with the same characteristic (e.g., color). In this subtask, the agent clarifies ambiguous instructions from the user. If an instruction does not contain enough information to identify any object on the table. NICO will, therefore, ask the user to specify and include more detailed information, which we call Unspecified Object Instruction (UOI, a_5). If the user's instructions contain multiple names or colors, NICO will announce its inability to point to multiple objects at once. We term this action Ambiguous Instruction Recognition (AIR, a_6). When the robot receives unambiguous instructions, it will confirm the valid instruction. We term this action Unambiguous Instruction Recognition (UIR, a_7). After this, NICO will detect the relevant objects using the camera command Image Call (IC, a_8). Table 4.1 shows all ambiguous and unambiguous instructions in this subtask.

Subtask 3: Confirming user requests. Upon receiving results from OD, the robot confirms the user's request. The main challenge of this subtask is matching an unambiguous instruction with the environment, so the model can find a corresponding dialog action (refer to Table 4.2 for a list of all dialog actions). NICO tackles this

[1] https://www.blender.org/
[2] https://www.coppeliarobotics.com/

Table 4.1: Instruction utterances in human-robot dialog. The parts left and right of the slash can be substituted for each other. An ambiguous instruction refers to multiple objects. An unambiguous instruction refers to one specific object.
Show me: SM, Where is: WI, Point to: PT, Where are: WA.

Unambiguous instructions	`SM/ WI/ PT the [name].`
	`SM/ WI/ PT the [color] object/ [name].`
	`SM/ PT the [position] object/ [name].`
	`SM/ PT the [color] object/ [name] on the [position].`
Ambiguous instructions	`SM/ PT the object on the table.`
	`SM/ WA the [color1] and [color2] objects.`
	`SM/ WA the [name1] and [name2].`

matching problem with one of the following approaches. If the user asks for the position, he replies with the name. We term this action Name Confirmation (NC, a_9). If the user's instructions contain a name and/or a color, NICO replies with the position. We term this action Position Confirmation (PC, a_{10}). If the user requests an object by its color, name, and position, NICO replies with the name. We term this action Object Identification (OI, a_{11}). If the user's request contains ambiguous information in relation to the environment, so that the matching task cannot be fulfilled, the robot should recognize this ambiguity. An action we term Scene Ambiguity Recognition (SAR, a_{12}). The robot can also automatically detect if an image capture failure within the camera occurred, henceforth called Visual Recognition Failure (VRF, a_{13}).

Subtask 4: Issuing *Arm_action_calls*. After getting the confirmed information, the user either gives a negative or an affirmative answer. For the negative response, the robot will ask for the reason and guide the user to restart the conversation. When receiving an affirmative response, the robot will call for the arm motor to execute the specific action.

4.3.2 Benchmark Dataset

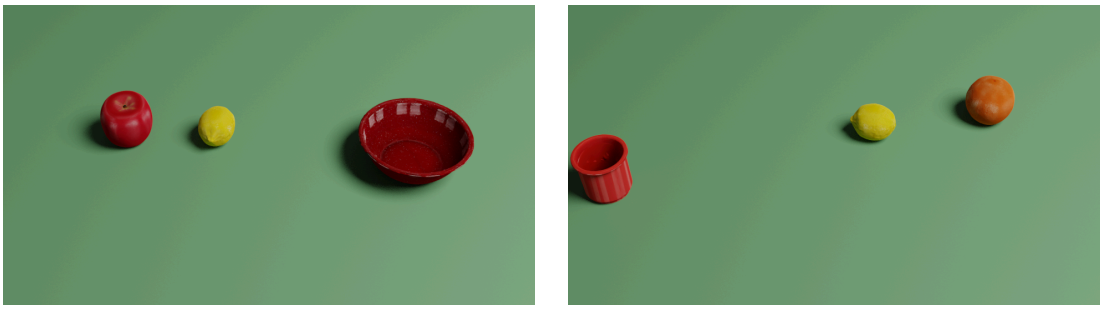
Visual Scenes

We employed a subset of everyday objects from the Yale-CMU-Berkeley (YCB) benchmark [30], focusing on those commonly encountered in domestic and daily environments. We use 28 objects each for both, the training and test set, where each set contains 9 colors (red (8 objects), blue (3 objects), yellow (5 objects), brown (2 objects), green (1 objects), orange (3 objects), black (2 objects), white (3 objects), purple (1 objects)). We selected three different objects out of 28 objects for every scene. Following the combination formula, we generate 3276 scenes for

Table 4.2: Dialog actions in human-robot dialog.

Tasks	**NO.**	**Content**
Subtask 1	$a_1 \sim a_4$	`[Greeting], what can I do for you.`
Subtask 2	a_5, (UOI)	`Please give me a specific target instruction.`
	a_6, (AIR)	`I cannot point to multiple things at once.`
	a_7, (UIR)	`Ok, let me check.`
	a_8, (IC)	`image_call`
Subtask 3	a_9, (NC)	`Do you mean the [name]?`
	a_{10}, (PC)	`Do you mean the one on the [position]?`
	a_{11}, (OI)	`Ok, let me show you the [name].`
	a_{12}, (SAR)	`There is more than one object you ask for.`
	a_{13}, (VRF)	`I cannot see anything.`
Subtask 4	a_{14}	`arm_action_call`
	a_{15}	`Am I wrong, or do you want to change your mind?`
	a_{16}	`It is fine, you can try again by saying hello.`
	a_{17}	`Sorry, I am still learning.`
	a_{19}	`Do you want to try again, start by saying hello.`
	a_{20}	`You are welcome, See you.`

every set. An ambiguous scene means objects have the same color. An unambiguous scene means the objects have three different colors, as shown in Figure 4.2. For each set, there are 1136 ambiguous scenes and 2140 unambiguous scenes. The test and training sets contain the same number of objects and the same color balance to let the percentage of ambiguous scenes be the same (34.7%).



(a) (b)

Figure 4.2: Example scenes in visually grounded human-robot dialog. (a) **Ambiguous scene**, (b) **Unambiguous scene**.

Conversations

Based on the task that we define in Section 4.3.1 and the visual scenes, we generate a dialog instance dataset, which is inspired by Bordes and Weston's study [25, 304]. We use 15 unambiguous instruction utterances (see Table 4.1) and create 15 templates for every visual scene. Every template also includes some ambiguous instructions. Overall, we have 147,420 dialog instances for both the training and test sets. For the training set, there are 117,936 dialog instances for training and 29,484 dialog instances for validation.

4.4 Proposed Method: HYNA

We proposed HYbrid Neural Architecture (HYNA), which contains six components: automatic speech recognition (ASR), text-to-speech (TTS) synthesis, object detection (OD), Robotic Planner (RP), Human-robot **i**nteraction Policy (HiPolicy), and multimodal dialog state tracker (M-DST), as shown in Figure 4.3. The cycle begins when the user starts with a greeting. The very first action of the architecture is to process the user input with its ASR. The ASR result is then fed into the M-DST, which predicts the dialog action (Table 4.2 states all the dialog actions used in this study). In addition to the user's utterance, the M-DST potentially has two other input modalities, depending on the situation. One is from the OD containing the scene's visual information. The other is its previous dialog action. Based on the classified action, the HiPolicy determines NICO's behavior. The behavior entails detecting the object with its camera, pointing to a target, or forming appropriate verbal responses. The main contribution of this contribution is that the M-DST can handle the ambiguity in users' linguistic instructions and environmental contexts. The OD and M-DST are both neural models trained on our datasets. In this chapter, we mainly focus on OD, M-DST, and HiPolicy.

4.4.1 Object Detection (OD)

Our object detection is realized with the single-stage object-detector RetinaNet [164]. The neural architecture backbone of RetinaNet is formed by a feature pyramid network (FPN) and a connected deep residual network. In essence, it constructs a semantic feature map at multiple scales, thereby compensating for the convolutional neural networks (CNNs)'s low resolution on high-level feature maps. The FPN regression and classification subnetworks perform the actual object detection. The RetinaNet is trained on a corpus of one thousand images like the one shown in Figure 4.4. The images are generated using Blender and automatically annotated with bounding boxes. During the interaction, the OD is invoked by the predicted dialog action to perceive environment information. The RetinaNet subsequently detects the objects and returns the object names, positions, and colors. These attributes can then be used to formulate appropriate responses through the M-DST and HiPolicy.

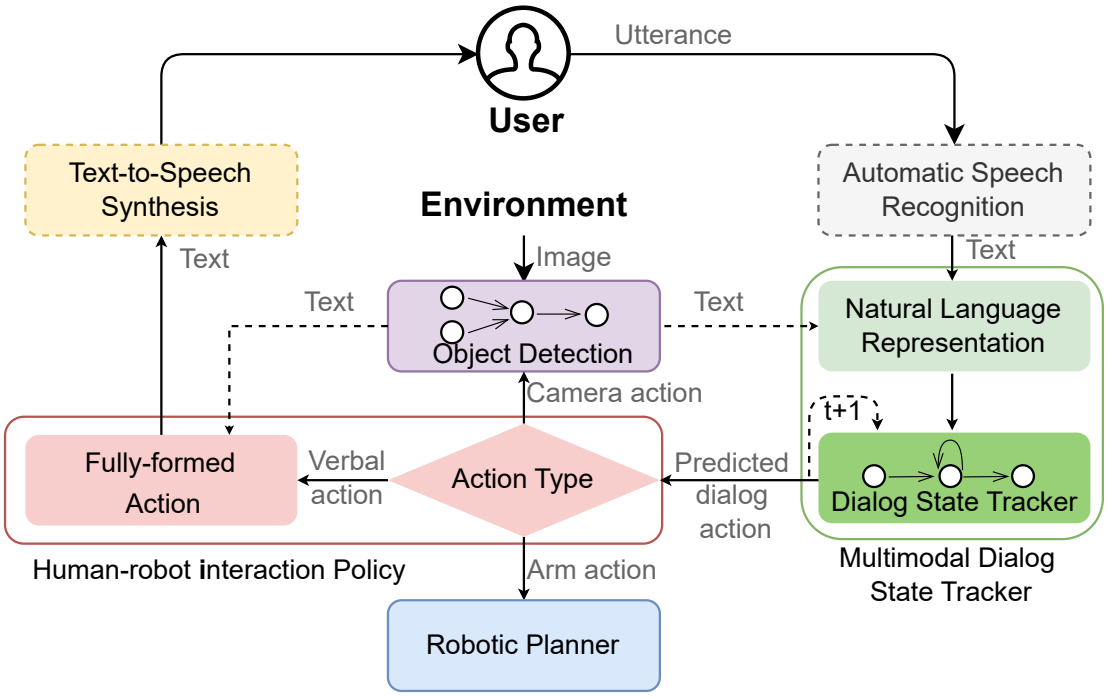


Figure 4.3: HYbrid Neural Architecture (HYNA). The main components, Object Detection (OD) and the Multimodal Dialog State Tracker (M-DST), are both trainable deep neural networks (DNNs), while the Human-Robot Interaction Policy (HiPolicy) is rule-based. The auxiliary components, Automatic Speech Recognition (ASR), Text-to-Speech (TTS) synthesis, and the Robotic Planner (RP), are pretrained modules. The dashed arrows denote conditional inputs.

4.4.2 Multimodal Dialog State Tracker (M-DST)

Multimodal Feature Representation

As shown in Figure 4.3, HYNA is based on multimodal audio-visual input, consisting of the user's utterance and an image of the environment. The M-DST receives purely textual input (see Figure 4.4, image I and utterances $U = [u_1, \ldots, u_n]$).

Therefore, the representation of these two modalities for M-DST is essential. As reviewed in Section 2.2.3, many methods can be chosen for this purpose. However, given the need for multimodal fusion and the successful application of these methods in other natural language processing tasks, we opt for simple representations: bag-of-words (BoW) and a pretrained utterance embedding (UE) method.

For the BoW representation, we first build a vocabulary dictionary from the training set. We then construct the representation vector by summing the individual one-hot vectors corresponding to the words in this dictionary. For the UE representation, we compute the mean vector of the word vector representations obtained from word to vector (Word2Vec). The Word2Vec model [194] was pretrained using the Skip-gram method on the Google News corpus.

These two representation methods are applied to both utterances and image descriptions. Moreover, each dialog instance d_j combines user utterances, image information and labeled dialog actions ($d_j = [(u_1, a_1), \ldots, (I, a_i), \ldots, (u_n, a_n)]$, i is the position of an image reading within a dialog instance, n is the number of conversational turns). There is a total of m dialog instances ($D = [d_1, \ldots, d_m]$).

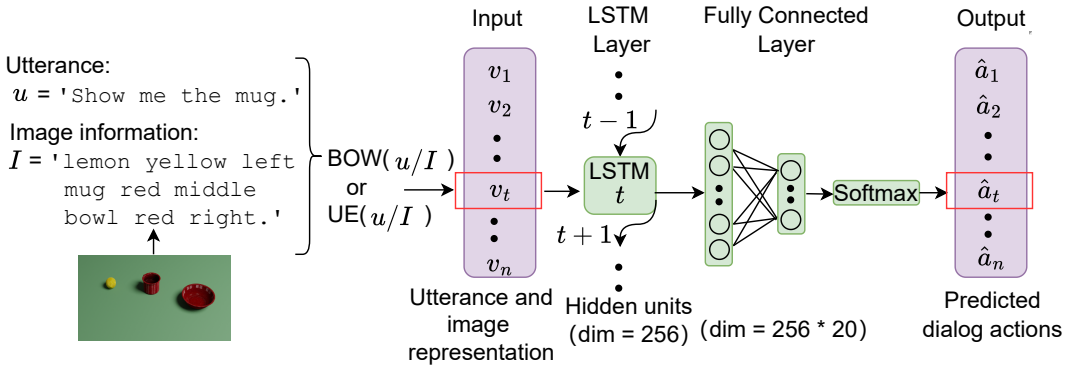


Figure 4.4: Multimodal dialog state tracker (M-DST). We use the BOW or UE to represent utterances and image information. These are fed as input, v_t, at time step t, into the network. The output $\hat{a}_t$ at time step t denotes the predicted dialog action. n is the number of conversational turns of the person as well as of the agent.

LSTM Network

We train a RNN, specifically an LSTM network [104], for visual and textual processing. As illustrated in Figure 4.4, the LSTM network consists of an input layer, an LSTM layer, a fully connected layer (FCL), a softmax function, and an output layer.

The user's utterances U and the image I are both represented using either BoW or UE. Additional variants incorporate the previously predicted dialog action (PDA) as a third input. Each utterance u_i in the dialog instance, its corresponding image information I, and its PDA a_{i-1} are concatenated to form a feature vector. These feature vectors are fed into the `long short-term memory (LSTM)` network. The output of this `LSTM` is then passed to the `FCL`, followed by the application of the `softmax function`. The output of the `softmax function` consists of the predicted dialog actions ($\hat{a}_1, \ldots, \hat{a}_n$).

4.4.3 Human-Robot Interaction Policy (HiPolicy)

In the previous section, we explained how to train an LSTM network for the M-DST. Additionally, a rule-based policy is required for knowledge extraction and decision-making. We designed HiPolicy to serve as this rule-based policy. Based on the dialog action predicted by the M-DST, HiPolicy determines the robot's behavior. Possible robot behaviors include requesting the camera to capture an image for OD, requesting the RP to execute the predicted action (e.g., pointing

to the target object), and generating appropriate verbal responses for the TTS synthesis.

4.5 Experiments and Results

4.5.1 Human-Robot Dialog Action Prediction

In this experiment, we aim to identify the optimal training strategy for our proposed method, HYNA, by quantitatively evaluating different variants. The primary idea behind HYNA is to integrate the visual modality into a dialog state tracker to develop a multimodal tracker for visually grounded Human-Robot Dialog. Accordingly, the architecture's core components are OD and M-DST. We employ the state-of-the-art (SOTA) OD model, RetinaNet, to convert the visual modality into text descriptions. After fine-tuning exclusively on our visual scenes, the extraction achieves 100% accuracy. Consequently, our quantitative evaluation focuses on comparing different variants of the M-DST component.

Experimental setup

During training, each dialog instance constituted an individual minibatch. For parameter updates, we employed non-truncated backpropagation through time (BPTT). As the task is formulated as a multi-class classification problem, we used `categorical cross-entropy (CCE)` to compute the loss. For parameter optimization, we chose `adaptive delta (AdaDelta)` [332] to minimize the loss function.

We vary the multimodal feature representations to evaluate their impact on the performance of the M-DST. We selected BoW and UE as our representation methods. The conversational context is another critical factor in dialog. Therefore, we augment the model configuration with the previously predicted dialog action (PDA) to create new variants.

Overall, we evaluated six M-DST variants: BoW features only, UE only, a combination of BoW and UE, and each of these three configurations with the addition of the previous system action. For the distinction, we denote them as M-DST_{BOW}, M-$\text{DST}_{\text{BOW+PDA}}$, M-$\text{DST}_{\text{UE}}$, M-$\text{DST}_{\text{UE+PDA}}$, M-$\text{DST}_{\text{BOW+UE}}$, and M-$\text{DST}_{\text{BOW+UE+PDA}}$. Each configuration was trained nine times with different sampling orders over 30 epochs to reduce variance and mitigate potential bias.

Results

Subtask 1 (*Opening greetings*) and subtask 4 (*Issuing Arm_action_calls*) are both essential parts of our visually grounded human-robot dialog. Since they are not the focus of our research, we define them in a simple fashion, using the same data in training and test sets. The M-DST successfully predicts the correct dialog action with an accuracy of 100% for all action classes belonging to these subtasks.

Table 4.3: HYNA variants: Average accuracy in subtask 2 and subtask 3 (Human-Robot Dialog). Unspecified Object Instruction (UOI), Ambiguous Instruction Recognition (AIR), Unambiguous Instruction Recognition (UIR), Image Call (IC), Name Confirmation (NC), Position Confirmation (PC), Object Identification (OI), Scene Ambiguity Recognition (SAR), Visual Recognition Failure (VRF).

Labeled actions	Mean accuracy(%)±Standard deviation					
	M-DST$_{\text{BOW}}$	M-DST$_{\text{BOW+PDA}}$	M-DST$_{\text{UE}}$	M-DST$_{\text{UE+PDA}}$	M-DST$_{\text{BOW+UE}}$	M-DST$_{\text{BOW+UE+PDA}}$
UOI	100±0	100±0	100±0	100±0	100±0	100±0
AIR	100±0	100±0	73.98±12.19	77.50±10.45	100±0	100±0
UIR	100±0	100±0	98.33±0.85	99.58±0.63	100±0	100±0
IC	100±0	100±0	98.72±1.94	99.91±0.26	100±0	100±0
NC	78.24±5.84	78.96±6.08	98.09±2.18	99.99±0.02	95.51±3.93	97.96±3.29
PC	100±0	99.94±0.18	97.76±1.54	99.48±1.52	100±0	96.64±9.44
OI	99.56±0.75	100±0	89.24±12.99	98.15±2.64	99.81±0.55	97.63±4.40
SAR	91.65±6.3	90.98±7.84	55.79±22.32	42.70±15.85	74.17±7.47	67.40±9.97
VRF	100±0	100±0	97.38±6.32	99.83±0.48	100±0	100±0

Table 4.3 only shows the results for subtask 2 (*Issuing user requests*) and subtask 3 (*confirming user requests*). Mean and standard deviation of the accuracy are computed from 9 runs. The challenge of subtask 2 is to deal with an ambiguous instruction from the user, and the challenge of subtask 3 is to correctly combine the user's instruction and image input.

Action AIR responds to a type of ambiguous instructions (see table 4.1). Compared with variants M-DST$_{\text{BOW}}$ and M-DST$_{\text{BOW+UE}}$ that can predict the dialog action with 100% accuracy, variant M-DST$_{\text{UE}}$ cannot properly react to a user's unspecific instruction, achieving only an accuracy of 73.98%. Action SAR responds to an ambiguous scene. The variant M-DST$_{\text{BOW}}$ performs better than M-DST$_{\text{UE}}$ and M-DST$_{\text{BOW+UE}}$ in situations when SAR is the expected dialog action. In unambiguous situations when action NC is expected, utterance embedding cannot help the model handle ambiguous situations, but it helps the model perform well in unambiguous situations. Comparing variant M-DST$_{\text{BOW}}$ with M-DST$_{\text{BOW+UE}}$, a striking difference is that the addition of UE helps in NC but has a negative effect in SAR.

4.6 Discussion

The results indicate that using BoW to represent the inputs for the M-DST model is a worthwhile option to tackle the visually grounded human-robot dialog challenge. For the BoW, we found that one problem is the multiple meanings of the word *orange*, which can be either a color or an object name, but is represented the same via BoW. For the UE in subtask 2, M-DST may predict a wrong dialog action, e.g., for *Show me the [name1] and [name2]*, it incorrectly predicts UIR. The reason is that the UE uses Word2Vec, which does not have a representation of stop words like *and*. Furthermore, the pretraining of the UE on a non-related corpus could explain

some issues. For the action SAR in subtask 3, the predicted dialog action is NC, which means the model misunderstood the user's instruction for a position. These limitations of our model are related to the simple utterance representations, which might be remedied in future work by more sophisticated sentence embeddings.

4.7 Chapter Summary

In this chapter, we explored the solution for our **Objective 1**:

> Develop a hybrid neural architecture that integrates visual content into a conversational agent, enabling a domestic robot to resolve linguistic and visual ambiguities in situated HRI.

We contribute a novel method HYNA for visually grounded human-robot conversations that integrates visual content into a conversational agent. The results demonstrate that our proposed model can resolve both linguistic and visual ambiguities in situated HRI. HYNA was inspired by Hybrid Code Networks [306] and supported by the fundamentals that described in Section 2.2.3.

Our quantitative experiments show that HYNA, using a simple bag-of-words (BoW) representation, outperforms HYNA with a pretrained Word2Vec representation, UE. Notably, the model generalizes to objects in the test set that were never observed during training, indicating its ability to handle unseen objects, provided that the object detector (OD) can recognize them. Moreover, although the number of dialog actions is fixed, these actions depend on the interaction scenario rather than the model architecture, allowing the system to be adapted to a wide range of scenarios and application domains. The limitation of HYNA is that ill-formed utterances - where user language does not strictly follow grammatical rules - were not included during training, which limits the robustness of the model. Additionally, the objects in the tabletop scene are fixed; introducing greater variability, such as a random number of objects, could further improve robustness and applicability.

Despite these limitations, HYNA leveraged SOTA visual and language models at the time to propose a novel hybrid architecture, which was valuable in its historical context. However, recent advances in foundation models, particularly large language models and large vision-language models, have fundamentally reshaped research on dialog and multimodal dialog systems. We acknowledge that many of the limitations of our hybrid neural architecture are addressed by these newer models, which incorporate world knowledge, advanced reasoning capabilities, language mastery, efficiency, and multi-task learning. Consequently, we shift our approach from training dialog systems based on predefined scripts to leveraging foundation models directly. On the other hand, we observe a mismatch in object-perceiving level between current embodied agent research and how humans interact with objects in real-world environments. We focus on this mismatch in the next chapter.

Chapter 5

Object State-Sensitive Agent Task Planning

5.1 Introduction

Object attributes such as color, shape, and size play important roles in shaping the diverse states objects can exhibit, significantly impacting our daily interactions. Understanding and managing these states is crucial, as overlooking them can result in unforeseen consequences. Humans intuitively rely on their understanding of object states and common sense to interact with everyday objects. However, integrating state-sensitive knowledge poses a significant challenge for robotic systems. Enumerating and integrating this nuanced understanding into robotic systems remains a complex endeavor.

Recent approaches that leverage large language models (LLMs) and large vision-language models (VLMs) [342] have shown promising results in tasks that require human-level commonsense knowledge. Such approaches use commonsense reasoning ability of LLMs to interpret natural language goals [345], such as '`put one apple into the fridge`.' According to those goals, LLM generates the suggested action plan, 'next actions: pick up the apple, move to the kitchen, open the fridge, ...' However, the object's physical state (e.g., intact apple, sliced apple, etc.) is crucial yet less considered in the task planning [257]. To the best of our knowledge, there is hardly any research that leverages a pretrained neural network (e.g., LLMs or VLMs) to address the integration of object states into planning tasks for household robots. To address this gap, we introduce an Object State-Sensitive Agent (OSSA), which utilizes commonsense reasoning of pretrained neural networks for robotic task planning.

We pose several real-world challenges in OSSA task planning. First, in order to solve the tasks, the agent not only involves identifying different objects in the scene but also distinguishes between their *states*. For example, in the `clear the table` [213, 99] task, a robot needs to be able to distinguish between *whole* and

sliced fruit, and between *clean* and *dirty* plates. This is a challenge because existing state-of-the-art (SOTA) object detection models often fail at differentiating between objects in different states. A plan lacking state-sensitive awareness may lead to unexpected results.

Second, the agent needs to employ commonsense reasoning for taking *state-sensitive* actions that correspond to the object states of various scenarios instead of asking the users for an exhaustive design or an intervention. For example, whole fruit goes directly into the fridge or to the cupboard, while sliced fruit can either go into the fridge or be discarded into the trash bin if they are regarded as leftovers. One way to solve the commonsense reasoning problem is using rule-based symbolic approaches [213], but by design, they do not generalize to situations outside their rule base, i.e., they cannot handle new objects and states. Commonsense reasoning with a data-driven model trained on a large dataset that generalizes well (e.g., an LLM [291]) is, therefore, a more viable choice.

Third, the robot needs to identify cases where common sense should not dominate. For example, a robot has to take into account the preferences of the user when handling specific objects in specific states [309]. In the table clearing example, different users might handle the leftover food differently, e.g., some might prefer to discard the leftovers, while others are more frugal and prefer to keep the leftovers for the next meal. The robot, therefore, needs to detect situations where different user preferences come into play, and has to ask the user for clarification instead of arbitrarily choosing one on its own. Existing approaches, however, do not take user preferences into account [239].

In this chapter, we contribute an Object State-Sensitive Agent (OSSA) for the problem of state-sensitive instruction following.

(1) We propose two architectural paradigms: a modular system consisting of an object detection module and an LLM, and a monolithic VLM-only to caption and reason universally for OSSA.

(2) We contribute a robotic task planning benchmark[1] based on 40 cluttered tabletop clearing scenes with 184 objects in total, each paired with three distinct instruction commands.

(3) We demonstrate that both approaches are sensitive to object states (e.g., clean or dirty) and user preferences, generating accurate and state-aware action sequences on our proposed benchmark.

The remainder of this chapter is structured as follows. Section 5.2 focuses on bridging gaps in robotic task planning. Section 5.3 outlines the task definition and provides a summary of the dataset. The architecture of OSSA is detailed in Section 5.4. Section 5.5 presents and analyzes the experimental results, while Section 5.6 discusses these findings. Finally, Section 5.7 summarizes this chapter and suggests directions for future work.

[1] `https://github.com/Xiao-wen-Sun/OSSA.git`

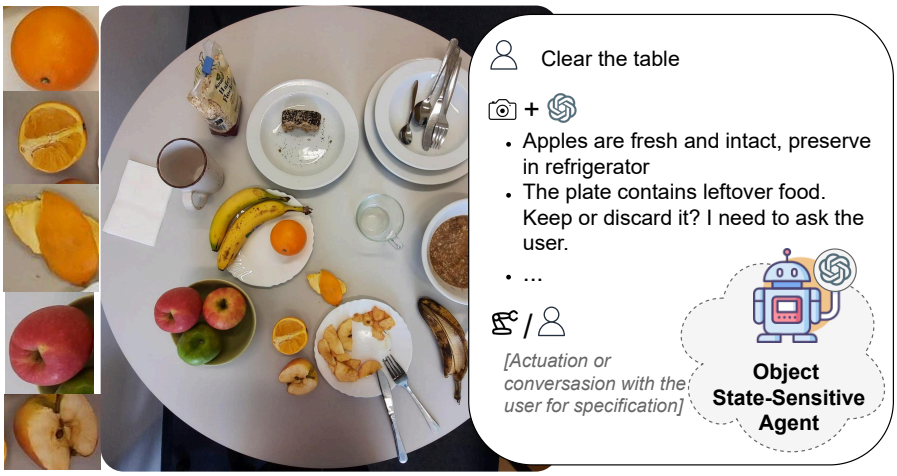


Figure 5.1: Diverse object states in a robotic task planning scene. For example, orange, half-orange, and orange peel; clean napkin and dirty napkin; banana and banana peel. Based on commonsense knowledge, the agent sorts the objects (discards the banana peel in the trash bin; keeps the bananas in the cupboard). However, the robot is not able to decide how to deal with the leftover food because different people may have different preferences regarding leftover food (e.g., half orange and half bread).

5.2 Bridging Gaps in Robotic Task Planning

5.2.1 Vision and Language Models for Task Planning in Robotics

Recently, VLMs [291, 295] and LLMs [342] have transformed robotic task planning, with the integration of LLMs' commonsense knowledge into planning for embodied agents, garnering growing interest [301, 113, 28, 258, 236, 344, 257, 345]. However, these methods assume a singular state for objects in their plans, which may not suit the variability of the domestic settings we focus on. The most widely used visual perception methods, such as OWL-ViT [198], OWL-V2 [197], YOLO [35], and CLIP [230], are not trained to distinguish the object states, e.g., half apple, apple, dirty plate, and clean plate. Therefore, these approaches are not sensitive to object states [347].

One of the closely related studies to ours is TidyBot [309], a system that enables a domestic cleaning robot to adapt to individual user preferences in managing objects, recognizing that each item may require a unique approach based on personal taste. They leverage the summarization capabilities of LLMs to get the user's preference for sorting the objects, e.g., '`put light-colored clothes in the drawer and dark-colored clothes in the closet`'. In this study, we take advantage of commonsense knowledge from LLMs to sort the objects and detect when human preference needs to be involved in automation tasks.

In other research, ViLa [109] utilizes GPT-4V(vision) as VLM to do task planning, which is also close to our focus. However, the difference is that they focus on long-horizon robotic tasks, where the manipulation actions and goals can be derived from the user's instructions. We focus on automation tasks where the robot manipulation goals stem from visual perception. Then our model performs reasoning and generates a manipulation plan, according to the visual perception.

5.2.2 Benchmarks for Task Planning in Household Robots

Gutman et al. [99] use the `clear the table` as an experiment task to explore the users' preference for different autonomy levels of an assistive robot. The results show that the participants prefer the highly autonomous assistant. Other research uses `table clearing task` to study the robot to understand and execute humans' natural language instructions [213]. However, overall, their studies do not mention that the objects from the same category may be in a different state, which would change the robot's action.

From the computer vision (CV) research perspective, two datasets [119] focus on object state recognition in cooking, e.g., whole, peeled, floured, etc. The Object State Detection Dataset [90] involves the object states: open, closed, empty, containing something liquid, containing something solid, plugged, unplugged, folded, and unfolded. However, it does not consider the leftover food in our daily life. From the perspective of robot kinematics research, many benchmarks [27, 352, 117, 282] and methods have been proposed to enable robots to execute manipulation tasks, e.g., picking, wiping, dragging, pushing, pouring, and placing.

We focus on automated task planning and research how to utilize these enabled robot skills to complete certain complex tasks. For example, the robot, according to various object states, generates plans for a robot low-level executor that receives a specification of how to 'pick' and where to 'place' the target.

5.3 Multimodal Dataset for Robotic Task Planning

5.3.1 Task Definition

In our study, we identify two types of object states about leftovers: "containing leftover food", where containers like plates or bowls hold liquid or semi-fluid contents that are not directly manipulable by the robot (e.g., a bowl filled with leftover soup), and "leftover food", referring to food remains that have been sliced or peeled, which robots can handle.

According to these specified leftover types, we consider three different common scenarios:

- T1) The instruction is `clear the table` without specifying what to do with

the leftover food. In this scenario, the robot is supposed to generate a manipulation plan for all the objects except those classified as "leftover food" or "containing leftover food". In this case, the expected behavior of the robot is to ask the user for handling specifications about the leftovers.

- T2) The instruction is `clear the table and keep all the leftover food`. In this scenario, if the object state is "leftover food" or "containing leftover food", the robot should store it in the fridge.
- T3) The instruction is `clear the table and discard all the leftover food`. In this scenario, if the object state is "leftover food", the robot is supposed to throw it into the trash bin.

However, if the object state is "containing leftover food", the robot is supposed to grab the container (not the soup directly) and pour the contents into the trash bin before putting the container into the dishwasher.

5.3.2 Benchmark Dataset

Figure 5.2: Example scenes in robotic task planning. (a) A scene from daily life. (b) A scene generated by a diffusion model.

To quantitatively evaluate the methods that we propose, we built a benchmark dataset, which is composed of 40 scenes involving 184 objects. First, we sourced 27 scenes from our daily lives. Second, to create balanced data, we use a diffusion model [242] to generate an additional 13 scenes, as shown in Figure 5.2. Figure 5.3 shows all the objects that are involved in this dataset and the proportion among them. The same object may be in different states at different times. The object states that we consider, and their proportion are shown in Figure 5.4.

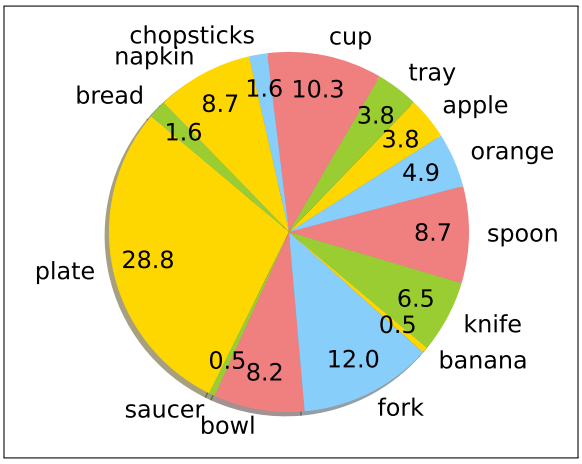


Figure 5.3: Object distribution in robotic task planning(%).

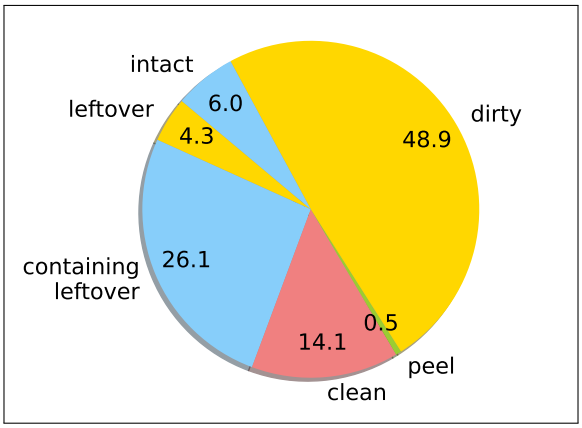


Figure 5.4: Object state distribution in robotic task planning(%).

Annotation Rules

The object manipulation plan format, outlined in Section 5.4.2, was employed for annotating the dataset. We asked two individuals to label the dataset according to set rules to reduce bias. In the end, they collaborated to resolve discrepancies and finalize the annotations.

Object name: use the object name as a JSON key. When more than one item from the same object category is in the scene, add the number behind the name (e.g., 'plate 1', 'plate 2'); **Color**: the object color (e.g., 'white', 'silver', 'orange', 'red'); **Size**: the object size, 'small', 'medium', 'big'; **Shape**: the object shape, 'elongated', 'irregular', 'oval', 'round', 'spherical', 'cylindrical', 'rectangle'; **Container**: use 'true' or 'false' to label the object as a container or not; **State**: If the object is a container, we label it with three states: 'clean', 'dirty', or 'containing leftover food'. If the object is not a container but edible, we label it with three states: 'intact', 'peel', or 'leftover food'. If the object is not a container and inedible (fork, knife, spoon), we label it with two states: 'dirty' or 'clean'; **Destination**: we use four places: 'trash bin', 'fridge', 'cupboard', and 'dishwasher'; **Grasping**

type: we set two types of grasp action: 'top grasp' or 'edge grasp'; **Placing type**: In most cases, the robot places the object in the destination. But in the special case that the object is a container that contains leftover food, the robot should pour the leftover food into the trash bin.

Evaluation Metric

In this study, we aim to generate a manipulation plan for the objects for the robot's low-level executor. The main evaluation metric is accuracy. There are five parts of the output needed for low-level execution. We evaluate our proposed methods' performance in those five parts: 1) State Detection Accuracy (**StaA**), 2) Ambiguous Detection Accuracy (**AmbA**), 3) Destination Generation Accuracy (**DesA**), 4) Grasping Type Generation Accuracy (**GraA**), and 5) Placing Type Generation Accuracy (**PlaA**). Finally, we calculate one overall accuracy, representing how many objects are being predicted correctly, Completion Accuracy (**ComA**).

5.4 Proposed Method: OSSA

5.4.1 OSSA Architecture

We introduce an OSSA that can perceive fine-grained scene information (object name and state, etc.) and generate an appropriate object manipulation plan for the robot's low-level executor according to those visual perceptions. The pseudocode of the system control flow is provided in **Algorithm 1**. OSSA generates object manipulation plans when an utterance (U) and an image of the table (I) are given. A chat system is for ambiguous situations. If ambiguities arise (e.g., encountering a peeled pear on the table), the robot will ask the user for disambiguation. Lastly, a low-level executor will execute language-conditioned instructions, such as picking and placing, to fulfill the tasks. We assume that the low-level executor has these skills, which can be acquired via machine learning (ML) [218] approaches (e.g., reinforcement learning (RL) [129, 16] and Imitation Learning [117]) or Direct Hard-coding.

5.4.2 Experimental Approaches

To explore the best way to employ commonsense knowledge [345] and reasoning [113] capabilities of pretrained neural networks (e.g., LLMs or VLMs) for OSSA, we propose two categories of methods for object manipulation plan generation. The input of the method is a scene (I) and a user's utterance (U). We prompt the LLM to generate structured responses in a JSON-like format, 'object name': { 'color', 'size', 'shape', 'container', 'state', 'grasping type', 'destination', 'placing type' }. 'Color', 'size', and 'shape' are the object attributes that are visible in the scene. The object's physical 'state' is visible in the scene. However, reasoning a human-defined object state requires commonsense knowledge. 'Grasping type' is an action that informs the robot how to grasp the target at the initial place. 'Destination' is

Algorithm 1 System control flow

```
r ← Robot.Initialize()
commands = NULL
while True do
    U ← r.GetUserUtterance()
    I ← r.GetImageofTable()
    {OMP_1, ..., OMP_n} ← r.Ossa(U, I)
    for OMP_i in {OMP_1, ..., OMP_n} do
        if OMP_i.destination is "ambiguity" then
            AI_Response ← r.ChatSystem(OMP_i)
            U ← r.GetUserUtterance()
            OMP_i ← r.ChatSystem(U)              # ChatSystem revise OMP_i
            commands.append(OMP_i)
        else
            commands.append(OMP_i)
        end if
    end for
    if commands not NULL then
        r.LowLevelExecutor(commands)
        commands = NULL
    end if
end while
```

a place that informs the robot where the target should be stored. 'Placing type' is an action that informs the robot what it should do at the destination. Those three items are from commonsense knowledge reasoning.

For object state recognition, a CV system dense captioning [124] can localize salient regions in images and use natural language to describe them. The natural language description of a salient region includes the condition of the object in the region. Therefore, SOTA dense captioning can be utilized in our study. However, the output of the dense captioning model (DCM) does not include 'grasping type', 'placing type', and 'destination'; another module with commonsense knowledge and reasoning abilities is needed to generate them. GPT-4 [321, 215] can not only be a visual perception model but also perform as a VLM. In this study, we utilize both of the abilities of GPT-4. We propose two methods: i) a modular model consisting of a vision processing module (dense captioning model) and a natural language processing model (LLM), and ii) a monolithic VLM-based model.

Modular Method

A modular pipeline [28, 344, 236, 258, 257, 113, 309] combines two models, e.g., vision and language. First, a vision detection model (e.g., YOLO [35], ViLD [94], OWL-ViT [198], OWL-V2 [197], etc.) detects the objects in the scene. Second, an LLM processes the user's instruction and the output of the vision model. Those

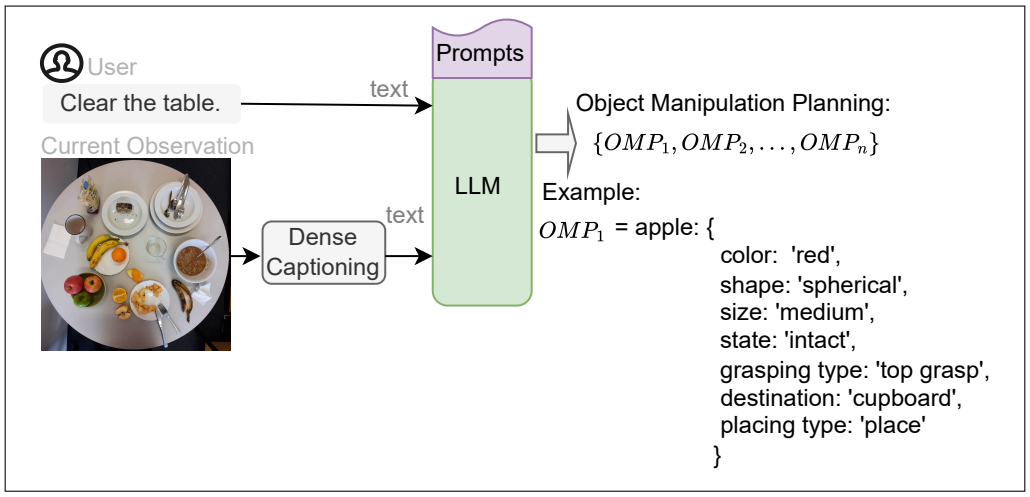


Figure 5.5: Modular method: $\text{OSSA}_{\text{LLM-DCM}}$ represents the modular model that combines a prompt large language model (LLM) and a dense captioning model (DCM)

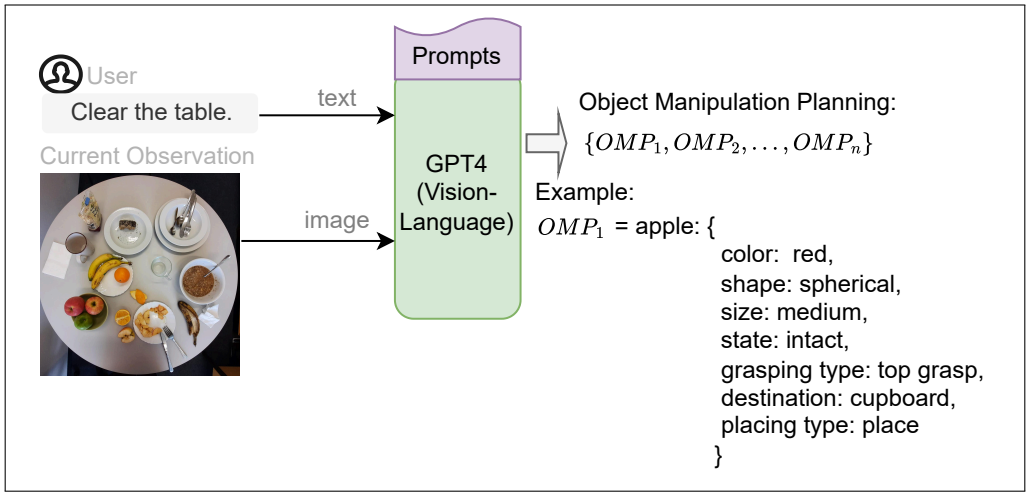


Figure 5.6: Monolithic Method: OSSA_{VLM} represents only a VLM.

vision models are trained to detect the abstract object category. Therefore, we cannot use their methods directly.

The DCM does not only localize salient regions in images but also uses natural language to describe them [124]. Instead of object detection models, we use a DCM to get the description of salient regions of a scene as text. As Figure 5.5 shows, a prompt LLM processes a user instruction and dense caption. The quality of the dense captions is an important factor influencing the model's performance. We choose the SOTA DCM GRiT [308]. Additionally, GPT-4V was also successfully used for image or video understanding [163, 291, 287]. In this method, we use both GPT-4V and GRiT to generate the dense caption for a given image.

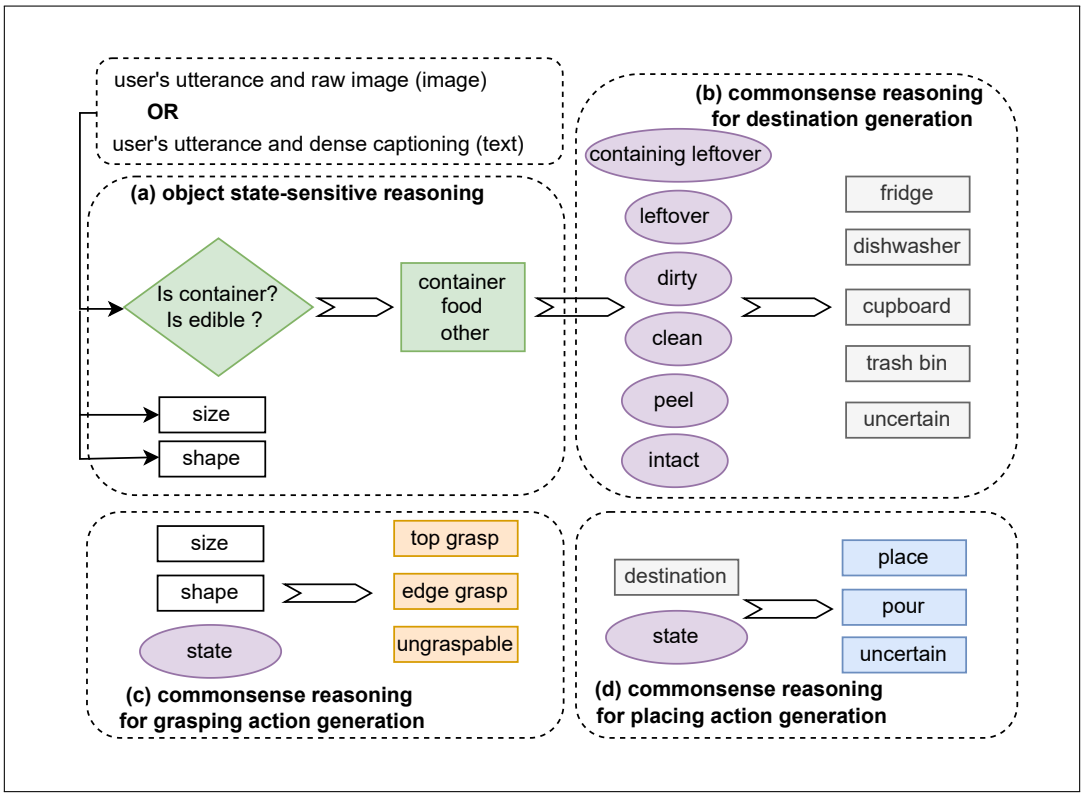


Figure 5.7: Chain-of-thought (CoT) for OSSA. (a) The pretrained model (e.g., LLM or VLM) utilizes commonsense knowledge to reason about the object state; (b) According to the object's state and the user's preference, the model generates a destination for the object. (c) According to the object's state, shape, and size, the model generates a grasping action for the object. (d) According to the object's state and destination, the model generates a placing action for the object.

Monolithic Method

As shown in Figure 5.6, the user's instruction U and image of a table I are the input of a monolithic model. The model's output is object manipulation plans: $\{\text{OMP}_1, \text{OMP}_2, \ldots, \text{OMP}_n\}$. Yang et al. [321, 215] are given preliminary explorations with GPT-4V(ision). GPT-4V has also shown remarkable results for game action prediction in the simulation environment [88, 70]. The approach VILA by Hu et al. [109] unveils the capability of GPT-4V for long-horizon robotic planning to generate a sequence of actionable steps. The difference from those studies is that we leverage GPT-4V to recognize the object state and generate a manipulation plan for an object according to its state.

Prompt Formulations for Object State Recognition and Reasoning

LLMs and VLMs are pretrained on massive datasets from a diverse set of sources, which allows them to acquire human-level commonsense knowledge for handling object states. To unleash their full potential, however, it is necessary to devise

appropriate prompts and guide them; otherwise, their generation cannot be directly used for robot tasks [28, 239]. In this study, we utilize an LLM or a VLM as a high-level planner for the robot's low-level executor. Figure 5.7 shows a schematic chart of the chain-of-thought (CoT) [299, 322] that both models share, which guides them to recognize the objects' states and plan the objects' destinations, grasping types, and placing types.

5.5 Experiments and Results

In our experiments, we aim to evaluate the proposed modular and monolithic methods on object state-sensitive tasks. First, we investigate the best combination of the two methods for object state-sensitive recognition performance. Second, beyond recognizing object states, we further explore effective object state-sensitive methods for more complex tasks, specifically generating an object manipulation plan consisting of 'grasping type', 'destination', and 'placing type'.

5.5.1 Object State Recognition

Experimental Setup

In the modular method, we use two different visual models to generate the description of the image. One is the GPV-4V(ision)[2] as a visual model. The other one is GRiT, a DCM. We leverage an LLM (GPT-4[3]) to process the image description (text) from the visual model to reason about the object states. In the monolithic method, we utilize GPV-4V(ision) as a VLM to directly recognize the object state. We set up two methods in zero-shot or few-shot settings, respectively.

Therefore, we get six different variants: OSSA-LLM-GRiT with the zero-shot setting, OSSA-LLM-GRiT with the few-shot setting, OSSA-LLM-GPT-4V with the zero-shot setting, OSSA-LLM-GPT-4V with the few-shot setting, OSSA-VLM with the zero-shot setting, and OSSA-VLM with the few-shot setting. For the distinction, we denote them as $\text{OSSA}_{\text{LLM(Z)+GRiT}}$, $\text{OSSA}_{\text{LLM(F)+GRiT}}$, $\text{OSSA}_{\text{LLM(Z)+GPT-4V}}$, $\text{OSSA}_{\text{LLM(F)+GPT-4V}}$, $\text{OSSA}_{\text{VLM(Z)}}$, and $\text{OSSA}_{\text{VLM(F)}}$, respectively.

Results

As Figure 5.8 shows, the monolithic method, $\text{OSSA}_{\text{VLM(F)}}$, achieves the highest performance compared to the other two combinations. $\text{OSSA}_{\text{LLM(F)+GPT-4V}}$ performs worse than OSSA_{VLM} but better than $\text{OSSA}_{\text{LLM+GRiT}}$. However, this combination is the most expensive in this experiment because it calls two pretrained models for each plan generation. $\text{OSSA}_{\text{LLM+GRiT}}$ performs worse than the other two combinations. Nevertheless, its advantage is that the agent does not require an extra object detection model to determine the object's location.

[2]gpt-4-vision-preview
[3]gpt-4-0125-preview

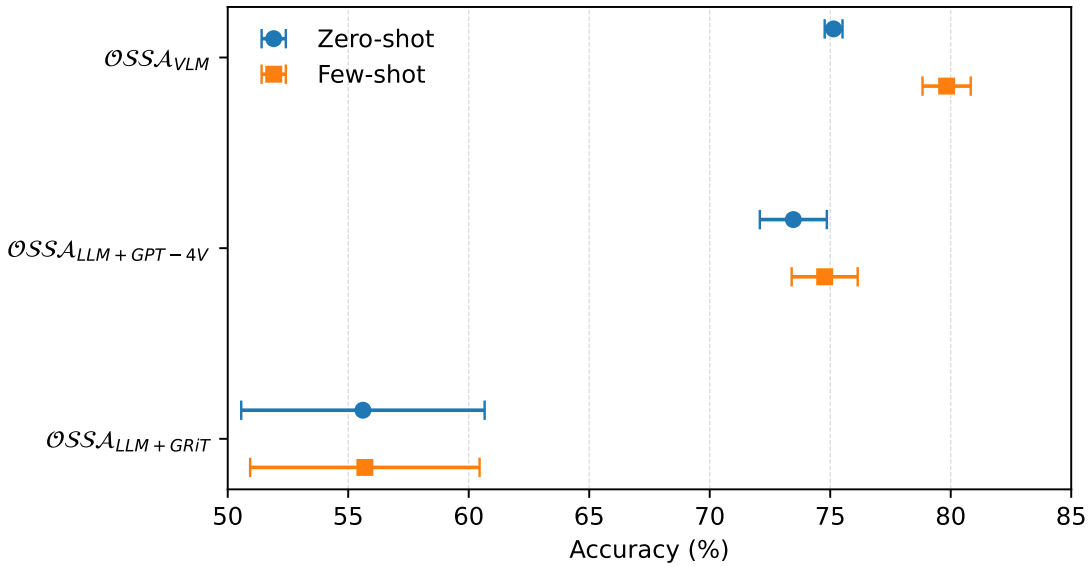


Figure 5.8: OSSA variants (with zero-shot or few-shot): Average accuracy of object state recognition. (Mean±Standard Deviation)

Overall, few-shot prompts enhance the models' performance, especially for the monolithic method. From the results of OSSA_{VLM} and $\text{OSSA}_{\text{LLM+GPT-4V}}$, we conclude that GPT-4 performs well as a VLM for planning multimodal tasks. Potential visual information may be lost when images are converted into text descriptions. For example, spatial reasoning and detailed understanding of object attributes are no longer possible.

5.5.2 Object Manipulation Planning

Experimental Setup

While the previous experiment demonstrated GPT-4V's superior efficiency as a VLM, the $\text{OSSA}_{\text{LLM+GPT-4V}}$ proved less performant and more costly than the OSSA_{VLM}. Furthermore, unlike $\text{OSSA}_{\text{LLM+GPT-4V}}$, it does not provide object location information. Consequently, we excluded this combination from the present experiment. We now evaluate the two remaining model combinations: $\text{OSSA}_{\text{LLM+GRiT}}$ and OSSA_{VLM}. Our evaluation also considers the influence of prompting by testing both zero-shot and few-shot settings. Therefore, we assess a total of four variants across the three tasks defined in Section 5.3.1.

Results

Table 5.1 presents the performance of the four variants on three different task types. The columns State Recognition Accuracy (StaA), Ambiguity Detection Accuracy (AmbA), Destination Generation Accuracy (DesA), Grasping Type Generation Accuracy (GraA), and Placing Type Generation Accuracy (PlaA) each

Table 5.1: Average accuracy in object manipulation plan generation. State Recognition Accuracy (StaA), Ambiguous Detection Accuracy (AmbA), Destination Generation Accuracy (DesA), Grasping Type Generation Accuracy (GraA), Placing Type Generation Accuracy (PlaA), and Completion Rate (ComR). T1 is designed with ambiguous intent; tasks T2 and T3 are designed as concrete controls with no intended ambiguity. A dash "-" indicates no ambiguity was detected.

Task	Method	Average accuracy(%)±Standard deviation					
		State	Ambiguity	Destination	Grasping	Placing	Completion
T1	$OSSA_{LLM(Z)+GRiT}$	51.09±0.94	65.78±12.54	50.37±2.62	85.83±1.44	90.74±2.72	21.74±1.09
	$OSSA_{LLM(F)+GRiT}$	50.54±0.55	95.30±0.26	57.00±1.47	**92.10**±1.73	96.41±0.66	26.27±0.83
	$OSSA_{VLM(Z)}$	71.10±1.33	92.97±2.33	83.38±0.19	89.51±0.57	**97.70**±0.76	52.91±1.14
	$OSSA_{VLM(F)}$	**74.36**±1.78	**97.37**±0.07	**84.81**±1.29	83.90±1.74	93.64±0.84	**53.09**±1.12
T2	$OSSA_{LLM(Z)+GRiT}$	50.18±1.13	–	59.20±2.52	88.45±0.38	90.31±3.72	22.65±0.83
	$OSSA_{LLM(F)+GRiT}$	50.36±0.63	–	46.76±3.26	**92.83**±2.15	**97.84**±0.03	21.38±1.75
	$OSSA_{VLM(Z)}$	68.67±1.32	–	81.17±0.90	90.20±1.50	94.99±2.73	48.63±1.91
	$OSSA_{VLM(F)}$	**72.55**±0.97	–	**85.59**±1.01	83.72±1.64	95.48±1.34	**54.18**±1.11
T3	$OSSA_{LLM(Z)+GRiT}$	50.72±0.83	–	57.54±2.44	84.26±2.70	92.85±1.26	22.83±1.44
	$OSSA_{LLM(F)+GRiT}$	50.18±0.31	–	58.13±4.45	**92.06**±1.67	90.97±0.65	25.18±1.25
	$OSSA_{VLM(Z)}$	70.00±0.39	–	74.80±2.36	91.43±2.81	**94.80**±1.19	47.26±1.92
	$OSSA_{VLM(F)}$	**72.19**±0.87	–	**77.85**±2.08	84.88±0.87	91.19±1.87	**47.45**±0.83

measure performance on a single planning subtask. The final column, Completion Rate (ComR), represents the overall success rate for generating a complete object manipulation plan.

Across all three tasks, the monolithic method $OSSA_{VLM(F)}$ performs best in state recognition, ambiguity detection, and destination generation. Few-shot prompts also enhance model performance in these tasks. For grasping type generation (GraA), the modular model, $OSSA_{LLM(F)+GRiT}$, performs best. We conclude that the pretrained model reasons more effectively from textual descriptions than from images, as the latter requires interpreting object size and shape. We notice outliers in the placement action generation, for which we cannot offer a clearer explanation. Nevertheless, for all variants in the placing action generation (PlaA), performance remains above 90 %. Based on the completion rate (ComA), $OSSA_{VLM(F)}$ outperforms the other variants.

5.6 Discussion

The findings of our OSSA evaluation demonstrate its capability for fine-grained scene perception, identifying object categories and states. Based on this perception and linguistic commands, OSSA generates appropriate object manipulation plans for a robot's low-level executor.

We propose two approaches: (1) a modular approach consisting of a vision mod-

ule (GRiT or GPT-4V) and a language module (GPT-4), and (2) a monolithic approach that relies solely on a vision-language model (VLM), namely GPT-4V. We quantitatively evaluate object state recognition and manipulation planning to demonstrate the state sensitivity of these approaches.

We further analyze the main findings from each experiment. A comparison of object state recognition accuracy across the two experimental settings reveals noticeable differences. Specifically, performance is slightly better in the pure object state recognition task than in the object manipulation planning task. This suggests that models dedicated to a single task of state recognition outperform those required to handle multiple tasks simultaneously. Accordingly, we conclude that when pretrained models (e.g., LLMs or VLMs) are tasked with increased reasoning demands, their performance tends to degrade. This observation aligns with human cognitive behavior, where focusing on a single task generally yields better performance than multitasking.

The analysis of OSSA_{VLM} in object manipulation planning shows that the overall success rate for generating a complete manipulation plan (ComR) is significantly lower than the accuracy achieved in individual reasoning components (StaA, AmbA, DesA, GraA, and PlaA). Nevertheless, the model demonstrates stable and reasonably strong performance under both zero-shot and few-shot prompting conditions. This performance gap can be attributed to the variability in object categories and state requirements, which poses challenges for the model's generalization capabilities.

Overall, we observe that the VLM (GPT-4V) provides more concrete and informative visual understanding than the SOTA dense captioning model (GRiT). As a result, the monolithic GPT-4V approach outperforms the pipeline-based modular approach that combines GRiT with GPT-4.

5.7 Chapter Summary

In this chapter, we explored the solution for our **Objective 2**:

> Design and evaluate effective approaches to leverage foundation models for enabling a robot to recognize objects, interpret user instructions, and generate contextually appropriate object manipulation plans at the object state-level, while aligning with commonsense knowledge and user preferences. These approaches are expected to enable the domestic robot to complete long-horizon tasks autonomously and remain robust in the face of ambiguous, vague, and uncertain situations.

Advances in foundation models, particularly LLMs and large VLMs, have fundamentally reshaped research on dialog and multimodal dialog systems. These models possess extensive world knowledge, advanced reasoning capabilities, strong

language mastery, high efficiency, and multi-task learning abilities. Motivated by these strengths, we leverage such models to develop a method termed OSSA, which perceives fine-grained scene information (e.g., object categories and states) and generates appropriate object manipulation plans for a robot's low-level executor based on linguistic commands and visual perceptions. For task formulation, we distinguish between specific and ambiguous instructions. When an instruction is ambiguous, the proposed method is expected to identify missing information and proactively request further clarification from the user.

We explore two approaches: (1) a modular approach consisting of a vision module (GRiT or GPT-4V) and a language module (GPT-4), and (2) a monolithic approach that relies solely on a vision-language model, namely GPT-4V. Quantitative evaluations of these approaches demonstrate that both are capable of enabling an embodied agent to perform object state-sensitive tasks. Based on the final performance in object state recognition and object manipulation planning, the monolithic approach outperforms the modular approach.

However, each approach has its own advantages and limitations. The modular approach benefits from explicit visual perception, as the vision model (GRiT) provides object localization information, despite its lower overall performance. In contrast, a key limitation of the monolithic approach (GPT-4V) is that it is not trained to generate object bounding boxes; therefore, an additional object detection model is required to obtain object locations. In the Chapter 7, we focus on how to adapt a monolithic model to better support object state-sensitive tasks, specifically in object state identification and affordance reasoning.

Chapter 6

Multimodal Language Processing for Cognitive Systems

6.1 Introduction

In this chapter, we extend the evaluation of foundation models, specifically large language models (LLMs), to the user age-sensitive and multimodal processing capabilities for human-robot interaction (HRI). Recently, the LLMs reviewed in Section 2.3 (e.g., the GPT family of models [232], LaMDA [277], BLOOM [149], LLaMA [278], etc.) have revolutionized research across many fields, especially HRI. Although some existing studies have explored the benefits of incorporating LLMs into HRI [289], they do not consider the relationship between the user's age and the robot's state. For example, Zhang et al. [336] investigated the potential of LLMs as zero-shot human models in HRI. Similarly, GPT-3 has been integrated with the Aldebaran Pepper and NAO robots, enabling participants to chat with them [19]. Their goal is to examine the possibilities and limitations of LLM in an alive HRI system. However, these studies overlook the fact that diverse user groups may expect different robot behaviors, even when performing the same task.

To address this limitation, we propose an adaptive, ROS-based framework with an open-source code base[1] for HRI with diverse user groups, as illustrated in Figure 6.1. This framework extends the PyCRAM language [132] with an interrupt client and recovery behavior, enabling the robot to interact with users via natural language speech. Users can interrupt the robot at two levels: minor and major. A minor interruption signals a change to the execution plan, while a major interruption commands a complete stop.

To meet these requirements and explore the LLM's capability for this framework, we propose a dialog bridge, a concept that leverages the language processing power of LLMs to interpret multimodal inputs from the robot's symbolic planner and the

[1] `https://github.com/TPekarekRosin/UHH_UB_AgeAwareHRI.git`

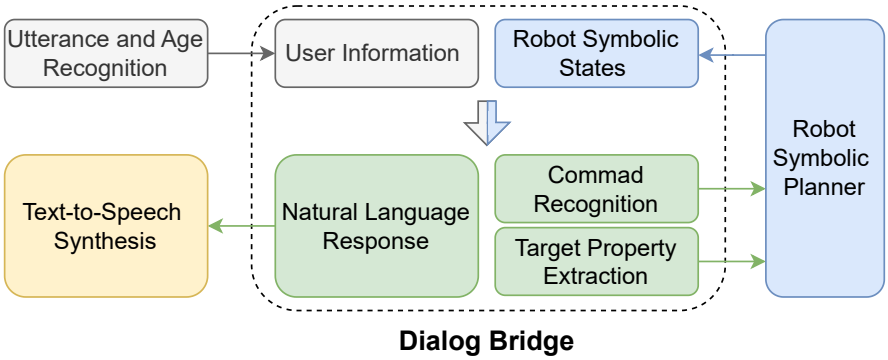


Figure 6.1: Within our adaptive ROS-based framework, the user interacts with the robot via natural language. A combined speech recognition model identifies the user's age group and transcribes their utterance. This information is then processed by a dialog bridge, which integrates it with the robot's internal states. This dialog bridge has three key functions: generating responses to the user, recognizing actionable commands, and extracting target properties for the robot's planning and execution. Figure adapted and rearranged from [221].

user. This method is elaborated on in Section 6.2. We then present the experiments, results, and discussion in Section 6.3, followed by the conclusion and summary in Section 6.4.

6.2 Proposed Method: Dialog Bridge

The dialog bridge is a mechanism that connects the user and the robot. As shown in Figure 6.2, a prompted LLM serves as this dialog bridge, processing information and generating responses between the two. Each turn of the message process can be formally represented:

$$R, C, P = LLM(U, A, S|prompts) \tag{6.1}$$

where U denotes the user's utterance; A denotes the user's age group; S denotes the robot's symbolic states ('step', 'interruptable', 'move_arm', 'move_base', 'current_location', and 'destination_location'); R denotes the response to the user; C denotes the commands (minor and major) to the robot, the minor commands include: 'bring_me', 'setting_breakfast', 'replace_object', and 'change_location', the major command is 'stop'; P denotes the target properties of the object ('type', 'size', and 'color'). While the robot can be interrupted by the user at any time, some of the atomic actions of the robot need to be finished before plan changes can be implemented (e.g., opening the cabinet to look for an object). We use the boolean 'interruptable' variable to inform the dialog bridge of these specific actions. However, major interruptions are the exception, and the robot immediately stops the current step, which is why the interaction needs to be restarted after a major interruption occurs.

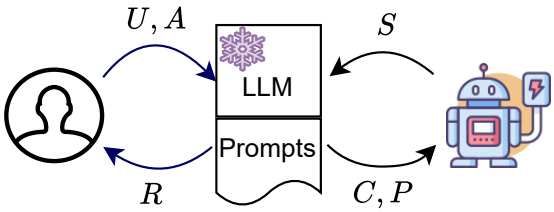


Figure 6.2: Dialog bridge. A mechanism connects the user and the robot by processing the user's utterances (U), user's age (A), turning them into a command (C) with extracted target properties (P) for the robot, as well as monitoring the internal state (S) of the robot and generating an appropriate response (R) to the user.

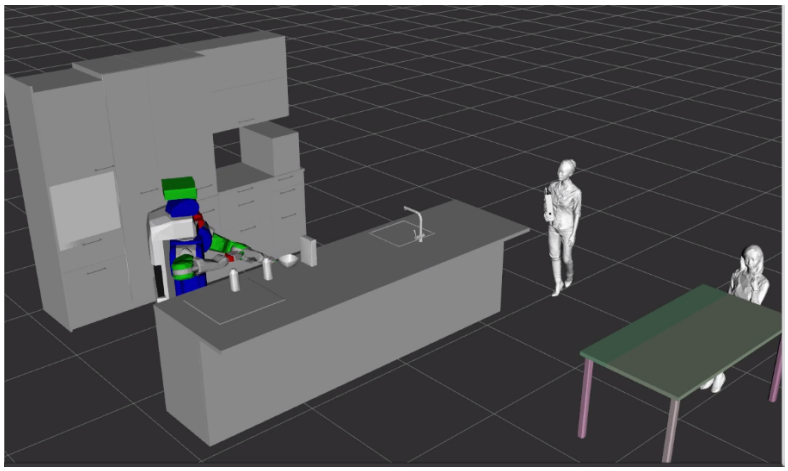


Figure 6.3: Experimental scenario. The kitchen simulation environment includes a service robot and two users.

6.3 Experiments and Results

We construct a simulated virtual kitchen environment comprising a robot and two users, as shown in Figure 6.3. The robot follows the user's requests to serve them in their daily lives. We aim to explore the LLM's capability to recognize commands and object properties from the user's input and to provide age-appropriate feedback based on the robot's symbolic state. For instance, the command, `Please bring me a cup instead of a bowl,` requires the LLM to identify a 'replace_object' action (a minor interruption) and 'cup' as the replacement for 'bowl.' For older users, the LLM initially generates detailed feedback, including information about the robot's navigation between locations and its arm movements during execution. Thus, the robot is more vocal about its actions overall. For younger users, the feedback is initially reduced to simple confirmations, and the robot displays a higher level of autonomy. We selected two GPT-3.5 models (gpt-3.5-turbo-1106 and gpt-3.5-turbo-0125) for their cost efficiency.

Table 6.1: Dialog Bridge: Average accuracy in Cooper(%)±Standard Deviation. 'Add object' refers to object requests, and 'Delete object' refers to object changes. Type is the object identifier. Color, size, and location are additional properties.

LLM	Command	Add object				Delete object			
		Type	Color	Size	Location	Type	Color	Size	Location
ChatGPT	**81.57**	**89.08**	86.60	68.40	84.53	**83.12**	**85.54**	**83.77**	99.96
(gpt-3.5-turbo-1106)	±0.003	±0.004	±0.003	±0.001	±0.001	±0.003	±0.001	±0.002	±0.000
ChatGPT	80.93	82.43	**88.56**	**69.28**	**85.48**	76.48	84.75	82.04	99.99
(gpt-3.5-turbo-0125)	±0.003	±0.002	±0.003	±0.003	±0.004	±0.004	±0	±0.002	±0.000

6.3.1 Multimodal Dataset for Command and Object Properties Recognition

To demonstrate the performance of our proposed dialog bridge in multimodal language processing, we constructed a benchmark, **Co**mmand and Object Prop**e**rties Recognition (Cooper), in a kitchen scenario. Cooper comprise five objects ('milk', 'bowl', 'cereal', 'spoon', and 'cup') with four colors ('green', 'blue', 'red', 'white'), three sizes ('small', 'normal', 'big'), and three locations ('countertop', 'dishwasher', 'cabinet'). We collect ten template instructions to request an object (e.g. `Bring me the small red cup.`) and generate 800 instructions for the command 'bring_me' by combining different object attributes. Interruptions to replace an object can either be expressed in a single sentence containing all necessary information (e.g. `Bring me a cup instead of a bowl.`), or require extracting the object from the context of previous instructions (e.g. `Bring me the bowl instead.`). In this module test, we only consider the first situation. We collect 15 template instructions for replacing one object with another, resulting in 1770 instructions by combining different object attributes. Additionally, we collect 41 variants to request breakfast preparation from the robot, leading to 2611 generated instructions for the benchmark dataset overall.

6.3.2 Results

Table 6.1 shows the performance of two GPT-3.5 models in Cooper. The models perform similarly on the benchmark dataset across three experiments, achieving overall above-average accuracies and low standard deviations, which indicates consistent performance. Because our instructions for replacing objects do not include location details, the accuracy of deleting object locations is nearly 100%. Overall, we conclude that GPT-3.5-Turbo-1106 outperforms GPT-3.5-Turbo-0125. Therefore, it will be used for the overall evaluation of the system.

Having evaluated the dialog bridge in isolation, we subsequently integrated and assessed it within the proposed adaptive ROS-based framework using the PR2 robot as our platform. Because the dialog bridge handles the flow of information between multiple modules, speech recognition, age recognition, and the robot planner, live

evaluation was complicated. Instead, we documented the framework's behavior over 150 trials. We observed that the LLM was able to distinguish the user's age and generate differentiated responses at the beginning of each interaction. However, this behavior diminished after several iterations due to the LLM's memory capacity. As new conversation turns were appended to the history, the LLM tended to forget the initial prompt.

6.4 Chapter Summary

In this chapter, we explored the solution for our **Objective 3**:

> Develop and formalize a mechanism leveraging foundation models for age-sensitive and multimodal processes in HRI. This mechanism is expected to connect users and the robot, enabling the robot to behave differently by default for different user groups while allowing users to modify their preferences.

We proposed a dialog bridge that connects the user and the robot within the adaptive ROS-based framework. A prompted LLM serves as this dialog bridge, processing information and generating responses between the two. We designed the Cooper task for quantitative and individual evaluation. The results show that the pretrained foundation model GPT-3.5 can extract the command and target properties for the robot.

The trial evaluation of the ROS-based framework demonstrates that the LLM was able to distinguish the user's age and generate differentiated responses at the beginning of each interaction. However, this behavior diminished after several iterations due to the LLM's memory limitations. As new conversation turns were appended to the history, the LLM tended to forget the initial prompt. While the system performed as intended in the majority of cases, the most common causes for unsuccessful interactions were: 1) the LLM misclassifying a request for object replacement, resulting in it fetching two objects instead of one; 2) the LLM receiving faulty transcriptions from the automatic speech recognition (ASR); and 3) the limitation of the LLM's context window.

In the future, incorporating the system into real-world settings through user studies with diverse age groups will further validate our simulation results and ensure the robustness of the framework across various environments.

Chapter 7

State-Sensitive Vision-Language Model for Affordance Reasoning

7.1 Introduction

Object- and region-level image understanding [179, 31, 41] is a form of image processing where a model does not just analyze a holistic image, but instead identifies, segments, and interprets distinct objects or regions within it. The approaches have become a crucial research area in computer vision (CV) for robotics, as they enable robots to interpret and interact with their surroundings effectively. This capability is particularly vital for applications like robotic manipulation, human-robot interaction (HRI), and autonomous driving, where precise localization and comprehension of objects within a scene are essential. In recent studies, large vision-language models (VLMs) [294, 97] have been successfully adapted for dense prediction tasks in image understanding, such as image captioning, visual question answering (VQA) [8], and referring expression comprehension (REC) [131, 190], yielding state-of-the-art (SOTA) results. However, there are two principal limitations of VLMs for robotics.

First, existing VLMs are predominantly limited to recognizing object categories and tend to overlook the nuanced states of individual objects [166, 37, 337, 293]. As a result, most downstream tasks in robotics can only be defined on a category level. However, understanding object states is more essential for affordance reasoning in robotic manipulation. Moreover, state-level affordance reasoning is significantly more challenging than category-level reasoning because object states are dynamic [262, 92, 319]. For instance, two dirty plates may require different grasping strategies depending on the specific locations of the dirt. Similarly, a knife may be entirely clean or dirty, or its handle and blade may exhibit different cleanliness conditions. For example, if a knife is accidentally dropped into food, it may have a clean blade but a contaminated handle. A human would readily infer that, in such a scenario, grasping the knife by the blade is more appropriate [178, 86].

Second, VLMs typically tokenize object locations (commonly represented as bounding box coordinates) as normal text symbols, starting and ending with the special tokens, such as ⟨box⟩ x_1, y_1, x_2, y_2 ⟨/box⟩, during large-scale training. Those approaches [166, 38, 37, 12, 338, 226, 325, 339, 293] treat continuous numbers as discrete symbols, which is an abstraction of visual reality. It can be termed as pixel-to-sequence (Pix2Seq) prediction. However, object coordinate prediction is inherently a regression task. To address this mismatch, NExT-Chat [335] proposed the pixel-to-embedding (Pix2Emb) method, which introduces a box decoder for continuous location outputs as embeddings. The key distinction between Pix2Seq and Pix2Emb methods concerns the decoding of numerical information. However, its performance on REC remains slightly below that of Shikra-7B [38], which is based on the Pix2Seq method, despite both using a similar detection-training data recipe.

To further explore the numerical regression task in VLMs, we propose a less intrusive alternative strategy. Specifically, during the training phase, we utilize the output of the box encoder to calculate an auxiliary regression loss for the VLM fine-tuning in an object state-sensitive task, while continuing to standardize sequence prediction in the inference phase.

Our primary contribution in this chapter is summarized as follows:

(1) We propose StateVLM[1], a novel model capable of perceiving fine-grained scene information (i.e., objects and their states). To address the limitation of standard VLMs that tokenize object locations as discrete text symbols, StateVLM introduces an auxiliary regression loss to emphasize location prediction during fine-tuning.

(2) We demonstrate that this auxiliary regression loss improves the VLM's convergence during fine-tuning, thereby encouraging the intermediate layers of the VLM to learn numerical features. The experiments were conducted on a classical region-level image understanding task, REC. For this task, we used the following datasets: RefCOCO, RefCOCO+, and RefCOCOg.

(3) We provide an open-source benchmark dataset, Object State Affordance Reasoning, focused on object state identification and affordance reasoning. It comprises both complex and simple scenes, totaling 1,172 scenes, containing 7,746 individual objects.

The remainder of this chapter is structured as follows. Section 7.2 focuses on bridging gaps in VLMs for object understanding and reasoning. Section 7.3 outlines the task definition and provides a summary of the dataset. The architecture and training procedure of StateVLM are detailed in Section 7.4. Section 7.5 presents and analyzes the experimental results, while Section 7.6 discusses these findings, highlighting the role of the auxiliary loss in improving performance. Finally, Section 7.7 concludes the chapter and suggests directions for future work.

[1] `https://github.com/Xiao-wen-Sun/StateVLM_V0.git`

7.2 Bridging Gaps in VLMs for Object Understanding and Reasoning

7.2.1 VLMs for Referring Expression Comprehension

Referring expression comprehension (REC) [131, 190] is a fundamental region-level image understanding task that seeks to identify and localize a target object described by a natural language expression within a given scene. In recent years, SOTA VLMs have achieved remarkable performance on this task [166, 38, 37, 12, 338, 226, 325, 339, 293], significantly surpassing traditional non-VLM approaches [292, 60, 40, 82, 320, 176]. This improvement can largely be attributed to the integration of visual representations or embeddings into large language models (LLMs). Due to their training on massive and diverse text corpora, LLMs exhibit strong linguistic interpretation abilities and extensive commonsense reasoning, both of which are crucial for advancing embodied artificial intelligence (AI) and robotic perception. Nevertheless, these models demand substantial computational resources and prolonged multi-stage training processes, which present notable practical challenges.

The computational resources and training times of the most representative VLMs for the REC task are summarized in Table 7.1. Notably, most of these models are designed for general-purpose multimodal understanding rather than being explicitly optimized for REC. Depending on the number of transformer layers in the LLM and the scale of the visual backbone, they typically contain either 7B(illion) or 13B parameters. Furthermore, the models adopt several different training configurations. First, they are trained on computational clusters of varying sizes, such as 8, 32, or 256×A100 GPUs. Second, they employ various training strategies, including one-stage, two-stage, or multi-stage training. Third, training durations are reported inconsistently, with some measured in wall-clock time and others in training steps.

A rigorous scientific conclusion is drawn under the premise that a single variable is controlled [307]. Therefore, it is difficult to determine whether simply combining two different types of data distributions (discrete text and continuous location) leads to low-efficiency training for bounding box coordinate prediction.

7.2.2 Affordance Reasoning: From Object Categories to Properties

Object affordance refers to the range of actions an agent can perform with an object, as perceived through its properties [86, 212]. A robust understanding of object categories (e.g., ‘a cup’, ‘a plate’, and ‘an apple’) provides a foundational basis for such reasoning. Building on this, a model can leverage finer-grained attributes to infer an object’s potential affordances. Object attributes, as generalizable properties, are central to this reasoning process. For instance, instead of rely-

Table 7.1: Summary of VLM configurations capable of performing the REC task. * refers to different resources to be used in the second stage. × states that this stage is not included. Pix2Seq and Pix2Emb stand for pixel-to-sequence and pixel-to-embedding.

VLMs Architecture	VLMs Name	VLMs Size	Training GPUs	Training Stage 1	Training Stage 2
Pix2Seq	Shikra [38]	~13B	8×A100	~100 hours	~20 hours
	Ferret [325]		8×A100	~125 hours	×
	Ferret (v2) [339]		8×A100	N/A	N/A
	SPHINX-1K [166]		32×A100	~125 hours	~38 hours*
	SPHINX-2K [166]		32×A100	~250 hours	~38 hours*
	VistaLLM [226]		32×A100	~72 hours	~30 hours
	CogVLM-G [293]		256×A100	~120,000 steps	~60,000 steps
	Shikra [38]	~7B	8×A100	~100 hours	~20 hours
	MiniGPT (v2) [37]		8×A100	~90 hours	~20 hours*
	Qwen-VL [12]		N/A	~50,000 steps	~19,000 steps
	LLaVA-G [338]		N/A	~10,000 steps	~8,000 steps
	VistaLLM [226]		32×A100	~48 hours	~22 hours
	Ferret [325]		8×A100	~60 hours	×
	Ferret (v2) [339]		8×A100	N/A	N/A
Pix2Emb	NExT-Chat [335]	~7B	8×A100	~59 hours	~10 hours

ing solely on category recognition, Yang et al. [319] developed a robotic grasping method based directly on object attributes. Similarly, Attr-POMDP [318] presents an attribute-guided formulation of a partially observable Markov decision process for task disambiguation. In the Octopi system [330], the authors selected hardness, roughness, and bumpiness as key physical attributes for physical reasoning. Another prominent direction in affordance reasoning, particularly for robotic manipulation, involves physically grounded methods [110, 157, 111]. PhyGrasp [96] is designed to accurately assess the physical properties of object parts to determine optimal grasping poses.

These finer-grained properties are complex, and some of their values are dynamic for any given object. Our method centers around understanding an object's state, the set of its property values at a given time, and leverages the commonsense knowledge embedded in VLMs to reason about affordances. This new approach of predicting object affordance by understanding the state of the object has been less explored so far.

7.2.3 The Missing Link Between Object State and Manipulation

Previous dataset challenges have primarily focused on object state recognition and classification [89]. For example, the cooking state recognition challenge [119] focused on 18 types of objects and their corresponding states, such as diced, julienned, or creamy paste. Another dataset [262] classified whether a container was full or half empty. Furthermore, the Object State Detection Dataset [90] includes a larger number of objects and state classes.

Another line of the dataset focuses on detecting or anticipating object state changes. For example, a subset of Ego4D [92] addresses object state change classification and detection. Ego4D-OSCA [189] contributes a dataset for anticipating object state changes from Ego4D video sequences. The OSCaR dataset [208] focuses on object state captioning and state change representation. State-aware Keypoint Trajectories (SKT) [158] provides a synthetic dataset encompassing a broad spectrum of garment configurations, ranging from flat to deformed and folded states. However, none of these datasets explicitly focus on how an object's state influences its manipulation.

Overall, further investigation is needed to determine whether simply combining discrete and continuous data distributions leads to inefficient training for bounding-box coordinate prediction in VLMs. Developing a multimodal dataset is essential for examining the potential of VLMs for object-state-level affordance reasoning in object manipulation.

7.3 Multimodal Dataset for Object State Affordance Reasoning

Robotic task planning can be decomposed into macro- and micro-planning. The macro-level involves task planning, where the robot identifies objects and reasons their destinations. The micro-level involves motion planning for grasping and placing the object, as shown in Figure 7.1. As noted in the previous Chapter 5, we focused on the macro-level task. A limitation of monolithic approaches is object localization. To investigate this, we design a novel benchmark for object state localization studies.

7.3.1 Task Definition

Object Detection and Object State Identification Similar to the REC [131] task, the user gives explicit instructions about a target, and the models confirm the request and predict the location of the referring object. We adopt the form for the conversational sample, as shown in Table 7.2 object detection.

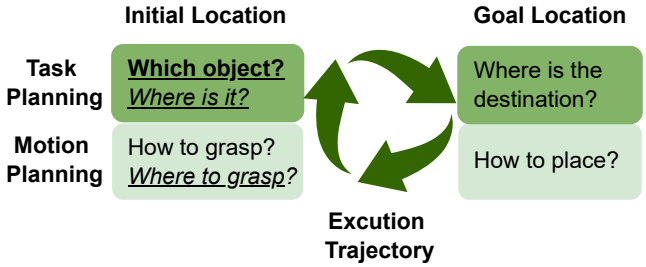


Figure 7.1: The decomposition of robotic task planning into macro- and micro-levels. The macro-level involves task planning, identifying objects, and their destinations. The micro-level involves motion planning, grasping, and placing the objects. While an object's state can affect all steps, we focus on its impact on localization, specifically determining the target object, its position, and the grasping point.

Table 7.2: Dialog templates in object state affordance reasoning. The placeholder [target] is instantiated with the state of objects present in the corresponding scenes, such as 'empty plate', 'dirty plate', or 'bowl with noodles'. $\langle$box$\rangle x_1, y_1, x_2, y_2 \langle$/box$\rangle$ denotes the bounding box coordinates, where x_1, y_1 is the top-left corner, and x_2, y_2 is the bottom-right corner.

Task Types	**Conversations**	
	Roles	**Content**
Object detection and object state identification	User:	"Show me the [target]."
	StateVLM:	"response: Here is the [target]. $\langle$box$\rangle x_1, y_1, x_2, y_2 \langle$/box$\rangle$."
	User:	"Where is the [target]?"
	StateVLM:	"response: Here is the [target]. $\langle$box$\rangle x_1, y_1, x_2, y_2 \langle$/box$\rangle$."
	User:	"Where is the location of the [target]?"
	StateVLM:	"response: Here is the location of the [target]. $\langle$box$\rangle x_1, y_1, x_2, y_2 \langle$/box$\rangle$."
Affordance reasoning	User:	"Hand me the [target]."
	StateVLM:	"response: Sure. I will hand you the [target]. $\langle$box$\rangle x_1, y_1, x_2, y_2 \langle$/box$\rangle$."
	User:	"Pass me the [target]."
	StateVLM:	"response: Alright, I will pick up the [target] for you. $\langle$box$\rangle x_1, y_1, x_2, y_2 \langle$/box$\rangle$."
	User:	"Give me the [target]."
	StateVLM:	"response: Okay, I will give you the [target]. $\langle$box$\rangle x_1, y_1, x_2, y_2 \langle$/box$\rangle$."

Affordance Reasoning: Object Grasp Prediction In everyday kitchen contexts, we usually refer to an object rather than its container. During physical interaction, the decision to grasp the container or its contents is determined by the properties and state of the contents. For instance, we grasp a bowl to pass noodles, but grasp an apple directly to pass it, even when it is contained within a bowl. For the purpose of this physical commonsense, we construct the second object grasp prediction task. It is the same as the previous sample (see Table 7.2 affordance reasoning).

(a) (b) (c)

Figure 7.2: Example scenes in object state affordance reasoning: (a) **Simple scenes** are defined as those with only one object from each category. (b) and (c) **Complex scenes** are defined as those with multiple objects from each category in various states. The red box marks the ideal grasp region; this is an example of how an object's state affects its affordance, as the area covered with food should be avoided when grasping.

7.3.2 Benchmark Dataset

We consider two levels of scene complexity: simple and complex. As illustrated in Figure 7.2, the 'Simple Scene' contains a limited number of objects and presents no ambiguous cases (see Figure 7.2a). In contrast, the 'Complex Scene' features a higher object count (see Figure 7.2b and c), including multiple instances of the same object category in different states, which are prone to causing ambiguity.

Visual Scenes

We employ the stable-diffusion model [243] to generate a total of 1,172 images, comprising 780, 267, and 125 images for templates (a), (b), and (c), respectively, as illustrated in Figure 7.2. From each scene, we selected 11.2% of the images as test samples. The final dataset comprises 7,746 individual objects, of which 11.17% belong to the test set. These objects were in different states within their respective categories, as shown in Figure 7.3.

Annotation Rules

When assessing the state of kitchen tools, multiple dimensions must be considered, including contents, usage, hygiene, physical condition, position, accessibility, and operational status. This dataset challenge specifically focuses on `clearing the table` after a meal. Our scope within this task is defined as follows: containers are analyzed based on their contents, specifically whether they are empty or contain food, and what type of food they contain. Cutlery and napkins are evaluated based on hygiene and readiness for usage; they are categorized simply as clean or dirty. For object state identification, we generated two distinct expressions for each object based on the ambiguous opposites and synonyms listed in Table 7.3. Overall, we obtained 25,401 expressions and 76,203 dialog instances.

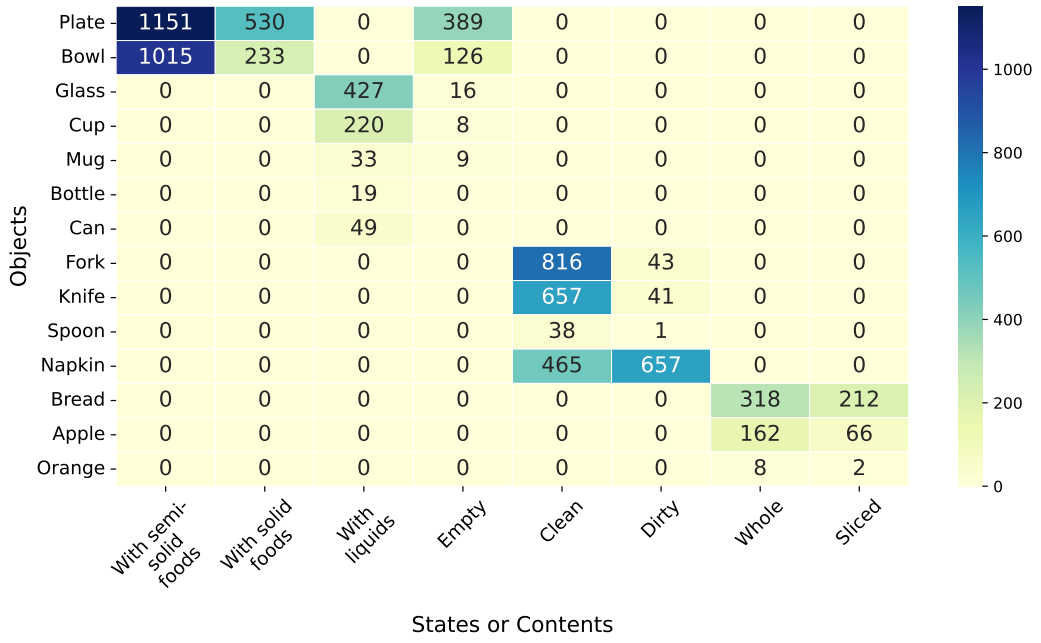


Figure 7.3: Object statistics in object state affordance reasoning. Semi-solid foods usually include sauces, pasta, soup, noodles, and creams. Solid foods typically refer to various states of apples, bread, and oranges. Liquids encompass a variety of drinks, beverages, and juices.

Table 7.3: Ambiguous opposites and synonyms in object state affordance reasoning: Tableware states reflect the shifting semantics of use and cleanliness.

Object Category	Physical Condition	Semantic Interpretation
Plate, Bowl Cup, Mug, Glass	Empty, no residue	clean / unused
	Empty with little residue	used / dirty
	With / holding contents	used / dirty
Bottle	Cap open	possibly used (depends on liquid level)
	Cap closed / unopened	possibly unused (depends on liquid level)
Spoon, Fork, Knife	No residue	clean / unused
	Has residue	dirty / used
Napkin	Has residue Unfolded	used / dirty
	No residue Folded neatly	clean / unused

Evaluation Metric

Our main challenges lie in accurate localization and bounding box prediction of object states and grasp regions. Following prior studies [335, 325, 38, 293], we adopt Generalized Intersection Over Union (GIoU) [240] as our evaluation metric. GIoU is the extension of the standard Intersection over Union (IoU).

The standard IoU is a widely used measure in CV for tasks such as object detection and segmentation. Let $B_p \subset \mathbb{R}^2$ denote the predicted bounding box and $B_g \subset \mathbb{R}^2$ the corresponding ground-truth box. The intersection and union are defined as

$$I = B_p \cap B_g, \qquad U = B_p \cup B_g.$$

The standard IoU is

$$\mathrm{IoU}(B_p, B_g) = \frac{|I|}{|U|},$$

where $|\cdot|$ denotes the area measure. Standard IoU is bounded in [0, 1], which does not capture spatial discrepancies when the boxes do not overlap.

Therefore, Generalized IoU introduces an additional penalty term that accounts for the normalized area outside the union of two bounding boxes but within their smallest enclosing box. Let C denote the smallest enclosing box that contains both B_p and B_g. The GIoU is defined as

$$\mathrm{GIoU}(B_p, B_g) = \mathrm{IoU}(B_p, B_g) - \frac{|C| - |U|}{|C|}.$$

GIoU is bounded in [-1, 1] and reduces to IoU when the predicted box B_p exactly coincides with the ground-truth box B_g.

7.4 Proposed Method: StateVLM

7.4.1 Motivation and Overall Structure

In the previous Chapter 5, we presented the Object State-Sensitive Agent (OSSA) to address task planning challenges. However, the high-performance monolithic approach exemplified by GPT-4V has notable limitations: it does not predict object locations, and its closed-source nature, combined with reliance on extensive GPU cluster training, precludes targeted fine-tuning. Conversely, small-scale and open-weight models have been achieving performance comparable to GPT-4V while allowing fine-tuning with limited computational resources. The proliferation of small-scale and open-weight VLMs means practitioners must choose from a wide range of models trained on diverse datasets with varying strategies, leading to significant performance disparities and making it difficult to identify the optimal candidate for downstream tasks.

As established in the previous section, one of the key limitations of VLMs is their training strategies on numerical prediction tasks. However, commonsense reasoning abilities are equally critical for success in robotic domains. Among the available models, MiniCPM-V [323] stands out as a compelling choice for our investigation: first, it is designed for edge devices and offers strong commonsense reasoning capabilities, which are essential for robotic tasks. Second, it lacks inherent object detection functionality. These characteristics make it an ideal testbed for isolating

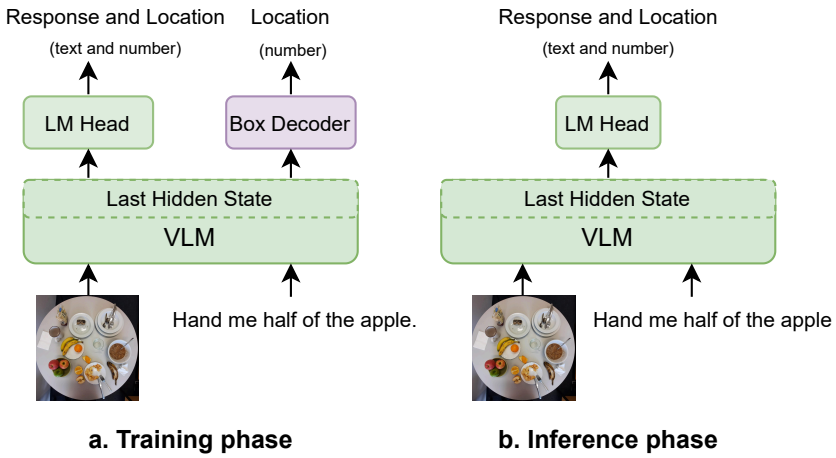


Figure 7.4: The overall framework of StateVLM. During the **training phase** (a), a specially designed box decoder converts sequence predictions into embedding predictions for calculating an auxiliary regression loss. In the **inference phase** (b), however, the model continues to rely on standard sequence-based text and number predictions.

and validating our hypothesis regarding numerical challenges in affordance reasoning.

To this end, we propose StateVLM, a state-sensitive vision-language model that uses MiniCPM-V as its backbone. Inspired by NExT-Chat [335] findings, and given the fundamental challenge of retraining a VLM for regression, we propose a less intrusive alternative. Specifically, in the training phase, we utilize the output of the box encoder to calculate an auxiliary regression loss for VLM fine-tuning in text-conditioned detection, as shown in Figure 7.4a. In the inference phase, we continue to use the standard VLM format, as shown in Figure 7.4b.

7.4.2 StateVLM Modules

BackBone

We adopt MiniCPM-V 2.6 [323] as our backbone, which comprises three key modules: a visual encoder, a compression layer, and an LLM. The visual encoder leverages an adaptive visual encoding approach, SigLip-400M [334], to tokenize the image inputs into visual tokens. The compression layer uses a perceiver-resampler structure [12] with a single cross-attention layer to compress the large number of visual tokens into a fixed set of 96 tokens. Finally, the compressed visual tokens and the text input are fed into the LLM, a causal language model (Qwen2-7B [273]), which uses a specific linear head to predict the next word.

Box Decoder

Building on the backbone network, we propose a box decoder module, a lightweight feed-forward linear head to transform the sequence output into continuous values. Specifically, we first perform global average pooling over the temporal (sequence) dimension using an **AdaptiveAvgPool1d** layer, reducing each hidden-state sequence to a single vector. The resulting vector is flattened and passed through a two-layer multilayer perceptron (MLP) with a **ReLU** non-linearity. The MLP maps the LLM hidden size to a lower-dimensional intermediate representation and then to a 4-dimensional output corresponding to the four continuous numbers, which are the predicted location coordinates $[x_1, y_1, x_2, y_2]$.

7.4.3 Training Objectives

The overall training objective of StateVLM is a weighted combination of the *causal language modeling (CLM)* loss and *auxiliary regression loss (ARL)*:

$$\mathcal{L}_{\text{CLM+ARL}} = \alpha \mathcal{L}_{\text{CLM}} + \beta \mathcal{L}_{\text{ARL}},$$

where α and β are the weights of these two losses. Based on the initial values of $\mathcal{L}_{\text{CLM}}$ and $\mathcal{L}_{\text{ARL}}$, α was set to 0.2 and β to 0.8 to balance them and ensure they play equivalent feedback roles in model tuning.

Causal Language Modeling

The backbone, Qwen2-7B, is trained autoregressively using the CLM objective, predicting each token conditioned on all previous tokens via a causal self-attention mask.

Given a token sequence $[x_1, \ldots, x_T]$ and vocabulary size V, with model logits $z_t \in \mathbb{R}^V$ at timestep t, the *CLM loss* is

$$\mathcal{L}_{\text{CLM}} = -\sum_{t=1}^{T-1} \log \frac{\exp(z_{t,x_{t+1}})}{\sum_{v=1}^{V} \exp(z_{t,v})},$$

where $z_{t,v}$ is the predicted logit for the token v and x_{t+1} is the corresponding ground-truth token.

Auxiliary Regression Loss

To enhance numerical reasoning and improve bounding box prediction, we incorporate an *ARL*. This loss jointly supervises both the location and shape of predicted bounding boxes by combining *Least Absolute Deviations (L1) loss* [29] and *GIoU loss* [240].

Let $B_p \subset \mathbb{R}^2$ denote the predicted bounding box and $B_g \subset \mathbb{R}^2$ the corresponding ground-truth box. The *L1 loss* is computed as

$$\mathcal{L}_{L1} = \|B_p - B_g\|_1,$$

and the *GIoU loss* as

$$\mathcal{L}_{\text{GIoU}} = 1 - \text{GIoU}(B_p, B_g).$$

The total bounding box loss is a weighted sum of the two components:

$$\mathcal{L}_{\text{ARL}} = \gamma \mathcal{L}_{L1} + \delta \mathcal{L}_{\text{GIoU}}.$$

$\gamma = 0.2$ and $\delta = 0.8$ follows the ratio utilized in DETR [31] and NExT-Chat [335].

7.4.4 Implementation

We implement StateVLM using PyTorch[2] and the HuggingFace libraries[3]. The model is initialized with weights from the pretrained MiniCPM-V 2.6. In our experiments, we perform both full fine-tuning of the entire model and parameter-efficient fine-tuning via LoRA [107]. All experiments are conducted on four NVIDIA A100 GPUs (80GB each). We employ DeepSpeed[4] to optimize distributed training and use an automatically warmed-up learning rate of 1e-6 for both training regimes. The batch size is set to 24 for full fine-tuning and 48 for LoRA fine-tuning.

7.5 Experiments and Results

In our experiments, we aim to explore our earlier hypothesis that the current training paradigm for numerical tasks may lead to inefficient learning. We also seek to investigate whether a small, well-annotated dataset can unlock the potential of the VLM for object state affordance reasoning.

7.5.1 Referring Expression Comprehension

Experimental Setup

To verify our hypothesis that the current training paradigm for numerical tasks may lead to inefficient learning, we conduct comparative experiments on the REC task. We use the standard REC datasets: RefCOCO, RefCOCO+ [131], and RefCOCOg [190]. RefCOCO and RefCOCO+ contain train, validation, testA, and

[2] `https://pytorch.org/`
[3] `https://huggingface.co/`
[4] `https://www.deepspeed.ai/`

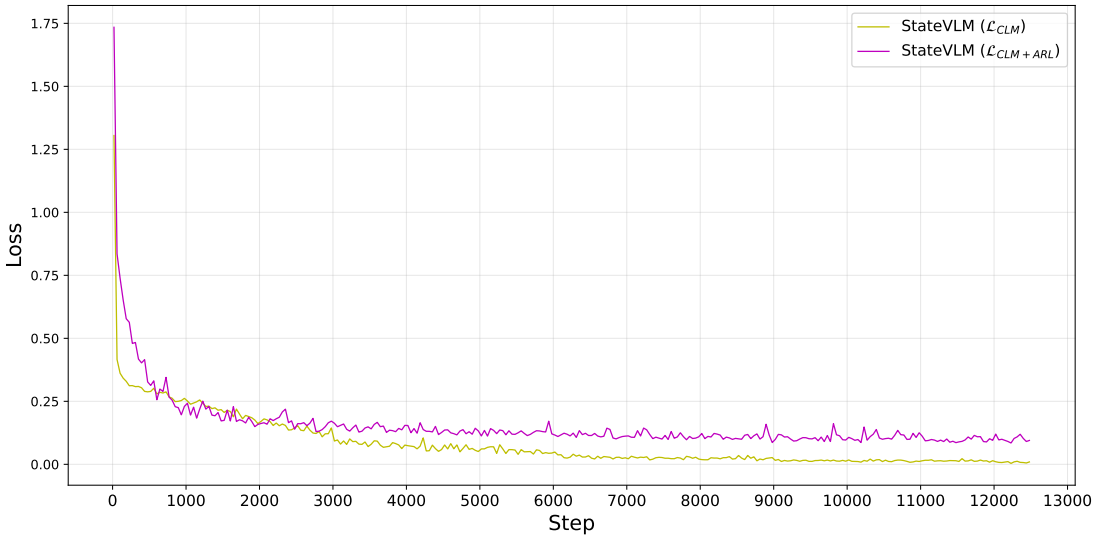
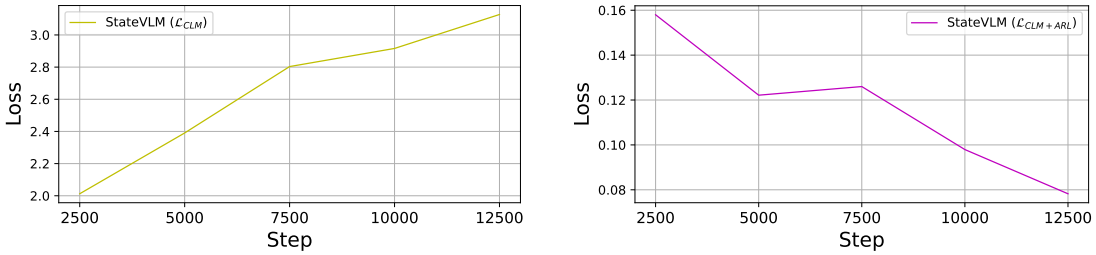


Figure 7.5: StateVLM: Training loss progression over steps.



(a) (b)

Figure 7.6: StateVLM: Validation loss progression over steps.

testB splits, while RefCOCOg contains train, validation, and test splits. We fine-tune our models on the training split and report performance on the corresponding validation and test sets (i.e., testA and testB for RefCOCO and RefCOCO+, and the test set for RefCOCOg.

We first assess the performance of our backbone model, MiniCPM-V, as a baseline. Then, we conduct two full fine-tuning experiments: one in which the model is trained solely with the standard *CLM* objective and another in which it is trained using the *CLM* combined with the *ARL* objective described in Section 7.4.3. For the distinction, we denote them as StateVLM ($\mathcal{L}_{\text{CLM}}$) and StateVLM ($\mathcal{L}_{\text{CLM+ARL}}$), respectively.

Results

We observe that the training loss varies with the number of steps for each model. As shown in Figure 7.5, the loss gradually decreases as the number of training steps increases. StateVLM ($\mathcal{L}_{\text{CLM}}$) and StateVLM ($\mathcal{L}_{\text{CLM+ARL}}$) show a similar trend (see

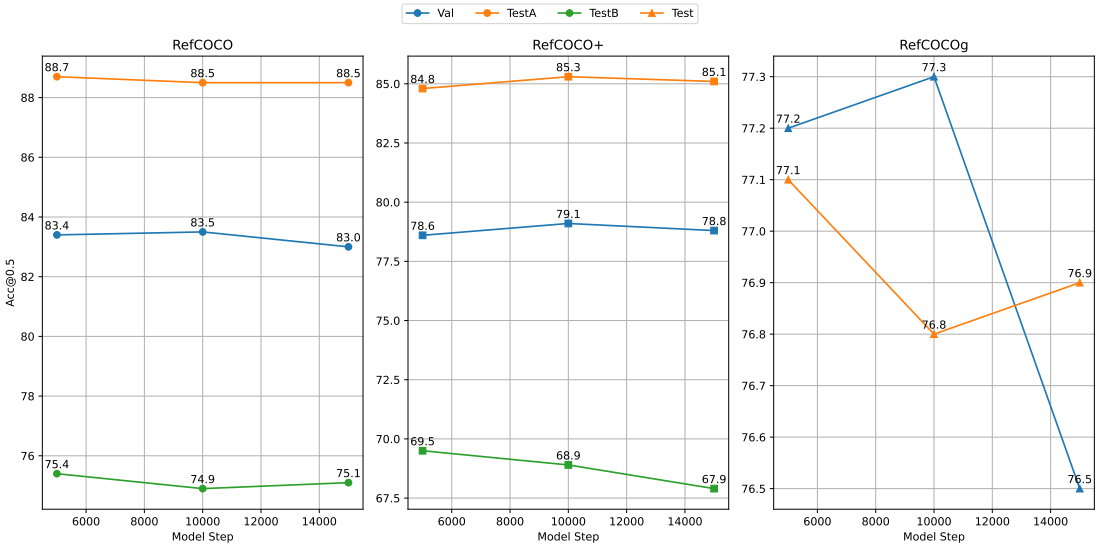


Figure 7.7: StateVLM ($\mathcal{L}_{\text{CLM}}$) performance changes over training steps.

Figure 7.6b). However, the validation loss increases with the number of training steps for StateVLM ($\mathcal{L}_{\text{CLM}}$) (see Figure 7.6a). We suspect that the model begins to overfit after 5000 steps of training.

Furthermore, we evaluated StateVLM ($\mathcal{L}_{\text{CLM}}$) and StateVLM ($\mathcal{L}_{\text{CLM+ARL}}$) on the validation and test splits of RefCOCO, RefCOCO+, and RefCOCOg. We used a fixed random seed for evaluation to ensure deterministic model outputs. Consequently, all runs yield identical results, and no deviation metrics are applicable.

As illustrated in Figure 7.7, the performance of StateVLM ($\mathcal{L}_{\text{CLM}}$) remains stable between 5,000 and 10,000 steps and decreases at 15,000 steps, indicating that it reaches its maximum performance at 5,000 steps. We compared the average performance of the two models at 5,000 and 10,000 training steps on the RefCOCO, RefCOCO+, and RefCOCOg splits in Figure 7.8. In contrast, StateVLM ($\mathcal{L}_{\text{CLM+ARL}}$) performance improves significantly from 5,000 to 10,000 steps and is better than StateVLM ($\mathcal{L}_{\text{CLM}}$)

Therefore, we continue training the model and, to better monitor its performance, evaluate it every 2,500 steps starting from 15,000. The model's performance generally improves over the training steps on RefCOCO and RefCOCO+, as shown in Figure 7.9. However, performance decreases at 17,500 steps, increases at 20,000 steps, and decreases again at 22,500 steps on RefCOCOg. Considering these fluctuations, we select the model at 22,500 steps as our best-performing model. We report the best performing models of StateVLM ($\mathcal{L}_{\text{CLM}}$ and $\mathcal{L}_{\text{CLM+ARL}}$) in Table 7.4, alongside SOTA models on the REC task.

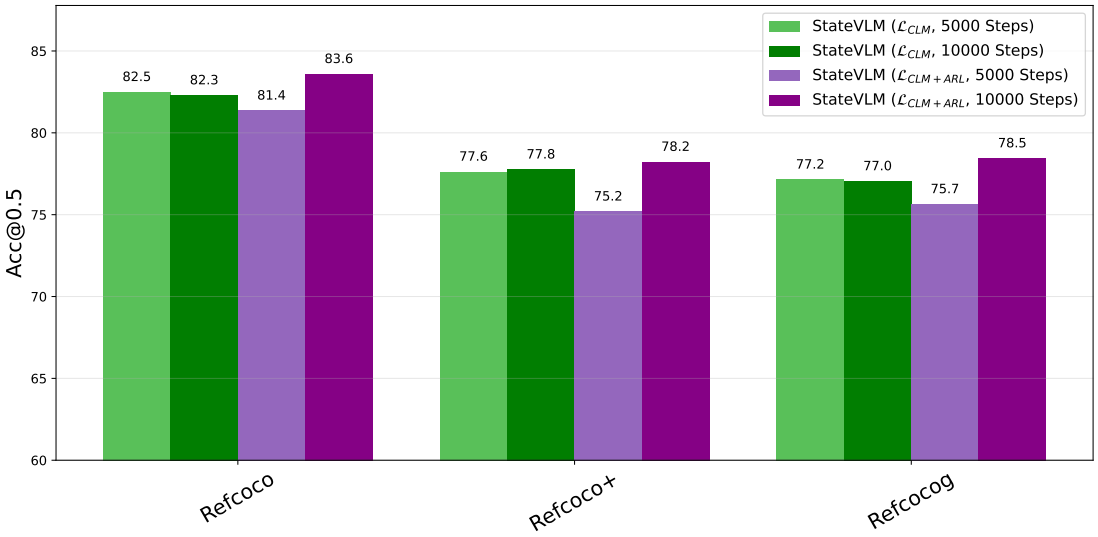


Figure 7.8: StateVLM ($\mathcal{L}_{\text{CLM}}$) and StateVLM ($\mathcal{L}_{\text{CLM+ARL}}$): Comparative performance over training steps

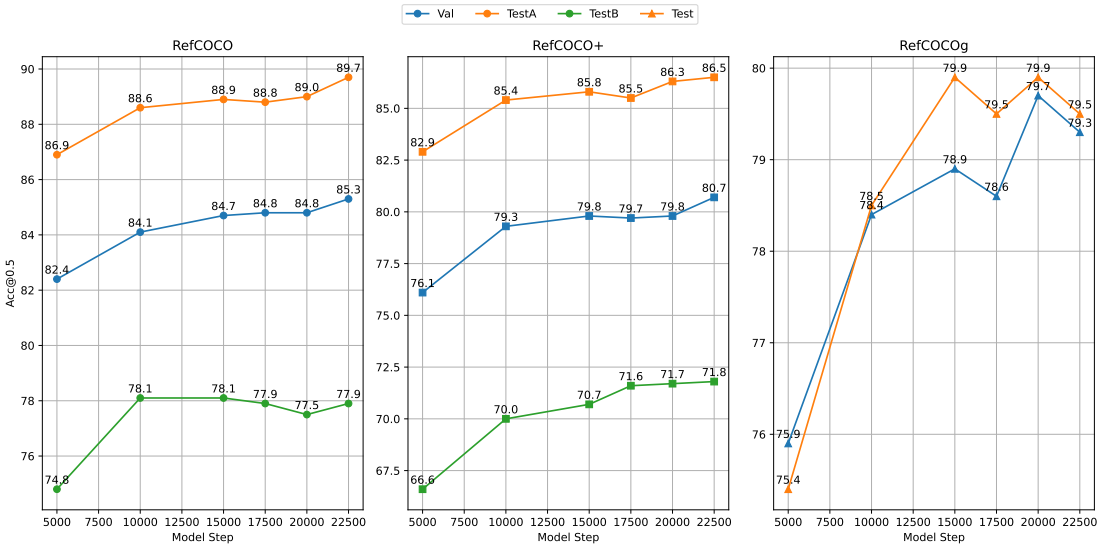


Figure 7.9: StateVLM ($\mathcal{L}_{\text{CLM+ARL}}$) performance changes over training steps.

Ablation Study

We conducted several ablation studies to identify the components essential for achieving the final performance.

Box Decoder Design: One way is to reduce the two linear layers to a single linear layer. Additionally, we tried different activation functions within the MLP, testing **GELU** and **Sigmoid**.

Text Prompts: We tested two different text prompts to understand the effect of text content on model performance:

Table 7.4: StateVLM and baselines performance comparison in REC. The evaluation metric is **Acc@0.5**. * refers to the specialist or fine-tuned methods.

VLMs Type	Methods	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
non-VLM	MAttNet* [329]	76.4	80.4	69.3	64.9	70.3	56.0	66.7	67.0
	OFA-L [292]	80.0	83.7	76.4	68.3	76.0	61.8	67.6	67.6
	TransVG* [60]	81.0	82.7	78.4	64.8	70.7	56.9	68.7	67.7
	UNITER* [40]	81.4	87.0	74.2	75.9	81.5	66.7	74.0	68.7
	VILLA* [82]	82.4	87.5	74.8	76.2	81.5	66.8	76.2	76.7
	UniTAB* [320]	86.3	88.8	80.6	78.7	83.2	69.5	80.0	80.0
	G-DINO-L* [176]	**90.6**	**93.2**	**88.2**	**82.8**	**89.0**	**75.9**	**86.1**	**87.0**
Pix2Seq-13B	Shikra [38]	87.8	91.1	81.8	82.9	87.8	74.4	82.6	83.2
	Ferret [325]	89.5	92.4	84.4	82.8	88.1	75.2	85.8	86.3
	Ferret (v2) [339]	92.6	**95.0**	88.9	87.4	92.1	81.4	89.4	90.0
	SPHINX [166]	89.2	91.4	85.1	82.8	87.3	76.9	84.9	83.7
	SPHINX-1K [166]	91.1	92.7	86.7	86.6	91.1	80.4	88.2	88.4
	SPHINX-2k [166]	91.1	92.9	87.1	85.5	90.6	80.5	88.1	88.7
	VistaLLM [226]	89.9	92.5	85.0	84.1	90.3	75.8	86.0	86.4
	CogVLM-G [293]	**92.8**	94.8	**89.0**	**88.7**	**93.0**	**83.4**	**89.8**	**90.8**
Pix2Seq-7B	Shikra [38]	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
	MiniGPT (v2) [37]	88.1	91.3	84.3	79.6	85.5	73.3	84.2	84.31
	Qwen-VL [12]	88.6	92.3	84.5	82.8	88.6	76.8	86.0	86.3
	LLaVA-G [338]	89.2	—	—	81.7	—	—	84.8	—
	VistaLLM [226]	88.1	91.5	83.0	82.9	89.8	74.8	83.6	84.4
	Ferret [325]	87.5	91.4	82.5	80.9	87.4	73.1	83.9	84.8
	Ferret (v2) [339]	**92.8**	**94.7**	**88.7**	**87.4**	**92.8**	**79.3**	**89.4**	**89.3**
Pix2Emb-7B	NExT-Chat [335]	85.5	90.0	77.9	77.2	84.5	68.0	80.1	79.8
Pix2Seq-7B	Baseline (MiniCPM-V)	16.2	19.6	13.4	12.4	15.7	9.9	7.2	6.7
	StateVLM ($\mathcal{L}_{\text{CLM}}$)	83.4	88.7	75.4	78.6	84.8	69.5	77.2	77.1
	StateVLM ($\mathcal{L}_{\text{CLM + ARL}}$)	85.3	89.7	77.9	80.7	86.5	71.8	79.3	79.5

- Simple Prompt: "A chat between a human and an AI that understands visuals. Follow instructions."

- Concrete Prompt: "A chat between a human and an AI that understands visuals. In images, [x, y] denotes points: top-left is [0, 0], bottom-right is [1, 1]. Increasing x moves right; y moves down. A bounding box is defined as $[x_1, y_1, x_2, y_2]$ where $(x_1, y_1]$ is the top-left corner and (x_2, y_2) is the bottom-right corner. Follow instructions."

Overall, the combination of the MLP architecture that we presented, along with the simple prompt, achieved the best performance.

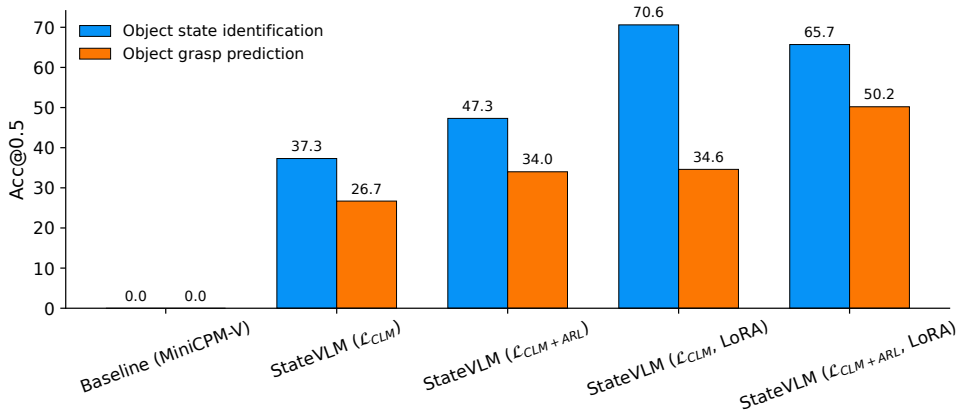


Figure 7.10: StateVLM and baselines: Comprehensive performance comparison.

7.5.2 Object State Affordance Reasoning

Experimental Setup

To investigate the potential of VLMs for object state affordance reasoning, we introduce a small, well-annotated dataset, as detailed in Section 7.3. We first evaluate our proposed dataset using the baseline model, MiniCPM-V, and then evaluate it on our fine-tuned models: StateVLM ($\mathcal{L}_{\text{CLM}}$) and StateVLM ($\mathcal{L}_{\text{CLM + ARL}}$). Subsequently, we further fine-tune these models by the LoRA strategy, obtaining the models StateVLM ($\mathcal{L}_{\text{CLM}}$, LoRA) and StateVLM ($\mathcal{L}_{\text{CLM + ARL}}$, LoRA).

Results

Figure 7.10 presents qualitative results on the test set for the proposed object state identification and object grasp prediction tasks. The baseline model, MiniCPM-V, fails to perform these tasks, as it is not trained for object detection or object state sensitivity.

After full fine-tuning on REC tasks with an auxiliary regression loss, StateVLM ($\mathcal{L}_{\text{CLM + ARL}}$), and without, StateVLM ($\mathcal{L}_{\text{CLM}}$), both models are able to predict object locations. We further evaluate these models on the proposed tasks. StateVLM ($\mathcal{L}_{\text{CLM}}$) and StateVLM ($\mathcal{L}_{\text{CLM + ARL}}$) achieve scores of (37.3, 47.3) and (26.7, 34.0) on object state identification and object grasp prediction, respectively. Applying LoRA fine-tuning further improves performance. Overall, StateVLM ($\mathcal{L}_{\text{CLM + ARL}}$, LoRA) achieves the best results. However, StateVLM ($\mathcal{L}_{\text{CLM}}$, LoRA) performs best on single-object state identification.

7.6 Discussion

The first finding is that our fine-tuning strategy, which incorporates an auxiliary regression loss and leverages limited computational resources, can efficiently adapt

a VLM for object localization on the public REC challenge. The second finding is that a small-scale and well-annotated dataset can unlock the potential capabilities of VLMs for object state-sensitive reasoning on our formulated object state affordance reasoning challenge.

To address the mismatch between the continuous nature of bounding boxes and the simple tokenization approach used in Pix2Seq models, we propose StateVLM, which introduces an auxiliary regression loss during fine-tuning. We also contribute an object state affordance reasoning dataset to enhance the state sensitivity of VLMs.

The final performance of StateVLM on the REC challenge is reported in Table 7.4. Although StateVLM does not reach the performance of SOTA Pix2Emb models, this is due to differences in computational resources and dataset scale. To quantitatively evaluate our fine-tuning strategy, we compare the baseline, the standard StateVLM ($\mathcal{L}$CLM), and StateVLM with the auxiliary regression loss ($\mathcal{L}$CLM+ARL). On average, the latter achieves a 2% performance improvement over the former.

StateVLM ($\mathcal{L}_{\text{CLM+ARL}}$) outperforms NExT-Chat on the RefCOCO+ task and achieves comparable performance on the RefCOCO and RefCOCOg tasks, while using only half of the computational resources required by NExT-Chat. Furthermore, the performance of StateVLM on our proposed object state affordance reasoning task demonstrates that the auxiliary regression loss is effective when combined with LoRA. Overall, we propose an alternative strategy for teams with limited computational resources, focusing on numerical reasoning tasks in VLMs through full fine-tuning and LoRA, particularly at the object state-level.

Regarding object state-level perception and interaction, the performance of StateVLM demonstrates the potential capabilities of VLMs in this setting. However, further refinement is still required. In practical applications, it is often necessary to predict the precise segmentation region of an object rather than relying solely on bounding box localization.

7.7 Chapter Summary

In this chapter, we explored the solution for our **Objective 4**:

> Establish an efficient strategy for modifying general models, large VLMs, to perform object localization and, further, affordance reasoning at the object state-level, all within constrained computational resources and using open-source or small-scale general models.

We proposed StateVLM, which leverages an efficient fine-tuning strategy incorporating an auxiliary regression loss while requiring limited computational resources. This approach serves as an alternative to the Pix2Emb and Pix2Seq methods for numerical tasks, retaining the generative nature of VLMs while explicitly speci-

fying object bounding boxes during training. The promising results on the REC task demonstrate that both full fine-tuning and LoRA-based training are effective. Furthermore, the results of the object state affordance reasoning challenge indicate that the latent capabilities of vision-language models can be unlocked using a small-scale yet well-annotated dataset.

Despite these promising outcomes, there remains a considerable gap between the proposed approach and real-world embodied agents. First, affordance is a broad and multifaceted concept, and we believe that more extensive and carefully annotated datasets are required to advance this field. Although we leveraged a diffusion model to generate the dataset, human verification is still necessary. Second, our evaluation focused on distinguishing objects in different states, and we did not investigate ambiguous or vague instructions in depth. Addressing these challenges will require the development of novel approaches that extend beyond efficient fine-tuning strategies.

Chapter 8

Conclusion

In this chapter, we first summarize the contributions of this thesis in Section 8.1. Second, we provide an in-depth discussion of the research questions and the corresponding problem-solving approaches in Section 8.2. Third, we acknowledge the challenges that persist 8.3. Fourth, we outline directions for future research in Section 8.4. Finally, we evaluate our overall contribution to the field in Section 8.5.

8.1 Thesis Summary

This thesis focuses on enhancing the communicative and perceptual capabilities of embodied agents in domestic environments. To achieve this, it explores training deep neural networks (DNNs) (e.g., long short-term memory (LSTM)) and leveraging pretrained foundation models (e.g., large language models (LLMs), vision-language models (VLMs)), with a specific focus on Uncertainty, Vagueness, and Ambiguity (UVA-phenomena) We first reviewed the state-of-the-art (SOTA) in the foundations of conversation, communication, and perception for embodied agents. Then, motivated by these foundations, we structured our goal into four objectives. For each objective, a specific research question is posed and addressed through a novel methodological contribution. In what follows, we summarize our contributions from two perspectives: **synthesis and foundations**, and **methodological advancements**.

Synthesis and Foundations First, the widely used dialog system categories are typically based on technology: rule-based, statistically data-driven, end-to-end neural, and hybrid. However, there are two problems with these categories. First, deep learning (DL) models can be applied as individual components within a modular dialog system, such as natural language understanding (NLU), dialog manager (DM), or natural language generation (NLG), or they can be used to construct an end-to-end neural dialog system. Second, there is flexibility in combining multiple technologies to develop a dialog system for a specific application, resulting in a hybrid dialog system. However, such practical combinations do not always align

with the original definition of a hybrid dialog system as introduced in research on BeRP [128], SCREEN [303], and HIS [326]. Therefore, we proposed and reviewed dialog systems from an architectural perspective and discussed the appropriate concept of a hybrid dialog system. In addition, we analyzed the dominant technology, transformer-based language models, LLMs, with respect to their training objectives.

Second, communication is a broader concept that encompasses the exchange of information across different modalities. We analyzed vision-language datasets (VLDs) from an interactive perspective, as well as the training objectives of general VLMs and the ways in which these models are integrated with vision and language components. We further investigated the transition from individual task-specific DNNs models to general foundation models.

Last, we identified three key challenges in the field of embodied agents:

- **C1**: The conceptual problem of UVA-phenomena remains poorly defined, characterized by inconsistent terminology, uneven emphasis, and a lack of a unified conceptual framework.
- **C2**: The representational problem lies in the fact that current embodied agent studies operate primarily at the object category level during interaction, whereas humans interact with everyday objects at the state level.
- **C3**: The methodological problem of existing general VLMs is that they are trained mainly for token classification tasks, whereas object localization, one of the core requirements for embodied agents, is a regression task.

Methodological Advancements We clarify the conceptual problem and establish the necessity of object state-level representations in Section 3.5. Based on this foundation, we propose four objectives with corresponding research questions and develop three methods, each addressing different combinations of the identified conceptual, representational, and methodological problems.

We identified gaps in Sections 4.2, 5.2, 6.1, and 7.2 between the SOTA and our objectives. We need to contribute benchmarks aligned with our objectives and then use them to support the proposed methodologies. A high-level summary of these datasets, covering their targeted challenges, primary use, target models, modalities, scene sizes, and language instance criteria, is provided in Table 8.1, and the corresponding proposed methodologies are shown in Table 8.2.

This thesis was initiated before the pivotal shift in which transformer-based foundation models came to dominate the fields of dialog systems and human-robot interaction (HRI). Overall, we proposed four distinct neural network methods. These methods can be categorized into three types: task-specific deep learning models trained from scratch, prompted foundation models, and fine-tuned foundation models.

The first method, HYbrid Neural Architecture (HYNA), enhances robot communication by integrating visual context into an end-to-end neural dialog system.

Table 8.1: Summary of four datasets for multimodal perception, reasoning, and communication.

Dataset	Human-Robot Dialog	Robotic Tasking Planning	Cooper	Object State Affordance Reasoning
Targeted Challenges (8.1)	C1	C1 and C2	-	C1, C2, and C3
Primary Use	Training DNNs from Scratch	Promting Pretrained LLMs and VLMs	Promting Pretrained LLMs	Fine-tuning Pretrained VLMs
Target Models	RetinaNet and LSTM	GRiT, LLM and VLM	LLM	VLM
Modalities	Vision and Language	Vision and Language	Language and Auxiliary Context	Vision and Language
Tasks	Visual Perception and Dialog	Visual Perception, Reasoning, Planning, and Dialog	Multimodal Processing and Reasoning, Dialog	Referring Expression Comprehension and Dialog
Scenes	3,276	40	-	1,172
Language Instances	147,420	120	2,611	76,203

This integration enables visually grounded human-robot dialog capable of resolving both linguistic and visual ambiguities in situated HRI. HYNA demonstrates how to train an individual deep neural network (DNN) tailored to a HRI scenario. As transformer-based foundation models became the mainstream approach in these domains, the second method, Object State-Sensitive Agent (OSSA), was developed as a task-planning agent that observes objects and reasons at the object state level by adapting pretrained foundation models. OSSA explores how to leverage chain-of-thought (CoT) to prompt foundation models for robotic tasks. The third method, Dialog Bridge, connects the user and the robot and serves as a core component of an adaptive ROS-based framework. It tackles the challenges of leveraging LLMs for multimodal processing. Dialog Bridge demonstrates how to employ CoT prompting to adapt foundation models for robotic tasks. The fourth method, State-sensitive Vision-Language Model (StateVLM), addresses the limitations of VLMs in numerical reasoning tasks, while maintaining a continued focus on object state-level observation and reasoning. StateVLM demonstrates how to fine-tune open-weight, small-scale foundation models under limited computational resources for robotic applications.

In the next section, we provide a detailed analysis of the research questions, discuss the limitations and progress in problem-solving, and outline directions for future research.

8.2 Addressing the Research Question

Our first objective is to **develop a hybrid neural framework for visually grounded human-robot dialog that integrates visual content into a con-**

Table 8.2: Summary of four methodologies evaluated against core desiderata for advanced embodied agents. ✔ indicate that a desideratum is explicitly addressed; ● indicates partially support; and ✘ indicates no support.

Core Desiderata	HYNA (Ch. 4)	OSSA (Ch. 5)	Dialog Bridge (Ch. 6)	StateVLM (Ch. 7)
Visual Perception	✔	✔	✘	✔
Object State-Sensitive	✘	✔	✘	✔
Uncertainty Awareness	✘	●	✘	●
Vagueness Awareness	✘	●	✘	✘
Ambiguity Awareness	✔	✔	✘	✔
Communication	✔	✔	✔	✔
Interaction and Adaptation	✔	✔	✔	✔
Scalability and Generalization	✔	✔	✔	✔
Trustworthy	✔	✔	✔	✔

versational agent, enabling a domestic robot to resolve linguistic and visual ambiguities in situated HRI. As demonstrated in Chapter 4, the proposed HYNA can resolve such ambiguities in situated HRI. We evaluated different variants of the core component of HYNA, multimodal dialog state tracker (M-DST). The results indicate that using a simple representation method, bag-of-words (BoW), for multimodal inputs to the M-DST model is an effective approach for the visually grounded human-robot dialog task. A comparison of the variant using only BoW against the variant using both BOW and utterance embedding (UE) reveals a key difference: while the addition of UE improves performance on name confirmation, it negatively impacts scene ambiguity recognition. Testing and evaluation on unseen scenes confirm that the system meets the desiderata of scalability and generalization. Finally, the deployment of HYNA on the physical robot Neuro-Inspired COmpanion (NICO) demonstrates that it fulfills the desiderata of communication, interaction, adaptation, and trustworthiness.

Our second objective is to **design and evaluate effective approaches that leverage foundation models to enable a robot to recognize objects, interpret user instructions, and generate contextually appropriate manipulation plans at the object-state level, while remaining aligned with commonsense knowledge and user preferences. These approaches are expected to allow a domestic robot to autonomously complete long-horizon tasks and to remain robust when faced with ambiguous or vague situations.** As exemplified in Chapter 5, we propose OSSA, a task-planning agent equipped with pretrained neural networks. We explore two architectural paradigms: (i) a modular approach with separate perception and reasoning modules, consisting of a pretrained vision processing module (dense captioning model (DCM)) and a

natural language processing module (LLM); and (ii) a monolithic approach, which consists solely of a VLM. We conduct a quantitative evaluation on tabletop scenarios in which the task is to *clear the table*. The command *clear the table* itself represents a long-horizon task, as it does not specify how leftover food should be handled. As such, the command is unclear, vague, and ambiguous. In contrast, the two more concrete commands, *clear the table and keep all the leftover food* and *clear the table and discard all the leftover food*, do not involve uncertainty or vagueness. Overall, the monolithic approach outperforms the modular approach. However, each approach has its own advantages and limitations. The modular approach benefits from explicit visual perception, as the vision model (GRiT) provides object localization information, despite its lower overall performance. In contrast, a key limitation of the monolithic approach (GPT-4V) is that it is not trained to generate object bounding boxes; therefore, an additional object detection model is required to obtain object locations.

Our third objective is to **develop and formalize a mechanism leveraging foundation models for age-sensitive and multimodal processes in HRI. This mechanism is expected to connect users and the robot, enabling the robot to behave differently by default for different user groups while allowing users to modify their preferences.** As shown in Chapter 6, the dialog bridge is a core component of an adaptive ROS-based framework. It connects the user and the robot by processing user information and the robot's symbolic states to generate natural language responses for the user, as well as to recognize commands and target properties for the robot. The individual evaluation of the dialog bridge shows that the prompted LLM (GPT-3.5) can successfully extract commands and properties from the user's input. System-level trials demonstrate that the LLM is able to distinguish the user's age and generate differentiated responses at the beginning of each interaction. However, this behavior diminishes after several iterations due to the LLM's memory limitations. As new conversation turns are appended to the interaction history, the LLM tends to forget the initial prompt. Therefore, an improved memory mechanism must be investigated to address this limitation.

Our fourth objective is to **establish an efficient strategy for modifying monolithic general models, VLMs, to perform object localization and, further, affordance reasoning at the object state-level, all within constrained computational resources and using open-source or small-scale general models.** As substantiated in Chapter 7, StateVLM is capable of perceiving fine-grained scene information, including objects and their states. We evaluate the influence of this auxiliary regression loss through extensive experiments on classical region-level image understanding tasks, specifically referring expression comprehension (REC), using the RefCOCO, RefCOCO+, and RefCOCOg datasets. The results demonstrate that incorporating an additional box decoder transforms sequence prediction into regression-based prediction, and that leveraging the auxiliary regression loss effectively facilitates the fine-tuning of VLMs. Furthermore, low-rank adaptation (LoRA) adaptations modify and unlock state-sensitive knowl-

edge embedded within VLMs, increasing model performance from 47.3% and 34.0% to 65.7% and 50.2% on object state identification and object grasp prediction, respectively. However, further refinement is required. In practical applications, it is often necessary to predict precise object segmentation regions rather than relying solely on bounding-box localization.

Although OSSA and StateVLM have not been implemented on a physical robot, we believe that these methods align with the desiderata of communication, interaction, adaptation, scalability, generalization, and trustworthiness.

8.3 Remaining Challenges

Overall, we discussed the concepts of UVA-phenomena and demonstrated the importance of object state-level in HRI. We developed a scenario that incorporates these aspects and, grounded in this scenario, proposed four corresponding approaches for the embodied agent domain. Nonetheless, a tripartite distinction framework could be further generalized to account for a broader range of UVA-phenomena, and additional theoretical investigation is required to fully articulate its scope. For a comprehensive framework that handles UVA-phenomena, while supporting fine-grained perceptual reasoning, the availability of physical data is crucial.

A further limitation lies in the difficulty of efficiently acquiring the relevant physical data, particularly the fine-grained data describing interactions between human and object states. Although we employed diffusion models to generate small-scale and well-annotated datasets, a larger and more diverse dataset will ultimately be necessary to enable full automation.

8.4 Future Directions

Building on the findings of this thesis and the preceding discussion, several promising directions are recommended for future research to further advance the communication, autonomy, and adaptability of domestic robots.

User Preference Adaptation and Atmospheric Sensitivity. For domestic robots, the ability to learn and adapt to the preferences of family members is essential. Additionally, robots should be capable of perceiving and distinguishing the surrounding social or emotional atmosphere, whether the owner is enjoying a meal, hosting guests, experiencing loneliness, or undergoing emotional conflict.

Haptic Sensor Fusion. While this thesis focused on vision-language fusion, it did not incorporate the haptic modality [80], which plays a crucial role in grasping. Many object properties can only be captured through tactile sensing. For a truly intelligent household robot, future research should aim to integrate touch, language, and vision into a unified multimodal framework.

Dexterous Grasping. Further advancing domestic robotics, dexterous grasping represents another essential direction. It pertains to physical execution, which remains highly challenging both algorithmically and hardware-wise. Consequently, further research into dexterous grasping, combined with a unified multimodal framework, is critical for enabling more realistic and robust manipulation in household environments.

8.5 Final Remarks

It is far too early for robots to achieve human-level capabilities across all aspects. However, we have taken an important first step in that direction. In this thesis, we enhance the intelligence of embodied agents along two dimensions:

Communication: we have explored agents to communicate with users with greater depth and robustness by distinguishing between Uncertainty, Vagueness, and Ambiguity. We considered that a mismatch between the user's intention (expressed through language) and the robot's environment (perceived through vision) can give rise to ambiguities, wherein multiple interpretations or actions are possible. In addition, the user and the robot may lack a shared, common sense threshold for interpreting a concept or executing a task, resulting in vagueness. The robot's inability to effectively manage these ambiguities and instances of vagueness constitutes epistemic uncertainty. To mitigate this form of uncertainty, we enhance the robot's reasoning capabilities by training DNNs or by leveraging pretrained foundation models. Furthermore, we made the robot capable of communicating with the user to request supplementary information to address aleatoric uncertainty, to resolve vagueness by establishing the appropriate threshold, and to clarify ambiguities when necessary.

Visual Perception: we have explored agents to detect, reason, and interact with the world at a finer granularity, specifically at the object state-level. Visual perception serves as a critical source of information for a robot's communication, reasoning, and task execution. However, increasing the granularity of perceptual information introduces more ambiguities and vagueness, thereby necessitating more advanced intelligence to mitigate epistemic uncertainty. We presented two diagrams outlining approaches that leverage SOTA pretrained models to perform task planning at the object state-level. These approaches were quantitatively evaluated on a representative task, and the results indicate that the monolithic approach (VLM) outperforms the modular approach (an LLM combined with a DCM). Furthermore, we explored modification strategies for monolithic methods (VLMs) in the context of object localization and affordance reasoning. Our proposed strategy provides a principled compensatory mechanism for its inherent limitations, and we anticipate that it can be generalized to a broader class of numerical tasks for a VLM.

Bibliography

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76(C):243–97, 2021.

[2] Josh Abramson, Arun Ahuja, Iain Barr, Arthur Brussee, Federico Carnevale, Mary Cassin, Rachita Chhaparia, Stephen Clark, Bogdan Damoc, Andrew Dudzik, et al. Imitating interactive intelligence. *arXiv preprint arXiv:2012.05672*, 2020.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Muhannad Alomari and Kais Dukes. Extended train robots, 2016.

[5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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*, pages 6077–6086, 2018.

[6] Pierre Andrews and Suresh Manandhar. A SVM cascade for agreement/disagreement classification. In Isabelle Tellier and Mark Steedman, editors, *Machine Learning for NLP*, volume 50, pages 89–107, France, 2009. ATALA.

[7] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. PaLM 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[8] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433. IEEE, December 2015.

[9] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[10] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[11] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[12] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[13] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[14] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[15] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[16] Hadi Beik Mohammadi, Mohammad Ali Zamani, Matthias Kerzel, and Stefan Wermter. Mixed-reality deep reinforcement learning for a reach-to-grasp task. In *International Conference on Artificial Neural Networks*, pages 611–623. Springer, 2019.

[17] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

[18] Bin Bi, Chenliang Li, Chen Wu, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. PALM: Pre-training an autoencoding&autoregressive language model for context-conditioned generation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 8681–8691, 2020.

[19] Erik Billing, Julia Rosén, and Maurice Lamb. Language models for human-robot interaction. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 905–906, Stockholm, Sweden, 2023. Association for Computing Machinery.

[20] Daniel G. Bobrow, Ronald M. Kaplan, Martin Kay, Donald A. Norman, Henry Thompson, and Terry Winograd. GUS, a frame-driven dialog system. *Artificial Intelligence*, 8(2):155–173, 1977.

[21] Dan Bohus and Eric Horvitz. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 1–8, 2010.

[22] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[23] Richard A. Bolt. "Put-That-There": Voice and gesture at the graphics interface. *ACM SIGGRAPH Computer Graphics*, 14(3):262–270, July 1980.

[24] Grady Booch, Robert A Maksimchuk, Michael W Engle, Bobbi J Young, Jim Connallen, and Kelli A Houston. Object-oriented analysis and design with applications. *ACM SIGSOFT Software Engineering Notes*, 33(5):29–29, 2008.

[25] Antoine Bordes, Y-Lan Boureau, and Jason Weston. Learning end-to-end goal-oriented dialog. In *International Conference on Learning Representations*, 2017.

[26] Hayet Brabra, Marcos Báez, Boualem Benatallah, Walid Gaaloul, Sara Bouguelia, and Shayan Zamanirad. Dialogue management in conversational systems: A review of approaches, challenges, and opportunities. *IEEE Transactions on Cognitive and Developmental Systems*, 14(3):783–798, 2022.

[27] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. RT-1: Robotics transformer for real-world control at scale. *Robotics: Science and Systems XIX*, 2023.

[28] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, et al. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning (CoRL)*, pages 287–318. PMLR, 2023.

[29] J.P. Brooks and J.H. Dulá. The L1-norm best-fit hyperplane problem. *Applied Mathematics Letters*, 26(1):51–55, 2013.

[30] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. Benchmarking in manipulation research: The YCB object and model set and benchmarking protocols. *arXiv Preprint arXiv:1502.03143*, 2015.

[31] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[32] Justine Cassell, Timothy Bickmore, Mark Billinghurst, Lee Campbell, Kenny Chang, Hannes Vilhjálmsson, and Hao Yan. Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 520–527, New York, NY, USA, 1999. Association for Computing Machinery.

[33] Roberta Catizone, Andrea Setzer, and Yorick Wilks. Multimodal dialogue management in the COMIC project. In *Proceedings of the 2003 EACL Workshop on Dialogue Systems: Interaction, Adaptation and Styles of Management*, 2003.

[34] Guan-Lin Chao and Ian Lane. BERT-DST: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. In *Proceedings of the Interspeech*, pages 1468–1472, 2019.

[35] Annie S. Chen, Archit Sharma, Sergey Levine, and Chelsea Finn. You only live once: single-life reinforcement learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 14784–14797, Red Hook, NY, USA, 2022. Curran Associates Inc.

[36] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. TOUCHDOWN: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019.

[37] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[38] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[39] Yangyi Chen, Xingyao Wang, Hao Peng, and Heng Ji. SOLO: A single transformer for scalable vision-language modeling. *Transactions on Machine Learning Research*, 2024.

[40] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal image-TExt representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer International Publishing, 2020.

[41] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021.

[42] Fenghua Cheng, Xue Li, Haoyang Wu, Jiangcheng Sang, and Wenqi Zhao. Recent advances on multi-modal dialogue systems: A survey. In *International Conference on Advanced Data Mining and Applications*, pages 33–47. Springer, 2024.

[43] Wei-Lin Chiang, Joseph E. Gonzalez, Dacheng Li, Zhuohan Li, Zi Lin, Ying Sheng, Ion Stoica, et al. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. *See https://lmsys.org/blog/2023-03-30-vicuna/*, 2023.

[44] Eugenio Chisari, Jan Ole Von Hartz, Fabien Despinoy, and Abhinav Valada. Robotic task ambiguity resolution via natural language interaction. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 14821–14827, 2025.

[45] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, 2014. Association for Computational Linguistics.

[46] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, et al. PaLM: scaling language modeling with pathways. *The Journal of Machine Learning Research*, 24(1):11324–11436, January 2023.

[47] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In Ulrike von Luxburg, Isabelle Guyon, Samy Bengio, Hanna Wallach, and Rob Fergus, editors, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, pages 4302–4310, Red Hook, NY, USA, 2017. Curran Associates, Inc.

[48] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Pre-training transformers as energy-based cloze models. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 285–294, Online, 2020. Association for Computational Linguistics.

[49] Philip R. Cohen, Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen, and Josh Clow. QuickSet: multimodal interaction for distributed applications. In *Proceedings of the Fifth ACM International Conference on Multimedia*, pages 31–40, New York, NY, USA, 1997. Association for Computing Machinery.

[50] Robert Csordás, Kazuki Irie, Jürgen Schmidhuber, Christopher Potts, and Christopher D. Manning. MoEUT: Mixture-of-Experts universal transformers. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37, pages 28589–28614, British Columbia, Canada, 2024. Curran Associates, Inc.

[51] Heriberto Cuayáhuitl, Simon Keizer, and Oliver Lemon. Strategic dialogue management via deep reinforcement learning. In *The first-ever Deep Reinforcement Learning Workshop at NIPS 15*, Montréal, Canada, December 2015.

[52] Heriberto Cuayáhuitl, Steve Renals, Oliver Lemon, and Hiroshi Shimodaira. Human-computer dialogue simulation using hidden markov models. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005*, pages 290–295. IEEE, 2005.

[53] Yongcheng Cui, Ying Zhang, Cui-Hua Zhang, and Simon X Yang. Task cognition and planning for service robots. *Intelligence & Robotics*, 5(1):119–142, 2025.

[54] Damai Dai, Chengqi Deng, Chenggang Zhao, Rx Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. DeepSeekMoE: Towards ultimate expert specialization in Mixture-of-Experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1280–1297, 2024.

[55] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 49250–49267, Red Hook, NY, USA, 2023. Curran Associates Inc.

[56] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 326–335, 2017.

[57] Ana Davila, Jacinto Colan, and Yasuhisa Hasegawa. LLM-based ambiguity detection in natural language instructions for collaborative surgical robots. *arXiv preprint arXiv:2507.11525*, 2025.

[58] Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. Talk the walk: Navigating new york city through grounded dialogue. In *Proceedings of the Visual Learning and Embodied Agents in Simulation Environments Workshop, ECCV 2018*, Munich, Germany, September 2018.

[59] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. GuessWhat?! visual object discovery through multi-modal dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017.

[60] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021.

[61] Li Deng and Dong Yu. Deep convex net: A scalable architecture for speech pattern classification. In *Proceedings of the Interspeech*, pages 2285–2288, 2011.

[62] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *International Conference on Learning Representations*, volume 36, pages 10088–10115, New Orleans, LA, USA, 2023. Curran Associates, Inc.

[63] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Association for Computational Linguistics*, volume 1, pages 4171–4186, 2019.

[64] Mirko Di Lascio, Manuela Sanguinetti, Luca Anselma, Dario Mana, Alessandro Mazzei, Viviana Patti, Rossana Simeoni, et al. Natural language generation in dialogue systems for customer care. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 112–117, Bologna, Italy, March 2020. CEUR Workshop Proceedings.

[65] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *Advances in Neural Information Processing Systems*, 37:52545–52567, 2024.

[66] Haiwen Diao, Xiaotong Li, Yufeng Cui, Yueze Wang, Haoge Deng, Ting Pan, Wenxuan Wang, Huchuan Lu, and Xinlong Wang. Evev2: Improved baselines for encoder-free vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21014–21025, October 2025.

[67] Fethiye Irmak Dogan, Weiyu Liu, Iolanda Leite, and Sonia Chernova. Semantically-driven disambiguation for human-robot interaction. *arXiv preprint arXiv:2409.17004*, 2024.

[68] 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. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[69] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, Hawaii, USA, 2023. JMLR.org.

[70] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. Agent AI: Surveying the horizons of multimodal interaction. *CoRR*, abs/2401.03568, 2024.

[71] Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. *Computer Speech & Language*, 59:123–156, 2020.

[72] Haihong E, Peiqing Niu, Zhongfu Chen, and Meina Song. A novel bi-directional interrelated model for joint intent detection and slot filling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5467–5471, Florence, Italy, July 2019. Association for Computational Linguistics.

[73] Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[74] Lee D. Erman, Frederick Hayes-Roth, Victor R. Lesser, and D. Raj Reddy. The hearsay-II speech-understanding system: Integrating knowledge to resolve uncertainty. *ACM Computing Surveys (CSUR)*, 12(2):213–253, 1980.

[75] Junming Fan and Pai Zheng. A vision-language-guided robotic action planning approach for ambiguity mitigation in human-robot collaborative manufacturing. *Journal of Manufacturing Systems*, 74:1009–1018, 2024.

[76] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[77] Huawen Feng, ZekunYao, Junhao Zheng, and Qianli Ma. Training large language models for retrieval-augmented question answering through backtracking correction. In *The Thirteenth International Conference on Learning Representations*, 2025.

[78] Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 7348–7363, Toronto, Canada, July 2023. Association for Computational Linguistics.

[79] Mary Ellen Foster. Interleaved preparation and output in the COMIC fission module. In *Proceedings of the Workshop on Software*, pages 34–46, USA, 2005. Association for Computational Linguistics.

[80] Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024.

[81] Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, et al. Embodied AI agents: Modeling the world. *arXiv preprint arXiv:2506.22355*, 2025.

[82] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020.

[83] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12462–12469. IEEE, 2024.

[84] Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. Dialog state tracking: A neural reading comprehension approach. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 264–273, Stockholm, Sweden, September 2019. Association for Computational Linguistics.

[85] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910, Punta Cana, Dominican Republic (Online), 2021. Association for Computational Linguistics.

[86] James J. Gibson. The theory of affordances. In Robert Shaw and John Bransford, editors, *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, pages 67–82. Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey, 1977.

[87] Rahul Goel, Shachi Paul, and Dilek Hakkani-Tür. HyST: A hybrid approach for flexible and accurate dialogue state tracking. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019-September:1458–1462, 2019.

[88] Ran Gong, Qiuyuan Huang, Xiaojian Ma, Yusuke Noda, Zane Durante, Zilong Zheng, Demetri Terzopoulos, Li Fei-Fei, Jianfeng Gao, and Hoi Vo. MindAgent: Emergent gaming interaction. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3154–3183, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[89] Filippos Gouidis, Konstantinos E. Papoutsakis, Theodore Patkos, Antonis A. Argyros, and Dimitris Plexousakis. Exploring the impact of knowledge graphs on zero-shot visual object state classification. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, volume 2, pages 738–749, 2024.

[90] Filippos Gouidis, Theodore Patkos, Antonis Argyros, and Dimitris Plexousakis. Detecting object states vs detecting objects: A new dataset and a quantitative experimental study. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, volume 5, pages 590–600, 2022.

[91] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[92] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[93] Alex Graves and Alex Graves. Long short-term memory. *Supervised Sequence Labelling with Recurrent Neural Networks*, pages 37–45, 2012.

[94] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022.

[95] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[96] Dingkun Guo, Yuqi Xiang, Shuqi Zhao, Xinghao Zhu, Masayoshi Tomizuka, Mingyu Ding, and Wei Zhan. PhyGrasp: Generalizing robotic grasping with physics-informed large multimodal models. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 14915–14922, 2025.

[97] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. RegionGPT: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13806, 2024.

[98] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In *Advances in Neural Information Processing Systems*, volume 31, pages 676–686, Montréal, Canada, 2018. Curran Associates, Inc.

[99] Dana Gutman, Samuel A. Olatunji, Noa Markfeld, Shai Givati, Vardit Sarne-Fleischmann, Tal Oron-Gilad, and Yael Edan. Evaluating levels of automation with different feedback modes in an assistive robotic table clearing task for eldercare. In *Applied Ergonomics*, volume 106, page 103859, 2023.

[100] Zellig S. Harris. Distributional structure. In Henry Hiż, editor, *Papers on Syntax*, pages 3–22. Springer Netherlands, Dordrecht, 1981.

[101] Matthew Henderson. Machine learning for dialog state tracking: A review. In *Proceedings of the First International Workshop on Machine Learning in Spoken Language Processing*, volume 54, 2015.

[102] Matthew Henderson, Blaise Thomson, and Steve Young. Word-based dialog state tracking with recurrent neural networks. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 292–299. Association for Computational Linguistics, 2014.

[103] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *Conference on Neural Information Processing Systems Workshop*, 2015.

[104] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[105] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, et al. CogVLM2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

[106] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

[107] Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[108] Haimin Hu and Jaime F. Fisac. Active uncertainty reduction for human-robot interaction: An implicit dual control approach. In Steven M. LaValle, Jason M. O'Kane, Michael Otte, Dorsa Sadigh, and Pratap Tokekar, editors, *Algorithmic Foundations of Robotics XV*, pages 385–401, Cham, 2023. Springer.

[109] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of GPT-4V in robotic vision-language planning. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[110] Siyuan Huang, Haonan Chang, Yuhan Liu, Yimeng Zhu, Hao Dong, Abdeslam Boularias, Peng Gao, and Hongsheng Li. A3VLM: Actionable articulation-aware vision language model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270, pages 1675–1690, Munich, Germany, 06–09 Nov 2024. PMLR.

[111] Siyuan Huang, Iaroslav Ponomarenko, Zhengkai Jiang, Xiaoqi Li, Xiaobin Hu, Peng Gao, Hongsheng Li, and Hao Dong. ManipVQA: Injecting robotic affordance and physically grounded information into multi-modal large language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7580–7587. IEEE, 2024.

[112] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser: Composable 3d value maps for robotic manipulation with language models. In *Conference on Robot Learning (CoRL)*, Atlanta, USA, 2023. PMLR.

[113] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, et al. Inner Monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning (CoRL)*, volume 205, pages 1769–1782. PMLR, 2022.

[114] Xin Huang, Chor Seng Tan, Yan Bin Ng, Wei Shi, Kheng Hui Yeo, Ridong Jiang, and Jung-jae Kim. Joint generation and Bi-Encoder for situated interactive multimodal conversations. In *AAAI 2021 DSTC9 Workshop*, 2021.

[115] Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. Look before you leap: An exploratory study of uncertainty analysis for large language models. *IEEE Transactions on Software Engineering*, 51(2):413–429, 2025.

[116] David A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[117] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-Z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning (CoRL)*, pages 991–1002. PMLR, 2022.

[118] Frederick Jelinek, John D Lafferty, and Robert L Mercer. Basic methods of probabilistic context free grammars. In Pietro Laface and Renato De Mori, editors, *Speech Recognition and Understanding: Recent Advances, Trends and Applications*, pages 345–360. Springer, Berlin, Heidelberg, 1992.

[119] Ahmad Babaeian Jelodar and Yu Sun. Joint object and state recognition using language knowledge. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3352–3356, 2019.

[120] Younghoon Jeong, Se Jin Lee, Youngjoong Ko, and Jungyun Seo. TOM: End-to-end task-oriented multimodal dialog system with GPT-2. In *AAAI 2021 DSTC9 Workshop*, 2021.

[121] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Computer Vision*, pages 4904–4916. PMLR, 2021.

[122] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

[123] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

[124] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. DenseCap: Fully convolutional localization networks for dense captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016.

[125] Michael Johnston, Srinivas Bangalore, Gunaranjan Vasireddy, Amanda Stent, Patrick Ehlen, Marilyn Walker, Steve Whittaker, and Preetam Maloor. MATCH: An architecture for multimodal dialogue systems. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 376–383, Philadelphia, USA, 2002. Association for Computational Linguistics.

[126] Michael I. Jordan. Serial order: A parallel distributed processing approach. In John W. Donahoe and Vivian Packard Dorsel, editors, *Neural-Network Models of Cognition*, volume 121 of *Advances in Psychology*, pages 471–495. North-Holland, 1997.

[127] Daniel Jurafsky and James Hugh Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. Pearson, Cambridge, MA, 3rd edition, 2025.

[128] Daniel Jurafsky, Chuck Wooters, Gary Tajchman, Jonathan Segal, Andreas Stolcke, Eric Fosler, and Nelson Morgan. The berkeley restaurant project. In *International Conference on Spoken Language Processing (ICSLP)*, volume 94, pages 2139–2142, 1994.

[129] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning (CoRL)*, pages 651–673. PMLR, 2018.

[130] Ulas Berk Karli and Tesca Fitzgerald. Extended abstract: Resolving ambiguities in LLM-enabled human-robot collaboration. In *2nd Workshop on Language and Robot Learning: Language as Grounding*, 2023.

[131] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 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, October 2014. Association for Computational Linguistics.

[132] Gayane Kazhoyan and Michael Beetz. Programming robotic agents with action descriptions. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 103–108. IEEE, 2017.

[133] Christopher Kennedy. Ambiguity and vagueness: An overview. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*, volume 1, pages 507–535. De Gruyter Berlin, Berlin, 2011.

[134] Matthias Kerzel, Fares Abawi, Manfred Eppe, and Stefan Wermter. Enhancing a neurocognitive shared visuomotor model for object identification,

localization, and grasping with learning from auxiliary tasks. *IEEE Transactions on Cognitive and Developmental Systems*, 14(4):1331–1343, 2022.

[135] Matthias Kerzel, Erik Strahl, Sven Magg, Nicolás Navarro-Guerrero, Stefan Heinrich, and Stefan Wermter. NICO-Neuro-Inspired COmpanion: A developmental humanoid robot platform for multimodal interaction. In *IEEE International Symposium on Robot and Human Interactive Communication*, pages 113–120, 2017.

[136] Akib Mohammed Khan, Alif Ashrafee, Reeshoon Sayera, Shahriar Ivan, and Sabbir Ahmed. Rethinking cooking state recognition with vision transformers. In *Proceedings of International Conference on Computer and Information Technology (ICCIT)*, pages 170–175. IEEE, 2022.

[137] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[138] Satwik Kottur and Seungwhan Moon. Overview of situated and interactive multimodal conversations (SIMMC) 2.1 track at dstc 11. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 235–241, 2023.

[139] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Punta Cana, Dominican Republic (Online), 2021. Association for Computational Linguistics.

[140] Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 582–595, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[141] Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. CLEVR-Dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 582–595, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[142] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, page 84–90, Nevada, USA, 2012. Curran Associates, Inc.

[143] Taku Kudo and John Richardson. 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*, pages 66–71, Brussels, Belgium, 2018. Association for Computational Linguistics.

[144] Minae Kwon, Erdem Biyik, Aditi Talati, Karan Bhasin, Dylan P Losey, and Dorsa Sadigh. When humans aren't optimal: Robots that collaborate with risk-aware humans. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 43–52, 2020.

[145] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

[146] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: scaling reinforcement learning from human feedback with ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024.

[147] Séverin Lemaignan, Mathieu Warnier, E. Akin Sisbot, Aurélie Clodic, and Rachid Alami. Artificial cognition for social human-robot interaction: An implementation. *Artificial Intelligence*, 247:45–69, 2017.

[148] Oliver Lemon. Learning what to say and how to say it: Joint optimisation of spoken dialogue management and natural language generation. *Computer Speech & Language*, 25(2):210–221, 2011.

[149] Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8608–8621, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.

[150] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021.

[151] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[152] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Computer Vision*, pages 19730–19742. PMLR, 2023.

[153] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Computer Vision*, pages 12888–12900. PMLR, 2022.

[154] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[155] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking LLMs for embodied decision making. *Advances in Neural Information Processing Systems*, 37:100428–100534, 2024.

[156] Xiaodong Li, Guohui Tian, and Yongcheng Cui. Fine-Grained task planning for service robots based on object ontology knowledge via large language models. *IEEE Robotics and Automation Letters*, 9(8):6872–6879, 2024.

[157] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. ManipLLM: Embodied multimodal large language model for object-centric robotic manipulation. In *Computer Vision and Pattern Recognition*, pages 18061–18070, 2024.

[158] Xin Li, Siyuan Huang, Qiaojun Yu, Zhengkai Jiang, Ce Hao, Yimeng Zhu, Hongsheng Li, Peng Gao, and Cewu Lu. SKT: integrating state-aware keypoint trajectories with vision-language models for robotic garment manipulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2025*, pages 18828–18833, Hangzhou, China, 2025. IEEE.

[159] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137, Cham, 2020. Springer, Springer International Publishing.

[160] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state-of-the-art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1587–1606, 2025.

[161] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Peter Florence, and Andy Zeng. Code as policies: Language model

programs for embodied control. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500, 2022.

[162] Kaiqu Liang, Zixu Zhang, and Jaime F Fisac. Introspective planning: Aligning robots' uncertainty with inherent task ambiguity. *Advances in Neural Information Processing Systems*, 37:71998–72031, 2024.

[163] Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, et al. MM-VID: Advancing video understanding with GPT-4V(ision). *arXiv preprint arXiv:2310.19773*, 2023.

[164] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

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

[166] 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. SPHINX: A mixer of weights, visual embeddings and image scales for multi-modal large language models. In *European Conference on Computer Vision (ECCV), LXII*, pages 36–55, Berlin, Heidelberg, 2024. Springer-Verlag.

[167] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

[168] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[169] Bing Liu and Ian Lane. Dialog context language modeling with recurrent neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5715–5719. IEEE, 2017.

[170] Bing Liu and Ian Lane. End-to-end learning of task-oriented dialogs. In Silvio Ricardo Cordeiro, Shereen Oraby, Umashanthi Pavalanathan, and Kyeongmin Rim, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 67–73, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics.

[171] Dongyang Liu, Renrui Zhang, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, Wenqi Shao, Chao Xu, Conghui He, Junjun He, Hao Shao, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. SPHINX-X: scaling data and parameters for a family of multi-modal large language models. In *Proceedings of the 41st International Conference on Machine Learning*, number 1314 in Proceedings of Machine Learning Research, pages 32400–32420. JMLR.org, 2024.

[172] Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. MOKA: Open-vocabulary robotic manipulation through mark-based visual prompting. In *Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[173] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, volume 36, pages 34892–34916, New Orleans, LA, USA, 2023. Curran Associates, Inc.

[174] Mingjiang Liu, Chengli Xiao, and Chunlin Chen. Perspective-corrected spatial referring expression generation for human-robot interaction. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(12):7654–7666, 2022.

[175] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan Yuille. CLEVR-Ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4185–4194, 2019.

[176] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

[177] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[178] Zhijian Liu, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. Physical primitive decomposition. In *Proceedings of the European Conference on Computer Vision (ECCV), Part XII*, pages 3–20, September 2018.

[179] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[180] Ryan Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65, 2017.

[181] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. DeepSeek-VL: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[182] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, volume 32, pages 13–23, Red Hook, NY, USA, 2019. Curran Associates Inc.

[183] Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jiawen Liu, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-InternVL: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24960–24971, 2025.

[184] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, 2015. Association for Computational Linguistics.

[185] Maria Lymperaiou and Giorgos Stamou. A survey on knowledge-enhanced multimodal learning. *Artificial Intelligence Review*, 57(10):284, 2024.

[186] F. Mairesse, M. Gasic, F. Jurcicek, S. Keizer, B. Thomson, K. Yu, and S. Young. Spoken language understanding from unaligned data using discriminative classification models. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4749–4752, 2009.

[187] François Mairesse and Steve Young. Stochastic language generation in dialogue using factored language models. *Computational Linguistics*, 40(4):763–799, December 2014.

[188] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999.

[189] Victoria Manousaki, Konstantinos Bacharidis, Filippos Gouidis, Konstantinos Papoutsakis, Dimitris Plexousakis, and Antonis Argyros. Anticipating object state changes. *arXiv preprint arXiv:2405.12789*, 2024.

[190] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2016.

[191] Michael McTear. Challenges and future directions. In Michael McTear, editor, *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*, pages 159–183. Springer International Publishing, Cham, Switzerland, 2021.

[192] Michael McTear. *Introducing Dialogue Systems*, pages 11–42. Springer International Publishing, Cham, 2021.

[193] Marcelo Mendoza and Juan Zamora. Identifying the intent of a user query using support vector machines. In Jussi Karlgren, Jorma Tarhio, and Heikki Hyyrö, editors, *String Processing and Information Retrieval*, pages 131–142, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[194] Tomas Mikolov, Kai Chen, G.S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Workshop Track Proceedings of the International Conference on Learning Representations*, volume 2013, 2013.

[195] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 2, pages 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

[196] Tomáš Mikolov. *Statistical language models based on neural networks*. Ph.d. thesis, Brno University of Technology, Faculty of Information Technology, 2012.

[197] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 72983–73007, Red Hook, NY, USA, 2023. Curran Associates Inc.

[198] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *17th European Conference on Computer Vision, Part X*, pages 728–755, Berlin, Heidelberg, 2022. Springer-Verlag.

[199] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[200] Jahangir Moini, Oyindamola Akinso, Katia Ferdowsi, and Morvarid Moini. The communication process. In *Health Care Today in the United States*, pages 467–484. Elsevier Science & Technology Books, Amsterdam, 2022.

[201] Seungwhan Moon, Satwik Kottur, Paul Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. Situated and interactive multimodal conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1103–1121, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[202] Carlos Fernando Morales and Fernando De la Rosa. Ambiguity analysis in learning from demonstration applications for mobile robots. In *2013 16th International Conference on Advanced Robotics (ICAR)*, pages 1–6. IEEE, 2013.

[203] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. In Greg Kondrak and Taro Watanabe, editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 462–472, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[204] Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Multi-domain dialog state tracking using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 2, pages 794–799, Beijing, China, July 2015. Association for Computational Linguistics.

[205] Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 1777–1788, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[206] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *European Conference on Computer Vision*, pages 336–352. Springer, 2020.

[207] Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. Collaborative dialogue in minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, 2019.

[208] Nguyen Nguyen, Jing Bi, Ali Vosoughi, Yapeng Tian, Pooyan Fazli, and Chenliang Xu. OSCaR: Object state captioning and state change representation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3565–3576, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[209] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[210] Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. Recent advances in deep learning based dialogue systems: a systematic survey. *Artificial Intelligence Review*, 56(4):3055–3155, August 2022.

[211] Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial Intelligence Review*, 56(4):3055–3155, 2023.

[212] Donald A. Norman. *The psychology of everyday things.* Basic books, New York, 1988.

[213] Daniel Nyga, Subhro Roy, Rohan Paul, Daehyung Park, Mihai Pomarlan, Michael Beetz, and Nicholas Roy. Grounding robot plans from natural language instructions with incomplete world knowledge. In *Conference on Robot Learning (CoRL)*, pages 714–723. PMLR, 2018.

[214] Alice H. Oh and Alexander I. Rudnicky. Stochastic language generation for spoken dialogue systems. In *Proceedings of ANLP/NAACL Workshop on Conversational Systems*, volume 3, pages 27–32, USA, 2000. Association for Computational Linguistics.

[215] OpenAI. GPT-4V(ision) system card. Technical report, OpenAI, 2023.

[216] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Zhang, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, New Orleans, United States (Hybrid), 2022. Curran Associates, Inc.

[217] Sharon Oviatt, Björn Schuller, Philip R. Cohen, Daniel Sonntag, Gerasimos Potamianos, and Antonio Krüger, editors. *The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations*, volume 14. Association for Computing Machinery and Morgan & Claypool, New York, NY, USA, 2017.

[218] Ozan Özdemir, Matthias Kerzel, Cornelius Weber, Jae Hee Lee, and Stefan Wermter. Language model-based paired variational autoencoders for robotic language learning. *IEEE Transactions on Cognitive and Developmental Systems*, 15(4):1812–1824, 2023.

[219] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan

Tur, and Dilek Hakkani-Tur. TEACh: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025, 2022.

[220] Nita Patil, Ajay Patil, and B. V. Pawar. Named entity recognition using conditional random fields. *Procedia Computer Science*, 167:1181–1188, 2020.

[221] Theresa Pekarek Rosin, Vanessa Hassouna, Xiaowen Sun, Luca Krohm, Henri-Leon Kordt, Michael Beetz, and Stefan Wermter. A framework for adapting human-robot interaction to diverse user groups. In *International Conference on Social Robotics*, pages 24–38. Springer, 2024.

[222] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, 2014.

[223] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[224] Michael Phi. Illustrated guide to LSTM's and GRU's: A step by step explanation, September 2018. Medium article, published in Data Science (TDS Archive).

[225] Pradip Pramanick, Chayan Sarkar, Snehasis Banerjee, and Brojeshwar Bhowmick. Talk-to-Resolve: Combining scene understanding and spatial dialogue to resolve granular task ambiguity for a collocated robot. *Robotics and Autonomous Systems*, 155:104183, 2022.

[226] Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks, master of many: Designing general-purpose coarse-to-fine vision-language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14076–14088, 2024.

[227] Kun Qian, Satwik Kottur, Ahmad Beirami, Shahin Shayandeh, Paul Crook, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. Database search results disambiguation for task-oriented dialog systems. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1158–1173. Association for Computational Linguistics, July 2022.

[228] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. AffordanceLLM: Grounding affordance from vision language models.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7587–7597, 2024.

[229] Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. End-to-end task-oriented dialogue: A survey of tasks, methods, and future directions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5925–5941, Singapore, December 2023. Association for Computational Linguistics.

[230] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[231] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. Technical report, OpenAI, San Francisco, CA, USA, 2018.

[232] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[233] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741, New Orleans, LA, USA, 2023. Curran Associates, Inc.

[234] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):5485–5551, 2020.

[235] Owen Rambow, Srinivas Bangalore, and Marilyn Walker. Natural language generation in dialog systems. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 1–4, USA, 2001. Association for Computational Linguistics.

[236] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. SayPlan: Grounding large language models using 3D scene graphs for scalable task planning. In *Conference on Robot Learning (CoRL)*, 2023.

[237] Suman Ravuri and Andreas Stolcke. Recurrent neural network and LSTM models for lexical utterance classification. In *Proceedings of the Interspeech*, pages 135–139, 2015.

[238] Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997.

[239] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, et al. Robots that ask for help: Uncertainty alignment for large language model planners. *Conference on Robot Learning (CoRL)*, 2023.

[240] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[241] Verena Rieser, Oliver Lemon, and Simon Keizer. Natural language generation as incremental planning under uncertainty: Adaptive information presentation for statistical dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(5):979–994, 2014.

[242] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[243] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA, June 2022. IEEE Computer Society.

[244] Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, New Orleans, Louisiana, USA, 2018. AAAI Press.

[245] Tara N. Sainath, Brian Kingsbury, George Saon, Hagen Soltau, Abdelrahman Mohamed, George Dahl, and Bhuvana Ramabhadran. Deep convolutional neural networks for large-scale speech tasks. *Neural networks*, 64:39–48, 2015.

[246] Victor Sanh, Albert Webson, Colin Raffel, et al. Multitask prompted training enables zero-shot task generalization. *International Conference on Learning Representations*, 2022.

[247] Ruhi Sarikaya, Geoffrey E. Hinton, and Anoop Deoras. Application of deep belief networks for natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):778–784, 2014.

[248] Lisa Scherf, Lisa Alina Gasche, Eya Chemangui, and Dorothea Koert. Are you sure? - multi-modal human decision uncertainty detection in human-robot interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 621–629, 2024.

[249] Jetze Schuurmans and Flavius Frasincar. Intent classification for dialogue utterances. *IEEE Intelligent Systems*, 35(1):82–88, January 2020.

[250] Ransalu Senanayake. The role of predictive uncertainty and diversity in embodied AI and robot learning. In Paulo Shakarian and Hua Wei, editors, *Metacognitive Artificial Intelligence*. Cambridge University Press, New York, 9 2025.

[251] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1715–1725, Berlin, Germany, 2016. Association for Computational Linguistics.

[252] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 1577–1586, Beijing, China, July 2015. Association for Computational Linguistics.

[253] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[254] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated Mixture-of-Experts layer. In *International Conference on Learning Representations*, 2017.

[255] Shingo Shimoda, Lorenzo Jamone, Dimitri Ognibene, Takayuki Nagai, Alessandra Sciutti, Alvaro Costa-Garcia, Yohei Oseki, and Tadahiro Taniguchi. What is the role of the next generation of cognitive robotics? *Advanced Robotics*, 36(1-2):3–16, 2022.

[256] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10740–10749, 2020.

[257] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. ProgPrompt: program generation for situated robot task planning using large language models. *Autonomous Robots*, 47(8):999–1012, August 2023.

[258] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. LLM-Planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023.

[259] Arpita Soni, Sujatha Alla, Suresh Dodda, and Hemanth Volikatla. Advancing household robotics: Deep interactive reinforcement learning for efficient training and enhanced performance. *Journal of Electrical Systems*, 20(3s):1349–1355, 2024.

[260] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado, May 2015. Association for Computational Linguistics.

[261] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. In Peter Willett, editor, *Document Retrieval Systems*, pages 132–142. Taylor Graham Publishing, United Kingdom, 1988.

[262] Josua Spisak, Matthias Kerzel, and Stefan Wermter. Clarifying the half full or half empty question: Multimodal container classification. In *International Conference on Artificial Neural Networks*, pages 444–456. Springer, 2023.

[263] Fabio Strazzeri and Carme Torras. Topological representation of cloth state for robot manipulation: Deriving the configuration space of a rectangular cloth. *Autonomous Robots*, 45(5):737–754, 2021.

[264] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[265] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020.

[266] Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. Executing instructions in situated collaborative interactions. 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)*, pages 2119–2130, Hong Kong, China, November 2019. Association for Computational Linguistics.

[267] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-End memory networks. In *Advances in Neural Information Processing Systems*, Cambridge, MA, USA, 2015. MIT Press.

[268] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. 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)*, pages 5100–5111, Hong Kong, China, November 2019. Association for Computational Linguistics.

[269] Grace Tang, Swetha Rajkumar, Yifei Zhou, Homer Rich Walke, Sergey Levine, and Kuan Fang. Kalie: Fine-tuning vision-language models for open-world manipulation without robot data. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9507–9515. IEEE, 2025.

[270] Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*, 2023.

[271] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[272] Google DeepMind Team. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv::2403.08295*, 2024.

[273] Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024.

[274] Christopher Tegho, Paweł Budzianowski, and Milica Gašić. Uncertainty estimates for efficient neural network-based dialogue policy optimisation. In *Bayesian Deep Learning Workshop at the Thirty-first Conference on Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, CA, USA, 2017.

[275] Dung Thai, Sree Harsha Ramesh, Shikhar Murty, Luke Vilnis, and Andrew McCallum. Embedded-state latent conditional random fields for sequence labeling. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 1–10, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[276] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR, 2020.

[277] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

[278] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMa: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[279] Susanne Trick, Dorothea Koert, Jan Peters, and Constantin A Rothkopf. Multimodal uncertainty reduction for intention recognition in human-robot interaction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7009–7016. IEEE, 2019.

[280] Luigi Troiano, Cosimo Birtolo, and Roberto Armenise. Modeling and predicting the user next input by bayesian reasoning. *Soft Computing*, 21(6):1583–1600, 2017.

[281] David Tuggy. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, 4(3):273–290, 1993.

[282] Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. PlanBench: an extensible benchmark for evaluating large language models on planning and reasoning about change. In *Advances Neural Information Processing Systems*, volume 36, pages 38975–38987. Curran Associates, Inc., 2024.

[283] Robert Van Rooij. Vagueness and linguistics. In Giuseppina Ronzitti, editor, *Vagueness: A Guide*, pages 123–170. Springer, Dordrecht, 2011.

[284] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 6000–6010. Curran Associates, Inc., 2017.

[285] Oriol Vinyals and Quoc V. Le. A neural conversational model. In *Proceedings of the 31st International Conference on Machine Learning (ICML)-Deep Learning Workshop*, volume 37, 2015.

[286] Wolfgang Wahlster. *SmartKom: Foundations of Multimodal Dialogue Systems (Cognitive Technologies)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[287] Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. GPT-4V(ision) for robotics: Multimodal task planning from human demonstration. *IEEE Robotics and Automation Letters*, 9(11):10567–10574, 2024.

[288] Marilyn Walker, John Aberdeen, Julie Boland, Elizabeth Bratt, John Garofolo, Lynette Hirschman, Audrey Le, Sungbok Lee, Shrikanth Narayanan, Kishore Papineni, et al. DARPA communicator dialog travel planning systems: the june 2000 data collection. In *7th European Conference on Speech Communication and Technology (Eurospeech 2001), 2nd INTERSPEECH Event*, pages 1371–1374, 2001.

[289] Chao Wang, Stephan Hasler, Daniel Tanneberg, Felix Ocker, Frank Joublin, Antonello Ceravola, Joerg Deigmoeller, and Michael Gienger. LaMI: Large language models for multi-modal human-robot interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–10, New York, NY, USA, 2024. Association for Computing Machinery.

[290] Fangju Wang, Shaidah Jusoh, and Simon X. Yang. A collaborative behavior-based approach for handling ambiguity, uncertainty, and vagueness in robot natural language interfaces. *Engineering Applications of Artificial Intelligence*, 19(8):939–951, 2006.

[291] Jiaqi Wang, Zihao Wu, Yiwei Li, Hanqi Jiang, Peng Shu, Enze Shi, et al. Large language models for robotics: Opportunities, challenges, and perspectives. *Journal of Automation and Intelligence*, 4(1):52–64, 2025.

[292] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[293] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, et al. CogVLM: Visual expert for pretrained language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 121475–121499, Red Hook, NY, USA, 2024. Curran Associates, Inc.

[294] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and Jifeng Dai. VisionLLM: large language model is also an open-ended decoder for vision-centric tasks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 61501 – 61513, Red Hook, NY, USA, 2023. Curran Associates Inc.

[295] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20:447–48, 2023.

[296] Xinlong Wang, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, et al. Multimodal learning with next-token prediction for large multimodal models. *Nature*, jan 2026.

[297] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics.

[298] Jason Wei, Maarten Bosma, Vincent Zhao, et al. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

[299] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-Thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

[300] Joseph Weizenbaum. ELIZA-a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[301] Stefan Wermter, Günther Palm, Cornelius Weber, and Mark Elshaw. *Towards Biomimetic Neural Learning for Intelligent Robots*, pages 1–18. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[302] Stefan Wermter and Ron Sun. An overview of hybrid neural systems. In Stefan Wermter and Ron Sun, editors, *Hybrid Neural Systems*, pages 1–13, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[303] Stefan Wermter and Volker Weber. SCREEN: learning a flat syntactic and semantic spoken language analysis using artificial neural networks. *Journal of Artificial Intelligence Research*, 6(1):35–85, January 1997.

[304] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

[305] Lance Whitney. Anthropic's claude 3 chatbot claims to outperform ChatGPT, Gemini, 2024.

[306] Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and

reinforcement learning. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 665–677, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[307] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. *Experimentation in Software Engineering*, volume 236. Springer, Berlin and Heidelberg, Germany, 2012.

[308] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. GRiT: A generative region-to-text transformer for object understanding. In *European Conference on Computer Vision*, pages 207–224, Cham, 2024. Springer Nature Switzerland.

[309] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. TidyBot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102, 2023.

[310] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[311] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. DeepSeek-VL2: Mixture-of-Experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

[312] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024.

[313] Xuan Xiao, Jiahang Liu, Zhipeng Wang, Yanmin Zhou, Yong Qi, Shuo Jiang, Bin He, and Qian Cheng. Robot learning in the era of foundation models: a survey. *Neurocomput.*, 638(C), July 2025.

[314] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[315] Jinxuan Xu, Shiyu Jin, Yutian Lei, Yuqian Zhang, and Liangjun Zhang. RT-Grasp: Reasoning tuning robotic grasping via multi-modal large language

model. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7323–7330, 2024.

[316] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, pages 483–498, Online, 2021. Association for Computational Linguistics.

[317] Natsuki Yamanobe, Weiwei Wan, Ixchel G Ramirez-Alpizar, Damien Petit, Tokuo Tsuji, Shuichi Akizuki, Manabu Hashimoto, Kazuyuki Nagata, and Kensuke Harada. A brief review of affordance in robotic manipulation research. *Advanced Robotics*, 31(19-20):1086–1101, 2017.

[318] Yang Yang, Xibai Lou, and Changhyun Choi. Interactive robotic grasping with attribute-guided disambiguation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8914–8920. IEEE, 2022.

[319] Yang Yang, Houjian Yu, Xibai Lou, Yuanhao Liu, and Changhyun Choi. Attribute-based robotic grasping with data-efficient adaptation. *IEEE Transactions on Robotics*, 40:1566–1579, 2024.

[320] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. UniTAB: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer Nature Switzerland, 2022.

[321] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The Dawn of LMMs: Preliminary explorations with GPT-4V(ision). In *arXiv preprint arXiv:2309.17421*, 2023.

[322] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc.

[323] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Chi Chen, Haoyu Li, Weilin Zhao, et al. Efficient GPT-4V level multimodal large language model for deployment on edge devices. *Nature Communications*, 16(1):5509, 2025.

[324] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[325] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2024.

[326] Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech and Language*, 24(2):150–174, 2009.

[327] Dian Yu, Michelle Cohn, Yi Mang Yang, Chun Yen Chen, Weiming Wen, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, Giritheja Sreenivasulu, Sam Davidson, Ashwin Bhandare, and Zhou Yu. Gunrock: A social bot for complex and engaging long conversations. 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): System Demonstrations*, pages 79–84, Hong Kong, China, 2019. Association for Computational Linguistics.

[328] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

[329] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. MAttNet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.

[330] Samson Yu, Kelvin Lin, Anxing Xiao, Jiafei Duan, and Harold Soh. Octopi: Object property reasoning with large tactile-language models. In *Robotics: Science and Systems*, Delft, Netherlands, July 2024.

[331] Weizhe Yuan, Zhengbao Chen, and Yue Zhang. RRHF: Rank responses to align language models with human feedback without tears. In *Advances in Neural Information Processing Systems*, volume 36, pages 10935–10950. Curran Associates, Inc., 2023.

[332] Matthew D. Zeiler. ADADELTA: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[333] Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. CLIP2: Contrastive language-image-point pretraining from real-world point cloud data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15244–15253, 2023.

[334] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[335] Ao Zhang, Yuan Yao, Wei Ji, Zhiyuan Liu, and Tat-Seng Chua. NExT-Chat: an lmm for chat, detection, and segmentation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 60116–60133, Vienna, Austria, 2024. JMLR.org.

[336] Bowen Zhang and Harold Soh. Large language models as zero-shot human models for human-robot interaction. In *2023 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 7961–7968. IEEE, 2023.

[337] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, et al. LlaVA-Grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024.

[338] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, and Jainwei Yang. LLaVA-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision, Part XLIII*, pages 19–35, Milan, Italy, 2024. Springer-Verlag.

[339] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, and Yinfei Yang. Ferret-v2: An improved baseline for referring and grounding with large language models. In *First Conference on Language Modeling*, 2024.

[340] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[341] Xiaoying Zhang, Baolin Peng, Jianfeng Gao, and Helen Meng. Toward self-learning end-to-end task-oriented dialog systems. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 516–530, Edinburgh, UK, 2022. Association for Computational Linguistics.

[342] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[343] Xinyan Zhao, Bin He, Yasheng Wang, Yitong Li, Fei Mi, Yajiao Liu, Xin Jiang, Qun Liu, and Huanhuan Chen. UniDS: A unified dialogue system

for chit-chat and task-oriented dialogues. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 13–22, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[344] Xufeng Zhao, Mengdi Li, Cornelius Weber, Muhammad Burhan Hafez, and Stefan Wermter. Chat with the Environment: Interactive multimodal perception using large language models. In *Conference on Intelligent Robotsand Systems (IROS)*, pages 3590–3596, 2023.

[345] Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc.

[346] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *The Thirteenth International Conference on Learning Representations*, 2025.

[347] Hongkuan Zhou, Xiangtong Yao, Yuan Meng, Siming Sun, Zhenshan Bing, Kai Huang, and Alois Knoll. Language-conditioned learning for robotic manipulation: A survey. *arXiv preprint arXiv:2312.10807*, 2023.

[348] Jie Zhou and Wei Xu. End-to-End learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 1127–1137, 2015.

[349] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[350] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. JudgeLM: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*, 2025.

[351] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China, August 2021. Chinese Information Processing Society of China.

[352] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

Appendix A

Nomenclature

AdaDelta adaptive delta. 64,

AI artificial intelligence. 4, 7, 17, 47, 91,

ANN artificial neural network. 13, 17,

ASR automatic speech recognition. 14, 16, 19, 61, 87,

BART Bidirectional and Auto-Regressive Transformers. 29, 57,

BERT Bidirectional Encoder Representations from Transformers. 22, 28, 29, 40, 41, 57,

BoW bag-of-words. 8, 21, 62, 63, 64, 65, 112

BPE byte-pair encoding. 22,

BPTT backpropagation through time. 64,

CAI constitutional AI. 31,

CBOW continuous bag-of-words. 21,

CCE categorical cross-entropy. 64,

CLIP Contrastive Language-Image Pre-training.

CLM causal language modeling. 30, 99,

CNN convolutional neural network. 16, 22, 42, 61,

COCO Common Objects in Context. 37, 57,

Cooper **Co**mmand and Object Prop**e**rties Recognition. 86,

CoT chain-of-thought. 46, 77, 111

CV computer vision. 4, 70, 74, 89, 97,

DBN deep belief network. 16,

DCM dense captioning model. 74, 75, 77, 112, 115

DDN dynamic decision network. 13,

DL deep learning. 4, 5, 13, 16, 18, 21, 109

DM dialog manager. 14, 16, 17, 52, 57, 109

DNN deep neural network. 4, 13, 18, 19, 109, 110

DPM dialog policy model. 16,

DPO direct preference optimization. 31,

DST dialog state tracker. 16,

ELMo embeddings from language models. 21,

FPN feature pyramid network. 61,

GloVe Global Vectors for Word Representation. 21,

GPT Generative Pre-trained Transformer. 22, 57,

GRU gated recurrent unit. 19, 25, 26,

HiPolicy Human-robot **i**nteraction Policy. 8, 61, 63,

HIS hidden information state.

HRI human-robot interaction. XIII, 2, 4, 6, 8, 36, 39, 48, 49, 50, 51, 53, 55, 57, 58, 66, 83, 89, 110, 111, 112

HYNA HYbrid Neural Architecture. 8, 61, 62, 64, 66, 110, 111, 112

KS knowledge source. 14, 16, 17, 19,

LLM large language model. 4, 6, 7, 17, 23, 28, 31, 32, 42, 44, 45, 46, 47, 48, 52, 67, 69, 73, 74, 75, 76, 77, 80, 83, 84, 85, 87, 91, 99, 109, 110, 111, 113, 115

LoRA low-rank adaptation. 31, 113

LSTM long short-term memory. 19, 21, 25, 26, 63, 109

M-DST multimodal dialog state tracker. 8, 61, 62, 63, 64, 65, 112

MDP Markov decision process. 17,

ML machine learning. 8, 14, 15, 16, 17, 73,

MLM masked language modeling. 28, 29, 30, 40,

MLP multilayer perceptron. 99, 103, 104,

MTP multi-token prediction. 30,

NICO Neuro-Inspired COmpanion. 8, 55, 56, 58, 59, 61, 112

NLG natural language generation. 14, 17, 57, 109

NLP natural language processing. 4, 12, 21, 22, 26, 40,

NLU natural language understanding. 14, 15, 16, 47, 48, 57, 109

NSP next sentence prediction. 28,

OD object detection. 8, 61, 64,

OSSA Object State-Sensitive Agent. 8, 67, 68, 73, 79, 81, 97, 111, 112, 114

PALM Pre-training an Autoencoding & Autoregressive Language Model. 29,

PaLM Pathways Language Model. 46,

PCFG probabilistic context-free grammar. 15,

Pix2Emb pixel-to-embedding. 90, 106,

Pix2Seq pixel-to-sequence. 90, 106,

POMDP partially observable Markov decision process.

PrefixLM prefix language modeling. 30, 42,

QLoRA quantized low-rank adaptation. 31,

REC referring expression comprehension. 9, 37, 89, 90, 91, 93, 100, 102, 105, 106, 107, 113

RefCOCO Referring Expressions in COCO. 37,

RefCOCO+ variant of Referring Expressions in COCO Plus. 37,

RefCOCOg Google version of Referring Expressions in COCO. 37,

RL reinforcement learning. 17, 31, 57, 73,

RLAIF reinforcement learning from AI feedback. 31,

RLHF reinforcement learning from human feedback. 31,

RNN recurrent neural network. 16, 19, 23, 26, 28, 31, 57, 58, 63,

RRHF rank responses to align human feedback. 31,

RSFT rejection sampling fine-tuning. 31,

RTD replaced token detection. 29,

Seq2Seq sequence-to-sequence. 13, 17, 19, 27,

SIMMC Situated and Interactive Multimodal Conversations. 37, 55, 57,

SL supervised learning.

SOP sentence order prediction. 28,

SOTA state-of-the-art. 14, 23, 28, 31, 41, 64, 66, 68, 74, 75, 80, 89, 91, 102, 106, 109, 110, 115

StateVLM State-sensitive Vision-Language Model. 111, 113, 114

SVM support vector machine. 13, 16, 17,

T5 Text-to-Text Transfer Transformer. 22, 29,

TF-IDF term frequency-inverse document frequency. 21,

TTS text-to-speech. 14, 19, 61,

UE utterance embedding. 8, 62, 63, 64, 65, 66, 112

UVA-phenomena Uncertainty, Vagueness, and Ambiguity. 3, 8, 9, 36, 48, 51, 52, 53, 109, 110, 114

VLA vision-language-action. 46, 53,

VLD vision-language dataset. 36, 110

VLM vision-language model. 4, 5, 6, 7, 9, 36, 40, 42, 44, 45, 46, 47, 52, 67, 69, 70, 73, 74, 75, 76, 77, 78, 80, 89, 90, 91, 92, 93, 97, 98, 105, 106, 109, 110, 111, 113, 114, 115

VQA visual question answering. 36, 55, 57, 89,

Word2Vec word to vector. 21, 62, 65, 66,

Appendix B

Publications Originating from this Thesis

B.1 Conferences

[1] **Xiaowen Sun**, Cornelius Weber, Matthias Kerzel, Josua Spisak, Stefan Wermter. *Uncertainty, Vagueness, and Ambiguity in Human-Robot Interaction: Why Conceptualization Matters*. In *Interactive AI for Human-Centered Robotics Workshop at Human-Robot Interaction 2026*.

[2] **Xiaowen Sun**, Xufeng Zhao, Jae Hee Lee, Wenhao Lu, Matthias Kerzel, Stefan Wermter. *Details Make a Difference: Object State-sensitive Neurorobotic Task Planning*. In *Proceedings of the 33rd International Conference on Artificial Neural Networks (ICANN 2024)*, pp. 261–275.

[3] Theresa Pekarek Rosin, Vanessa Hassouna, **Xiaowen Sun**, Luca Krohm, Henri-Leon Kordt, Michael Beetz, Stefan Wermter. *A Framework for Adapting Human-Robot Interaction to Diverse User Groups*. In *Proceedings of the 2024 International Conference on Social Robotics (ICSR 2024)*, pp. 24–38.

[4] **Xiaowen Sun**, Cornelius Weber, Matthias Kerzel, Tom Weber, Mengdi Li, Stefan Wermter. *Learning Visually Grounded Human-Robot Dialog in a Hybrid Neural Architecture*. In *Proceedings of the 31st International Conference on Artificial Neural Networks (ICANN 2022)*, pp. 258–269.

Table B.1: Statement of Author Contributions.

Paper NO.	**Dataset Details**	**Conception**	**Implementation**	**Writing**
[1]	No dataset was used.	X.S. proposed the idea; C.W., M.K., and J.S. participated in discussions.	No code was released.	X.S. wrote the draft; all authors reviewed and edited the manuscript.
[2]	X.S. designed the dataset; X.S. and W.L. annotated it; M.K. contributed a portion of the images.	X.S. proposed the idea; X.S., X.Z., and J.L. participated in discussions.	X.S. designed, implemented, and evaluated the method, and conducted the analysis.	X.S. wrote the draft; all authors reviewed and edited the manuscript.
[3]	X.S. designed and annotated the dataset for the dialog bridge; T.R. and V.H. contributed the remaining components.	T.R. and V.H. proposed the initial idea; T.R., V.H., and X.S. participated in discussions.	X.S. designed, implemented, evaluated, and analyzed the dialog bridge; T.R. and V.H. contributed the remaining components.	T.R. and X.S. wrote the draft; all authors reviewed and edited the manuscript.
[4]	X.S. designed and annotated the dataset; M.K. contributed the images.	X.S. proposed the idea; C.W. and M.K. participated in discussions.	X.S. designed, implemented, evaluated, and analyzed the method.	X.S. wrote the draft; all authors reviewed and edited the manuscript.

Appendix C

Acknowledgements

Finally, I would like to express my heartfelt gratitude to all those who made this research possible.

I am deeply grateful to my supervisor, Professor Wermter, for his rigorous guidance and unwavering support throughout my doctoral studies. I would like to thank my three advisors: Dr. Weber for his kindness, Dr. Kerzel for his encouragement, and Dr. Lee for his constructive criticism. I also thank the members of my thesis committee for their time, effort, and constructive comments. My thanks also go to Katja, Claudia, Erik, and Reinhard for their professional assistance.

Thank you to my colleagues: Alexander Sutherland, Burak Can Kaplan, Connor Gäde, Dr. Dennis Becker, Dr. Di Fu, Dr. Huanjian Fang, Hassan Ali, Dr. Johannes Twiefel, Jan-Gerrit Habekost, Julia Gachot, Kun Chu, Marie Sophie Bauer, Dr. Mengdi Li, Mostafa Kotb, Dr. Ozan Özdemir, Dr. Hugo Carneiro, Tom Weber, Dr. Philipp Allgeuer, Dr. Kyra Ahrens, Sergio Lanza, Theresa Pekarek-Rosin, Wenhao Lu, and Dr. Xufeng Zhao. You are all professional scientists with fascinating souls. I hope we can develop lasting friendships even after we are no longer colleagues.

Thank you to my friends: Huarui, Dr. Lijuan Sun, Benjamin, Hong Vi, Dr. Jianzhi Lyu, and Vatstal. Your kindness and companionship have warmed my heart, and I hope you feel the same.

Thank you to my family: my parents, Dexia and Jinju, and my uncles, Dewei and Deli. Your unconditional support gave me the courage to explore the world and myself. I am especially grateful to my grandparents, Lizheng and Lan, who gave me a wonderful childhood. Although grandparents have passed away, you live on in my heart through wonderful memories that will last forever.

Lastly, and most importantly, thank you to my partner, Josua. Thank you for your companionship and for your understanding and guidance regarding my emotions. Thank you to your parents, Peter and Judith, for their trust and open-minded acceptance.

I want to gratefully acknowledge support from the China Scholarship Council (CSC) for my living expenses in Germany and the German Research Foundation DFG under the project CML (TRR 169) for my conference traveling.

In the end, I also want to thank myself for my Curiosity, Optimism, and Persistence.

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.

25.07.2026 Xiaowen Sun

Ort, Datum Unterschrift

Erklärung zur Veröffentlichung

Ich stimme der Einstellung der Dissertation in die Bibliothek des Fachbereichs Informatik zu.

25.07.2026 Xiaowen Sun

Ort, Datum Unterschrift