Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FACULTY
OF BUSINESS ADMINISTRATION

Leveraging Machine Learning for Modelling Unstructured Data in Marketing

Cumulative Dissertation

to obtain the academic degree of a

„Doctor rerum oeconomicarum (Dr. rer. oec.)“

according to doctoral degree regulations 2024

at the Faculty of Business Administration

Moorweidenstr. 18

20148 Hamburg / Germany

of Universität Hamburg

submitted by: Keno Tetzlaff

born on 16.10.1994 in Bremen (Germany)

Hamburg, 08.02.2025

Chairperson: Prof. Dr. Jan Recker

First Examiner: Prof. Dr. Mark Heitmann

Second Examiner: Prof. Dr. Jochen Hartmann

Date of Disputation: 14.07.2025

TABLE OF CONTENTS

I. SYNOPSIS

SYNOPSIS

Leveraging Machine Learning for Modelling Unstructured Data in Marketing

by Keno Tetzlaff

Unstructured data has always been an integral part of marketing. Advertising on billboards relies on visual and textual elements, jingles on the radio or brand mentions in songs are audio-based, and TV spots are a combination of all these elements. Thus, marketing researchers have been very interested in understanding how these affect consumer responses and which elements are the most important adjustable drivers (e.g., Edell and Burke 1987; MacInnis et al. 1991; Olney et al. 1991). With the rise of the Internet and social media, the volume of unstructured data has increased. Additional use cases for marketers and applied researchers in the field have emerged. In this cluttered digital environment, consumers are faced with information overload, and marketers need to optimize all communication to stand out amidst the noise.

Although unstructured data provides a wealth of information, the sheer amount available makes it difficult to analyze it at scale and extract structured features. Initially, marketing researchers were restricted to smaller sample sizes and manually coded features in their analyzes, but the advent of machine learning (ML) technologies has enabled larger datasets and novel insights. Today, the intersection of ML and unstructured data provides an enormous pool of opportunities for research and continues to shape the agendas of leading journals (Balducci and Marinova 2018; Grewal et al. 2021). Marketing research evolves with technological possibilities and applies ML methods such as text (e.g., Netzer et al. 2012; Timoshenko and Hauser 2019), image (e.g., Troncoso and Luo 2023; Dzyabura et al. 2023), a combination

thereof (e.g., Hartmann et al. 2021; Chang et al. 2023), or most recently generative artificial intelligence (AI) (e.g., Reisenbichler et al. 2022; Li et al. 2024).

The essays in this dissertation revolve around the ongoing research stream at the intersection of quantitative marketing and ML. Three papers leverage a diverse set of ML tools to analyze and model unstructured data for marketing. What unifies all essays is the integration of state-of-the-art ML techniques with well-established methods and marketing theories. Specifically, the dissertation explores the applicability of image classification to marketing tasks and provides usage guidelines, analyzes video marketing in the context of stimulation theory across online and offline channels, and suggests a novel approach toward information dissemination in marketing research. Several extensive datasets (field, survey, experiments) obtained from industry partners enable academic rigor and allow for applicability to practical marketing. These help to showcase how innovative methodological development (i.e., ML) can lead to new insights from unexplored unstructured data sources and enhance practical decision-making in data-driven marketing. The first essay conducts comparative studies on image classification. The second essay combines all modes of data in the analysis of video ads, while the third proposes a novel approach for the extraction of information from academic articles. Together, they demonstrate the value and of possibilities for the application of ML methods in marketing.

The first essay *"The Language of Images: Performance of Image Classification Paradigms in Marketing"* compares 9 methods for image classification across 18 datasets, 2 training data sizes (large vs. limited), and 3 types of tasks (what, who, how). The essay summarizes current automated image classification applications in marketing. It challenges the predominant use of older, popular methods (CNNs) and highlights conceptual differences to two more recent, language-inspired modelling

approaches (vision transformers, vision language models) and an ensemble that combines the two methods. The results of this large-scale comparison (i.e., 3,150 experiments) suggest that vision transformers (i.e., ConvNeXt, ViT) outperform CNNs by +10 ppts on average and show less variation in their performance across datasets. However, more nuanced results show the value of vision language models for marketing-relevant applications – that is, complex, perceptual tasks with limited training data. For researchers willing to invest more effort, it appears that an ensemble model can combine the best of both worlds and deliver consistent top performance. Lastly, the essay provides decision support for applied researchers, as it discusses relevant trade-offs across different tasks (e.g., effort, interpretability, performance)

The second essay *"Valuable visual variety? How temporal dynamics of visual features in video ads affect ad effectiveness"* explores the impact of visual variety on ad effectiveness in more than 3,000 video ads across online (Facebook) and offline (Television) channels. In an increasingly stimulating world, advertisers need to balance a fine line to stand out from the noise without overstimulating their prospective customers. This essay develops a scalable, multi-dimensional measure of semantic visual variety by comparing the similarity of images. In a multi-method approach, it leverages various ML methods to control for further stimulation from the videos and relates the developed measure to consumer behavior. The results suggest that the relation of visual variety and consumer behavior follows an inverted-U shape across channels. This relation is mediated by both stimulation and comprehension where advertisers need to strike the right balance. Consequently, this essay provides guidelines for visual variety optimization and code to calculate the measure for videos.

The third essay *"Methods in Marketing: Leveraging Artificial Intelligence for Academic Information Dissemination"* proposes a novel approach to reviewing

literature and disseminating information in marketing research. In contrast to the other essays, the focus of this essay lies solely on text analysis. In a dynamic field such as marketing, it is imperative to follow substantive and methodological developments. However, an ever-increasing number of publications and the frequent sprouting of novel (AI) methods make tracking these developments a daunting task. This essay develops a scalable process for literature review that combines a few-shot classifier (Tunstall et al. 2022) with generative AI to extract structured information from more than 21,000 peer-reviewed marketing articles. Using this information, it provides insights into the development of the marketing research domain from 2010 to 2023 and identifies increasing complexity in marketing research (see https://www.berd-nfdi.de/analytics-portal-new/ for a web interface to explore slightly different data based on the developed extraction approach). The version included in this paper leverages an extended and updated dataset (analyzing 3x the amount of articles) but is based on the version accepted for EMAC 2024.

In summary, this dissertation contributes to an increasing stream of literature at the intersection of ML and its application to practical, data-driven marketing challenges. It includes a large-scale, comparative study that provides marketers with guidelines for the use of automated image classification, applies a multi-method approach to understand visual variety in video ads, and develops a novel process for information dissemination in marketing research. Together, these essays shed light on some of the exciting possibilities for the application of ML to marketing topics and their value.

Table 1 Overview of Dissertation Projects

Title	Authors	Status	Research Objective	Methodology	Main Results
The Language of Images: Performance of Image Classification Paradigms in Marketing	Tetzlaff, Witte, Hartmann, Heitmann (2024)	*International Journal of Research in Marketing* (2nd round)	Compare 3 image classification paradigms for marketing and aid decision making of applied researchers	Comparative study using 3,150 unique experiments (method-dataset-training data size-hyperparameter tuning combinations) to assess methods under a variety of settings. Across 3 task types, the best performing methods and trade-offs are evaluated (e.g., effort vs. interpretability)	CNNs prevail in marketing research, but recent language-inspired modeling (vision transformers, vision language models) could improve image classification accuracy across applications. On average, +10ppts with less performance variation
Valuable visual variety: How temporal dynamics of visual features in video ads affect ad effectiveness	Tetzlaff[1], Krugmann[1], Kinzinger[1], Hartmann[1] (2024)	*Journal of the Academy of Marketing Science* (3rd round)	Explore the optimal level of visual variety in video advertising across media channels (social media vs. TV) to optimize ad effectiveness	More than 3,000 video ads from two media channels (TV, Facebook – field and survey data partially including actual purchases), one lab experiment. Mix of state-of-the-art image mining, text classification, and audio analysis machine learning tools combined with established empirical models (OLS, neg. bin. Regression, mediation)	We find an inverted U-shaped relationship across platforms and dependent variables. The optimum level of visual variety to drive ad effectiveness is mediated by both stimulation and comprehension. It is important to strike a balance for sufficient stimulation without undermining comprehension
Methods in Marketing: Leveraging Artificial Intelligence for Academic Information Dissemination	Witte, Tetzlaff, Reisenbichler, Heitmann (2025)	*Working paper* (Accepted for EMAC 2024 – updated underlying data thereafter)	Automatically assess method usage in empirical marketing papers at scale	Few-shot text classification and conversational AI (GPT4) for information retrieval from more than 21,000 full-text articles published in marketing journals. Information includes methods used, constructs, data source, and others.	Complexity in marketing increases over time. Since 2010 the number of articles published per year and the number of methods used per article have increased. Our tool helps researchers to easily find relevant information.

[1] Equal contribution

References

Balducci, Bitty and Detelina Marinova (2018), "Unstructured data in marketing," *Journal of the Academy of Marketing Science*, 46 (4), 557–590.

Chang, Hannah H., Anirban Mukherjee and Amitava Chattopadhyay (2023), "More Voices Persuade: The Attentional Benefits of Voice Numerosity," *Journal of Marketing Research*, 60 (4), 687–706.

Dzyabura, Daria, Siham El Kihal, John R. Hauser and Marat Ibragimov (2023), "Leveraging the Power of Images in Managing Product Return Rates," *Marketing Science*, 42 (6), 1125– 1142.

Edell, Julie A. and Marian Chapman Burke (1987), "The Power of Feelings in Understanding Advertising Effects," *Journal of Consumer Research*, 14 (3), 421–433.

Grewal, Rajdeep, Sachin Gupta and Rebecca Hamilton (2021), "Marketing Insights from Multimedia Data: Text, Image, Audio, and Video," *Journal of Marketing Research*, 58 (6), 1025–1033.

Hartmann, Jochen, Mark Heitmann, Christina Schamp and Oded Netzer (2021), "The Power of Brand Selfies," *Journal of Marketing Research*, 58 (6), 1159–1177.

Li, Peiyao, Noah Castelo, Zsolt Katona and Miklos Sarvary (2024), "Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis," *Marketing Science*, 43 (2), 254–266.

MacInnis, Deborah J., Christine Moorman and Bernard J. Jaworski (1991), "Enhancing and Measuring Consumers' Motivation, Opportunity, and Ability to Process Brand Informa- tion from Ads," *Journal of Marketing*, 55 (4), 32–53.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg and Moshe Fresko (2012), "Mine Your Own Business: Market-Structure Surveillance Through Text Mining," *Marketing Science*, 31 (3), 521–543.

Olney, Thomas J., Morris B. Holbrook and Rajeev Batra (1991), "Consumer Responses to Advertising: The Effects of Ad Content, Emotions, and Attitude toward the Ad on Viewing Time," *Journal of Consumer Research*, 17 (4), 440–453.

Reisenbichler, Martin, Thomas Reutterer, David A. Schweidel and Daniel Dan (2022), "Fron- tiers: Supporting Content Marketing with Natural Language Generation," *Marketing Sci- ence*, 41 (3), 441–452.

Timoshenko, Artem and John R. Hauser (2019), "Identifying Customer Needs from User-Generated Content," *Marketing Science*, 38 (1), 1–20.

Troncoso, Isamar and Lan Luo (2023), "Look the Part? The Role of Profile Pictures in Online Labor Markets," *Marketing Science*, 42 (6), 1080–1100.

Tunstall, Lewis, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat and Oren Pereg (2022), "Efficient Few-Shot Learning Without Prompts," , https://arxiv.org/abs/2209.11055.

II. SELF-DECLARATION

Table 2 Self-declaration

Title	Authors	Work Steps (Participation)
The Language of Images: Performance of Image Classification Paradigms in Marketing	Tetzlaff, Witte, Hartmann, Heitmann	• Identification of research gap and literature review • Conceptualization and execution of empirical analyses and result preparation • Report writing and multiple overhauls within the revision process for the *International Journal of Research in Marketing*
Valuable visual variety: How temporal dynamics of visual features in video ads affect ad effectiveness	Tetzlaff[1], Krugmann[1], Kinzinger[1], Hartmann[1]	• Identification of research gap and literature review • Conceptualization and execution of empirical analyses and result preparation • Report writing and overhaul within the revision process for the *Journal of the Academy of Marketing Science*
Methods in Marketing: Leveraging Artificial Intelligence for Academic Information Dissemination	Witte, Tetzlaff, Reisenbichler, Heitmann	• Identification of research gap and literature review • Conceptualization of methodological approach, empirical analyses, and result preparation • Report writing for submission at the *European Marketing Conference 2024*

[1] Equal contribution

III. DISSERTATION PROJECTS

A. The Language of Images: Performance of Image Classification Paradigms in Marketing

Authors

Keno Tetzlaff

Maximillian Witte

Jochen Hartmann

Mark Heitmann

Year

2024

Bibliographic Status

Under review at

International Journal of Research in Marketing

2nd round

The Language of Images: Performance of Image Classification Paradigms in Marketing

Keno Tetzlaff[1]

Maximilian Witte[2]

Jochen Hartmann[3]

Mark Heitmann[4]

Funding:

This work was funded by the German Research Foundation (DFG) research grant number 460037581.

[1]Keno Tetzlaff is Doctoral Student at the University of Hamburg, Hamburg Business School, Moorweidenstrasse 18, 20148 Hamburg, Germany. keno.tetzlaff@uni-hamburg.de.

[2]Maximilian Witte is Doctoral Student at the University of Hamburg, Hamburg Business School, Moorweidenstrasse 18, 20148 Hamburg, Germany. maximilian.witte@uni-hamburg.de.

[3]Jochen Hartmann is Professor of Digital Marketing at the Technical University of Munich, TUM School of Management, Arcistrasse 21, 80333 Munich, Germany. jochen.hartmann@tum.de.

[4]Mark Heitmann is Professor of Marketing & Customer Insight at the University of Hamburg, Hamburg Business School, Moorweidenstrasse 18, 20148 Hamburg, Germany. mark.heitmann@uni-hamburg.de.

The Language of Images: Performance of Image Classification Paradigms in Marketing

Abstract

Images say more than a thousand words. But can marketing leverage language modeling concepts to study image content? Marketing utilizes image classification for studying how advertising, social media, or e-commerce images relate to consumer perceptions and economic outcomes. To accomplish this, marketing publications apply variants of convolutional neural networks that analyze local image patterns by examining neighboring pixels. Recent transformer architectures, inspired by language modeling, study images more holistically by learning relationships across distant image parts. Even newer vision language models based on generative AI advances create detailed text interpretations of what they 'see' in images. We study the benefits of these advances based on 18 marketing-related datasets that cover what and who is visible, as well as how images are perceived. On average, language modeling concepts improve image classification accuracy by more than 10 percentage points. When training data is abundant, transformer architectures perform best and also most consistently across datasets. Performance of vision language models varies. Relative to alternatives, these models perform strongest with limited training data and for complex tasks focused on how images are perceived. Combining them with transformer-inspired architectures as a multi-paradigm ensemble achieves the best of both worlds, with the highest and most consistent performance across all tasks and datasets we study.

Keywords: generative AI, computer vision, image mining, machine learning, image classification, marketing insight

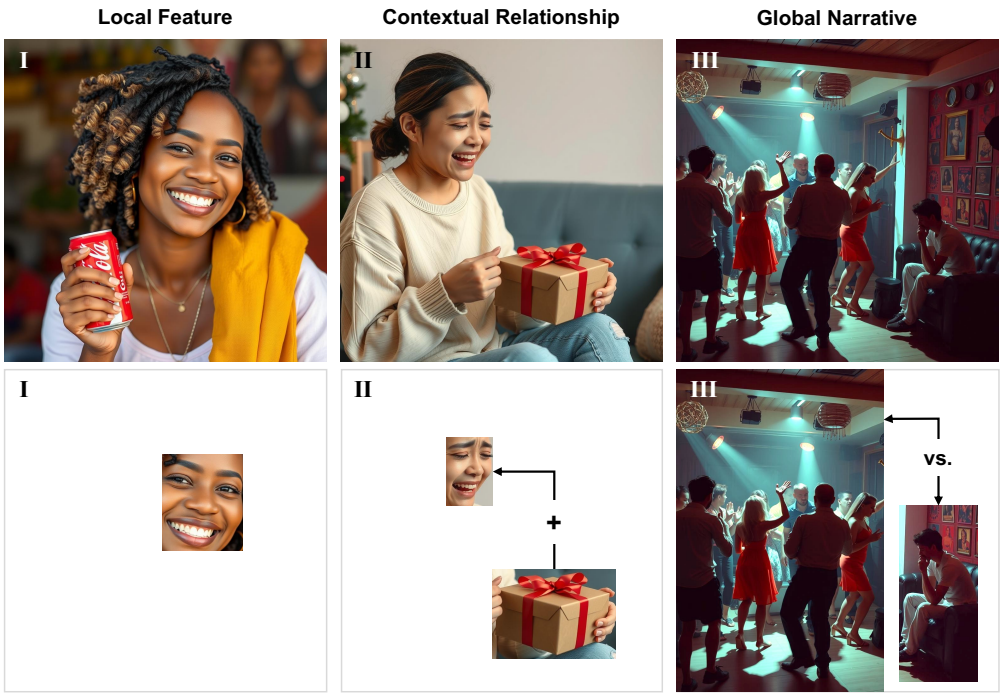
Images attract attention and convey rich information. They can say more than a thousand words, which makes them a crucial element in modern-day marketplace interactions. Consumers post images that contain brand content (Beichert et al. 2023), marketers distribute a wealth of advertising images (e.g., Pieters et al. 2010; Rietveld et al. 2020; Lee 2021) and crowdsource designs or use generative AI to create multiple image versions. This proliferation of image content needs new ways of assessment to understand how to select appropriate image content. Recently, methods have become available that study images similar to language. In this research, we discuss and assess empirically when these are particularly useful and how they perform in terms of classification accuracy.

Image analysis has become popular in marketing research. Notably, some applications go beyond image classification (see Dzyabura et al. 2022 for a comprehensive review of image analytics in marketing). For example, Gorn et al. (1997) analyze how color relates to consumer responses of print ads. Others assess the impact of mere image presence on social media engagement (Li and Xie 2020) or for predicting restaurant survival (Zhang and Luo 2022). However, the vast majority of image analysis in marketing classifies images into various categories and uses these classifications in downstream econometric analysis. For example, Lee (2021) classifies the emotionality of brand-posted images and analyzes its effect on the average basket size of consumers. Other research classifies facial profile images to study how ethnicity relates to responses on online platforms (e.g., Zhang, Mehta, Singh and Srinivasan 2021; Gunarathne et al. 2022). Image classification is also used to classify relevant frames in crowdsourcing videos and predict the success of crowdsourcing campaigns (Li et al. 2019). It helps further to assess the impact of specific types of emotions displayed in online streaming (Lin et al. 2021) and TV broadcasting (Bharadwaj et al. 2022) on viewer response and sales, or to identify how the attractiveness of the seller relates to their success on e-commerce platforms (Peng et al. 2020).

These types of analysis are possible because increasingly sophisticated image classification methods derive structured data from unstructured image content. However, different images

and different image tasks require different capabilities. Consider three conceptual scenarios: A marketer might want to classify customer emotions in user-generated social media posts to extract consumer sentiment from image content. Some cases are straightforward, like a person displaying an unambiguous smile (see Figure 1, panel I). A reliable extraction of local image features and neighboring smile-related pixels can suffice. However, marketing imagery often involves more nuance, such as analyzing how consumers interact with products (see Figure 1, panel II). Understanding contextual relationships between multiple elements can improve conclusions – like a person's facial expression while unboxing a gift. This can make those image classification models more attractive that learn global connections between image parts, even if they are distant. What appears as sadness when examining the face alone can be reinterpreted as joy when considered alongside the gift being received, much like how understanding a sentence requires connecting words that may be far apart.

Figure 1: Image classification scenarios with increasing difficulty



Notes: Upper panel shows original images (FLUX-generated). The lower panel indicates key image regions and relationships needed for correct classification.

Even more challenging are situations requiring a deep contextual understanding of a vi-

sual narrative that marketing is often interested in, e.g., when studying multiple nuanced emotions. The correct interpretation often depends on more complex relations among multiple focal points (see Figure 1, panel III). For instance, the classification of an event photo can yield very different results depending on whether the focus is on multiple celebrating individuals, the isolated individual, or the relation between all of them (e.g., majority celebrating, individual thinking vs. individual not integrated). Recently, automated, detailed text descriptions have become available that can capture the global narrative of an image. Translating visual content into natural language might be beneficial to marketing applications, where context and subtle emotional cues can drive consumer perceptions.

Clearly, the three examples vary in complexity. More broadly, marketers can be interested in *what* is displayed (brand, gift, indoor vs. outdoor setting), *who* is visible (one or multiple individuals, gender, ethnicities), or *how* images are likely perceived and which emotions the images might elicit from their viewers. The latter is a more subjective image-related task, requiring multiple and more costly annotations to train a classification model (Schamp et al. 2024). These annotations can include several possible emotions a brand's content evokes and nuanced multi-class 'how' tasks with many classes (Dzyabura et al. 2022)). On the other hand, determining whether or not a product is visible in a social media post (a binary 'what' task; Hartmann et al. 2021) typically requires just one or two reliable annotations per image.

This creates several important trade-offs: While large training datasets help models learn robust patterns, obtaining high-quality annotations for perception-based tasks is costly and time-consuming. The challenge intensifies when moving beyond binary classification (e.g., presence vs. absence of a brand) to multi-class problems (e.g., distinguishing between multiple emotional states or brand attributes), as each additional class reduces the number of observations available per category. To which extent attainable accuracy differs has not been studied so far.

To date, peer-reviewed marketing publications have applied different versions of convolutional neural networks (CNNs). These include popular open-source models such as Inception

(Troncoso and Luo 2023), ResNet (e.g., Zhang, Mehta, Singh and Srinivasan 2021; Dzyabura et al. 2023), VGG (e.g., Hartmann et al. 2021; Zhang, Lee, Singh and Srinivasan 2021), and Xception (Bharadwaj et al. 2022). All of these models share the conceptual idea to group pixels across multiple levels of abstraction to identify relevant features for the classification task. This might not be ideal for the image examples in panels II and III of Figure 1 that marketing is also interested in and where relational aspects and image narratives are more important.

More recently, entirely different classes of models have emerged that follow successful principles from language modeling in general and text classification in particular. These include vision transformers like ViT (Dosovitskiy et al. 2020). Vision transformers convert images into smaller patches that are processed as if they were words in a sentence, allowing the model to capture both local features and global relationships within images. For example, the model can learn that distant features like the individual and the gift in panel II are semantically related, much like distant words could relate in a sentence (so-called long-range dependencies; Guo et al. 2022).

Even more recently, vision language models like Phi-3 (Abdin et al. 2024), also known as multi-modal large language models, extend these capabilities further. They combine generative language modeling with image analytics. They are trained on large-scale, multi-modal datasets where text descriptions are paired with images (Bordes et al. 2024). Among other things, this allows researchers to interact with visual data through simple text prompts, e.g., by asking "What do you expect a person seeing the image to think and feel?". The written output can serve as an input to image classification to, for example, understand a consumer's emotional response. At the same time, the text output can provide rich diagnostic insight.

Can the application of these two language-inspired approaches – vision transformers and vision language models – meaningfully improve image classification accuracy for marketing applications? Encouraging benchmark studies in computer science suggest vision trans-

formers can outperform CNNs (e.g., Maurício et al. 2023). However, these datasets do not resemble the full scope of tasks and dataset characteristics that marketing is interested in. Most prominently, benchmarks in computer science test large datasets with many training examples and relatively simple questions like *what* is visible in an image (e.g., objects, products; Deng et al. 2009). These results provide little guidance about the actual benefits in applied marketing settings when training data is costly and often more limited, other classifications tasks are of interest, and numbers of classes vary.

To our best knowledge, there are no published marketing applications of vision transformers, so their actual advantages on marketing tasks remain unknown. Despite their conceptual appeal, the image classification potential of vision language models remains entirely unexplored. More broadly, it is not clear how these different architectures would perform across tasks, datasets, and sample sizes, making reasonable expectations for different application settings impossible. Research on text classification suggests architectural rather than incremental improvements provide practically relevant leaps in performance (Hartmann et al. 2023). In this research, we study whether similar advantages in image classification can be observed and when these are strongest.

Choosing the most appropriate image classification methods is complex for marketing researchers, as they are ultimately interested in substantive problems and downstream analysis. Image classification is only one, often preliminary, step in their analytical pipeline. Exploring a complex technical solution space of thousands of alternative options is time-consuming and nontrivial to explore. Similar to other areas of method choice, e.g., on mediation and moderation (Zhao et al. 2010) or text classification (Hartmann et al. 2019, 2023; Krugmann and Hartmann 2024), guidance is useful to understand what works when. For example, Dzyabura et al. (2023) find that image analysis helps online retailers improve return rates, and report that an increase of only 8.76 percentage points in accuracy results in more than 1% profit improvement. This is in line with previous studies that suggest substantial economic and econometric consequences of suboptimal method choice (Hartmann et al. 2019;

Jaidka et al. 2020; Frankel et al. 2022). The present research reviews past practices of image classification in marketing. It conceptually compares models applied outside of marketing to the most recent vision language models, not yet applied for image classification. It tests these alternatives empirically, and proposes practical, evidence-based decision guidelines for marketing applications.

Based on more than 3,100 experiments across 18 unique datasets, we find converging evidence that treating images more like language benefits classification accuracy across applications. On average, processing image patches like words in a sentence with vision transformers outperforms traditional CNNs by more than 10 percentage points, suggesting that marketing scholars can benefit from transitioning to these models. It is worth noting that vision transformer performance also varies less across applications and datasets than that of CNNs. Vision language models prove particularly valuable for complex perceptual tasks with limited training data – precisely the conditions many marketing researchers are most interested in. These models achieve competitive performance while offering higher interpretability through their conversational interface. However, their performance is less consistent than that of vision transformers. Combining the strengths of both approaches – the holistic pattern recognition of vision transformers with the natural language understanding of vision language models – achieves the best of both worlds. Even a relatively simple neural network ensemble delivers consistent top performance across all marketing applications we study. We discuss these findings in terms of practical implementation considerations to help marketing researchers choose the right approach for their specific needs.

The remainder of this paper is structured as follows. First, we systematically review existing marketing applications of image classification and then discuss the conceptual differences between the image classification paradigms. In our empirical assessment, we compile datasets, tasks, training data sample sizes, and number of classes that collectively reflect the range of marketing applications we can so far observe. We test the highest performing and most popular exemplars of CNNs and vision transformer and compare these to vision

language models and an ensemble. We study classification accuracy as the most prevalent performance metric and compare accuracy within and across applications. We conclude with practical considerations and specific recommendations to help marketing scholars choose the right image classification paradigm.

I. IMAGE CLASSIFICATION IN MARKETING RESEARCH

How does marketing research apply automated image classification? To obtain a comprehensive overview of peer-reviewed applications, we systematically analyzed the full text of all articles published between 2012 and October 2024 in all marketing-related journals included in all major journal rankings, reviewing nearly 25,000 articles in a semi-automated process.[1] This surfaced a total of 45 peer-reviewed marketing publications that apply automated image classification (see Web Appendix A for a full overview of all image classification applications in marketing). Overall, publications span five distinct research domains with different research questions and fundamental variations in execution and data characteristics.

First, marketing research studies *advertising & promotional effectiveness*. It leverages image classification to understand what attributes of visual content drive engagement. For instance, Hartmann et al. (2021) investigate the impact of *what* is visible in brand-related social media images by training VGG-16 (a CNN) on 19,949 manually annotated images to classify the image perspective (ego vs. third person). They apply 5-fold cross-validation and achieve above 90% accuracy allowing them to study a large image dataset. They find that ego perspectives, i.e., brand selfies, drive higher brand engagement than facial third-person images, i.e., consumer selfies. Other research investigates the relationship of *promotional effectiveness* and *who* is displayed in an image. For example, Hu et al. (2019) train a model from scratch based on a large dataset of 26,523 images and find that the inclusion of a face drives perceived deal quality of promotional images.

[1] FT50, UT Dallas, ABCD, Cnrs, EJL, ABS, Den, Scimago

Second, marketing research on *consumer insights & behavior* also builds on image classification of who appears on an image. For instance, Chuah and Yu (2021) study how a broader range of eight emotions expressed by different humanoid robots (e.g., anger, contempt, fear, surprise) relate to social media sentiment with a small training sample of only 223 images. Others study the impact of expected viewer judgments and *how* images are perceived. For example, Troncoso and Luo (2023) classify perceived job fit based on freelancer images that they classify with multiple CNN methods (Inception, VGG, ResNet). They find that better subjective fit of personal images drives hiring outcomes.

How consumers perceive images is also often studied in a third group of research centered around *brand management*. For instance, Liu et al. (2020) fine-tune a CNN to determine whether brands are perceived as glamorous, fun, healthy, or rugged based on user-generated image content on social media (Flickr). They find that such 'how' classifications provide reliable inferences about consumer brand perceptions.

As in the other three research domains, VGG is also a popular method choice in *product design & management*. This includes predictions of perceived aesthetics of car designs in Burnap et al. (2023) (i.e., how designs are perceived) that allow marketers to identify promising designs produced by a generative AI. Since few actual car designs are available for training a prediction model and many ratings per design are needed to compute meaningful averages, training data only contains 203 images in this case. Others, extract many image classes. For example, Dzyabura et al. (2023) classify fashion retailing images into 15 categories and find that product return rates are a function of image content.

Fifth and similar to research on advertising and promotional effectiveness, research on *media management* applies image classification to extract 'who' or 'what' objects are visible on images. For instance, Zhang et al. (2024) study Airbnb booking rates as a function of host's facial expressions based on a binary image-related task, whether host images feature an engaging smile. For this purpose, they collect a sizeable training dataset of 9,000 images to fine-tune a ResNet model, finding that host smiles drive booking rates by 3.5% on average.

Similarly, He et al. (2023) identify whether a bed- or living room is displayed in Airbnb photos. Results indicate that showcasing a living room in the background image of the listing is increasing booking rates.

In summary, these diverse marketing applications in these five areas differ in terms of the image-related task (what is visible, who is visible, and how the image is perceived), the size of the available annotated training data (ranging from a few hundred to many thousand observations), the classification method that is applied (training models from scratch or fine-tuning CNN models like ResNet, VGG, Inception), and the number of classes that they study (ranging from binary classifications to more than 99 classes; see Web Appendix A for a full overview). They also typically include various steps of hyperparameter selection. The most prevalent performance metric is accuracy, i.e., the share of classifications that is assigned correctly as evidenced by a validation dataset or n-fold cross-validation. Table 1 summarizes the scope of image classification in marketing. We will follow these types of applications, datasets, and tasks in our empirical comparison to obtain relevant insights on how CNNs compare to more recent alternatives in typical marketing research settings. Next, we will discuss these alternative image classification architectures in more detail.

II. PARADIGMS OF IMAGE CLASSIFICATION

Three main paradigms of image classification have emerged. While nearly all 45 existing peer-reviewed publications in marketing have relied on CNNs, two language-inspired models exist that follow very different paradigms: vision transformers and vision language models. Figure 2 summarizes these paradigms and highlights their main differences.

Convolutional Neural Networks

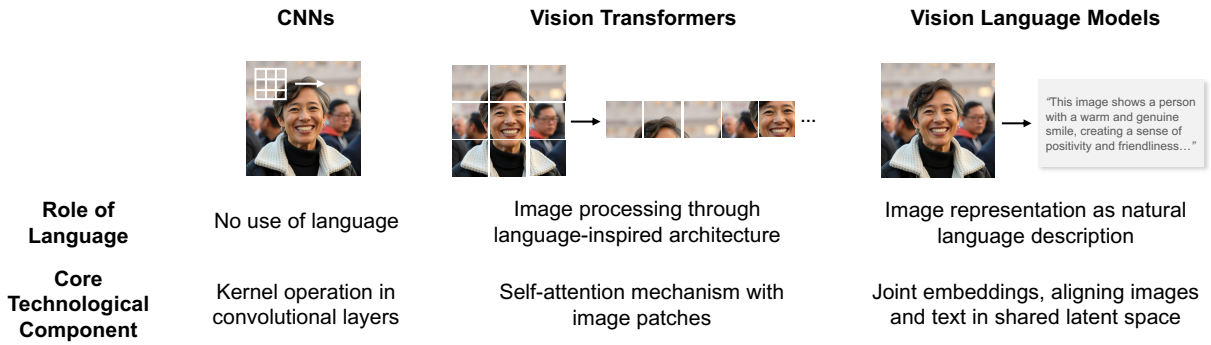
CNNs have the longest history in image classification. LeCun et al. (1998) introduced the idea to employ a sequence of kernel operations across convolutional neural network layers.

Table 1: Core categories for image classification and current marketing practices

Category	Current Marketing Practice
Research Domains & DVs	• *Advertising & promotional effectiveness:* Ad effectiveness, CTR, purchase intention, sales (e.g., Feng et al. 2021; Hartmann et al. 2021) • *Brand management:* Brand perception, hiring outcome (e.g., Klostermann et al. 2018; Lee 2021) • *Consumer insights & behavior:* Emotional response, engagement, viewer reaction (e.g., Peng et al. 2020; Chuah and Yu 2021; Zhang et al. 2023) • *Media management:* Booking rate, intention to watch, popularity, sales (e.g., Liu et al. 2018; Schwenzow et al. 2021) • *Product design & management:* Aesthetic appeal, crowdfunding success, product rating, return rates, sales (e.g., Burnap et al. 2023; Dzyabura et al. 2023)
Image-related Tasks	• *What* is on the image (e.g., Hartmann et al. 2021; Overgoor et al. 2022; Zhang and Luo 2022) • *Who* is on the image (e.g., Zhang, Mehta, Singh and Srinivasan 2021; Bharadwaj et al. 2022) • *How* is the image perceived (e.g., Liu et al. 2020; Lee 2021)
Dataset Characteristics	• *Number of training images:* Ranging from 203 to $> 500,000$ (Burnap et al. 2023; Zhang, Mehta, Singh and Srinivasan 2021) • *Number of classes:* Ranging from 2 to > 99 (Chang et al. 2023; Zhang et al. 2023)
Methods	• *Full-NN-Training:* e.g., 4-layer CNN, 7-layer CNN, trained from scratch (Hu et al. 2019; Peng et al. 2020) • *Transfer learning:* e.g., VGG, ResNet, pretrained on ImageNet (Dzyabura et al. 2023; Hartmann et al. 2021)
Method Training	• *Hyperparameter selection:* e.g., batch size, epochs, learning rate (Liu et al. 2020; Hartmann et al. 2021) • *Evaluation:* e.g., train-validation-test split, 5-fold cross validation (Liu et al. 2020; Hartmann et al. 2021) • *Metrics:* e.g., accuracy, AUC-ROC (Burnap et al. 2023; Zhang, Mehta, Singh and Srinivasan 2021; Troncoso and Luo 2023)

Notes: Analysis based on 45 publications in marketing-related journals (see Web Appendix A).

Figure 2: Overview of image classification paradigms



Each kernel slides across the input image, computing weighted sums of pixel values within a small local region at each position (e.g., a 3×3-kernel as indicated in Figure 2). This sliding kernel operation makes CNNs translation invariant – meaning they can detect patterns regardless of where they appear in the image – since the same kernel weights are applied at every spatial position. Using these shared weights across positions also makes CNNs much more parameter efficient than fully-connected networks that would require separate weights for each pixel position.

Lower convolutional layers are more directly related to (groups of) image pixels and learn low-level features in close proximity (e.g., characteristic patterns, edges of pixels). Higher-level layers then combine these low-level features to identify more abstract patterns (e.g., the smile of a person) (Chollet 2021). Based on a classification head, the CNN can identify which combination of low- and high-level features is most indicative for an image-related classification tasks (see Web Appendix B for more details). For example, facial features like a smile can be particularly useful to automatically assign expressed emotions to images. CNN models can learn such features in different image positions and orientations, as well as in different contexts. This makes them well suited to capture all smiles, no matter where and how they appear. This hierarchical feature extraction approach is likely well-suited when contextual visual information surrounding a target object is less important for a correct assessment. For example, when studying which room an Airbnb ad displays, individual

image features like a sink or a couch can provide reliable indications. However, classifying more nuanced emotions like compassion related to a complex group setting could benefit from taking more contextual detail into account.

Vision Transformers

Vaswani et al. (2017) introduced the transformer architecture based on self-attention mechanisms, which enable it to learn which combination of words in a longer sequence are most relevant to each other when processing sequential data. Unlike CNNs that analyze information through local receptive fields, transformers can process relationships between elements regardless of their distance in the input sequence. This capability is important in natural language processing. For example, to understand the sentence "I threw the ball to her", the model must connect the distant words describing who performed the action ("I") and who received it ("her"). Through self-attention, transformers learn which parts of an input sequence to emphasize or to ignore, enabling them to capture these long-range dependencies between features.

Dosovitskiy et al. (2020) propose to translate this language processing capability to image analysis. Their vision transformer treats images analogously to sentences – just as transformers process sequences of words, vision transformers process sequences of image patches, learning global dependencies in images (Dosovitskiy et al. 2020). Conceptually, this language-inspired approach promises to be particularly powerful for complex image classification tasks that require understanding relationships between distant objects or identifying patterns across the entire scene. For example, when studying brand personality based on user-generated image content, associations such as glamorous or fun might be of interest. Small reconfigurations of similar image objects can shift such nuanced associations. While a person wearing expensive jewelry can signal glamour, a person with a baby which wears the same jewelry might signal fun. Since such nuanced relations are of interest to many marketing applications, transformer models might yield better results than convolutional

architectures. However, the actual size of these possible benefits and whether these warrant transitioning to transformer methods is an open question that we address in this research.

Vision Language Models

Vision language models have recently attracted much attention due to their ability to process meaningful multi-modal data, i.e., providing detailed textual interpretations based on input images and text prompts (Bordes et al. 2024). These models forge a novel link between visual understanding and textual interpretation by mapping images to language. While their potential for image classification has yet to be explored, their ability to create detailed text descriptions offers new data and new classification potential. Specifically, these text descriptions can then be used as an input for a final classification layer, offering unique properties that are different from the image embedding representation of CNNs and vision transformers.

Unlike the pixel-level features used by CNNs and vision transformers, vision language models can better capture higher-order semantic meaning through their integration of visual and textual information. Marketing researchers can build on models that have been trained on vast amounts of paired visual-text data to deliver plausible and nuanced image interpretations that might grasp subtle contextual differences not attainable with the other architectures. When extracting nuanced emotions from sparse data, plausible image narratives can be more informative than the analysis of image features or relations between them. For example, the vision language model Phi-3 interprets panel III in Figure 1 as "the sitting person [...] appears possibly frustrated.". In contrast, a fine-tuned CNN (ResNet) simply predicts the emotion of the entire image as "joy".

Similar insights are more challenging to attain with CNNs and vision transformers. Accordingly, research has criticized them as black boxes (e.g., Rai 2020; Liu et al. 2023) requiring complex post-hoc interpretation algorithms, such as GradCAM (Selvaraju et al. 2020; Hartmann et al. 2021). Vision language models offer inherent interpretability through the

textual outputs.

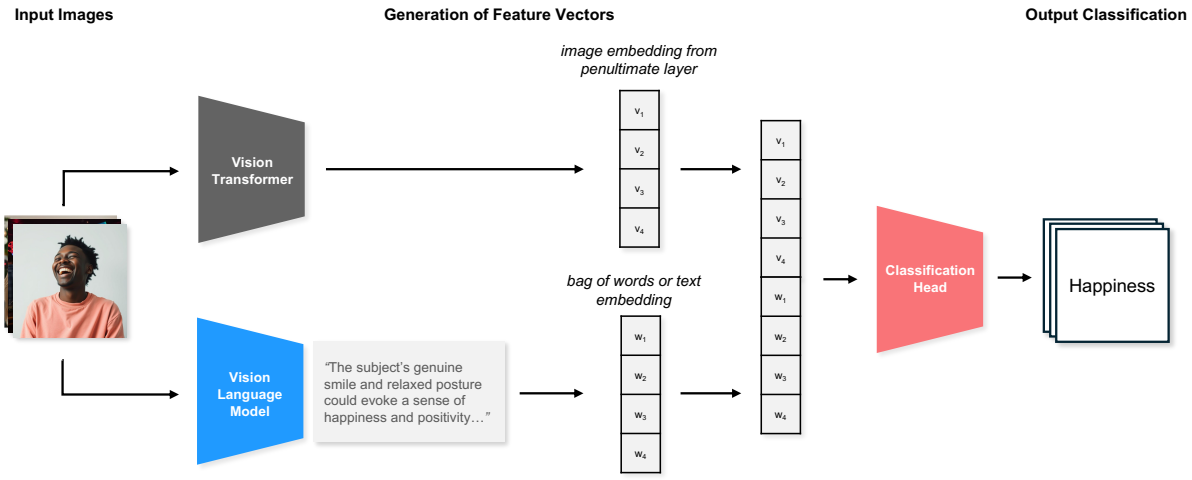
In terms of image classification, the dimensionality reduction of translating images into text could make vision language models achieve a performance that matches and perhaps exceeds current state-of-the-art image classification methods. Detailed text interpretations can be obtained from general-purpose models without fine-tuning or model training. This allows researchers to train models exclusively on text output, which operates on much lower dimensionality than high-dimensional image data. These capabilities can make vision language models particularly appealing for marketing questions centering around viewer perceptions, where gathering large annotated training datasets from multiple respondents from specific target groups is costly. Vision transformers and CNNs, on the other hand, often lack pretraining on the image-related tasks marketing is interested in. For example, for individual brands, product categories or research objectives, different taxonomies of brand personality can be of interest. This requires extensive fine-tuning and retraining for each brand personality study, including costly collection of labelled training data. Image classification with vision language models can build on text analysis and might simplify such fine-tuning.

Multi-Paradigm Ensemble

While vision language models offer unique advantages through their textual interpretations, relying solely on these representations may discard valuable pixel-level information. Since existing benchmarks suggest superior performance of vision transformers for objective 'what' image-related tasks, combining these models with vision language models is a natural extension. Such an ensemble of both model paradigms (i.e., vision transformers and vision language models) could allow a classification model to leverage both local feature extraction and the text descriptions of more complex contextual relationships.

Figure 3 illustrates the inference pass of such an ensemble architecture. The ensemble consists of a classification head trained on concatenated feature vectors of both models. For the vision transformer component, one can extract the high-dimensional embedding vector

Figure 3: Exemplary inference pass of the ensemble architecture



from the penultimate layer. The vision language model component generates a textual interpretation of the image, which can be converted into a numerical representation using bag-of-words methodology before feeding it to the classification head[2]. Note that such an ensemble requires fine-tuning at the level of the vision transformer as well as training the classification head. To avoid data leakage, it is advisable to employ different training datasets for both objectives and exclude both of them from the out-of-sample predictions for model assessment (see Web Appendix C for further details on the ensemble architecture and training process).

Implications for Different Image Classification Tasks

The fundamental architectural distinctions between CNNs, vision transformers, and vision language models lead to distinct approaches in processing visual information. Depending the selected image-related task of interest as well as dataset characteristics, these architectural properties can become helpful or detrimental to the classifiers' performance. As illustrated in Figure 1, image classification tasks can vary significantly in complexity. While CNNs

[2]Using embeddings from OpenAI (text-embedding-3-small) instead of the raw bag-of-words texts resulted in lower overall performance in our experiments.

likely excel at tasks requiring local feature detection in unambiguous images through their convolutional operations, more complex tasks may benefit from vision transformers' ability to capture contextual relationships between objects or vision language models' capacity to interpret global narratives based on different visual focal points.

When choosing an image classification model, the actual consequences of these conceptually different methodological choices remain an open question. Next, we will assess a broad range of marketing-relevant datasets to understand benchmark accuracy values and provide empirical evidence about what works when.

III. EMPIRICAL ANALYSIS OF IMAGE CLASSIFICATION METHODS

Datasets and Methodology

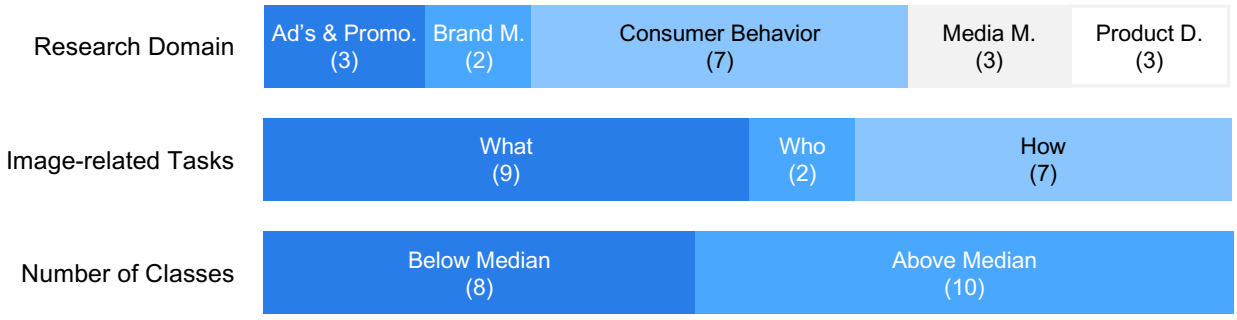
To systematically evaluate how different image classification paradigms perform across marketing applications, we compile 18 unique datasets spanning the types of questions, image-related tasks, and numbers of classes that have so far been studied in marketing research. We obtain publicly available datasets from marketing publications listed in Web Appendix D (e.g., Wang et al. 2020; Hartmann and Exner 2024) and complement these by public image datasets available on Kaggle that mirror typical image-related classification tasks in marketing.

Figure 4 summarizes the number of datasets related to each research domain, image-related tasks, as well as the number of classes. The datasets we were able to identify feature more image-related tasks containing 'what' and 'how' questions rather than 'who' questions. Most datasets also originate from consumer behavior research. Both reflect current publications in marketing. More importantly, the total compilation of all datasets contains representative examples of each research question and image-related tasks allowing us to assess the stability of results across applications.

In terms of the number of classes, we split datasets into a few (below median, i.e., less than 6) vs. many (above median, 6 or more) number of classes to gain an intuition about the impact of the training data size. Full information about each individual dataset and its characteristics is available in Web Appendix D and Web Appendix E.

According to section I, marketing research often uses several hundred to several thousand manually annotated training images for fine-tuning and training classification models. Since collecting training data can be costly, in particular when multiple labels are needed, we investigate the role of the training data size. For each dataset, we randomly sample 1,000 images per dataset (large training data) and also limit training data to only 200 images per dataset (limited training data).

Figure 4: Overview of datasets in our empirical analysis



Notes: The number of datasets is displayed in parentheses.

We compare representative methods from the three key paradigms of image classification. Specifically, we include all CNN architectures previously used in peer-reviewed marketing research, i.e., InceptionV3 (Szegedy et al. 2015), ResNet152 (He et al. 2015), VGG19 (Simonyan and Zisserman 2015), and Xception (Chollet 2017) as well as a CNN trained from scratch. For architectures with multiple versions (e.g., VGG16 and VGG19), we use the newest and largest variant to ensure fair comparison (see Web Appendix B for a more detailed description of all methods).

To test how language-inspired architectures perform, we include the original vision transformer (i.e., ViT; Dosovitskiy et al. 2020) that pioneered treating image patches as words and ConvNeXt (Liu et al. 2022). ConvNeXt builds upon the ResNet, but segments images

into patches and organizes them into sequences, allowing the model to process images in a manner analogous to how sentences are handled in transformer architectures. We add ConvNeXt because it has achieved good performance in benchmarks for other applications (Maurício et al. 2023). For a fair comparison across paradigms, we use ImageNet pre-training for both CNNs and vision transformers, following typical practice in marketing.

To evaluate the potential of vision language models for image classification, we include Microsoft's Phi-3 (Abdin et al. 2024), an open-source model that can provide detailed textual descriptions of images.[3] We use the vision language model to generate textual descriptions, which are subsequently transformed into numerical representations using the bag-of-words methodology. These representations serve as an input to a two-layer fully-connected neural network, which is trained separately for each dataset (Chollet 2021).

Lastly, we implement a multi-paradigm ensemble combining ConvNeXt's visual features with Phi-3's cross-modal understanding, as described in section II. We investigate this specific combination because ConvNeXt builds on ResNet and thereby includes conceptual advantages of both CNNs and vision transformers, while Phi is extending the model by an entirely different text-based approach. As illustrated in Figure 3, we extract the image embeddings of the ConvNeXt model and combine these with the bag-of-words representations of the Phi-generated text descriptions. The concatenated features are fed into a two-layer fully-connected neural network that serves as the classification head of the ensemble (Chollet 2021).

We conduct hyperparameter optimization for each method, dataset, and training data size combination to take the fact into account, that optimal settings can vary across paradigms and tasks. This results in a total of 3,150 experiments (18 datasets $\times$ 2 large vs. limited training data sizes $\times$ 9 methods $\times$ 10 hyperparameter optimization runs[4]) (see Web Ap-

[3]We deliberately used an open-source model instead of a closed-source model such as GPT-4o to ensure replicability and cost-efficient accessibility.

[4]The AIDA funnel car dataset contains less than 1,000 images and is therefore excluded from the large training data availability condition. This reduces the total number by 9 * 10 = 90 experiments.

pendix C and F for a full overview of all implementations in our empirical analysis). We focus on accuracy as a performance metric because it is the most widely reported metric across marketing publications. Since manually annotated observations are limited in some cases, we implement 10-fold cross-validation to provide a robust measure of model performance while reducing the risk of overfitting to a single validation dataset. Specifically, we split the training data into 10 equal-size subsets. We then run 10 iterations, training the model on 90% of the data (nine folds) and validating on the remaining 10% (one fold), ensuring each fold serves as the validation set exactly once. The average accuracy score across all 10 validation sets is the primary measure of performance across methods, tasks, and datasets that we report next.

Accuracy of Image Classification Paradigms

Average Performance Across Datasets

We first compute global accuracy scores across all datasets and tasks for both the large and limited training data scenarios. According to Figure 5 average accuracy differs strongly between traditional CNNs, vision transformers, and vision language models. Most notably, language-inspired vision transformers and the ensemble demonstrate superior performance for both data-rich (1,000 training data; left panel) and data-scarce applications (200 training data; right panel).

While the overall ranking of methods is similar, there are also important differences between data-rich and data-scarce applications. When large training data is available, differences between paradigms (ensemble, vision transformers, CNNs, and training from scratch) are more pronounced than differences within each paradigm. Specifically, which CNN or vision transformer is chosen has relatively small consequences on inferences of only a few percentage points. Conversely, differences between paradigms exceed double-digits in percent accuracy. This is similar to research on text classification, where architectural rather

than incremental innovations are the main performance driver (Hartmann et al. 2023).

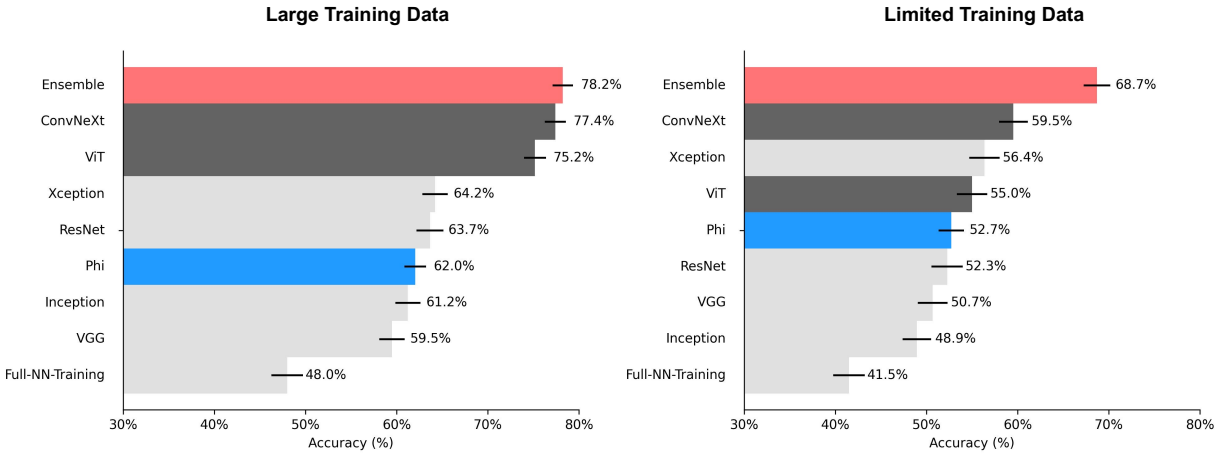
With large training data, the vision transformer ConvNeXt model emerges as the best individual method (77.4% accuracy), substantially outperforming all traditional CNNs. Among CNNs, Xception achieves the highest accuracy (64.2%). Full training of neural networks from scratch assigns more than 50% of all validation images to the wrong categories. On average and across all applications, the vision language model Phi has only moderate performance (62.0%) when used individually. However, the method ensemble of Phi and ConvNeXt achieves the best performance (78.2%) across all methods. It also exceeds the accuracy of ConvNeXt (77.4%) on its own, although the difference is small.

For limited training data, differences within paradigms, i.e., within the vision transformer models and within the CNN models, are larger, suggesting incremental differences between models can make a greater difference. As a consequence, the best performing CNN Xception even outperforms the lowest performing vision transformer ViT (56.4% vs. 55.0%). Despite the limitation to only 200 training images, ConvNeXt still classifies about 60% of all validation images correctly, making it the best individual method, also for limited training data. Conversely, lack of training data results in about 60% false labels for neural network training from scratch. Interestingly, while Phi again performs about average with limited training data, the difference between Phi and ViT is much smaller (52.7% vs. 55.0%). More importantly, combining Phi with ConvNeXt comes with a much larger benefit than we observed for large training data. For limited training data, the ensemble makes about 10% fewer errors than the next best model, ConvNeXt (68.7% vs. 59.5%). This suggests that Phi can extract relevant information that is otherwise not considered by the other visual classifiers. Specifically, combining different aspects of language-inspired processing – treating images as sequences of patches and generating natural language descriptions – appears to create complementary benefits.

Overall, the vision transformer ConvNeXt provides a sizeable improvement in attainable accuracy when compared to CNNs such as VGG that was used most frequently in prior

marketing research (difference about 18% for large and 11% for small training data). This suggests transitioning to transformers can have a sizeable impact on conclusions. Beyond that, the effort involved in training an ensemble pays off most strongly when training data is limited.

Figure 5: Average accuracy across all datasets across methods



Notes: Accuracy distribution for each method based on 10-fold cross-validation of best performing hyperparameter settings. Performances of CNNs are displayed in light gray, vision transformers in dark gray, vision language model in blue, and the multi-paradigm ensemble in red color.

Performance on Dataset Level

In contrast to other image benchmark studies that are often limited to a few illustrative datasets (e.g., Guo et al. 2017), the previous results were based on 18 different datasets with actual applications in marketing. While these average values provide reasonable expectations, it is not clear whether performance generalizes across individual applications. For example, the ensemble achieves the highest relative performance on average. However, individual applications might still fall well below the average accuracy of 78.2 or 68.7 percentage points. Therefore, we compare method performance across all individual datasets in Figure 6. To facilitate the interpretation of performance consistency, Figure 6 orders datasets by decreasing accuracy and groups each method paradigm following the color scheme of the aggregate results.

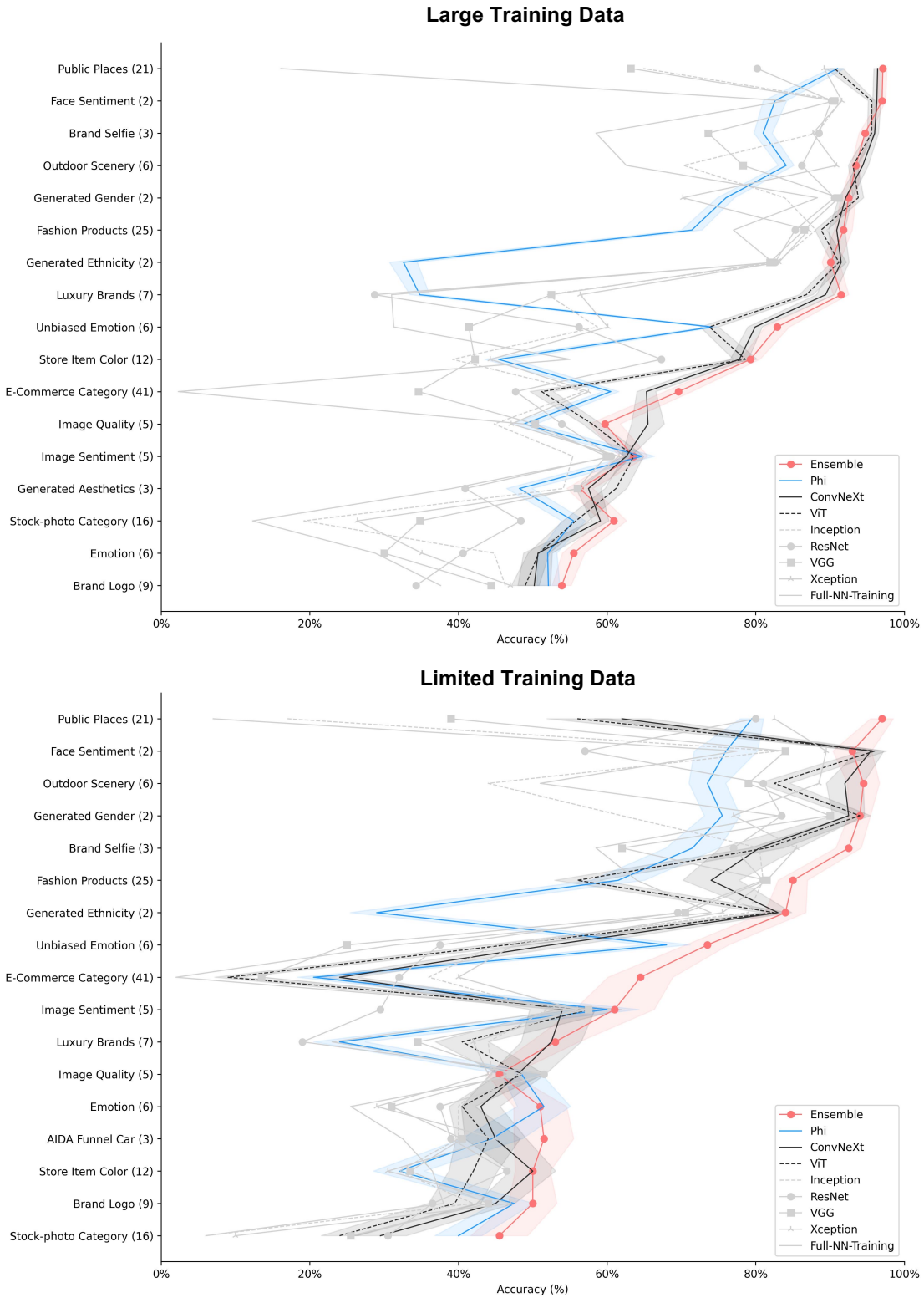
This comparison across datasets and paradigms results in several noteworthy conclusions. In relative terms, the ensemble of vision transformers and vision language models (red line) results in consistent top performance for both large and limited training data. This suggests that combining transformer-based visual processing with language understanding creates robust performance advantages across diverse marketing applications.

Absolute performance, however, varies strongly. While, for example, positive and negative facial emotions are classified correctly in above 90% of cases, only about 50% of brand logos are correctly identified on image content with the top-performing ensemble method. This suggests 'lighthouse' leaderboard illustrations with isolated applications can be misleading for individual marketing tasks. Even for ensembles, image classification accuracy can vary substantially and needs to be carefully assessed for each image-related task.

Traditional CNNs and training from scratch (light gray lines) show highly variable performance across datasets for both large and limited training data. For instance, ResNet's accuracy fluctuates by over 30 percentage points between different datasets and tasks under both data conditions. This volatility means marketing researchers cannot form reasonable expectations about the performance of CNNs. Consequently, comparing CNNs to alternative paradigms is highly recommended to ensure robust and informed decision-making.

For vision transformers (dark gray lines) relative performance variability is a function of the available training data. As we saw in the aggregate analysis, with few exceptions they perform en par and sometimes slightly better than the ensemble. However, performance varies strongly with limited training data, where ensemble performance is much less affected by differences between image-related tasks. Both aggregate performance and consistency of performance make the ensemble method an especially attractive choice when training data is limited – a common scenario in marketing applications.

Figure 6: Method performance across datasets



Notes: Number of classes is provided in parentheses to the right of the dataset name. Line shade indicates standard error for top-4 methods.

Interestingly, the vision language model Phi (blue line) can provide competitive performances for a few datasets – most importantly classifications of emotions and image sentiment. Performance is particularly strong for small datasets. This can be due to its large-scale pre-training on vast, multi-modal data, as well as its zero-shot natural language descriptions of the images that require no task-specific training data. In addition, many datasets for which the vision language model achieves a competitive performance contain a relatively large number of classes (e.g., 21, 16). On the other hand, performance of Phi alone varies most strongly. Frequently, it fails to match the accuracy of vision transformers and, in some cases, even underperforms CNNs, achieving correct label assignments in less than 30% of the samples – for instance, in the classification of products into e-commerce categories. This suggests the intuitive appeal and ease of application of vision language models should not be mistaken for consistent or reliable classification performance.

Next, we will investigate when each method works best by conducting an analysis of method performance depending on the number of classes and group of image-related tasks next.

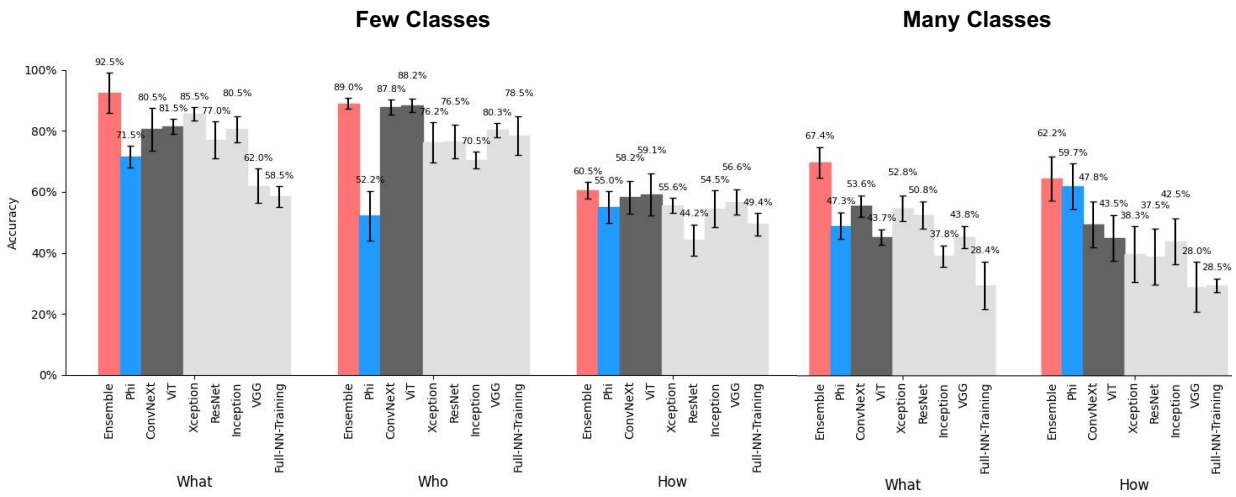
Performance Across Image-related Tasks and Number of Classes

According to the previous analysis, both ensembles and vision transformers result in similar relative performance for large training data. For larger training data, performance is consistently above and in a few cases en par with CNNs and vision language models on their own. For limited training data, performance is less consistent and therefore more interesting to investigate in more detail.

Therefore, Figure 7 displays the performance of each method trained on 200 samples split by image-related tasks and few. vs. many classes (above and below median) (see Web Appendix G for the same analysis for large training data). Vision transformers and stronger CNNs, i.e., ConvNeXt, ViT, ResNet, and Xception, achieve high accuracy values close to 80 percent for datasets containing an image-related task resembling a 'what' or 'who' question

and less than six classes, reflecting their strong capabilities for basic object and person recognition. These results seem reasonable, taking in consideration that these methods have been developed and tested with objective image-related tasks in mind (e.g., classification of individuals and objects shown on an image). In addition, the pre-training on the ImageNet dataset, which is a multi-object classification task, makes these models well suited to address 'what' and 'who' tasks. On their own, simple to apply vision language models do not represent a valid alternative for these objectives. This is different for perceptual 'how' tasks, where vision language models attain comparable performance. For many classes, ensembles perform best or en par across all three image-related tasks.

Figure 7: Accuracy depending on image-related task and number of classes (limited training data)



Notes: Error bars indicate standard error. Analysis of many classes limited to 'what' and 'how' tasks, since datasets with many classes and 'who' tasks were not available.

However, conclusions differ fundamentally when task complexity increases. As expected, classification accuracy declines with more classes. However, this effect is strongest for CNNs and vision transformers. For many classes, the ensemble outperforms both vision transformers and CNNs by more than 10% making the effort involved in ensemble training particularly useful when many classes shall be assigned. Interestingly, Phi on its own achieves comparable performance as the ensemble for 'how' questions with many classes. It also assigns 10%

more labels correctly than vision transformers and CNNs. Pre-trained knowledge about semantic relationships between visual concepts and natural language descriptions may have contributed to performance for these types of marketing applications.

These results suggest that language-inspired approaches – whether through ensembles or vision language models – offer particular advantages for the complex perceptual tasks common in marketing research, especially when faced with limited training data and multiple classification categories.

IV. DISCUSSION

Visual content is rapidly proliferating and increasing in importance for marketing research and practice. Some argue that computer vision enables a new frontier in marketing that facilitates a deeper understanding of visual communication and its effects (Grewal et al. 2021). While a plethora of image classification methods are available, marketing has lacked comprehensive guidance on their relative performance.

This research has provided an overview of automated image classification practices in marketing research. Across publications, it became evident that marketing research has primarily relied on CNNs that process images by analyzing local patterns of neighboring pixels. Marketing has not yet leveraged more recent paradigms that process images more like language – either by treating image patches as words in a sentence (vision transformers) or by generating natural language descriptions of image content (vision language models).

We conducted a large-scale empirical comparison of these three paradigms – CNNs, vision transformers, and vision language models – performing over 3,100 experiments on 18 distinct datasets that reflect typical marketing applications. Our results demonstrate clear benefits of language-inspired approaches. On average over all datasets, vision transformers that process image patches like words achieve accuracy improvements of more than 10 percentage points compared to CNNs by better capturing relationships between distant image regions. Vision language models perform less consistently. They show particular promise for complex

perceptual tasks with limited training data by leveraging their ability to interpret images through natural language. Most notably, combining both language-inspired paradigms in an ensemble that leverages both patch-based processing and natural language understanding achieves the highest and also remarkably most consistent relative performance across nearly all experiments.

These findings suggest that marketing researchers can significantly benefit from moving beyond traditional CNNs to embrace language-inspired approaches. Processing images more like language appears particularly valuable for more nuanced perceptual tasks common in marketing research, especially in data-scarce contexts with many classes. Our results provide clear guidance for selecting and implementing these more advanced approaches while maintaining practical applicability.

Choosing the right method is important because it helps to avoid inferior method choices that can result in suboptimal potential economic decisions and inflated standard errors, as well as biased estimates in econometric analysis (Hartmann et al. 2019; Jaidka et al. 2020; Frankel et al. 2022). To assist in method choice, Table 2 summarizes the trade-offs involved in choosing a classification method based on the associated average accuracy values that we have observed in this investigation. From left to right, researchers will start with a research question that defines the image-related classification task they are interested in. The collection of manually annotated training images can involve varying levels of effort – collection may be straightforward, such as quickly tagging different image perspectives (e.g., Hartmann et al. 2021), or more demanding, such as gathering multiple consumer responses to evaluate advertising effectiveness (e.g., Jansen et al. 2024). Similarly, the image-related task defines how many classes are of interest, e.g., nuanced emotions as in Chuah and Yu (2021), or binary sentiment as in Zhang et al. (2024). Depending on these decisions related to the research question and the amount of effort researchers can invest in data labelling, different methods perform best. These methods in turn come with different training effort and provide different levels of insight into the underlying classifications.

Table 2: Ranking of method performance to aid marketing researchers in selection process

			Available image classification paradigms with top performing methods				
			Multi-paradigm ensemble	**Vision language models**	**Vision transformers**		**Convolutional neural networks**
			Phi & ConvNeXt	Phi	ConvNeXt	ViT	Xception
Image-related task	*Required training data size*	*Number of classes to classify*	*High effort, low interpretability*	*Low effort, high interpretability*	*Medium effort, low interpretability*	*Medium effort, low interpretability*	*Medium effort, low interpretability*
What	Large	Many	1		2 (-1.8%)	3 (-5.5%)	
		Few	3 (-1.3%)		1	2 (-0.4%)	
	Limited	Many	1		2 (-13.8%)		3 (-14.8%)
		Few	1			3 (-11.0%)	2 (-7.0%)
Who	Large	Few	3 (-1.2%)		2 (-0.7%)	1	
	Limited	Few	1		3 (-1.2%)	2 (-0.8%)	
How	Large	Many	1	3 (-6.2%)	2 (-3.7%)		
		Few	3 (-1.3%)		1	2 (-0.9%)	
	Limited	Many	1	2 (-2.5%)	3 (-14.4%)		
		Few	1		3 (-2.3%)	2 (-1.4%)	

Notes: Difference in accuracy to top-performing method in percentage points in parentheses.

Table 2 provides evidence-based guidance, on the relative accuracy that is attainable for these different considerations. It highlights the relevant trade-offs involved in choosing the best method. For example, some methods perform very similar and differ only on a few percent accuracy. In these cases, even a suboptimal method can represent the better overall trade-off. The overview demonstrates that researchers can achieve strong performance by focusing on just a few key methods. Ensemble application requires fine-tuning of a vision transformer such as ConvNeXt and training of a classification model on top that integrated text and image features. Clearly, this represents the most effort out of all methods we have reviewed. Researchers willing to invest the effort in training the ConvNeXt-Phi ensemble can achieve a consistent high performance for seven out of ten application settings. In the three cases, where the ensemble is not top-ranked, the gap to the best performing method, i.e., ConvNeXt or ViT, remains small (less than 1.3% in all cases). This means focusing exclusively on ensembles comes with little risk in terms of suboptimal accuracy, making them the best choice for researchers that study many different image classification problems

and can leverage their effort of setting up a fine-tuning pipeline once and applying it to many problems.

Some researchers might refrain from choosing multi-paradigm ensembles because of the effort involved and expertise required. Perhaps image classification is only used in a robustness check or to provide complimentary evidence. Whenever image classification is only indirectly related to the main marketing problems, researchers might prefer to conduct simpler fine-tuning. In these cases, collecting training data pays off, in particular for 'what' tasks, where training data is easiest to collect. Whenever large training data is attainable, vision transformers and ConvNeXt in particular perform best and only slightly worse than the multi-paradigm ensemble. When training data is limited, the accuracy costs of not estimating the best possible ensemble model become larger. This is of particular concern for 'how' tasks, where data collection is particularly costly.

Finally, interpretability and diagnostic information about which elements contribute to certain classifications can be of particular interest (e.g., when understanding which image features result in specific emotions). Given the black-box nature of CNN and transformer-based methods, such objectives can be best achieved with Phi alone. While Phi never achieved top accuracy in any of our applications, the accuracy trade-off is smallest for complex scenarios involving many classes and 'how' tasks, especially with limited training data. On their own, they are particularly easy to apply since text output is readily available without training and standard text analysis can be applied for classification (see Hartmann et al. 2019). We expect that this small trade-off in accuracy compared to the ensemble will be attractive for many researchers to better understand the classification task and save effort of multiple fine-tuning exercises.

As vision language models continue to advance rapidly, we expect their performance – and by extension that of ensembles incorporating them – to improve further. At the same time, fine-tuning vision language models is currently challenging for most researchers and will likely remain so due to their more complex nature. This means image objects that

standard models do not know (e.g., logos of unknown brands, novel product designs) will not be included in the text descriptions limiting their application. This is another advantage of the multi-paradigm ensemble, which allows training specific content into the ConvNeXt model while still leveraging the detailed image narrative of vision language models for all remaining content.

Despite the breadth of datasets and methods included in our study, our investigation cannot cover all potential image classifications tasks marketers might become interested in. Other (novel) tasks might result in different levels of accuracy. However, in terms of relative performance, we have found remarkably consistent advantages of the ensemble and vision transformers. While performance assessments are always advisable, it is reasonable to expect similar top performance of the ensemble and vision transformers for other applications as well. Furthermore, this research has revealed predictable and theoretically plausible limitations. Specifically, when training data is limited, conventional CNNs, such as Xception, demonstrate comparable performance, particularly for 'what' tasks—those for which these models were originally designed. Conversely, vision language models should be considered a unified alternative for tasks involving large training datasets and 'how' tasks, which often require the interpretation of more complex content relationships. Since both vision language models and the ensemble have not been applied to image classification and vision transformers are new to marketing, we provide an end-to-end Python script, requiring little to no coding knowledge to leverage state-of-the-art image classification methods [URL BLINDED DURING PEER REVIEW].

None of these implementations require computational resources beyond what has been already applied in marketing. Specifically, VRAM memory can be an important constraint and none of our implementations require more than 16 GB, with the multi-paradigm ensemble requiring only 10 GB. In terms of computational resources, the different paradigms therefore do not differ in relevant ways. Only the ensemble requires training two models, which

increases the computational effort.[5] For example, fine-tuning ConvNeXt on the emotion dataset comprising six classes, with 900 training images and a batch size of 16, required an average of 9.18 minutes per fold when performed on a single NVIDIA A6000 RTX GPU with 48 GB of VRAM. Since these computing resources are easily available on cloud services, we consider these as well as the shorter computational times of the individual methods not a relevant concern for most practical and scientific applications.

In this research, we optimized learning rate, batch size, and number of epochs, the three most commonly adapted hyperparameters in the application of image classification. Our results are based on these optimal decisions and should be interpreted with this consideration in mind. Whenever manual hyperparameter tuning seems not feasible, we recommend using such automated procedures (e.g., Akiba et al. 2019). How to arrive at these values is documented in the supplementary code to this article.

During our empirical analysis, we implemented the ensemble using a two-step approach (i.e, first generating textual descriptions, extracting bag-of-words vectors, and training a neural network classification head). Arguably, this represents a relatively simple approach. We have experimented with more sophisticated language modeling and text embeddings, which have not improved performance. However, there is much potential in additional fine-tuning of vision language models and testing alternative prompting strategies. For example, vision language models could be used to generate a classification directly by prompting the model to output the most probable class given a pre-determined set of classes. We tested this approach anecdotally using Llama 3.2-Vision, a state-of-the-art open-source vision language model, on two datasets. Compared to our two-step approach of classification, the model did not reach the performance of lowest performing full neural network training. This suggests that while direct classification through prompting is theoretically appealing, these models

[5]To reduce the time required for training and inference of image classification models, researchers can use parallel computations using multiple GPUs. This approach is particularly advantageous for large-scale models and datasets, as it allows computational workloads to be distributed across GPUs (see Dehghani et al. 2023; Sreedhar et al. 2024 for information on applying parallel computations).

still benefit from the more structured two-step methods. However, various prompting and fine-tuning strategies could be tested that go beyond the scope of this investigation.

An important caveat when employing vision language model-based approaches to classification are tasks that could be related to societal bias. For example, in our study, descriptions and classifications of a person's ethnicity were more difficult to obtain, as some methods have barriers in place to prevent them from making judgements that could further societal bias (e.g., gender, age, ethnicity, beauty; Feuerriegel et al. 2024). This can make these models very sensitive to differences in prompts that researchers should carefully evaluate.

Beyond classification tasks, many marketing applications are also interested in object detection (e.g., logos and other marketing assets) and in studying the impact of presence, size, and location. On the surface, such tasks seem similar, as they also make predictions based on images. However, it requires different methods and training data and cannot be directly compared to the full image classification tasks that we have studied in this research. For example, Nanne et al. (2020) compare open-source methods, such as YOLO (Redmon et al. 2016), and commercial solutions, such as Clarifai, to understand what works best for object detection. While such comparisons provide valuable insights, the rapid pace of development in vision language models suggests exciting new possibilities. In the future, vision language models might address these tasks in a zero-shot manner simply by prompting them "What can you see on this image? Please provide the coordinates of the bounding box including the focal object."

Looking further ahead, future research can also explore vision language models for multi-image and video scenarios that recent technological advances enable (Cascio Rizzo et al. 2024; Li et al. 2024). As videos are essentially a sequence of images, similar methods can be applied, but additional considerations play a role (Li et al. 2019; Schwenzow et al. 2021). This expansion into video analysis represents a natural evolution of our findings, as the language-inspired approaches that proved valuable for single images may be even more powerful for understanding temporal sequences of visual content.

While new methods in computer vision will certainly emerge, many findings of our analysis will likely remain relevant. Specifically: (1) ensemble methods combining multiple paradigms help with performance across many use cases, (2) relative performance of vision transformers and the ensemble is more robust than that of CNNs or training from scratch, (3) compared to alternatives, readily interpretable vision language model classifications are most promising for datasets with limited training data and many classes.

Our results have revealed an encouraging message for marketing researchers: useful methods for various interesting image-related tasks are available and can be applied relatively easily. The emergence of language-inspired approaches has made sophisticated image analysis more accessible while often improving performance. Using these technologies will further expand our understanding of how image communication drives consumer behavior and market response. We hope that our findings encourage further marketing research to leverage the exciting potential associated with large-scale image data.

References

Abdin, Marah, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, ... and Xiren Zhou (2024), "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone," arXiv:2404.14219 [cs] http://arxiv.org/abs/2404.14219.

Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta and Masanori Koyama (2019), "Optuna: A Next-generation Hyperparameter Optimization Framework," https://arxiv.org/abs/1907.10902.

Alom, Md Zahangir, Tarek M. Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C. Van Esesn, Abdul A. S. Awwal and Vijayan K. Asari (2018), "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches," *arXiv:1803.01164 [cs]*, http://arxiv.org/abs/1803.01164.

Anand, Piyush and Vrinda Kadiyali (2024), "Frontiers: Smoke and Mirrors: Impact of E-cigarette Taxes on Underage Social Media Posting," *Marketing Science*, 43 (3), 479–487 https://pubsonline.informs.org/doi/10.1287/mksc.2022.0241.

Beichert, Maximilian, Andreas Bayerl, Jacob Goldenberg and Andreas Lanz (2023), "Revenue Generation Through Influencer Marketing," *Journal of Marketing*, 88 (4), 40–63 https://journals.sagepub.com/doi/10.1177/00222429231217471.

Bharadwaj, Neeraj, Michel Ballings, Prasad A. Naik, Miller Moore and Mustafa Murat Arat (2022), "A New Livestream Retail Analytics Framework to Assess the Sales Impact of Emotional Displays," *Journal of Marketing*, 86 (1), 27–47 https://journals.sagepub.com/doi/10.1177/00222429211013042.

Bordes, Florian, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, ... and Vikas Chandra (2024), "An Introduction to Vision-Language Modeling," arXiv:2405.17247 [cs] http://arxiv.org/abs/2405.17247.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, ... and Dario Amodei (2020), "Language Models are Few-Shot Learners," arXiv:2005.14165 [cs] http://arxiv.org/abs/2005.14165.

Burnap, Alex, John R. Hauser and Artem Timoshenko (2023), "Product Aesthetic Design: A Machine Learning Augmentation," *Marketing Science*, 42 (6), 1029–1056 https://pubsonline.informs.org/doi/full/10.1287/mksc.2022.1429.

Cascio Rizzo, Giovanni Luca, Jonah A. Berger and Mi Zhou (2024), "How Hand Movement Shapes Communication's Impact," https://papers.ssrn.com/abstract=4962927.

Chang, Hannah H., Anirban Mukherjee and Amitava Chattopadhyay (2023), "More Voices Persuade: The Attentional Benefits of Voice Numerosity," *Journal of Marketing Research*, 60 (4), 687–706 https://journals.sagepub.com/doi/pdf/10.1177/00222437221134115.

Chollet, François (2017), "Xception: Deep Learning with Depthwise Separable Convolutions," *arXiv:1610.02357 [cs]*, http://arxiv.org/abs/1610.02357.

Chollet, François (2021), *Deep Learning with Python* 2nd edn Manning Publications Co. https://www.manning.com/books/deep-learning-with-python-second-edition.

Chuah, Stephanie Hui-Wen and Joanne Yu (2021), "The future of service: The power of emotion in human-robot interaction," *Journal of Retailing and Consumer Services*, 61, 102551 https.//doi.org/10.1016/j.jretconser.2021.102551.

Dehghani, Mostafa, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, ... and Neil Houlsby (2023), "Scaling Vision Transformers to 22 Billion Parameters," arXiv:2302.05442 http://arxiv.org/abs/2302.05442.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei (2009), "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 https://doi.org/10.1109/CVPR.2009.5206848.

Dosovitskiy, A., L. Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, M. Dehghani, Matthias Minderer, G. Heigold, S. Gelly, Jakob Uszkoreit and N. Houlsby (2020), "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv:2010.11929 [cs]*, https://arxiv.org/abs/2010.11929.

Dzyabura, Daria and Renana Peres (2021), "Visual Elicitation of Brand Perception," *Journal of Marketing*, 85 (4), 44–66 http://journals.sagepub.com/doi/10.1177/0022242921996661.

Dzyabura, Daria, Siham El Kihal, John R. Hauser and Marat Ibragimov (2023), "Leveraging the Power of Images in Managing Product Return Rates," *Marketing Science*, 42 (6), 1125–1142 https://pubsonline.informs.org/doi/10.1287/mksc.2023.1451.

Dzyabura, Daria, Siham El Kihal and Renana Peres (2022), "Image Analytics in Marketing," in *Handbook of Market Research*, C.Homburg, M.Klarmann and A.Vomberg, eds Cham Springer International Publishing, 665–692.

Feng, Yang, Huan Chen and Qian Kong (2021), "An expert with whom i can identify: the role of narratives in influencer marketing," *International Journal of Advertising*, 40 (7), 972–993 https://doi.org/10.1080/02650487.2020.1824751.

Feuerriegel, Stefan, Jochen Hartmann, Christian Janiesch and Patrick Zschech (2024), "Generative AI," *Business & Information Systems Engineering*, 66 (1), 111–126 https://doi.org/10.1007/s12599-023-00834-7.

Frankel, Richard, Jared Jennings and Joshua Lee (2022), "Disclosure Sentiment: Machine Learning vs. Dictionary Methods," *Management Science*, 68 (7), 5514–5532 https://doi.org/10.1287/mnsc.2021.4156.

Ghose, Anindya, Panagiotis G. Ipeirotis and Beibei Li (2012), "Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content," *Marketing Science*, 31 (3), 493–520 https://doi.org/10.1287/mksc.1110.0700.

Gorn, Gerald J., Amitava Chattopadhyay, Tracey Yi and Darren W. Dahl (1997), "Effects of Color as an Executional Cue in Advertising: They're in the Shade," *Management Science*, 43 (10), 1387–1400 https://doi.org/10.1287/mnsc.43.10.1387.

Grewal, Rajdeep, Sachin Gupta and Rebecca Hamilton (2021), "Marketing Insights from Multimedia Data: Text, Image, Audio, and Video," *Journal of Marketing Research*, 58 (6), 1025–1033 https://doi.org/10.1177/00222437211054601.

Guan, Yue, Yong Tan, Qiang Wei and Guoqing Chen (2023), "When Images Backfire: The Effect of Customer-Generated Images on Product Rating Dynamics," *Information Systems Research*, 34 (4), 1641–1663 https://doi.org/10.1287/isre.2023.1201.

Gunarathne, Priyanga, Huaxia Rui and Abraham Seidmann (2022), "Racial Bias in Customer Service: Evidence from Twitter," *Information Systems Research*, 33 (1), 43–54 https://doi.org/10.1287/isre.2021.1058.

Guo, Jianyuan, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang and Chang Xu (2022), "CMT: Convolutional Neural Networks Meet Vision Transformers," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* New Orleans, LA, USA IEEE, 12165–12175 https://ieeexplore.ieee.org/document/9879701/.

Guo, Tianmei, Jiwen Dong, Henjian Li and Yunxing Gao (2017), "Simple convolutional neural network on image classification," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, 721–724 https://doi.org/10.1109/ICBDA.2017.8078730.

Hartmann, Jochen, Juliana Huppertz, Christina Schamp and Mark Heitmann (2019), "Comparing automated text classification methods," *International Journal of Research in Marketing*, 36 (1), 20–38 https://www.sciencedirect.com/science/article/pii/S0167811618300545.

Hartmann, Jochen, Mark Heitmann, Christian Siebert and Christina Schamp (2023), "More than a Feeling: Accuracy and Application of Sentiment Analysis," *International Journal of Research in Marketing*, 40 (1), 75–87 https://linkinghub.elsevier.com/retrieve/pii/S0167811622000477.

Hartmann, Jochen, Mark Heitmann, Christina Schamp and Oded Netzer (2021), "The Power of Brand Selfies," *Journal of Marketing Research*, 58 (6), 1159–1177 http://journals.sagepub.com/doi/10.1177/00222437211037258.

Hartmann, Jochen and Yannick Exner (2024), "GenImageNet," (accessed 2024-09-18), https://osf.io/8ctjy/.

He, Jiaxiu, Bingqing Li and Xin (Shane) Wang (2023), "Image features and demand in the sharing economy: A study of Airbnb," *International Journal of Research in Marketing*, 40, 760–780 https://doi.org/10.1016/j.ijresmar.2023.04.001.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren and Jian Sun (2015), "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, http://arxiv.org/abs/1512.03385.

Hu, Mantian (Mandy), Chu (Ivy) Dang and Pradeep K. Chintagunta (2019), "Search and Learning at a Daily Deals Website," *Marketing Science*, 38 (4), 609–642 https://doi.org/10.1287/mksc.2019.1156.

Huang, Li and Laura Pricer (2024), "How video conferencing promotes preferences for self-enhancement products," *International Journal of Research in Marketing*, 41 (1), 93–112 https://www.sciencedirect.com/science/article/pii/S0167811623000666.

Jaidka, Kokil, Salvatore Giorgi, H. Andrew Schwartz, Margaret L. Kern, Lyle H. Ungar and Johannes C. Eichstaedt (2020), "Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods," *Proceedings of the National Academy of Sciences*, 117 (19), 10165–10171 https://doi.org/10.1073/pnas.1906364117.

Jansen, Tijmen, Mark Heitmann, Martin Reisenbichler and David A. Schweidel (2024), "Automated Alignment: Guiding Visual Generative AI for Brand Building and Customer Engagement," *SSRN Electronic Journal*, https://www.ssrn.com/abstract=4656622.

Klostermann, Jan, Anja Plumeyer, Daniel Böger and Reinhold Decker (2018), "Extracting brand information from social networks: Integrating image, text, and social tagging data," *International Journal of Research in Marketing*, 35 (4), 538–556 https://doi.org/10.1016/j.ijresmar.2018.08.002.

Kornblith, Simon, Jonathon Shlens and Quoc V. Le (2019), "Do Better ImageNet Models Transfer Better?" *arXiv:1805.08974 [cs, stat]*, http://arxiv.org/abs/1805.08974.

Krugmann, Jan Ole and Jochen Hartmann (2024), "Sentiment Analysis in the Age of Generative AI," *Customer Needs and Solutions*, 11 (1), 3 https://doi.org/10.1007/s40547-024-00143-4.

LeCun, Y., L. Bottou, Y. Bengio and P. Haffner (1998), "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 86 (11), 2278–2324 https://doi.org/10.1109/5.726791.

Lee, Jeffrey K. (2021), "Emotional Expressions and Brand Status," *Journal of Marketing Research*, 58 (6), 1178–1196 https://doi.org/10.1177/00222437211037340.

Li, Bo, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu and Chunyuan Li (2024), "LLaVA-OneVision: Easy Visual Task Transfer," arXiv:2408.03326 [cs] http://arxiv.org/abs/2408.03326.

Li, Hanwei, David Simchi-Levi, Michelle Xiao Wu and Weiming Zhu (2023), "Estimating and Exploiting the Impact of Photo Layout: A Structural Approach," *Management Science*, 69 (9), 5209–5233 https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2022.4616.

Li, Xi, Mengze Shi and Xin (Shane) Wang (2019), "Video mining: Measuring visual information using automatic methods," *International Journal of Research in Marketing*, 36 (2), 216–231 https://doi.org/10.1016/j.ijresmar.2019.02.004.

Li, Yiyi and Ying Xie (2020), "Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement," *Journal of Marketing Research*, 57 (1), 1–19 https://doi.org/10.1177/0022243719881113.

Lin, Yan, Dai Yao and Xingyu Chen (2021), "Happiness Begets Money: Emotion and Engagement in Live Streaming," *Journal of Marketing Research*, 58 (3), 417–438 https://journals.sagepub.com/doi/pdf/10.1177/00222437211002477.

Liu, Haotian, Chunyuan Li, Qingyang Wu and Yong Jae Lee (2023), "Visual Instruction Tuning," arXiv:2304.08485 http://arxiv.org/abs/2304.08485.

Liu, Liu, Daria Dzyabura and Natalie Mizik (2020), "Visual Listening In: Extracting Brand Image Portrayed on Social Media," *Marketing Science*, 39 (4), 669–686 https://pubsonline.informs.org/doi/abs/10.1287/mksc.2020.1226.

Liu, Xuan, Savannah Wei Shi, Thales Teixeira and Michel Wedel (2018), "Video Content Marketing: The Making of Clips," *Journal of Marketing*, 82 (4), 86–101 https://doi.org/10.1509/jm.16.0048.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov (2019), "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692 [cs] http://arxiv.org/abs/1907.11692.

Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell and Saining Xie (2022), "A ConvNet for the 2020s," http://arxiv.org/abs/2201.03545.

Lu, Shasha, Li Xiao and Min Ding (2016), "A Video-Based Automated Recommender (VAR) System for Garments," *Marketing Science*, 35 (3), 484–510 https://doi.org/10.1287/mksc.2016.0984.

Lu, Shijie, Dai Yao, Xingyu Chen and Rajdeep Grewal (2021), "Do Larger Audiences Generate Greater Revenues Under Pay What You Want? Evidence from a Live Streaming Platform," *Marketing Science*, 40 (5), 964–984 https://doi.org/10.1287/mksc.2021.1292.

Maurício, José, Inês Domingues and Jorge Bernardino (2023), "Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review," *Applied Sciences*, 13 (9), 5521 https://www.mdpi.com/2076-3417/13/9/5521.

Mehta, Sachin and Mohammad Rastegari (2021), "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," *arXiv:2110.02178 [cs]*, (accessed 2023-04-04), https://arxiv.org/abs/2110.02178.

Nanne, Annemarie J., Marjolijn L. Antheunis, Chris G. van der Lee, Eric O. Postma, Sander Wubben and Guda van Noort (2020), "The Use of Computer Vision to Analyze Brand-Related User Generated Image Content," *Journal of Interactive Marketing*, 50, 156–167 https://doi.org/10.1016/j.intmar.2019.09.003.

Overgoor, Gijs, William Rand, Willemijn van Dolen and Masoud Mazloom (2022), "Simplicity is not key: Understanding firm-generated social media images and consumer liking," *International Journal of Research in Marketing*, 39 (3), 639–655 https://doi.org/10.1016/j.ijresmar.2021.12.005.

Peng, Ling, Geng Cui, Yuho Chung and Wanyi Zheng (2020), "The Faces of Success: Beauty and Ugliness Premiums in e-Commerce Platforms," *Journal of Marketing*, 84 (4), 67–85 https://doi.org/10.1177/0022242920914861.

Pieters, Rik, Michel Wedel and Rajeev Batra (2010), "The Stopping Power of Advertising: Measures and Effects of Visual Complexity," *Journal of Marketing*, 74 (5), 48–60 https://doi.org/10.1509/jmkg.74.5.048.

Rai, Arun (2020), "Explainable AI: from black box to glass box," *Journal of the Academy of Marketing Science*, 48 (1), 137–141 https://doi.org/10.1007/s11747-019-00710-5.

Redmon, Joseph, Santosh Divvala, Ross Girshick and Ali Farhadi (2016), "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788 https://doi.org/10.1109/CVPR.2016.91.

Rietveld, Robert, Willemijn Van Dolen, Masoud Mazloom and Marcel Worring (2020), "What you Feel, Is what you like Influence of Message Appeals on Customer Engagement on Instagram," *Journal of Interactive Marketing*, 49 (1), 20–53 https://doi.org/10.1016/j.intmar.2019.06.003.

Schamp, Christina, Jochen Hartmann and Dennis Herhausen (2024), "Bye-bye Bias: What to Consider When Training Generative AI Models on Subjective Marketing Metrics," *NIM Marketing Intelligence Review*, 16 (1), 42–48 https://doi.org/10.2478/nimmir-2024-0007.

Schwenzow, Jasper, Jochen Hartmann, Amos Schikowsky and Mark Heitmann (2021), "Understanding videos at scale: How to extract insights for business research," *Journal of Business Research*, 123, 367–379 https://doi.org/10.1016/j.jbusres.2020.09.059.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh and Dhruv Batra (2020), "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, 128 (2), 336–359 https://doi.org/10.1007/s11263-019-01228-7.

Simonyan, Karen and Andrew Zisserman (2015), "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv:1409.1556 http://arxiv.org/abs/1409.1556.

Sreedhar, Kavya, Jason Clemons, Rangharajan Venkatesan, Stephen W. Keckler and Mark Horowitz (2024), "Vision Transformer Computation and Resilience for Dynamic Inference," arXiv:2212.02687 http://arxiv.org/abs/2212.02687.

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens and Zbigniew Wojna (2015), "Rethinking the Inception Architecture for Computer Vision," *arXiv:1512.00567 [cs]*, http://arxiv.org/abs/1512.00567.

Troncoso, Isamar and Lan Luo (2023), "Look the Part? The Role of Profile Pictures in Online Labor Markets," *Marketing Science*, 42 (6), 1080–1100 https://pubsonline.informs.org/doi/10.1287/mksc.2022.1425.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin (2017), "Attention Is All You Need," arXiv:1706.03762 [cs] http://arxiv.org/abs/1706.03762.

Wang, Jing, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng and Shuqiang Jiang (2020), "LogoDet-3K: A Large-Scale Image Dataset for Logo Detection," *arXiv:2008.05359 [cs]*, http://arxiv.org/abs/2008.05359.

Wu, Yiqi, Xiaodan Hu, Ziming Fu, Siling Zhou and Jiangong Li (2024), "GPT-4o: Visual perception performance of multimodal large language models in piglet activity understanding," arXiv:2406.09781 [cs] http://arxiv.org/abs/2406.09781.

Yosinski, Jason, Jeff Clune, Yoshua Bengio and Hod Lipson (2014), "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems* Vol. 27 Curran Associates, Inc.

Zhang, Mengxia and Lan Luo (2022), "Can Consumer-Posted Photos Serve as a Leading Indicator of Restaurant Survival? Evidence from Yelp," *Management Science*, 69 (1), 25–50 https://doi.org/10.1287/mnsc.2022.4359.

Zhang, Shunyuan, Dokyun Lee, Param Vir Singh and Kannan Srinivasan (2021), "What Makes a Good Image? Airbnb Demand Analytics Leveraging Interpretable Image Features," *Management Science*, 68 (8), 5644–5666 https://doi.org/10.1287/mnsc.2021.4175.

Zhang, Shunyuan, Elizabeth M S Friedman, Kannan Srinivasan, Ravi Dhar and Xupin Zhang (2024), "Serving with a Smile on Airbnb: Analyzing the Economic Returns and Behavioral Underpinnings of the Host's Smile," *Journal of Consumer Research*,, ucae049 https://doi.org/10.1093/jcr/ucae049.

Zhang, Shunyuan, Kaiquan Xu and Kannan Srinivasan (2023), "Frontiers: Unmasking Social Compliance Behavior During the Pandemic," *Marketing Science*, 42 (3), 440–450 https://doi.org/10.1287/mksc.2022.1419.

Zhang, Shunyuan, Nitin Mehta, Param Vir Singh and Kannan Srinivasan (2021), "Frontiers: Can an Artificial Intelligence Algorithm Mitigate Racial Economic Inequality? An Analysis in the Context of Airbnb," *Marketing Science*, 40 (5), 813–820 https://doi.org/10.1287/mksc.2021.1295.

Zhao, Xinshu, John G. Lynch, Jr. and Qimei Chen (2010), "Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis," *Journal of Consumer Research*, 37 (2), 197–206 https://doi.org/10.1086/651257.

Zhou, Mi, George H. Chen, Pedro Ferreira and Michael D. Smith (2021), "Consumer Behavior in the Online Classroom: Using Video Analytics and Machine Learning to Understand the Consumption of Video Courseware," *Journal of Marketing Research*, 58 (6), 1079–1100 https://doi.org/10.1177/00222437211042013.

WEB APPENDIX

TABLE OF CONTENT

I: RELATED LITERATURE

Web Appendix A: Overview of automated image classification in marketing literature

Author	Year	Architecture	Research Question	DV	No. Samples	No. Classes
Anand and Kadiyali	2024	ResNet	How are e-cigarette taxes influencing posting behavior on social media?	Posting behavior	388,593	2
Bharadwaj et al.	2022	Xception	How do emotional displays during livestream retailing affect sales over time?	Sales	99,451	6
Burnap et al.	2023	VAE, GANs, VGG16	Can a model predict aesthetic scores and generate appealing product designs in the automotive industry?	Aesthetic scores	203	1
Chang et al.	2023	Google Cloud Vision	Does hearing different voices narrate a persuasive message in broadcast videos facilitate persuasion?	Pledged Amount, Number of backers, Project Success, ad effectiveness	11,801	2
Chuah and Yu	2021	Microsoft Azure Face API	How do emotional expressions of robots influence potential consumers' affective feelings on Instagram?	Affective feelings	223	8
Dzyabura et al.	2023	ResNet152, VGG19	How can prelaunch product images be used to predict individual item return rates for online fashion items?	Return rates	4,585	15
Dzyabura and Peres	2021	Clarifai Computer Vision API	How can a brand visual elicitation platform be used to understand consumer associations with brands?	Brand associations	4,743	20
Feng et al.	2021	ResNet50	How do social media influencers use narrative strategies in their posts, and how does this affect sponsorship disclosure effectiveness?	Sponsorship disclosure effectiveness, engagement	7,745	7
Ghose et al.	2012	MicrosoftVirtual Earth Interactive SDK	How can social media data be leveraged to improve product search engine ranking algorithms and provide consumers with the best value for their money?	Product ranking	1,497	> 99
Guan et al.	2023	MobileNet, MTCNN	What is the effect of customer-generated images (CGIs) on subsequent product ratings and post-purchase satisfaction?	Product ratings	1,650	2

Web Appendix A: Overview of automated image classification in marketing literature

Author	Year	Architecture	Research Question	DV	No. Samples	No. Classes
Gunarathne et al.	2022	Kairos API	What is the extent of business-to-customer racial bias (B2C bias) in corporate social media customer service, and how does it affect customer complaints?	Response Likelihood	110,533	6
Ha et al.	2021	ResNet50, AlexNet	How can visually mismatched content in brand-related posts on Instagram be characterized, and how effective is a machine-learning model in detecting such mismatches?	Visual information mismatch	3,169	10
Hartmann et al.	2021	VGG16	How do different types of brand-related selfie images impact sender engagement and brand engagement on social media platforms?	Engagement, CTR, purchase intention	19,949	3
He et al.	2023	AWS Rekognition	How do content and aesthetic features of background images on Airbnb affect the booking rate during a holiday period?	Booking rate	20,823	2
Henkel et al.	2020	RNN	Does augmenting service employees with AI-based emotion recognition software enhance their effectiveness in interpersonal emotion regulation (IER)?	Interpersonal emotion regulation (IER)	360	6
Hu et al.	2019	4-layer CNN	How does consumer behavior evolve over time on daily deal websites, specifically regarding clicking on newsletters and making purchases?	Probability of clicking, probability of purchasing	26,523	2
Huang and Pricer	2024	Google Cloud Vision	Is use of video conferencing platforms changing product consumption behavior?	Sales	66	7
Klostermann et al.	2018	Google Cloud Vision	How can brand-related images on social networking services be effectively analyzed to derive insights about consumer perceptions and experiences?	Brand perception	10,325	20
Lee	2021	Microsoft Azure Face API	How does brand emotionality on social media platforms relate to perceived brand status and median basket price?	Median Price	91,148	8

Web Appendix A: Overview of automated image classification in marketing literature

Author	Year	Architecture	Research Question	DV	No. Samples	No. Classes
Li et al.	2019	Imagga	What standard measures of visual information can be for analyzing videos in marketing and how do they affect funding outcome in crowdsourcing?	Funding outcome	6,822	4
Li et al.	2023	ResNet50	How do Airbnb property photos impact customers' renting decisions?	Booking rate	4,000	5
Li and Xie	2020	Google Cloud Vision	How does image content influence social media engagement for brands on Twitter and Instagram?	Engagement	4,537	10
Lin et al.	2016	SVM	How effective is an online face-recognition system in detecting fraud in e-commerce applications?	Face recognition precision	341	50
Lin et al.	2021	Pre-trained CNN	How does broadcaster emotion influence viewer engagement and financial returns in live streaming?	Engagement, amount tipped	1,450	7
Liu et al.	2018	NVISO	How effective are short video clips in promoting comedy movies?	intention to watch, box office success	1,271	1
Liu et al.	2020	BVLC CaffeNet	Can machine learning methods help brands measure their portrayal on social media using visual content?	Brand portrayal	16,360	4
Lu et al.	2016	SVM	How can garment retailers improve the shopping experience and increase sales using real-time in-store videos?	Hit-rate of recommendations and choices	540	3
Lu et al.	2021	Microsoft Azure	What is the impact of audience size on revenue in live streaming events under a pay-what-you-want (PWYW) pricing strategy?	Tipping revenue	2,222	2
Matz et al.	2019	Viola Jones Face-detection	How does individual image-person fit influence consumer responses to images in digital marketing?	Consumer responses	1,040	2
Nanne et al.	2020	YOLOv2, Clarifai, Google Cloud Vision	How can computer vision models be utilized to automatically analyze visual brand-related UGC on social media platforms?	Label accuracy	21,669	10

Web Appendix A: Overview of automated image classification in marketing literature

Author	Year	Architecture	Research Question	DV	No. Samples	No. Classes
Overgoor et al.	2022	RCNN, MSVO	How does the visual complexity of firm-generated imagery (FGI) on social media influence consumer liking?	Consumer liking	150,000	10
Peng et al.	2020	7-layer-CNN, RF	How does facial attractiveness of online sellers influence product sales on customer-to-customer e-commerce platforms?	Product sales	17,749	1
Rietveld et al.	2020	MSVO, Microsoft Azure Face API	How do visual emotional and informative appeals in brand-generated content on Instagram influence customer engagement?	Customer engagement (likes and comments)	46,900	2
Schwenzow et al.	2021	MTCNN, YOLOv3, ResNet152	How can businesses efficiently extract and analyze video-based variables to enhance research across disciplines?	Movie trailer likes	975	2
Soderlund et al.	2019	FaceReader 7.0	How does employee happiness portrayed in marketing materials impact consumer attitudes?	Consumer attitude toward service employee	4	1
Teixeira et al.	2012	Bayesian Neural Network Classifier	How do emotions in internet video advertisements influence viewer attention and retention?	Viewer attention, zapping behavior	500	2
Troncoso and Luo	2023	Face++, VGG-16, Inception, ResNet	How does the perceived job fit, inferred from profile pictures, impact hiring outcomes on freelancing platforms?	Perceived job fit, hiring outcome	3,000	2
Xiao and Ding	2014	Eigenface	How do faces in print advertisements influence viewer reactions on metrics important to advertisers?	Viewer reactions, purchase intention	148	12
Zhang et al.	2018	Face++	What are the antecedents influencing the perceived trust of guests towards hosts on Airbnb?	growth rate of bookings	3,801	7
Zhang, Mehta, Singh and Srinivasan	2021	ResNet50	How does Airbnb's smart-pricing algorithm impact the racial disparity in daily revenue earned by Airbnb hosts?	Daily revenue	800,000	3
Zhang, Lee, Singh and Srinivasan	2021	VGG16	How does image quality affect Airbnb property demand and what attributes make a good image for an Airbnb property?	Occupancy rate	510,000	2

Web Appendix A: Overview of automated image classification in marketing literature

Author	Year	Architecture	Research Question	DV	No. Samples	No. Classes
Zhang et al.	2023	Face++, FaceMaskDetection, MTCNN	How did people's shopping behaviors change with the onset of the COVID-19 pandemic based on their compliance with mask-wearing recommendations?	Customer behavior, transaction characteristics, diversity exploration	21,277	2
Zhang and Luo	2022	Clarifai Computer Vision API	Can consumer-posted photos on Yelp serve as leading indicators of restaurant survival, beyond traditional factors?	Restaurant survival	755,758	> 99
Zhang et al.	2024	ResNet50	Is showing a smile on Airbnb increasing rental sales?	Booking rate	9,000	2
Zhou et al.	2021	Microsoft Azure Face API	How does video content affect the consumption of online classroom courseware?	Video popularity, consumer behavior	771	8

Notes: Number of classes = 1 indicates image regression with a continuous outcome between 0 to 1.

II: IMAGE CLASSIFICATION METHODS

Web Appendix B: Method Summary

Chronologically, having a closer look at our included methods, Simonyan and Zisserman were among the first to develop a deep convolutional neural network. In their study from 2015, they introduce VGG-19, which is consists of 19 weight layers and leverages small 3×3 convolutional filters to enable increasing depth of the network. In the ILSVRC 2015, a VGG architecture was able to secure the second place (Alom et al. 2018).

Shortly after, He et al. (2015) introduced the ResNet family, which provide a framework to enable training of ultra-deep architectures. The largest version among several architectures within this family is the ResNet152, which we include in our benchmark. In contrast to earlier architectures (e.g., VGG-19) these Residual Networks (ResNets) address the vanishing gradient problem from increased depth that previous models had and led to winning the ILSVRC 2015 (Alom et al. 2018). He et al. (2015) managed to achieve this partly by including shortcut connections that skip one or more layers to add output from previous layers to later stages.

Similarly, an update to the Inception family was published in 2015. Szegedy et al. (2015) developed InceptionV3 with the goal in mind to utilize computational power more efficiently through strong regularization and factorized convolutions. For example, they introduced 1×1 convolutional blocks to increase efficiency of their model. So-called *bottlenecks* limit the number of features entering the following computationally expensive parallel blocks.

In 2017, Chollet (2017) developed a novel deep convolutional neural network architecture called Xception. Inspired by the Inception models, he uses depthwise separable convolutions only to replace the original Inception modules. In short, the architecture's structure can be described as a linear stack of these modules with residual connections (similar to ResNets). While Xception's number of parameters is identical to InceptionV3, Chollet's study indicates improved performance realized through more efficient use of model parameters instead of increased capacity.

While transformer models, such as RoBERTa, have taken the text analysis space by storm already (Hartmann et al. 2023), they are in a close battle with CNNs in the image classification space still. Since the introduction of the Vision Transformer and its attention mechanism in 2017 (Vaswani et al. 2017), several new methods have been developed. Some of these are pure transformers (e.g., the MobileViT by Mehta and Rastegari 2021) whereas others combine the best of CNNs and transformers into a single architecture (e.g., ConvNeXt by Liu et al. 2022). The attention mechanism and patch-wise processing introduced by Vaswani et al. (2017), however, remained at the core of transformer methods and enable their awareness for global (i.e., across the entire image) interdependencies and set them apart from classical CNNs.

The emergence of large language models (e.g., GPT, Llama) has led to large boom in AI application during the last couple of years. These models are often magnitudes larger than vision transformer models (for example OpenAI's popular GPT-3.5 (ChatGPT) containing 375 billion parameters compared to the largest version of RoBERTa containing 355 million parameters (Brown et al. 2020; Liu et al. 2019). This enables large language models to learn a vast understanding of numerous contexts and situations. Additionally, some newer models allow multiple data sources as input into the model (e.g., text and images). This enables users to leverage the flexibility of prompting such a vision language model and its free-text output with image data (Wu et al. 2024).

III: SETUP OF EMPIRICAL ANALYSIS

Web Appendix C: Machine Learning Pipeline

General pipeline To prepare the data for our experiments, we begin by drawing a random sample of $1,000$ observations without replacement for each of our 18 analyzed datasets[6]. Following that, we take a subset of 200 samples without replacement for the limited data condition.

The pipeline can be separated into three distinct sections: 1) hyperparameter search, 2) final training using 10-fold cross validation on $1,000$ and 200 observations, 3) model evaluation and documentation. We begin our experiments by testing different hyperparameter settings to not only provide information on the optimal method choice, but also on the respective empirically best initial hyperparameter settings for training. In each run of our pipeline, batch size, and learning rate are varied based on the extracted settings. We run grid-search varying the batch size between 16, 32, or 64, have a learning rate of $1e^{-2}$, $1e^{-3}$, $1e^{-5}$. We allow training for a maximum of 32 epochs, using an early stopping mechanism to stop the training process if no improvement in performance occurred over the last 3 epochs. Once all hyperparameters are set, we initialize the model architecture and start training. Depending on the respective dataset, we initiate the architectures for each experiment and match the correct number of classes by replacing the last dense layer only. After fine-tuning of each method-dataset combination, we measure the performance of the method for its best performing epoch. After determining the optimal hyperparameters for each method and dataset combination, we run 10-fold cross validation using the highest performing hyperparameter combination to robustly determine the performance of each method and dataset. We run this final training process using 10-fold cross validation on both the $1,000$ and 200 sample datasets. After each run of the 10 iterations per method-dataset combination the settings, time and performance are automatically documented. The model implementation

[6]The AIDA Car Funnel dataset contains only 559 images in total. Therefore, we include it only in the limited training data experimental setup

is intentionally simple and as close as possible to the original architectures to conduct a fair comparison and enable easy replication by other researchers. Several parameters were fixed across all methods. All pre-trained CNN and vision transformer methods were used with pre-trained ImageNet weights. Switching-out each model's last layer with respect to the number of classes that needed to be predicted for the dataset in question was the only adjustment we had to make. Images were resized to match the models' respective resolutions they were trained on. Lastly, we decided to use "Adam" as our optimizer for all methods. We ensure ease of implementation and replicability with reasonable computational for the chosen state-of-the-art methods, as we leverage the benefits of *transfer learning*.

Using transfer learning is in line with recent marketing research (e.g., Hartmann et al. 2021), and helps to obtain the best possible performance with reasonable computational requirements, as it leads to faster convergence (Kornblith et al. 2019). Transfer learning describes a training paradigm where in the first step a deep learning model is pretrained on open-domain data, and in a second step fine-tuned on a smaller dataset for a particular application. Specifically, we utilize weights pretrained on more than 1.2 million ImageNet images (Deng et al. 2009) to avoid re-training of fundamental object characteristics for each application. This allows to apply complete models with limited labelled training data as commonly available in marketing research. Generally, it can be stated that the performance of transferred weights is almost always better than random initialization, but their advantage diminishes with greater differences between the task at hand and the original training task (Yosinski et al. 2014). If available, we use the Huggingface implementation of each method (i.e., ConvNeXt (https://huggingface.co/facebook/convnext-base-384-22k-1k), ResNet (https://huggingface.co/microsoft/resnet-152), ViT (https://huggingface.co/google/vit-base-patch16-224-in21k)). For all other CNN methods we used the Keras Hub implementation, due to no Huggingface implementation being available (i.e., Inception (https://keras.io/api/applications/inceptionv3/), VGG (https://keras.io/api/applications/vgg/), Xception (https://keras.io/api/applications/

xception/)). Deviations in the process for Full-NN-Training, the vision language model Phi-3 and the multi-paradigm ensemble are described in the following sections.

Full-NN-training We implement the Full-NN-training method using the `Keras` framework in Python. This method is defined as a sequential model comprising the following six layers:

1. **2D Convolutional Layer**: Containing 64 neurons with a 3×3 kernel size, ReLU activation function, and an input shape corresponding to the dimensions of the input image.
2. **2D Max Pooling Layer**
3. **Flattening Layer**
4. **Dense Layer**: Fully connected layer, containing 64 neurons and ReLU activation function.
5. **Dense Layer**: Fully connected layer, containing 64 neurons and ReLU activation function.
6. **Output Layer**: The number of neurons corresponds to the number of classes in the dataset, and the layer uses a softmax activation function for binary/multi-class classification.

Unlike other models, this neural network is initialized with random weights and trained on all layers from scratch. Following the same methodology as the all-pretrained methods, we first optimize hyperparameters for each dataset and subsequently perform 10-fold cross-validation using the best-performing combination of hyperparameters.

Vision language model In contrast to the pipeline for CNNs and vision transformer methods (described in the previous paragraph), to use a vision language model for classification adapt the process to generate natural language text and convert it into numerical features. We use the Huggingface implementation of Microsoft's Phi-3 model (https://huggingface.co/microsoft/Phi-3-vision-128k-instruct). We generate text descriptions for each sample by prompting the model: "What could you expect a person seeing the image to think and feel? Please relate your answer to both the low-level features and high-level features of the image." We convert these text descriptions into numerical values using the bag-of-words methodology. To specify exactly, we use the Count Vectorizer provided by the Sklearn Package in Python (using unigrams with binary probabilistic mode). Alternatively, the text descriptions can be converted into numerical features using text embeddings (e.g., text-embedding-3-small by OpenAI). Lastly, we train a two-layer neural network as the classification head following the methodology of Chollet (2021). This neural network is fed with the bag-of-words representations of the text descriptions and labels of the dataset. As described by Chollet (2021) we initialize the network with a dense layer incorporating 256 neurons, a dropout layer of 50% and a final output layer using a softmax activation. Similar to the pipeline for all other methods, we search for the optimal hyperparameters of this classification head and run 10-fold cross validation per dataset using the hyperparameter settings with the highest performance.

Multi-paradigm ensemble The pipeline for the multi-paradigm ensemble integrates the ConvNeXt model with the vision language model Phi-3. At a high level, the outputs from both models are fed into a classification head, which consists of the same two-layer neural network as described in the previous section (i.e., dense layer using 256 neurons, dropout layer, and output layer). The fine-tuning process begins by training the vision transformer model on an auxiliary image dataset containing 500 images following the process described in the section regarding the general pipeline above. Note that none of these images overlap

with the validation datasets. Next, we combine the embeddings from the penultimate layer of the vision transformer with the numerical representations of textual descriptions generated by the vision-language model. These textual descriptions are converted into numerical form using the bag-of-words methodology (refer to the previous section for details). We then fine-tune the classification head using the training dataset, similarly testing all hyperparameter combinations and performing 10-fold cross-validation using the best-performing hyperparameter configuration.

During an inference pass, new images are passed through the fine-tuned ConvNeXt model to access the state of the penultimate visual embedding layer. Similarly, the image is passed through the vision language model to generate the text description. The text description is converted into numerical representation, both feature vectors are concatenated and fed into the classification head to predict the most probable class.

Web Appendix D: Dataset overview

Dataset	Research Domain	Image-related task	Number of images	Classes	Image dimensions	Source
AIDA Funnel Car	Advertising and Promotional Effectiveness	How (low, medium, or high AIDA)	559	3	341 × 283	Jansen et al. (2024)
Brand Logo	Brand Management	What (e.g., electronics, foods, sports)	158,652	9	482 × 396	Wang et al. (2020)
Brand Selfie	Consumer Insights and Behavior	What (brand-, consumer selfie, or packshot)	6,928	3	481 × 491	Hartmann et al. (2021)
E-Commerce Category	Product Design & Management	What (e.g., parfum, smartphones, tools)	117,590	41	710 × 710	E-commerce Products (n.d.)
Emotion	Consumer Insights and Behavior	How (e.g., anger, joy, surpise)	8,350	6	722 × 555	Panda et al. (2018)
Fashion Products	Product Design & Management	What (e.g., bags, shoes, watches)	44,441	45	60 × 80	Fashion Product Images (n.d.)
Outdoor Scenery	Media Management	What (e.g., buildings, forest, street)	17,034	6	150 × 150	Intel Scene Classification (n.d.)
Image Sentiment	Consumer Insights and Behavior	How (e.g., highly positive, neutral, negative)	12,550	5	465 × 400	crowdflower (n.d.)
Face Sentiment	Consumer Insights and Behavior	How (savory, unsavory)	12,420	2	295 × 350	Good guys - Bad guys (n.d.)
Image Quality	Advertising and Promotional Effectiveness	How (e.g., low, medium, very high)	10,073	5	512 × 384	Hosu et al. (2020)
Generated Aesthetics	Advertising and Promotional Effectiveness	How (low, medium, high)	10,320	3	512 × 512	Hartmann et al. (2024)
Generated Gender	Consumer Insights and Behavior	Who (female, male)	10,001	2	256 × 256	Generated.photos (n.d.)
Generated Ethnicity	Consumer Insights and Behavior	Who (e.g., asian, black, white)	10,001	4	256 × 256	Generated.photos (n.d.)
Luxury Brands	Brand Management	What (e.g., Chanel, Gucci, Versace)	2,184	7	150 × 150	Brand Styles (n.d.)
Public Places	Media Management	What (e.g., bathroom, conference room, kitchen)	9,456	22	700 × 700	Bylinkskii et al. (2015)
Store Item Colour	Product Design & Management	What (e.g., black, red, yellow)	6,239	12	224 × 224	StoreItemColour (n.d.)
Unbiased Emotion	Consumer Insights and Behavior	How (e.g., anger, fear, love)	3,045	6	837 × 659	Panda et al. (2018)
Stock-photo Category	Media Management	What (e.g., arts & culture, business, nature)	8,000	16	200 × 221	Unsplash Images (n.d.)

Note: Image Dimensions are measured as the average of all images in the dataset and displayed as width × height.

Web Appendix E: Exemplary images of the datasets

AIDA Car Funnel

Task: How
Classes: 3

Brand Logo

Task: What
Classes: 9

Brand Selfie

Task: What
Classes: 3

E-Commerce Category

Task: What
Classes: 41

Emotion

Task: How
Classes: 6

Face Sentiment

Task: How
Classes: 2

Fashion Products

Task: What
Classes: 25

Generated Aesthetics

Task: How
Classes: 3

Generated Ethnicity

Task: Who
Classes: 2

Generated Gender

Task. Who
Classes. 2

Image Quality

Task: How
Classes: 5

Image Sentiment

Task: How
Classes: 2

Luxury Brands

Task: What
Classes: 7

Outdoor Scenery

Task: What
Classes: 6

Public Places

Task: What
Classes: 21

Stock-photo Category

Task: What
Classes: 16

Store Item Color

Task: What
Classes: 12

Unbiased Emotion

Task: How
Classes: 6

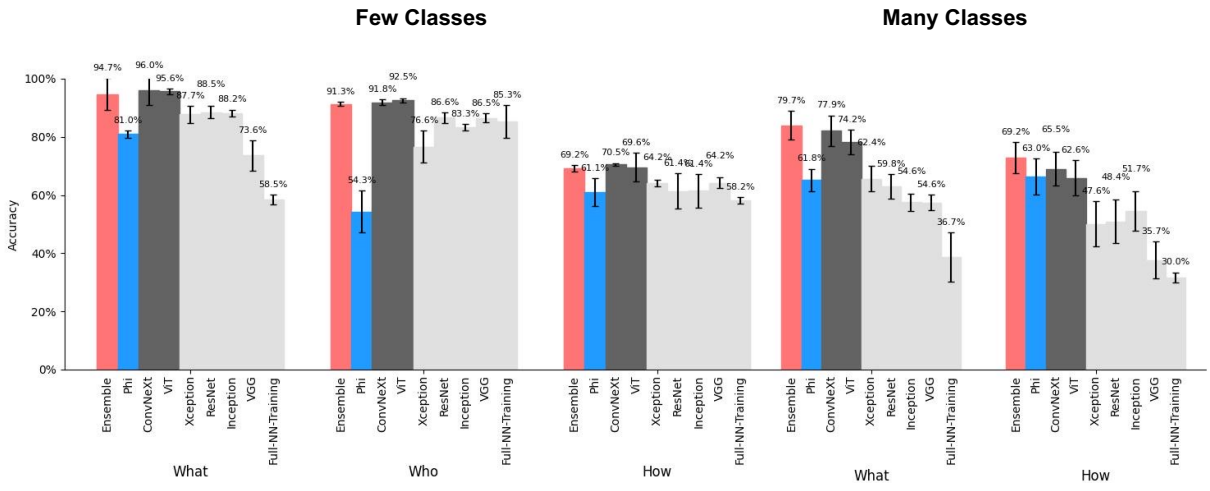
Web Appendix F: Overview of method architectures

Model	Architecture Group	Number of Parameters	Training strategy
Full-NN-Training	Convolutional Neural Network	50,473,157	Training from Scratch
ConvNeXt-Base	Vision Transformer	87,571,589	Transfer Learning
Multi-paradigm ensemble	Ensemble Method	1,666,565	Transfer Learning, Prompting, Training of Classification Head
Inception v3	Convolutional Neural Network	21,813,029	Transfer Learning
Phi-3	Vision Language Model	1,404,933	Prompting, Training of Classification Head
ResNet152	Convolutional Neural Network	58,154,053	Transfer Learning
ViT	Vision Transformer	85,802,501	Transfer Learning
VGG19	Convolutional Neural Network	20,149,829	Transfer Learning
Xception	Convolutional Neural Network	20,871,725	Transfer Learning

Note: The number of parameters is measured on a binary classification task.

IV: FURTHER ANALYSIS

Web Appendix G: Accuracy depending on image-related task and number of classes (large training data)



Note: Error bars indicate standard error.

When trained on large datasets containing 1,000 samples, vision transformers (i.e., ConvNeXt, ViT) and the multi-paradigm ensemble consistently outperform CNN-based models. For perception-based 'how' tasks, vision language models achieve performances comparable to vision transformers. These findings suggest that vision language models excel in tasks requiring a nuanced understanding of visual content that goes beyond objective categorization.

Additional References

Alom, Md Zahangir, Tarek M. Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C. Van Esesn, Abdul A. S. Awwal and Vijayan K. Asari (2018), "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches," *arXiv:1803.01164 [cs]*, http://arxiv.org/abs/1803.01164.

Brand Styles (n.d.), (accessed 2022-06-29), https://www.kaggle.com/datasets/olgabelitskaya/style-color-images.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Hearbert-Voss, A., Krueger, G., Hanighan, T., Child, R., Ramesh, A., Ziegler, D. M., au, J., ainter, C., . . . , Amodei, D. (2020), "Language Models are Few-Shot Learners" *arXiv:2005.14165*, https://doi.org/10.48550/arXiv.2005.14165

Bylinskii, Zoya, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva (2015), "Intrinsic and extrinsic effects on image memorability," *Vision Research*, 116, 165–78.

Cheetah, Hyena, Jaguar and Tiger (n.d.), (accessed 2022-06-29), https://www.kaggle.com/datasets/90cc4ccdd758880764780f90255c7f18790d1efd4a38494 7f578fd174955de69.

Chollet, Francois (2021). *Deep Learning with Python (2nd ed.).* Manning Publications Co. https://www.manning.com/books/deep-learning-with-python-second-edition

crowdflower (n.d.), "Image Sentiment Polarity," (accessed 2022-06-27), https://data.world/crowdflower/image-sentiment-polarity.

E-commerce Products (n.d.), (accessed 2022-06-28), https://www.kaggle.com/datasets/kennethrithvik/shopee.

*Fashion Product Images*n.d.), (accessed 2022-06-28), https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-small.

Feng, Yang, Huan Chen and Qian Kong (2021), "An expert with whom i can identify: the role of narratives in influencer marketing," *International Journal of Advertising*, 40 (7), 972-993.

Generated.photos (n.d.), "Image Datasets For Machine Learning — Generated.photos," accessed 2022-06-27), https://generated.photos.

Good guys-Bad Guys (n.d.), (accessed 2022-06-27), https://www.kaggle.com/gpiosenka/good-guysbad-guys-image-data-set.

Ha, Yui, Kunwoo Park, Su Jung Kim, Jungseock Joo and Meeyoung Cha (2021), "Automatically Detecting ImageText Mismatch on Instagram with Deep Learning," *Journal of Advertising*, 50 (1), 52-62.

Hartmann, J., Exner, Y. (2024)., "GenImageNet", (accessed 2024-06-18), https://osf.io/8ctjy/

Henkel, Alexander P., Stefano Bromuri, Deniz Iren and Visara Urovi (2020), "Half human, half machine - augmenting service employees with AI for interpersonal emotion regulation," *Journal of Service Management*, 31 (2), 247-265 Publisher.

Hosu, Vlad, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe (2020), "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, 29, 4041–4056 http://arxiv.org/abs/1910.06180.

Hu, Mantian (Mandy), Chu (Ivy) Dang and Pradeep K. Chintagunta (2019), "Search and Learning at a Daily Deals Website," *Marketing Science*, 38 (4), 609-642.

Intel Scene Classification (n.d.), (accessed 2022-06-27), https://www.kaggle.com/puneet6060/intel-image-classification.

Jansen, Tijmen, Heitmann, Mark, Reisenbichler, Martin, Schweidel, D. A. (2023), "Automated Alignment: Guiding Visual Generative AI for Brand Building and Customer Engagement" *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4656622

Jung, Soon-Gyo, Jisun An, Haewoon Kwak, Joni O. Salminen, and B. Jansen (2018), "Assessing the Accuracy of Four Popular Face Recognition Tools for Inferring Gender, Age, and Race," in *ICWSM*.

Mehta, Sachin and Mohammad Rastegari (2021), "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," *arXiv:2110.02178 [cs]*.

Liang, Lingyu, Luojun Lin, Lianwen Jin, Duorui Xie, and Mengru Li (2018), "SCUT-FBP5500: A Diverse Benchmark Dataset for Multi-Paradigm Facial Beauty Prediction."

Lin, Wen-Hui, Ping Wang and Chen-Fang Tsai (2016), "Face recognition using support vector model classifier for user authentication," *Electronic Commerce Research and Ap- plications*, 18, 71-82.

Liu, Yinhan, Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019), "RoBERTa: A Robustly Optimized BERT Pretraining Approach" *arXiv:1907.11692*, https://doi.org/10.48550/arXiv.1907.11692

Liu, Zhuang, Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S. (2022), "A ConvNet for the 2020s", *ArXiv:2201.03545 [Cs]*, http://arxiv.org/abs/2201.03545.

Liu, Ziwei, Ping Luo, Xiaogang Wang, and Xiaoou Tang (2015), "Deep Learning Face Attributes in the Wild," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 3730–3738.

Matz, Sandra C., Cristina Segalin, David Stillwell, Sandrine R. Müller and Maarten W. Bos (2019), "Predicting the Personal Appeal of Marketing Images Using Computational Methods," *Journal of Consumer Psychology*, 29 (3), 370-390.

Murray, Naila, Luca Marchesotti, and Florent Perronnin (2012), "AVA: A large-scale database for aesthetic visual analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2408–2415.

Panda, Rameswar, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K. Roy-Chowdhury (2018), "Contemplating Visual Emotions: Understanding and Overcoming Dataset Bias," in *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds., Cham: Springer International Publishing, 594–612.

Perez, Luis and Jason Wang (2017), "The Effectiveness of Data Augmentation in Image Classification using Deep Learning", *arXiv:1712.04621 [cs]*, http://arxiv.org/abs/1712.04621.

Sharma, Chhavi, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari and Bj orn Gamb ack (2020), "SemEval-2020 Task 8: Memotion Analysis- the Visuo-Lingual Metaphor!," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation Barcelona (online) International Committee for Computational Linguistics*, 759–773.

Soderlund, Magnus and Hanna Berg (2019), "Employee emotional displays in the extended service encounter: A happiness-based examination of the impact of employees depicted in service advertising," *Journal of Service Management*, 31 (1), 115-136.

Store Items Colour (n.d.), (accessed 2022-06-29), https://www.kaggle.com/datasets/imoore/6000- store-items-images-classified-by-color.

Tan, Mingxing, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard and Quoc V. Le (2019), "MnasNet: Platform-Aware Neural Architecture Search for Mobile," *arXiv:1807.11626 [cs]*, http://arxiv.org/abs/1807.11626.

Teixeira, Thales, MichelWedel and Rik Pieters (2012), "Emotion-Induced Engagement in Internet Video Advertisements," *Journal of Marketing Research*, 49 (2), 144-159.

Unsplash Images (n.d.), (accessed 2022-06-29), https://www.kaggle.com/datasets/prathameshbhalekar/imagescraped-from-unsplash.

Wang, Jing, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, Haishuai Wang and Shuqiang Jiang (2020), "Logo-2K+: A Large-Scale Logo Dataset for Scalable Logo Classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 6194–6201.

Wang, Jing, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng and Shuqiang Jiang (2020), "LogoDet-3K: A Large-Scale Image Dataset for Logo Detection," *arXiv:2008.05359 [cs]*, http://arxiv.org/abs/2008.05359.

Wu, Y., Hu, X., Fu, Z., Zhou, S., Li, J. (2024), "GPT-4o: Visual perception performance of multimodal large language models in piglet activity understanding" *arXiv:2406.09781*, https://doi.org/10.48550/arXiv.2406.09781

Xiao, Li and Min Ding (2014), "Just the Faces: Exploring the Effects of Facial Features in Print Advertising," *Marketing Science*, 33 (3), 338-352.

Xiao, Han, Kashif Rasul and Roland Vollgraf (2017), "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms".

Zhang, Le, Qiang Yan and Leihan Zhang (2018), "A computational framework for understanding antecedents of guests' perceived trust towards hosts on Airbnb," Decision Support Systems, 115, 105-116.

Özgenel, Ç.F. and Arzu Gönenç Sorguç (2018), "Performance Comparison of Pretrained Convolutional Neural Networks on Crack Detection in Buildings," *ISARC Proceedings*, 693–700.

B. Valuable visual variety? How temporal dynamics of visual features in video ads affect ad effectiveness

Authors Keno Tetzlaff
Jan Ole Krugmann
Leonard Kinzinger
Jochen Hartmann

Year 2024

Bibliographic Status Under review at
Journal of the Academy of Marketing Science
3rd round

Valuable visual variety: How temporal dynamics of visual features in video ads affect ad effectiveness

Keno Tetzlaff[1*]

Jan Ole Krugmann[2*]

Leonard Kinzinger[3*]

Jochen Hartmann[4*]

Declarations:

This work was funded by the German Research Foundation (DFG) research grant number 460037581.

The authors have no competing interests to declare that are relevant to the content of this article.

[1]Keno Tetzlaff is Doctoral Student at the University of Hamburg, Hamburg Business School, Moorweidenstrasse 18, 20148 Hamburg, Germany. keno.tetzlaff@uni-hamburg.de.

[2]Jan Ole Krugmann is Doctoral Student at the Technical University of Munich, TUM School of Management, Arcisstr. 21, 80333 Munich, Germany. jan.krugmann@tum.de.

[3]Leonard Kinzinger is Graduate Student at the Technical University of Munich, TUM School of Management, Arcisstr. 21, 80333 Munich, Germany. leonard.kinzinger@tum.de.

[4]Jochen Hartmann is Professor of Digital Marketing at the Technical University of Munich, TUM School of Management, Arcisstr. 21, 80333 Munich, Germany. jochen.hartmann@tum.de.

* All authors contributed equally

Valuable visual variety: How temporal dynamics of visual features in video ads affect ad effectiveness

Abstract

In today's digital marketing landscape, video ads play a critical role in capturing consumer attention amid widespread content saturation. This paper investigates the impact of *visual variety* — defined as the temporal variability of a video's visual features — on advertising effectiveness. Utilizing machine learning, we develop a multidimensional visual variety measure and apply it to over 3,000 video ads across two media channels (TV, social media) and various product categories. Three studies reveal an inverted U-shaped relationship between visual variety and advertising effectiveness. A controlled lab experiment reveals that this relationship is mediated by both stimulation and comprehension. Whereas too little visual variety can lack stimulation and too high visual variety can undermine comprehension, ads with medium visual variety strike an effective balance, being both simulating and comprehensible for consumers. We provide guidelines for advertisers on optimizing visual variety and introduce a web-based tool that calculates visual variety scores for videos.

Keywords: machine learning, video advertising, visual variety, visual analytics, cognitive load theory, optimum stimulation level theory

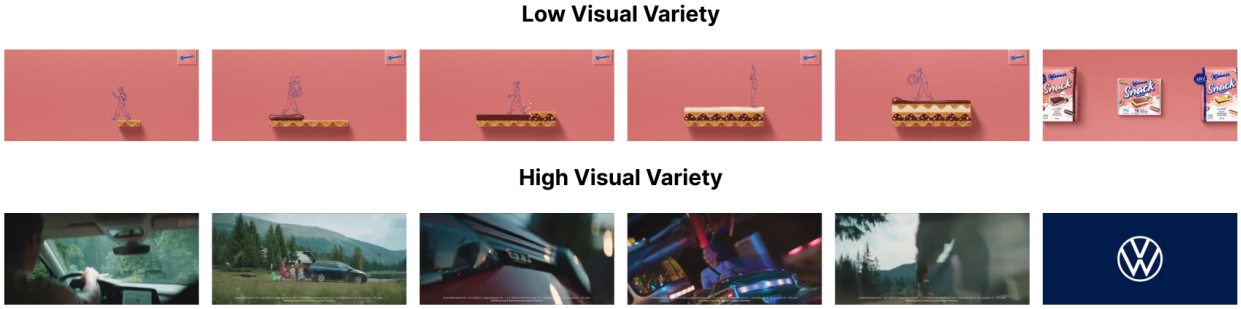
Introduction

In today's media landscape, stimulation from information and entertainment has become pervasive, with advertisers fiercely competing for the increasingly scarce resource of consumers' attention (Galloway 2022). In this "war for attention," advertisers must create content that is sufficiently intrusive to seize consumer attention (Lindsay 2024; Anderson and de Palma 2013; Galloway 2022). There has been a trend in video ads becoming visually more complex and showcasing greater variety (Lindsay 2024). This evolution reflects advertisers' efforts to capture and retain viewer attention in an increasingly saturated media environment. For example, the 1979 Coca-Cola ad famously known as "Hey Kid, Catch!" which aired during the Super Bowl, featured a simple, singular scene in a dimly lit stadium tunnel, creating a focused and static moment. Contrasting, modern advertisements like Coca-Cola's 2023 spot "Masterpiece" employ rapid shot changes, varying objects, vibrant colors, and dynamic action. Given the stark contrast between the simplicity of past advertisements and the high visual variety of today's marketing efforts, one must ask: does increased visual variety actually enhance advertising effectiveness? Or is there a point at which it becomes too overwhelming and incomprehensible, detracting from the ad's core message?

Little is understood about the effectiveness of visual variety in video ads, despite its potential importance. Research on video advertising has largely explored specific factors such as emotionality (e.g., Nikolinakou and King 2018; Olney et al. 1991), perceived speed through effects like slow motion or fast product animations (Jia et al. 2020; Shani-Feinstein et al. 2022; Stuppy et al. 2023), and the presence of specific objects (Tellis et al. 2019). To thoroughly explore the impact of visual design choices on video ad effectiveness, this research operationalizes *visual variety* using advanced machine learning methods. We define visual variety as the degree of visual stimulation derived from the range of visual features displayed (i.e., colors, patterns, shapes, objects) and the temporal dynamics of the video, including the

number, duration, and arrangement of shots[1]. Figure 1 illustrates visual variety along two example video ads, one with low and one with high visual variety. A video content creator can plan and adjust visual variety to make a video more or less stimulating for a consumer. This makes visual variety an important adjustable feature with practical implications.

Figure 1: EXAMPLE ADS WITH LOW AND HIGH LEVELS OF VISUAL VARIETY



Note: The images above represent exemplary frames from two video ads with different levels of visual variety (low visual variety for Manner ad: 2.4 vs. high visual variety for Volkswagen ad: 4.2). Visual variety is measured on a scale from 1 to 5, with 5 indicating the maximum. The frames for each ad are shown in their temporal sequence from left to right as they appear in the video.

Pursuing a multi-method approach, this research analyzes over 3,000 videos from TV ads (study 1), Facebook field ads (study 2), and experimental data (study 3), across a broad range of product categories (e.g., automotive, cosmetics, food, financial products, etc.). To our best knowledge, this is the first study to develop a methodology that integrates 18 distinct visual variety video features into a singular measure using machine learning techniques. Across three studies, our findings consistently indicate an inverted U-shaped relationship between visual variety and advertising effectiveness, suggesting there is an optimal level of visual variety that maximizes ad effectiveness, beyond which effectiveness declines. We identify stimulation and comprehension as underlying mechanisms driving this effect, grounded in the optimum stimulation level theory and cognitive load theory. These insights guide both researchers and practitioners in achieving the ideal balance between viewer stimulation and

[1]A shot is the smallest increment of a video, defined as an unbroken sequence of frames recorded from the same camera (Chasanis et al. 2009). Whereas a scene is defined as a "series of semantically correlated shots" which transport the narrative (Chasanis et al. 2009)

information comprehension to optimize video ad impact. We calculate exemplary budget saving potentials by optimizing visual variety, offer technical guidelines on adjusting visual variety during both pre- and post-production phases and introduce a web-based tool for video analysis, allowing content creators to evaluate their videos to achieve optimal visual variety [LINK BLINDED FOR PEER REVIEW].

Theoretical development

Visual variety in video ads

Video advertising remains a cornerstone of marketing strategies and enriches the advertising landscape with its dynamic and visually engaging content (Grewal et al. 2021)[2]. Videos are particularly adept at capturing the viewer's attention through their ability to stimulate the senses (Lang 2000). The effectiveness of video ads is evidenced by research showing that dynamic digital video ads can achieve click-through rates significantly higher than static banner ads. Pauwels et al. (2023) reveal that the click-through rates for sponsored brand videos can be up to 17.7 times higher compared to static images. But what drives successful video ads, particularly in terms of visual design decisions?

In cinematography, the level of visual variety represents a central design decision in visual storytelling, enhancing the stimulation and narrative comprehension (Brine 2020; Gleicher and Liu 2008). When designing video ads, content creators face two primary visual design decisions to optimize their ad's effectiveness. First, they must decide on what to display (i.e., people, objects, shapes, patterns, and colors) by defining the required scenes to tell the narrative. Pieters et al. (2010) define variations in colors, objects, shapes and patterns within static images as "feature complexity" and "design complexity". These concepts can be extended to video frames, where each frame can exhibit different levels of complexity. Second, they need to determine the temporal dynamics of these scenes, which include the

[2]Web Appendix A provides a comprehensive overview of related marketing studies on video ads.

number of shots, the duration of each shot, and their sequencing (Liu et al. 2018). Together, these two dimensions determine the overall visual variety of a video, specifically, variations in color, shapes, and objects, as well as the overall temporal composition of shots (e.g., variations within shots and transitions between shots). These visual variety design decisions finally shape viewers' perceptions, influencing advertising effectiveness.

Existing research tends to focus single elements of a video to predict advertising effectiveness, like narrative features (e.g., Toubia et al. (2021); Stayman and Batra (1991); Fossen and Schweidel (2019); Guitart and Stremersch (2021); Knight and Bart (2024)), ad emotionality (e.g., Akpinar and Berger (2017); Teixeira et al. (2012); Yang et al. (2022)), distinct visual elements (e.g., Tellis et al. (2019); Becker et al. (2023); Chang et al. (2022); McGranaghan et al. (2022)), or pixel-level variation between frames (Li et al. 2019). However, no study has comprehensively captured the full spectrum of visual variety decisions in video ads. In summary, while previous research has provided valuable insights into various aspects of video advertising effectiveness, a comprehensive understanding of the broad spectrum of visual variety and its underlying psychological mechanisms remain under-explored.

The effect of visual variety on advertising effectiveness

Variety demonstrates a dual role in marketing: depending on its use and intensity, it can either engage or overwhelm the viewer. Rafieian and Yoganarasimhan (2022) find that a unit increase in variety between ads in a browsing session results in up to 13% higher click-through-rates (CTR). Chang et al. (2022) discover that greater variety in narrators (i.e., more narrators) of crowdfunding videos significantly increases the chances of campaign success. According to Berger et al. (2021) and Toubia et al. (2021), topical variety in films and TV shows enhances viewer experience and leads to a more positive evaluation of the content. Welke and Vessel (2022) demonstrate that dynamic video stimuli (i.e., more variety), reduce boredom and improve aesthetic ratings.

On the contrary, too much variety can also overwhelm consumers and in turn negatively

affect advertising effectiveness. Jacoby (1977) and Lang (2000) argue that human cognitive capacity is limited, and excessive variety can impair information processing and decision-making. For example, Barnett and Cerf (2017) show that high complexity from visual information stimulation within movie trailers leads to decreased recall and lower ticket sales due to less efficient neural processing. These contrasting effects of variety on advertising effectiveness indicate a potential non-linear relationship that is closely linked to broader concept of stimulus complexity. Variety acts as a lever of stimulus complexity, either enhancing viewer stimulation or overwhelming them based on how it is applied and its intensity. This characteristic aligns with Berlyne's seminal aesthetic theory, which states that the most appealing level of stimulus complexity - and by extension variety - follows an inverted U-shape with an optimal level at medium complexity (Berlyne 1958). This theory is supported by several studies. For instance, Morrison and Dainoff (1972) observe that viewing time for magazine ads exhibits an inverted U-shaped relationship with perceived visual ad complexity. Similar patterns are noted in website design, where an optimal level of visual complexity enhances perceived aesthetics, attitudes towards the website, and purchase intentions (Geissler et al. 2006; Miniukovich and Marchese 2020; Im et al. 2021; Reinecke et al. 2013). For videos, Koh and Cui (2022) demonstrate that the number of views follows an inverted U-shape based on the complexity of the thumbnails used, and Tong et al. (2022) report that the background complexity in online live streams influences purchase intentions in a similar manner. Given this evidence, we anticipate that an optimal level of visual variety exists, which maximizes advertising effectiveness:

H1: The relationship between visual variety and advertising effectiveness follows an inverted U-shape.

Visual variety and stimulation

Optimum Stimulation Level (OSL) theory posits that individuals aim for an optimal level of environmental stimulation that allows for performance at the best level (Berlyne 1966).

In advertising, this manifests as a balance: too little stimulation leads to boredom and disengagement, while excessive stimulation causes over-stimulation and subsequent avoidance behaviors, such as ad skipping (Menon and Kahn 1995). Steenkamp and Baumgartner (1992) find that stimulus arousal and consumer response exhibit an inverted U-shaped relationship, highlighting the balance required for advertising stimulation levels. Menon and Kahn (1995) directly manipulate the variety of product choices available in retail settings to measure its impact on consumer stimulation and subsequent behavior. They find that increased variety in soft drink options enhances consumer stimulation, which, in turn, influenced their subsequent choice behaviors, reducing their need to seek variety elsewhere. Similarly, Atalay et al. (2023) draw on OSL theory to create more effective written marketing messages in the context of static social media ads (i.e., images paired with a caption). They find that stimulation from a moderately surprising syntax in these messages is more effective than too little or too much syntactic surprise. Menon and Kahn (1995) and Atalay et al. (2023) show that manipulating elements of complexity or variety not only influences consumer behavior, but also serves to disentangle stimulation as the underlying mechanism behind these effects. In conclusion, we predict the same mediation effect in the context of visual variety in video ads:

H2: Stimulation mediates the relationship between visual variety and consumer response.

(a) Visual variety has a positive linear relationship with stimulation.

(b) Stimulation has an inverted U-shaped relationship with advertising effectiveness.

Visual variety and comprehension

Building on the understanding that over-stimulation can significantly reduce consumer attention and engagement, Cognitive Load (CL) theory offers a valuable explanation for these effects. Developed by John Sweller in the late 1980s, CL theory posits that excessive cognitive demand can impede the effective processing of information and degrade overall problem-

solving capabilities (Sweller 1988). Research by Shin et al. (2020) extends this theory to visual stimuli, demonstrating that increased complexity in images increases the cognitive effort required to process and understand the presented information. This insight is critical for managing visual variety in video ads, where maintaining a balance between stimulation and comprehension is essential to prevent cognitive overload and optimize ad effectiveness. Further expanding on this concept, Glaser and Reisinger (2021) explore how product-story links in video ads can enhance brand attitudes, identifying comprehension as the key mechanism driving this effect. Their findings suggest that video ads with less complexity (i.e., where the connection between the product and the visual elements of the ad is straightforward) facilitate easier comprehension, thereby enhancing advertising effectiveness. Similarly, Brasel and Gips (2014) manipulate visual and verbal variety (i.e., complexity) of TV commercials and demonstrate that high levels of visual variety decrease ad comprehension, as measured by brand recall. We thus hypothesize:

H3: Comprehension mediates the relationship between visual variety and advertising effectiveness.

(a) Visual variety has a negative linear relationship with comprehension.

(b) Comprehension has a positive linear relationship with advertising effectiveness.

Building on the insights from OSL and CL theory and their application in advertising, it becomes evident that both stimulation and comprehension are potential mechanisms explaining the effect of visual variety on advertising effectiveness. While we hypothesize that stimulation and comprehension each mediate this relationship in different ways, it is also plausible that these mechanisms operate in parallel, each contributing to the overall advertising effectiveness:

H4: The effect of visual variety on advertising effectiveness is mediated in parallel by stimulation and comprehension.

Overview of studies

Adopting a multi-method strategy, this research leverages advanced machine learning methods to develop a compound measure for visual variety in videos (see Section Methodology). We quantitatively assess the impact of visual variety on video ad effectiveness across both traditional platforms (i.e., TV) and new-age digital platforms (i.e., Facebook). Table 1 provides an overview of the studies, their objectives and the methods applied. Study 1 investigates the influence of visual variety on click and sharing intentions, utilizing more than 27,000 consumer responses to a large-scale international survey on 910 real TV ads. Study 2 investigates whether the experimental results from Study 1 hold true in a real-world context by analyzing performance data from Facebook video ads. This data includes actual consumer behaviors, such as clicks and purchases, providing a practical validation and comparison of our initial findings. Lastly, Study 3 employs a controlled experimental setup to establish causality and explore the mechanisms through which visual variety affects advertising effectiveness, including potential mediation by stimulation and comprehension.

Methodology

Identification of features defining visual variety

To create a comprehensive set of features that explain visual variety, we focused on both the temporal dynamics and feature granularity of the videos. Temporal dynamics (see Figure 2) refer to the timing, sequence, and duration of shots within a video, capturing how visual features change over time. Feature granularity (see Figure 3) describes the layers of visual detail within a video, ranging from pixel-level details to broader, high-level object information. Together, they form a 4x4 matrix (see Figure 5), representing the interaction between temporal and content-based features across different levels of granularity. Alongside the number of shots and the average shot duration, this results in 18 extracted features that

Table 1: OVERVIEW OF STUDIES AND OBJECTIVES

	Data	Objective	Method
Study 1: Identify effect of visual variety on ad effectiveness in video advertising	• 910 TV ads across multiple brands and from almost 30 countries provided and rated in ad effectiveness (click and sharing intention) in an international consumer survey by an industry partner	• Introduce relevant controls to isolate effect of visual variety on advertising effectiveness • Test H1 and identify the effect of visual variety on click and sharing intentions for real TV ads evaluated in a survey	• OLS regression model
Study 2: Replicate findings for Facebook field data ads	• 2,182 Facebook video ads including real performance data on clicks and purchases across multiple brands provided by an industry partner	• Validate H1 and findings from Study 1 and test for differences to TV context • Control for potential country differences	• Negative binomial regression model
Study 3: Demonstrate causality in the lab and provide mechanism evidence	• Three self-generated video ads, varying in their visual variety only (low, medium, high) with identical product, brand and ad duration • 157 valid Prolific survey responses rating click intentions and evaluating stimulation and comprehension as potential mediators	• Isolate the effect of visual variety and test for causality • Test H2, H3 and H4 to identify the underlying mechanisms explaining the effect of visual variety on advertising effectiveness	• Between-subjects experiment design • Parallel mediation regression

were employed to train a compound measure using XGBoost (Chen and Guestrin 2016).

To accurately extract the 18 visual variety features, it is essential to segment the video into analyzable units, specifically shots and frames. A video consists of a series of frames grouped into shots, with the sequence of these shots collectively forming the complete video (Schwenzow et al. 2021). As a first preprocessing step in our feature extraction pipeline, we segmented the raw video into individual shots using the video splitting algorithm `PySceneDetect` (Brandon Castellano 2014). `PySceneDetect` is widely used in practice (Mourchid et al. 2019; Chen et al. 2024; Ye et al. 2018), and considered one of the best-performing video splitting algorithms (Wang et al. 2021; Gao et al. 2019). It uses rolling average of differences in HSL colorspace combined with thresholding to detect shot changes (Brandon Castellano 2014). For each detected shot, we selected three keyframes: the first, middle, and last frame (see Figure 2). This selection aims to capture the most representative moments within each shot. The keyframes extracted through this process are the basis for our subsequent feature extraction.

Temporal Dynamics. Temporal dynamics in videos refer to the variations in shot composition over time. These dynamics are critical for understanding how a video evolves, highlighting the continuity or abrupt changes in visual content. We extracted temporal dynamics at four different levels:

- **Within Shot:** This metric evaluates the average difference within a single shot, indicating the level of variation within individual shots.

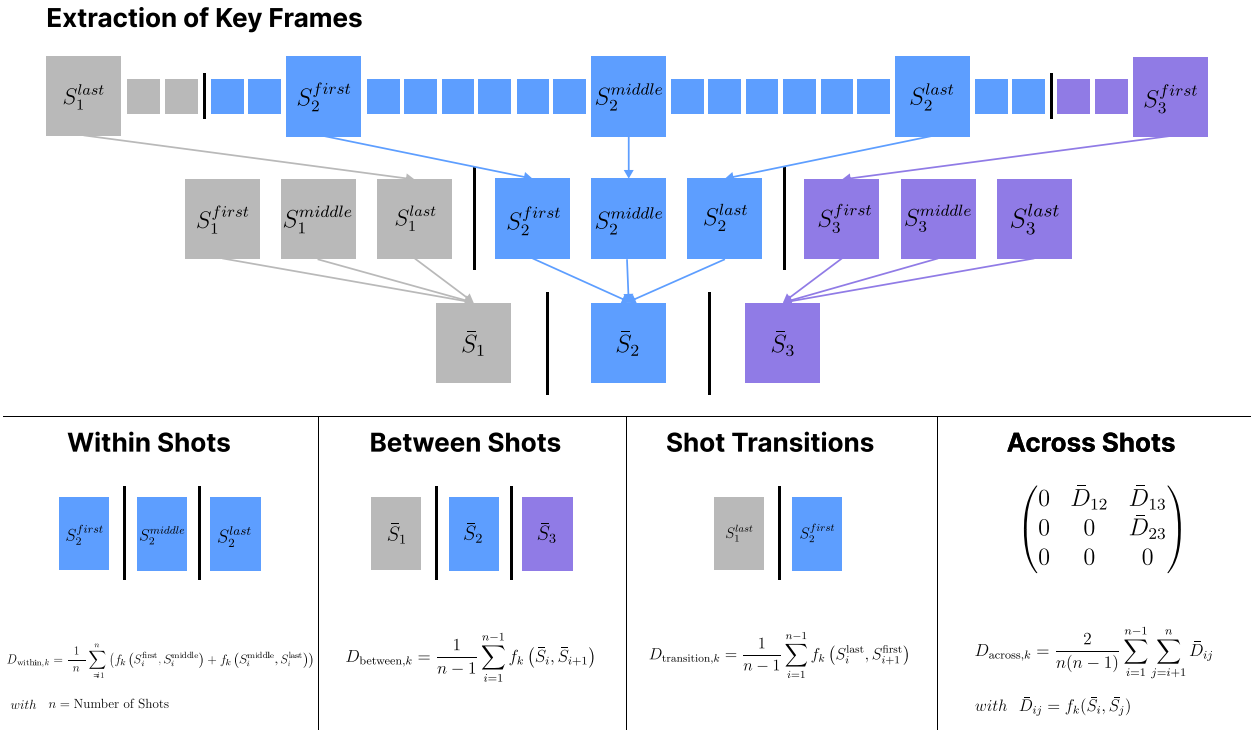
- **Between Shot:** This metric measures the average difference between neighboring shots in a video. It captures the extent of variation across feature granularity levels between adjacent shots.

- **Shot Transitions:** This metric assesses the average difference in transitions between shots and reflects the strength of transitions.

- **Across Shots:** This metric quantifies the average difference between each shot and

all other shots in the video. It provides a measure of how distinct each shot is relative to the entire video.

Figure 2: FEATURE CREATION TO CAPTURE TEMPORAL DYNAMICS



Note: Each color indicates different shots. Each box indicates frames. The first and last frame of each shot are selected with an offset of 2 frames. The function $f_k(S_i, S_j)$ computes the difference between two image frames. The function $f_k : ([0,255]^{H\times W\times C}, [0,255]^{H\times W\times C}) \to [0,1]$, where S_i and S_j are image frames of size $H \times W \times C$. To avoid any potential distortions caused by transitional effects, we apply an offset of two frames when selecting the first and last frames, which corresponds to a 0.067 second offset for a 30 FPS video.

Feature Granularity. Feature granularity defines the richness and detail of visual elements across multiple layers within a video. By evaluating visual features at various levels of detail, from pixel-level analysis to more abstract object representations, we capture the visual variety in content. This dimension of feature granularity focuses on *what* viewers see and is orthogonal to temporal dynamics, which define *when* and for *how long* each visual element or shot is viewed:

- **Pixel Level:** At the pixel level, frames are compared on a per-pixel basis using the Manhattan distance metric. This method captures fine-grained visual differences by quantifying the exact changes in color and intensity between corresponding pixels in

different frames. This approach is sensitive to small variations and helps capture subtle changes in texture, lighting, and shading across frames.

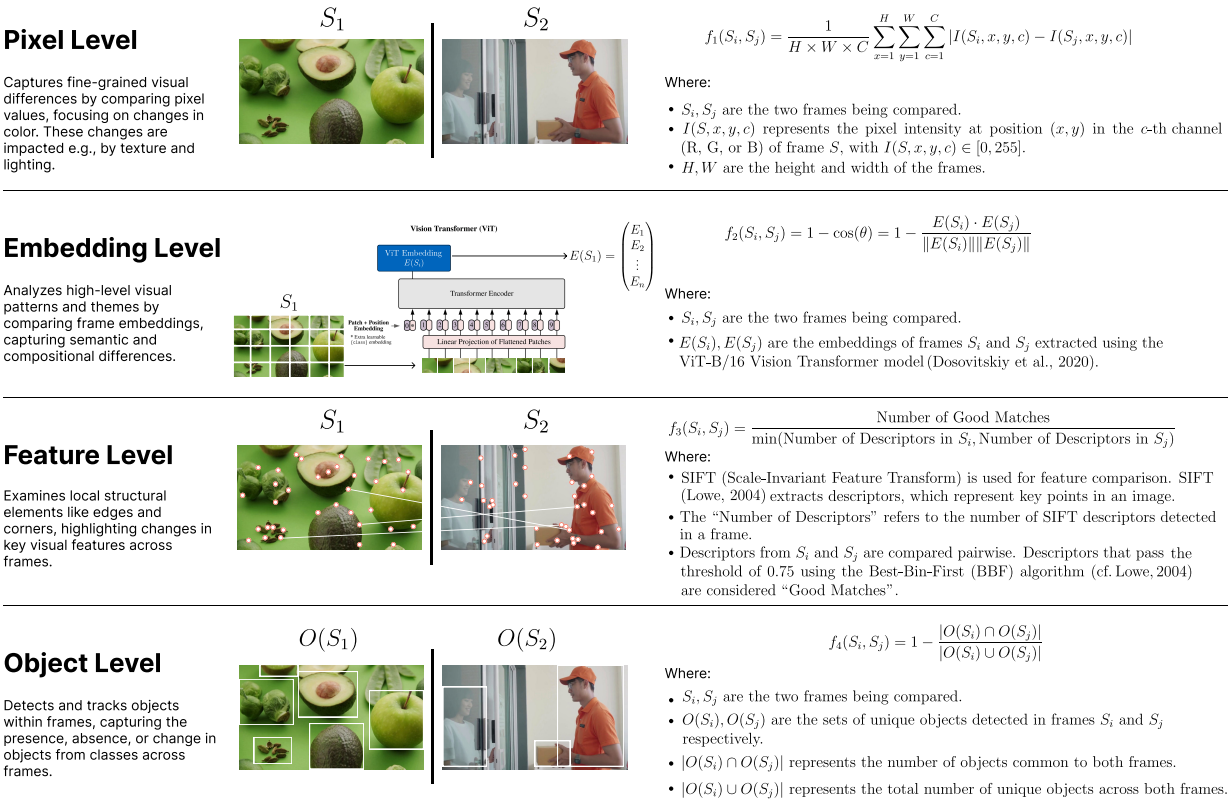
- **Embedding Level:** At the embedding level, we utilized the ViT-B/16 Vision Transformer to generate high-dimensional embeddings (dim=768) of each frame (Dosovitskiy et al. 2020).[3] Frames are then compared using cosine similarity between their embeddings. This approach captures more abstract and semantic differences by focusing on the higher-level visual patterns and concepts that the Vision Transformer identifies. The embedding-level comparison is particularly useful for recognizing changes in the overall scene composition or the presence of distinct visual themes.

- **Feature Level:** At the feature level, we employed the Scale-Invariant Feature Transform (SIFT) to extract local features from each frame (Lowe 2004). By comparing these local features across frames, we can measure differences in structure and key visual elements, even when there are variations in scale or rotation (Lowe 2004). This method captures the persistence or change of prominent features such as edges, corners, and textures, providing a detailed understanding of frame-level differences.

- **Object Level:** At the object level, we compared the objects identified in each frame using a pretrained object detection model, specifically YOLOv10, which was trained on the COCO dataset, which includes 80 distinct everyday objects such as person, car, and apple (Wang et al. 2024). This level captures the change in objects within the frames. By focusing on detected objects, this approach emphasizes changes that are most relevant to the viewer's understanding of the scene, such as the appearance of new objects or removal of existing objects (see Figure 3).

[3]The weights are available here: https://huggingface.co/google/vit-base-patch16-224.

Figure 3: FEATURE CREATION AT DIFFERENT FEATURE GRANULARITY



Note: The dots in the Feature Level frames represent SIFT descriptors. The lines between the dots represent good matches between two frames. The bounded boxes in the Object Level frames represent identified objects with a respective class.

Consolidation into a visual variety compound measure

To reliably predict a visual variety compound score for any video, we trained a machine learning model using the 18 extracted features. These features encompass both the fine-grained temporal dynamics (from local to global) and different levels of feature granularity (from low-level to high-level) that define visual variety (see Figure 5, feature 17 representing the number of shots and feature 18 the average shot duration, are both not depicted in the matrix). Due to the wide range of features, they were consolidated into a single compound measure through the machine learning model. The model was trained using a supervised learning framework, leveraging ground truth labels from 150 randomly selected and manually labeled videos. The labeling process was carried out by three research assistants blind to the

study's hypotheses, using a visual variety scale rated on a 5-point Likert scale. The scale included four items (items ranging from "not complex - complex", "not crowded - crowded", "no variety - variety", "simple - complex"), adapted from (Lee et al. 2018, Cronbach's alpha = .92).The labelled dataset served as input to train an XGBoost model, a robust and efficient gradient boosting algorithm (Chen and Guestrin 2016; Hartmann 2021). XGBoost is particularly well-suited for this task due to its ability to handle complex data structures, its built-in regularization to prevent overfitting, and its high computational efficiency (Chen and Guestrin 2016).

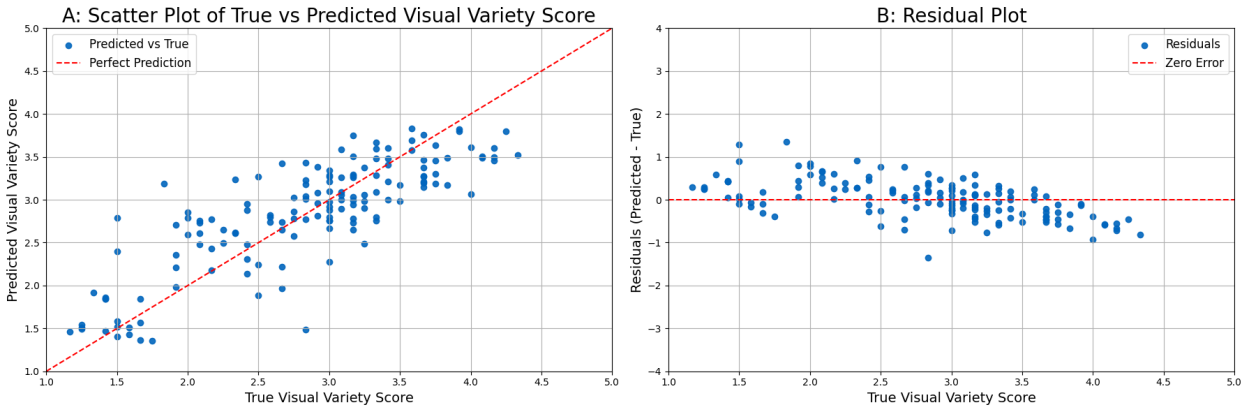
Table 2: EVALUATION MEASURES FOR XGBOOST VISUAL VARIETY MODEL

Pearson Correlation Coefficient (PCC)	R-squared (R2)	Mean Squared Error (MSE)
0.823	0.677	0.193

The XGBoost model demonstrates strong predictive performance, as evidenced by the Pearson Correlation Coefficient (PCC) of 0.823 (see Table 2), indicating a strong linear relationship between the predicted and true values (see Figure 4). The residual plot further supports this, showing consistent predictive accuracy across the scale of visual variety scores, with only a slight increase in residuals at higher scores. The average Mean Squared Error (MSE) on a 10-fold cross validation is 0.193 (see Table 2) and shows that the average squared difference between the predicted and actual values is low on unseen data, confirming the model's accuracy. This slight drop in performance for high visual variety scores may be attributed to the smaller number of data points in this range, but overall, the model maintains solid predictive power.

The SHapley Additive exPlanations (SHAP) values provide further insights into the contributions of each of the 18 extracted features to the prediction of visual variety (Lundberg and Lee 2017). Figure 5 illustrates the relative importance of each feature, as determined by the SHAP values. The top five features influencing visual variety, in decreasing order, are the *number of shots, mean shot duration, within shot embedding, transition embedding, and within shot pixel* (see Web Appendix B Figure W1 for detailed SHAP chart). The

Figure 4: XGBOOST PERFORMANCE FOR VISUAL VARIETY PREDICTION



Note: The figures were created using a 10-fold cross-validation with n=150. For each iteration, 9 parts are used to train the model while the remaining part is used to test the model. The scatter and residual plots depict the model's predictive performance, with predictions aggregated across all folds to provide a comprehensive view of accuracy and error.

distribution of SHAP values shows that all features are relevant for the prediction, with embedding and pixel-level features having a slightly greater impact compared to SIFT and object-level features (see Figure 5). This underscores the robustness of our feature set, which effectively captures both temporal dynamics at different feature granularity levels, allowing the model to accurately quantify visual variety across a diverse range of video content. The trained compound measure was used to predict the visual variety of all videos in the dataset, forming the basis for the subsequent econometric analyses.

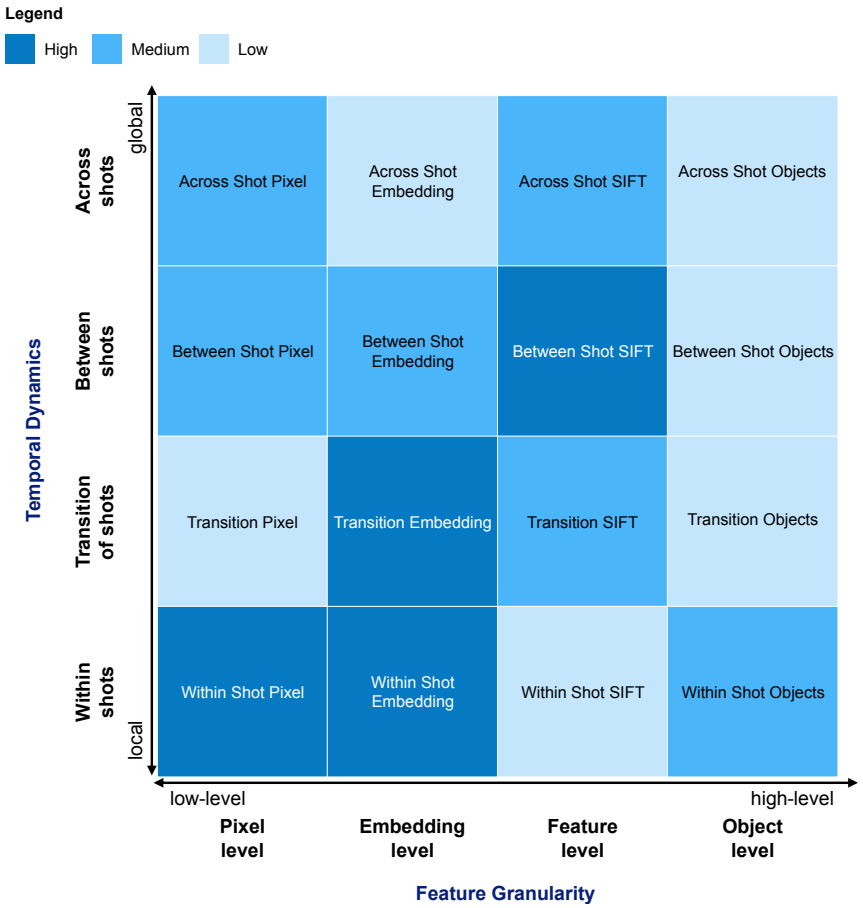
Study 1: Visual variety in TV ads

In study 1 we assess the relationship between visual variety and advertising effectiveness of 910 videos using survey data on real-world TV ads to investigate H1. We isolate the effect of visual variety by implementing numerous textual, audio and visual controls.

Data

We partnered with a leading market research and analytics company to obtain a dataset comprising survey responses from over 27,000 participants regarding their perception of multinational TV ads. They provided us with 1,049 TV advertising videos from almost 30

Figure 5: FEATURE MATRIX FOR VISUAL VARIETY



Note: The color grading of the matrix cells is based on the SHAP values as follows: Low ($< .03$), Medium ($.03 - .04$) and High ($> .04$). See Web Appendix B Figure W1 for detailed SHAP chart.

countries across the globe (i.e., US, Europe, Asia), featuring multiple brands and product categories. In a large international survey, our partner gathered information on consumer reactions to the advertising videos, including their purchase and sharing intention, resulting in a total of more than 200,000 observations. After exclusion of videos without multiple shots (i.e., 57 videos with 1 shot only) and list-wise deletion of videos with NA entries (i.e., 82 videos with NAs in regression-relevant variables) a set of 910 ads remains as the final dataset for analysis. Study 1 draws on this dataset to test the predicted inverted U-shaped effect of visual variety on advertising effectiveness.

Control variables

To accurately isolate the effect of visual variety on advertising effectiveness, we control for several factors that potentially also influence consumer response to video ads. Our approach involves a multi-model video analytics pipeline to extract a comprehensive set of visual, textual, and audio control variables. Table 3 provides and overview of these variables, along with their definitions. For detailed information on the extraction process and tools used, please refer to Web Appendix B.

Structural features. In terms of structural controls, we focus on the video duration to capture the total stimulus duration (e.g., MacInnis et al. 2002; Fossen and Schweidel 2017; McGranaghan et al. 2022).[4]

Audio features. Audio and narrative features significantly influence content success by shaping perceived emotionality and complexity (Berger et al. 2021). We control for key emotional indicators from text analysis, namely arousal and valence, which are well-documented in advertising research (e.g., Olney et al. 1991; Teixeira et al. 2012; Akpinar and Berger 2017). Additionally, we account for the audio's tempo, capturing its pace as either upbeat and dynamic or slow and relaxed (e.g., Teixeira et al. 2014). Drawing inspiration from Yang

[4]We do not control for the number of shots or the average shot duration as this part of our compound visual variety measure.

Table 3: VARIABLE OVERVIEW

Variable	Description
Dependent variables	
Purchase intention (TV)	Purchase intention is measured on a 5-point Likert scale for the TV ads.
Sharing intention (TV)	Sharing intention is measured on a 5-point Likert scale for the TV ads
Clicks (Facebook)	Clicks is measured as the actual number of clicks for the respective Facebook ad
Purchases (Facebook)	Purchases is measured as the actual number of purchases for the respective Facebook ad
Main variable	
Visual variety	Compound measure of visual variety based on the 18 extracted features, measured on a range of 1 to 5
Visual variety^2	Squared term of visual variety
Structural features	
Duration	Total length of a video ad, measured in seconds. We applied a natural logarithm transformation with a $log(x+1)$ adjustment to the variable
Audio features	
Arousal	Arousal level in ad audio measures the degree of excitement conveyed through the choice of words in the transcript of spoken content (Warriner et al. 2013)
Valence	Valence in ad audio measures the pleasantness or emotional appeal of the words used in the transcript, indicating the spectrum of emotions invoked by the language, ranging from unhappy to happy (Warriner et al. 2013)
Energy level	Score that indicates how energetic the audio of the ad is, taking into account spoken text and music (Yang et al. 2022)
Frequency variation	Score that indicates how strongly (%) the voice frequency (in Hz) varies compared to the ad's average frequency. We applied a natural logarithm transformation with a $log(x+1)$ adjustment to the variable
Tempo	Score that indicates how much tempo an audio has, taking beats per minute into account (Askin and Mauskapf 2017; Yang et al. 2022)
Transcript complexity	Transcript complexity is measured using the grade level score developed by Kincaid et al. (1975). This metric estimates the school grade level necessary for comprehension of a text and is applied to the transcribed audio of the video ads.
Visual features	
Face area	Average percentage of the screen that is covered by human faces. We applied a natural logarithm transformation with a $log(x+1)$ adjustment to the variable
Unique faces	Number of unique faces that have been identified over the entire video. We applied a natural logarithm transformation with a $log(x+1)$ adjustment to the variable
Brightness	Brightness refers to the perceived intensity or luminosity of an image and is calculated as the average across keyframes (first, middle, and last frame) of each video ad. It serves as a proxy for visually induced emotionality (Wilms and Oberfeld 2018)
Saturation	Saturation measures the intensity or purity of colors in an image and is calculated as the average across keyframes of each video ad. It serves as a proxy for the vibrancy of emotionality (Wilms and Oberfeld 2018)
Image aesthetics	NIMA quantifies the aesthetic quality of images by predicting a distribution of viewer ratings on a scale from 1 to 10 (Talebi and Milanfar 2017). We calculate the NIMA score as average across all keyframes of a video.
Entropy	Average change in image entropy between keyframes of detected shots across the video is used to measure the unpredictability or randomness in the visual content of video ads, relating to the perceived processing complexity (Wu et al. 2013; Overgoor et al. 2022).
Depth	Average image depth across the keyframes of a video ad indicates the capacity of the image to display a range of colors or shades of gray and is used as a proxy for image quality. Greater image depth allows for richer and more detailed color representation, enabling more nuanced images
Further controls	
Views (Facebook)	Number of views a specific video has accumulated over time as an indicator for its popularity. We applied a natural logarithm transformation with a $log(x+1)$ adjustment to the variable

et al. (2022), we employ a custom classifier to assess the energy levels in video soundtracks. To evaluate transcript complexity, we utilize the Flesch-Kincaid grade-level score (Kincaid et al. 1975), similar to approaches in prior studies (Berger et al. 2021). We also measure the variation in voice frequency throughout the ad, reflecting changes in vocal intensity.

Visual features. Given the impact of visual features on consumer reactions (Tellis et al. 2019; McGranaghan et al. 2022; Becker et al. 2023), we implement several visual controls to account for varying elements within the video ads. We compute the average entropy of scene frames to measure the perceived processing complexity of the visual content (Wu et al. 2013; Overgoor et al. 2022). To control for the effect of image quality and aesthetics on advertising effectiveness, we use average depth and neural image assessment (NIMA) scores, which measure the aesthetic appeal of key frames across all shots in a video ad (Talebi and Milanfar 2018*a*). To account for the emotional influence of frame compositions, we rely on the findings by Wilms and Oberfeld (2018), which suggest that brightness, saturation and hue of images significantly influence viewer emotions. Due to multicollinearity concerns, hue is excluded from our controls. Additionally, we measure the number of unique faces and calculate the average screen area they occupy, controlling for the influence of human presence on advertising effectiveness (Schwenzow et al. 2021). Table 4 provides summary statistics across all variables of the TV dataset.

Results

We estimate OLS regression models for the TV data's continuous rating scales, averaging all participants' responses on the video level. Both models are estimated with further controls and aim to assess the relationship between the dependent variables and our main independent variable *visual_variety*:

$$purchase_{ijk} = \beta_0 + \beta_1 visual\_variety_i + \beta_2 visual\_variety_i^2 + X_i\gamma + country_j + brand_k + \epsilon_i \quad (1)$$

where $i = 1, ..., N$ indicates the different ads and β_1 and β_2 measure the relationship between visual variety and the estimated *purchase intention*. X_i is the matrix of control variables; the fixed effects $country_j$ account for country-specific differences across advertising videos, and $brand_k$ accounts for brand fixed effects. ϵ_i is the error term. Analogous to Equation 1, we estimate a second OLS model for our second dependent variable *sharing intention*.

Table 5 shows our results. We find a consistent and significant inverted U-shape across both dependent variables (i.e., sharing intention and purchase intention)[5]. This confirms H1, that optimum levels of visual variety maximize advertising effectiveness as measured by purchase and sharing intentions in the TV advertising context. For sharing intention, model (1) indicates 3.1 as the optimum level of visual variety to generate viewer engagement (visual variety: 1.273, $p < .001$; visual variety^2: $-.199$, $p = .002$), and model (2) shows an optimum score of 3.25 for evoking purchase intention with TV ads (visual variety: 1.248, $p = .006$; visual variety^2: $-.192$, $p = .014$). Recall that our visual variety compound measure is bound by 1 and 5, where 5 indicates the maximum level of visual variety. All other control variables do not demonstrate a significant impact on advertising effectiveness.

Discussion

The results of study 1 provide robust evidence for the non-linear relationship between visual variety and the effectiveness of video ads, as measured by both purchase and sharing intentions. The results align with Berlyne's aesthetic theory, related studies, and hypothesis H1,

[5]Robustness checks confirm that the results remain consistent when excluding duration from the regression analysis, as the compound measure of visual variety already includes duration as a function of the two features: average shot duration and number of shots.

Table 4: SUMMARY STATISTICS FOR TV DATASET

Variable	Mean (SD)	Min	Max
Dependent variables			
Sharing intention	3.104 (0.435)	1.886	4.219
Purchase intention	3.194 (0.457)	2.020	4.233
Main independant variables			
Visual variety	3.059 (0.340)	1.546	3.846
Structural features			
Duration (in seconds)	29.801 (15.575)	5.960	173.480
Audio features			
Arousal	0.724 (0.142)	0.199	1.048
Valence	0.616 (0.146)	0.028	0.979
Energy level	0.273 (0.152)	-0.044	0.705
Frequency variation	0.260 (0.162)	0	0.568
Tempo	114.836 (13.835)	71.150	170.624
Transcript complexity	4.583 (6.303)	-15.700	91.000
Visual features			
Face area	0.048 (0.038)	0	0.268
Unique faces	1.614 (0.606)	0	3.091
Entropy	1.369 (1.764)	0.051	7.547
Depth	4.233 (2.136)	0.626	17.382
Brightness	0.465 (0.160)	0.061	0.936
Saturation	0.319 (0.124)	0.016	0.851
Image aesthetics	4.638 (0.280)	3.819	5.533

which suggest that a moderate level of visual variety maximizes advertising effectiveness. The identified optimal visual variety score for purchase intention of 3.25 is slightly above the average visual variety level of 3.1 observed in real-world TV ads. This represents a minor deviation of about 5% from the optimal level, indicating that TV ads are closely aligned with the optimal balance of visual variety (see Table 4 and Web Appendix C). Study 1 has been conducted in a more controlled setting, employing a large-scale international consumer survey to evaluate intentional behavior towards TV ads. To validate these insights in a more realistic settings and to understand how they translate into actual consumer behavior, study 2 examines real-world interaction data from Facebook video ads. This step is essential as it

Table 5: REGRESSION RESULTS FOR TV ADS DATASET

	Sharing intention	Purchase intention
	(1)	(2)
Visual variety		
Visual variety	1.273*** (.365)	1.248** (.453)
Visual variety^2	−.199** (.062)	−.192* (.078)
Structural features		
Duration	.001 (.001)	−.000 (.002)
Audio features		
Arousal	−.009 (.124)	.112 (.136)
Valence	−.078 (.107)	−.052 (.133)
Energy level	.024 (.105)	−.017 (.121)
Frequency variation	.090 (.090)	.093 (.112)
Tempo	−.001 (.001)	−.000 (.001)
Transcript complexity	.001 (.002)	.001 (.002)
Visual features		
Face area	.133 (.391)	.229 (.469)
Unique faces	−.004 (.003)	−.040 (.003)
Entropy	−.003 (.008)	−.007 (.010)
Depth	−.002 (.008)	.007 (.009)
Brightness	−.040 (.110)	−.014 (.137)
Saturation	−.015 (.142)	−.109 (.176)
Image aesthetics	−.001 (.062)	−.002 (.065)
Country fixed effects	Yes	Yes
Brand fixed effects	Yes	Yes
Observations	910	910
R^2	.95	.94

Note: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$
Intercept omitted from Table. Robust standard errors in parentheses.
All Variance Inflation Factor (VIF) values are below 1.5.

helps to demonstrate the validity of our results and ensure that our theoretical model holds up under practical conditions.

Study 2: Visual variety in real-world Facebook ads

Study 2 aims to extend the insights from study 1 by examining how visual variety influences advertising effectiveness in the dynamic and uncontrolled environment of real Facebook ad interactions. By leveraging field data, this study tests the generalizability of the inverted U-shaped relationship identified. In identical fashion to study 1, we extract visual variety scores for real Facebook ads and respective audio, textual and visual controls.

Data

The Facebook dataset consists of 2,766 video ads displayed on Facebook. An industry partner provided the Facebook ads along with actual performance metrics, including number of views, clicks, and purchases, as well as allocated budget. After exclusion of videos without multiple shots (i.e., 469 videos with 1 shot only) and list-wise deletion of videos with NA entries (i.e., 116 videos with NAs in regression-relevant variables) a set of 2,181 ads remains as the final dataset for analysis. In contrast to study 1, our primary dependent variable is the number of clicks as a direct measure of ad effectiveness. While we also consider purchases as dependent variable, purchases are influenced by other factors not captured in our data, such as landing page effectiveness, pricing strategy, and the user experience during shopping, which we are unable to observe and account for. This can skew the relevance of purchases as direct indicator of the relationship between visual variety and video ad effectiveness. Table 6 provides summary statistics for all variables of the Facebook ad dataset [6].

[6]We filter out static and minimally animated banner ads and solely focus on videos with more than one shot to ensure that our findings are not influenced by fundamental differences in video ad design between TV and Facebook ads.

Table 6: SUMMARY STATISTICS FOR FACEBOOK DATASET

Variable	Mean (SD)	Min	Max
Dependent variables			
Clicks	2,555.313 (13,930.41)	1.000	392,480
Purchases	92.006 (531.643)	0	12,937
Main independet variables			
Visual Variety	2.981 (0.378)	1.341	3.840
Audio features			
Frequency Variation	0.239 (0.190)	0	0.632
Unique Faces	1.482 (0.594)	0	3.135
Face Area	0.102 (0.072)	0	0.403
Arousal	0.639 (0.142)	0.245	1.068
Valence	0.485 (0.145)	0.085	0.976
Visual features			
Entropy	1.013 (0.999)	0.003	7.199
Energy level	0.270 (0.147)	-0.114	0.691
Depth	4.219 (2.061)	0.188	14.288
Brightness	0.615 (0.110)	0.243	0.966
Saturation	0.256 (0.082)	0.076	0.673
Image aesthetics	4.916 (0.279)	3.903	6.079
Transcript complexity	1.177 (6.002)	-15.700	46.100
Structural features			
Duration	24.814 (10.925)	5.040	209.680
Views	25,905 (1418.461)	0	39,033,135

Results

Due to overdispersion of our dependent variables, clicks and purchases (see Table 6), we estimate two negative binomial regression models, which result in better model fits than Poisson regressions (see also Akpinar and Berger 2017; Villarroel Ordenes et al. 2019; Hartmann et al. 2021). We include the same controls as in study 1, add the number of views as an additional control variable, and include fixed effects on brand and year-month level:

$$clicks_{ijk} = \beta_0 + \beta_1 visual\_variety_i + \beta_2 (visual\_variety_i)^2 + X_i\gamma + time_j + brand_k + \epsilon_i \quad (2)$$

where $i = 1, ..., N$ indicates the different ads and β_1 and β_2 measure the association of visual variety with the estimated count of clicks; the other X_i is the vector of control variables; $time_j$ represents year-month fixed-effects of each video; and $brand_k$ is the time-invariant brand fixed-effect. We estimate the same negative binomial regression model for purchases. All controls were z-standardized for increased interpretability. Table 7 shows our results.

Table 7: REGRESSION RESULTS FOR FACEBOOK ADS DATASET

	Clicks	Purchases
	(1)	(2)
Visual variety		
Visual variety	1.227* (.563)	−.730 (.835)
Visual variety^2	−.245* (.102)	.147 (.152)
Structural features		
Duration	.177** (.058)	.008 (.037)
Audio features		
Arousal	−.075* (.033)	.042 (.034)
Valence	−.002 (.026)	−.046 (.030)
Energy level	.040 (.045)	.023 (.046)
Frequency variation	−.041 (.027)	−.011 (.033)
Tempo	.096** (.033)	.003 (.039)
Transcript complexity	−.048 (.032)	.001 (.030)
Visual features		
Face area	.015 (.033)	−.038 (.036)
Unique faces	.057 (.040)	.040 (.041)
Entropy	−.072*** (.022)	−.009 (.022)
Depth	−.026 (.027)	.002 (.027)
Brightness	.012 (.028)	.004 (.033)
Saturation	−.027 (.033)	−.038 (.039)
Image aesthetics	−.022 (.026)	.003 (.030)
Further controls		
Views	2.189*** (.040)	2.563*** (.058)
Brand fixed effects	*Yes*	*Yes*
Time fixed effects	*Yes*	*Yes*
Observations	2,181	2,171
Pseudo R^2	0.143	0.257

Note: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$
Intercept omitted from Table. Robust standard errors in parentheses.
All Variance Inflation Factor (VIF) values are below 1.5.

Similar to study 1, we observe a consistent inverted U-shape for the visual variety of Facebook video ads and the number of clicks [7]. Model (1) indicates that optimal level of visual variety to maximize clicks is 2.5 (visual variety: 1.227, $p = .029$; visual variety^2: $-.245$, $p = .016$). Several other variables also significantly relate to clicks. First, as expected, views are highly predictive of the number of clicks (views (log): 2.289, $p < .001$). Second, the level of arousal in the audio negatively affects clicks (arousal: $-.075$, $p = .021$). Other controls, such as entropy, duration, and tempo, also contribute to click variations, though their effects are relatively minor (entropy: $-.072$, $p < .001$; duration: .177, $p = .002$; tempo: .096, $p = .004$), underlining that while these features influence clicks, visual variety remains a predominant factor. For the number of purchases (model (2)), visual variety and its squared term do not significantly link to actual purchasing behavior, highlighting a potentially different dynamic for this more complex consumer action compared to clicks (visual variety: $-.276$, $p = .382$; visual variety^2: .316, $p = .334$).

Discussion

Comparing the optimal level of visual variety on Facebook for maximizing clicks, which is 2.5, to the observed average level of 2.98 suggests that advertisers are utilizing visual variety levels that are approximately 20% higher than optimal for click generation (see Table 6 and Web Appendix C). This deviation implies that current marketing practices might not always align with maximizing clicks. While clicks are an essential metric for evaluating immediate ad performance, they primarily focus on sales-related outcomes. In contrast, some campaigns are designed with the long-term objective of building or strengthening a firm's brand identity, which aims more at awareness. Such campaigns might prioritize overall impressions or engagement, such as ad sharing, over direct clicks (Keller 1993; Keller and Lehmann 2006). Our findings from study 1 reveal an inverted U-shaped relationship between

[7] Robustness checks confirm that the results remain consistent when excluding duration from the regression analysis, as the compound measure of visual variety already includes duration as a function of the two features: average shot duration and number of shots.

visual variety and both sharing intentions and purchase intentions. Study 2 confirms this relationship only for clicks, an upper-funnel metric. Ad clicks and click-through rates (CTR) are widely recognized as key indicators of advertising effectiveness, and they are among the most commonly tracked metrics in online advertising due to their strong correlation with campaign success (Aribarg and Schwartz 2020; Johnson et al. 2017; Hartmann et al. 2021). However, this relationship is not confirmed for purchases in study 2, a lower-funnel metric, which suggests that additional visual or content-related stimuli encountered on the landing page could influence conversion rates and underscores the complexity of visual variety's influence across different consumer responses.

In conclusion, both study 1 and study 2 provide significant evidence supporting H1: an optimal (i.e., medium) level of visual variety maximizes advertising effectiveness, outperforming both low and high levels. Study 2 reinforces this finding with real-world data, enhancing the robustness and generalizability of our results. To mitigate potential endogeneity arising from managers adjusting visual variety based on expected customer engagement, we conduct a controlled lab experiment in study 3. This approach allows us to directly manipulate visual variety for highly comparable videos, thereby providing a clearer test of causality and mitigating concerns over the external validity of our results from previous studies. Additionally, we examine the underlying mechanisms through which visual variety influences advertising effectiveness, addressing H2, H3, and H4.

Study 3: Causality and mechanism evidence

Study 3 is designed to fulfill two primary objectives. First, to establish causality, we manipulate visual variety in self-produced video ads and observe its effects on advertising effectiveness. Second, we test the underlying processes behind the observed effects. We have hypothesized that visual variety drives advertising effectiveness through two countervailing effects. On the one hand, visual variety increases stimulation, which should drive advertising effectiveness, up to a point of over-stimulation (H2). On the other hand, too high visual

variety hinders ad comprehension, which negatively impacts effectiveness (H3). These effects are assessed in a parallel mediation to understand their combined effect (H4).

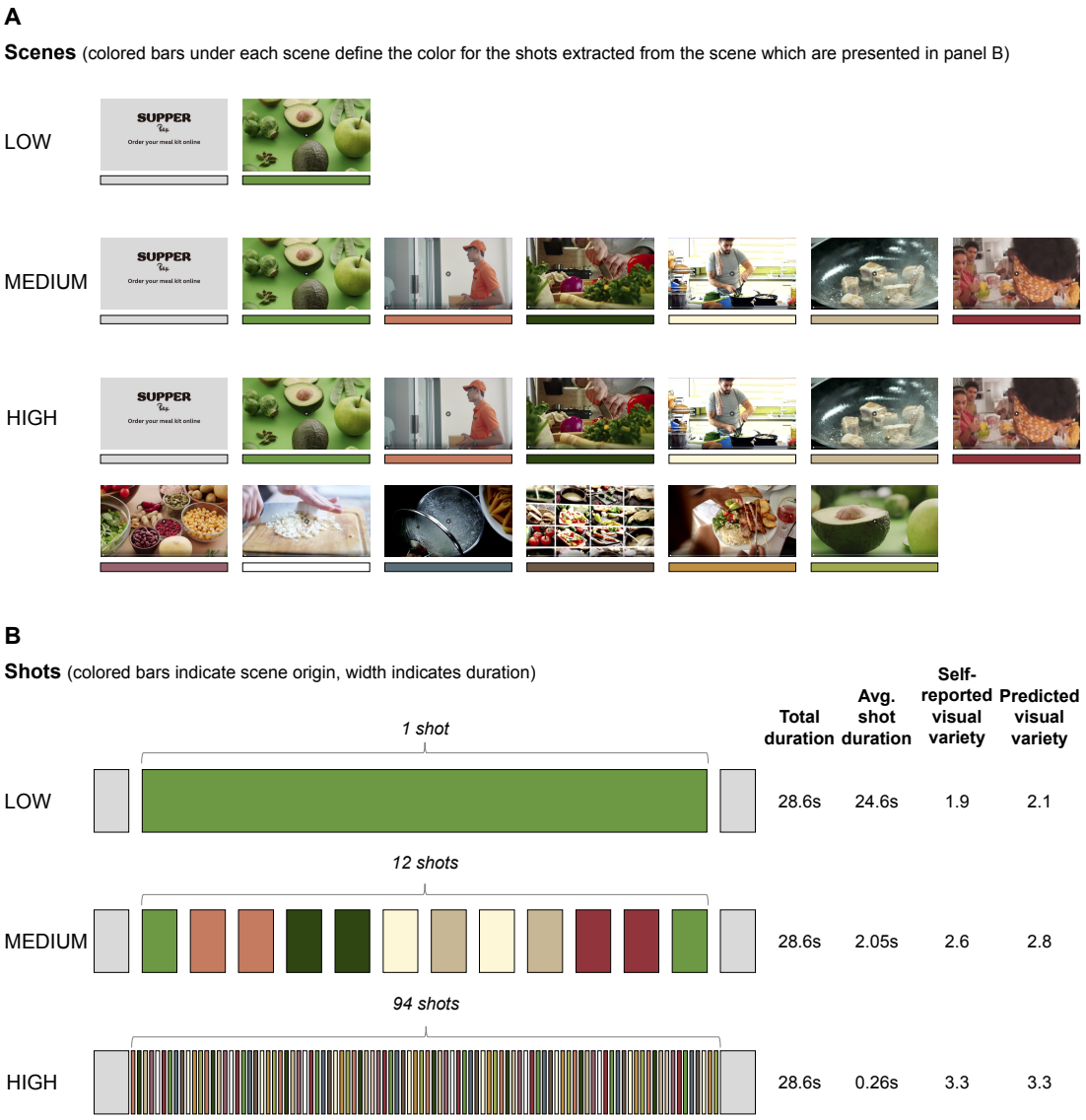
Methods

Our sample comprises 157 participants (52% female, average age: 38). Each participant was randomly assigned to one of three conditions in a 3 cell between-subject design on Prolific (visual variety: low, medium, high). Web Appendix D provides additional details on the experimental setup and exclusion criteria. All participants were shown a video ad for a fictitious meal-kit delivery service (brand name: SUPPER BOX). The video ads were crafted using professionally produced stock footage scenes obtained from Canva to ensure a high level of production quality across all conditions. Each ad maintained a consistent video duration, as well as identical opening and closing shots. We focused on the visual variety features of our compound measure that could be manipulated in post-production without altering the original content's professionalism[8]. In total, we manipulated 14 out of the 18 features of the compound visual variety measure (see Methodology chapter for detailed description of features). Our manipulation targeted (1) the number of shots, (2) average shot duration, (3) visual variety in transitions between shots, (4) visual variety between different shots, and (5) visual variety across all shots of a video. For (3), (4), (5) manipulations were effective at the pixel, embedding, feature, and object level.

Figure 6 illustrates the video composition for each condition. Panel A displays the stock-video scenes selected for each condition. The low condition comprises only two scenes: a static image of the fictitious brand with a call-to-action (i.e., "Order your meal kit online"), along with one additional scene. The medium and high conditions were crafted using six and twelve distinct stock-video scenes, in addition to the static brand image scene. This increase in the number of scenes directly drives visual variety across the video (5), as it allows for

[8]We refrained from manipulating within shot visual variety as adjusting pixels or objects in post-production requires advanced editing techniques and could compromise the professionalism of the selected scenes.

Figure 6: VISUAL COMPOSITION OF THREE EXPERIMENTAL CONDITIONS



Note: Excluding the first and last shot of each video, which is held constant at 2 seconds each across all three conditions.

greater variation in objects, colors, and patterns. The other dimension of visual variety were manipulated by dividing the scenes into shots (low: 3 shots, medium: 14 shots, high: 96 shots), as well as adjusting their duration and order. This strategic shot arrangement involved sequencing shots in the high condition to ensure that the colors, patterns, and objects between two consecutive shots exhibited maximal contrast, as demonstrated in panel B of Figure 6. The effectiveness of these manipulations was confirmed through our visual

variety pipeline, which quantified visual variety with values of 2.1 for low, 2.8 for medium, and 3.3 for the high visual variety conditions.

After viewing the video ad, participants rated their click intention on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree), adapted from Yoo (2007) (*"How likely is it that you will click on the video ad to be redirected to the product website?"*). To examine the underlying mechanisms, participants assessed their comprehension using a 7-point Likert scale (*"I was able to fully understand and follow the storyline of the video"*), which was adapted from the narrative understanding scale of Busselle and Bilandzic (2009). Participants reported their perceived level of stimulation (1 = not at all, 7 = extremely) using a single-item scale adopted from Berger et al. (2021) (*"How stimulating did you find the video?"*). Lastly, participants evaluated the perceived level of visual variety using the same scale as described in the Methodology chapter used to label the ground truth of visual variety.

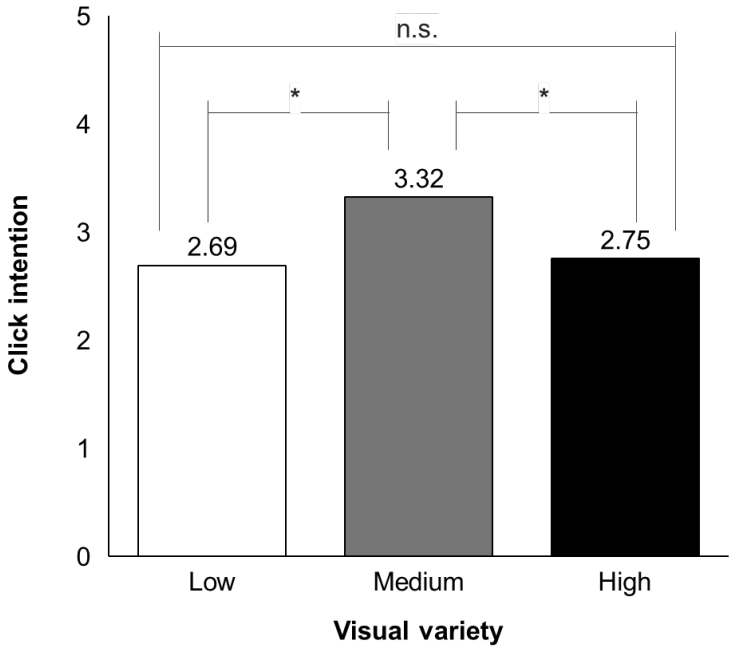
Results

Manipulation check. The comparison between perceived and calculated visual variety levels indicates a successful manipulation in our experiment. The video categorized as low with a calculated visual variety score of 2.1 received a mean perceived score of 1.9. Participants rated the medium condition video, which had a calculated visual variety of 2.8, as 2.6. The video in the high condition, with a calculated score of 3.3, was perceived as having a visual variety of 3.3.

Base effect. We conduct one-tailed t-tests to assess significant differences in click intentions across the three experimental conditions (Churchill and Iacobucci 2006). Our results indicate that mean click intention is significantly higher in the medium visual variety condition than in the low condition ($M_{medium} = 3.32$ vs. $M_{low} = 2.69$, $t(104) = -1.73$, $p = .043$, see Figure 7), and it is also higher compared to the high condition ($M_{medium} = 3.32$ vs. $M_{high} = 2.75$, $t(106) = 1.71$, $p = .045$, see Figure 7). These findings support our hypothesis that

visual variety impacts advertising effectiveness in an inverted U-shaped manner, providing robust evidence for hypothesis H1 and its causal nature.

Figure 7: EFFECT OF VISUAL VARIETY ON CLICK INTENTION

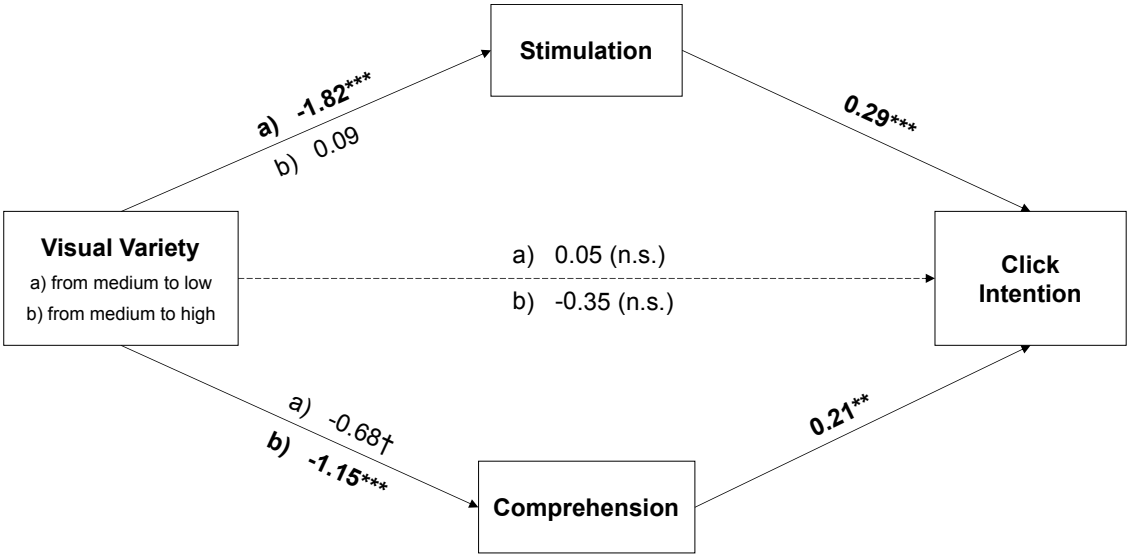


Note: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$

Mechanism evidence To explore whether stimulation and comprehension are driving the results, we perform a mediation analysis. Following the approach for multiple mediator analysis of Preacher and Hayes (2008) and similar to Berger et al. (2021) we use biased-corrected bootstrapping for our parallel mediation model (n = 5,000, 95% confidence interval (CI)). The mediation is successful if the confidence interval does not include zero (Preacher and Hayes 2008).

First, the mediation analysis, illustrated in Figure 8, provides substantial support for Hypothesis H2, confirming that stimulation mediates the relationship between visual variety and click intention for medium and low levels of visual variety (95% CIs: [−.922, −.197]). Specifically, we find that when visual variety decreases from a medium to a low level, stimulation notably decreases ($\beta_{stimulation_{a)}} = -1.817$, p < .001). Stimulation has a positive and

Figure 8: PARALLEL MEDIATION ANALYSIS



Note: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$

significant impact on click intention ($\beta_{stimulation_{click}} = 0.29$, p $< .001$), which aligns with H2a. Conversely, an increase from medium to high visual variety does not significantly enhance stimulation ($\beta_{stimulation_{b)}} = .087$, $p = .801$) and the mediation path becomes insignificant (95% CIs: [−.214, .233]). This outcome suggests that higher levels of visual variety do not further enhance viewer stimulation and thus click intention, indicating a limit to the benefits of increased visual variety. However, the expected negative impact of excessive stimulation on advertising effectiveness, indicative of a complete inverted U-shaped relationship (H2b), can not be confirmed as stimulation still shows a positive if non-significant relationship at high levels of visual variety. Second, The mediation analysis provides clear support for H3, that comprehension mediates the relationship between visual variety and advertising effectiveness. We observe that this mediation is significant for transitions from medium to high levels of visual variety (95% CIs: [−.524 to −.039]) and find that this shift (i.e., from medium to high levels of visual variety) leads to a substantial decrease in comprehension ($\beta_{comprehension_{b)}} = -1.15$, p $< .001$). While the transition from low to medium visual variety also results in a decrease in comprehension ($\beta_{comprehension_{a)}} = -.68$, p $= .050$), the effect is

not significant (95% CIs: [−.377, .011]), suggesting a more nuanced impact at lower levels of visual variety. In line with H3a, an increase in visual variety from a medium to a high level has a detrimental effect on comprehension. Additionally, the positive linear relationship between comprehension and click intention ($\beta_{comprehension_{click}} = .21$, p = .009), supports H3b, confirming that higher comprehension enhances advertising effectiveness. Finally, hypothesis H4 suggests that the effects of visual variety on advertising effectiveness are jointly mediated by stimulation and comprehension. Our parallel mediation analysis confirms this, identifying stimulation as the primary mechanism that explains the decreased click intentions when visual variety falls below the medium level, indicative of under-stimulation. Conversely, for levels of visual variety exceeding the medium, comprehension emerges as the critical factor, with diminished understanding leading to reduced click intentions. Collectively, these findings highlight that the medium level of visual variety optimally balances stimulation and comprehension, thereby maximizing advertising effectiveness.

Discussion

Study 3 confirms the results of the first two studies, that a medium level of visual variety in video ads drives the highest advertising effectiveness, measured through click intention. Our study provides causal evidence for our theory by clearly demonstrating that visual variety not only influences the effectiveness of ads, but that it does so through different mechanisms at different levels of visual variety. The mediation analysis reveals these mechanisms: stimulation significantly enhances click intentions when visual variety is moderate, effectively capturing viewer interest without overwhelming them. However, as visual variety increases beyond this optimal level, comprehension decreases, leading to reduced ad effectiveness. This study emphasizes the necessity for advertisers to find the right balance in visual variety to maximize both viewer stimulation and comprehension, ensuring the most effective visual communication.

General discussion

In an intensifying competition for consumer attention, advertisers need to strategically optimize their video ads to capture viewer interest amid widespread media saturation (Anderson and de Palma 2013; Galloway 2022; Lindsay 2024). One way to achieve this is through manipulating visual variety, which involves adjusting the composition of the visual elements such as colors, patterns, shapes, and objects, alongside the temporal dynamics across the video, including the number, duration, and arrangement of shots. But how does visual variety impact advertising effectiveness?

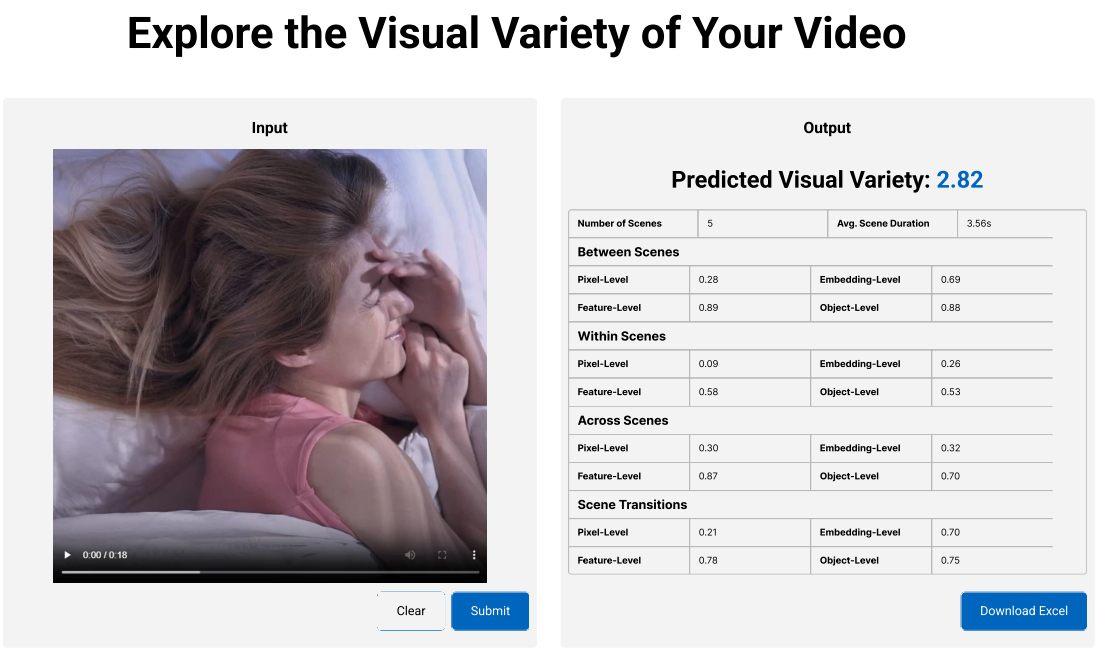
This research introduces a novel multi-faceted visual variety measure, spanning 18 features organized within a 4×4 matrix, focusing on varying feature granularity and temporal dynamics utilizing advanced machine learning techniques. Grounded in Berlyne's aesthetic theory, our studies reveal an inverted U-shaped relationship between visual variety and advertising effectiveness. Study 1 leverages a large dataset of real-world TV ads to show that an optimal level of visual variety relates to higher sharing and purchase intentions compared to both lower and higher levels of visual variety. These findings are substantiated by real-world performance data in study 2, where the number of clicks on Facebook video ads follows an inverted U-shaped relationship with visual variety. We use a broad set of controls, including factors like emotionality, visual quality, aesthetic appeal, and narrative complexity to further validate the robustness of the observed effects. Study 3 provides causal evidence for the relationship between visual variety and advertising effectiveness and demonstrates that stimulation and comprehension mediate this relationship in parallel. We find that insufficient stimulation from low visual variety decreases ad effectiveness, while excessive visual variety can undermine comprehension, which, in turn, reduces advertising effectiveness. This indicates that advertisers must carefully calibrate visual variety to maintain an optimal balance that avoids both under-stimulation and incomprehension.

Contributions and implications

This research offers several notable contributions. First, we demonstrate the practical value of machine learning in video analytics by developing and operationalizing a novel measure of visual variety. Our publicly available tool provides near-real-time visual variety predictions, enabling users to upload videos and instantly receive a compound visual variety score, along with detailed scores for all 18 individual features [LINK BLINDED FOR PEER REVIEW]. Figure 9 illustrates the interface of the visual variety extraction tool. This web-based interface introduces a new dimension of creativity to the video production process, enabling an iterative approach where content creators can make evidence-based adjustments to optimize visual variety for maximum advertising effectiveness. Managers can monitor and adjust visual variety in real-time during the campaign life-cycle. Additionally, we provide technical guidelines for practitioners on adjusting visual variety in both pre-production and post-production phases. During the storyboard and script development phases (pre-production), as well as during the process of video editing (post-production), creative teams can plan for the optimal level of visual variety. Web Appendix E provides detailed measures on how to manipulate visual variety in pre- and post-production processes.

Second, our findings offer practical insights into potential budget savings for practitioners by optimizing visual variety. Using financial performance data from our real-world Facebook dataset alongside the calculated optimal visual variety level of 2.5 (see study 2), we examine how optimal levels of visual variety decrease the cost per click (CPC) for advertisers. We conducted a post-estimation analysis using the negative binomial regression model from study 2 to estimate the number of clicks at the optimal visual variety level and at deviations of one standard deviation around this optimum. Due to the symmetry of the inverted U-shape, the deviations at the same level are identical. The analysis reveals substantial budget-saving opportunities for advertisers, as illustrated in Table 8. Specifically, we find that optimizing the visual variety level of videos that are one standard deviation above or below the optimal level can lead to potential CPC savings of 3.5%. These findings underscore

Figure 9: INTERFACE OF VISUAL VARIETY EXTRACTION TOOL



the importance of carefully calibrating visual variety, as even small adjustments could yield substantial financial savings, making this a critical consideration for maximizing advertising effectiveness in practice.

Third, to our best knowledge, our research provides the first empirical evidence of the multi-faceted nature of visual variety on advertising effectiveness. We thereby contribute to the literature investigating the impact of visual content on ad performance (e.g., (Teixeira et al. 2012; Rafieian and Yoganarasimhan 2022; McGranaghan et al. 2022). Beyond demonstrating the effect of visual variety on advertising effectiveness, we explain the underlying mechanisms and psychological processes that explain viewer responses, contributing to broader literature of consumer response to modern-day multimedia content in advertising (Grewal et al. 2021).

Limitations and future research

Several relevant research areas remain open for further investigation. First, the potential for endogeneity due to the complexities inherent in designing advertising campaigns must

Table 8: POST-ESTIMATION ANALYSIS OF COST-PER-CLICK SAVINGS FOR THE FACEBOOK DATASET

VOptimal visual variety for the Facebook dataset	2.504
Standard deviation of visual variety (SD)	0.378
Low visual variety (-1 SD)	2.126
High visual variety (+1 SD)	2.882
Clicks at deviation from optimal visual variety (+/- 1 SD)	32,127
Average cost-per-click (CPC) for low and high visual variety	$0.533
Clicks at optimal visual variety	33,284
CPC at optimal visual variety	$0.515
CPC saving potential	3.476%

Note: Average CPC is based on average clicks and cost across the entire Facebook dataset.
All other regression variables are held constant at their mean values.
Deviation of 2 SD from the optimum result in a saving potential of 13%.

be acknowledged. Advertisers may already be knowingly or unknowingly adjusting visual variety based on target audience, marketing objectives and past performance data, which could lead to a selection bias. For instance, more visually complex ads might be targeted at younger audiences on specific platforms to increase ad clicks. While our controlled lab experiment provides causal explanations for the observed effect, future research could further validate these findings in real-world settings across various platforms, thereby addressing endogeneity concerns and increasing external validity. Second, exploring differences between media channels presents another intriguing research avenue. Our findings suggest channel-specific optima for visual variety, with optimal levels identified at 3.25 for TV ads and 2.5 for Facebook video ads [9]. It would be beneficial to replicate these findings across platforms like Instagram, X, and TikTok in additional field studies. Additionally, the unique nature of these platforms, where content often blends polished advertisements with influencer-created or user-generated content, may reveal further nuances in how visual variety should be optimized. The platform-specific familiarity and interaction patterns, such as the habitual

[9]It should be noted that the optimal level of visual variety identified in the TV ad dataset reflects self-reported intentional behavior, whereas the Facebook ad dataset captures actual consumer behavior observed in the field

viewing of influencer content, could significantly influence the effectiveness of visual variety in ads, underscoring the need to tailor ads not only in appearance but in their integration with native content. Third, an investigation of heterogeneous treatment effects could strengthen the understanding of how visual variety impacts advertising outcomes. Heterogenous treatment effects are observable from two distinct perspectives. From the viewer's side, individual characteristics such as cultural background, age, or preferences could affect how visual content is perceived and its effectiveness. For example, Steenkamp and Baumgartner (1992) note that individuals with higher OSLs tend to seek more variety in consumer behaviors, suggesting a nuanced interplay between personal characteristics and the reaction to stimulation in advertising contexts. From the producer's side, the campaign objective (e.g., conversion vs. awareness), product category, and brand identity could influence how visual variety impacts advertising effectiveness. Lastly, the advancements in generative artificial intelligence tools offer an intriguing context for future research. It would be interesting to assess how semantic concepts like visual variety can be integrated into tools for automated video generation or re-composition, using cutting-edge technology from companies like Runway AI, Stability AI's Stable Diffusion Video, or OpenAI's Sora model.

Conclusion

In conclusion, our research quantifies and operationalizes visual variety in video ads through advanced machine learning techniques, uncovering its impact on ad effectiveness. We find an inverted U-shaped relationship between visual variety and advertising effectiveness, confirming that there exists an optimal level of visual variety that maximizes ad clicks. Stimulation and comprehension mediate the balance between under-stimulation through too low levels of visual variety and incomprehension through too high levels of visual variety. Thus, the strategic incorporation of visual variety emerges not only as a creative endeavor, but as a crucial element in enhancing the persuasive power of video ads. We hope that our publicly available visual variety extraction tool allows future researchers to expand their knowledge

of the relationship between visual variety and consumer engagement, fostering further innovation in advertisement design and effectiveness evaluation.

Declarations

Ethical Approval

Not applicable

Availability of supporting data

Data available on request from the authors.

Competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Funding

This work was funded by the German Research Foundation (DFG) research grant number 460037581.

Authors contributions

BLINDED FOR PEER REVIEW

Acknowledgments

Not applicable

References

Akpinar, Ezgi and Jonah Berger (2017), "Valuable Virality," *Journal of Marketing Research*, 54 (2), 318–330.

Anderson, Simon P. and André de Palma (2013), "Shouting to Be Heard in Advertising," *Management Science*, 59 (7), 1545–1556.

Aribarg, Anocha and Eric M. Schwartz (2020), "Native advertising in online news: Trade-offs between clicks and brand recognition," *Journal of Marketing Research*, 57 (1), 20–34.

Askin, Noah and Michael Mauskapf (2017), "What makes popular culture popular? product features and optimal differentiation in music," *American Sociological Review*, 82 (5), 910–944 https://doi.org/10.1177/0003122417728662.

Atalay, A. Selin, Siham El Kihal and Florian Ellsaesser (2023), "Creating Effective Marketing Messages Through Moderately Surprising Syntax," *Journal of Marketing*, 87 (5), 755–775.

Barnett, Samuel B and Moran Cerf (2017), "A Ticket for Your Thoughts: Method for Predicting Content Recall and Sales Using Neural Similarity of Moviegoers," *Journal of Consumer Research*, 44 (1), 160–181.

Becker, Maren and Maarten J. Gijsenberg (2023), "Consistency and commonality in advertising content: Helping or Hurting?" *International Journal of Research in Marketing*, 40 (1), 128–145.

Becker, Maren, Thomas P. Scholdra, Manuel Berkmann and Werner J. Reinartz (2023), "The Effect of Content on Zapping in TV Advertising," *Journal of Marketing*, 87 (2), 275–297.

Berger, Jonah, Yoon Duk Kim and Robert Meyer (2021), "What Makes Content Engaging? How Emotional Dynamics Shape Success," *Journal of Consumer Research*, 48 (2), 235–250.

Berlyne, D. E. (1958), "The influence of complexity and novelty in visual figures on orienting responses," *Journal of Experimental Psychology*, 55 (3), 289.

Berlyne, D. E. (1966), "Curiosity and Exploration," *Science*, 153 (3731), 25–33.

Brandon Castellano (2014), "PySceneDetect: Python and OpenCV-based scene cut/transition detection program & library.," (accessed 2024-08-15), https://github.com/Breakthrough/PySceneDetect.

Brasel, S. Adam and James Gips (2014), "Enhancing television advertising: same-language subtitles can improve brand recall, verbal memory, and behavioral intent," *Journal of the Academy of Marketing Science*, 42, 322–336 https://doi.org/10.1007/s11747-013-0358-1.

Brine, K. G. (2020), *The Art of Cinematic Storytelling: A Visual Guide to Planning Shots, Cuts, and Transitions* Oxford University Press.

Busselle, R. and H. Bilandzic (2009), "Measuring narrative engagement," *Media Psychology*, 12 (4), 321–347.

Chandrasekaran, Deepa, Raji Srinivasan and Debika Sihi (2018), "Effects of offline ad content on online brand search: insights from super bowl advertising," *Journal of the Academy of Marketing Science*, 46 (3), 403–430.

Chang, Hannah H., Anirban Mukherjee and Amitava Chattopadhyay (2022), "More Voices Persuade: The Attentional Benefits of Voice Numerosity," *Journal of Marketing Research*, .

Chang, Hannah H., Anirban Mukherjee and Amitava Chattopadhyay (2023), "More Voices Persuade: The Attentional Benefits of Voice Numerosity," *Journal of Marketing Research*, 60 (4), 687–706.

Chasanis, V. T., A. C. Likas and N. P. Galatsanos (2009), "Scene detection in videos using shot clustering and sequence alignment," *IEEE Transactions on Multimedia*, 11 (1), 89–100.

Chen, Tianqi and Carlos Guestrin (2016), "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 arXiv:1603.02754 [cs] http://arxiv.org/abs/1603.02754.

Chen, Tsai-Shien, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang and Sergey Tulyakov (2024), "Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers," arXiv:2402.19479 [cs] (accessed 2024-08-15), http://arxiv.org/abs/2402.19479.

Churchill, G. A. and D. Iacobucci (2006), *Marketing Research: Methodological Foundations* Vol. 199 New York Dryden Press.

Dall'Olio, Filippo and Demetrios Vakratsas (2023), "The Impact of Advertising Creative Strategy on Advertising Elasticity," *Journal of Marketing*, 87 (1), 26–44.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly et al. (2020), "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, .

Draganska, Michaela, Wesley R. Hartmann and Gena Stanglein (2014), "Internet versus Television Advertising: A Brand-Building Comparison," *Journal of Marketing Research*, 51 (5), 578–590.

Fossen, Beth L. and David A. Schweidel (2017), "Television Advertising and Online Word-of-Mouth: An Empirical Investigation of Social TV Activity," *Marketing Science*, 36 (1), 105–123.

Fossen, Beth L. and David A. Schweidel (2019), "Social TV, Advertising, and Sales: Are Social Shows Good for Advertisers?" *Marketing Science*, 38 (2), 274–295 Publisher: INFORMS.

Galloway, Scott (2022), "Attentive," (accessed 2023-09-27), https://www.profgalloway.com/attentive/.

Gao, Mingfei, Mingze Xu, Larry Davis, Richard Socher and Caiming Xiong (2019), "Start-Net: Online Detection of Action Start in Untrimmed Videos," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* Seoul, Korea (South) IEEE, 5541–5550 https://ieeexplore.ieee.org/document/9008394/.

Geissler, Gary L., George M. Zinkhan and Richard T. Watson (2006), "The influence of home page complexity on consumer attention, attitudes, and purchase intent," *Journal of Advertising*, 35 (2), 69–80.

Glaser, Matthias and Heribert Reisinger (2021), "Don't lose your product in story translation: How product–story link in narrative advertisements increases persuasion," *Journal of Advertising*, 51 (2), 188–205 https://doi.org/10.1080/00913367.2021.1973623.

Gleicher, Michael L. and Feng Liu (2008), "Re-cinematography: Improving the camerawork of casual video," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 5 (1), Article 2 https://doi.org/10.1145/1404880.1404882.

Grewal, Rajdeep, Sachin Gupta and Rebecca Hamilton (2021), "Marketing Insights from Multimedia Data: Text, Image, Audio, and Video," *Journal of Marketing Research*, 58 (6), 1025–1033.

Guitart, Ivan A. and Stefan Stremersch (2021), "The Impact of Informational and Emotional Television Ad Content on Online Search and Sales," *Journal of Marketing Research*, 58 (2), 299–320.

Hartmann, Jochen (2021), "Classification using decision tree ensembles," in *The Machine Age of Customer Insight* Emerald Publishing Limited, 103–117.

Hartmann, Jochen, Mark Heitmann, Christina Schamp and Oded Netzer (2021), "The Power of Brand Selfies," *Journal of Marketing Research*, 58 (6), 1159–1177.

Im, H., H. W. Ju and K. K. Johnson (2021), "Beyond visual clutter: the interplay among products, advertisements, and the overall webpage," *Journal of Research in Interactive Marketing*, 15 (4), 804–821.

Jacoby, Jacob (1977), "Information Load and Decision Quality: Some Contested Issues," *Journal of Marketing Research*, 14 (4), 569–573.

Jia, He (Michael), B. Kyu Kim and Lin Ge (2020), "Speed Up, Size Down: How Animated Movement Speed in Product Videos Influences Size Assessment and Product Evaluation," *Journal of Marketing*, 84 (5), 100–116.

Johnson, Garrett, Randall A. Lewis and Elmar Nubbemeyer (2017), "Ghost ads: Improving the economics of measuring online ad effectiveness," *Journal of Marketing Research*, 54 (6), 867–884.

Keller, Kevin Lane (1993), "Conceptualizing, measuring, and managing customer-based brand equity," *Journal of Marketing*, 57 (1), 1–22.

Keller, Kevin Lane and Donald R Lehmann (2006), "Brands and branding: Research findings and future priorities," *Marketing Science*, 25 (6), 740–759.

Kincaid, J. P., R. P. Jr. Fishburne, R. L. Rogers and B. S. Chissom (1975), "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Research Report RBR-8-75 Defense Technical Information Center Fort Belvoir, VA.

Knight, Samsun, Rocklage Matthew D. and Yakov Bart (2024), "Narrative reversals and story success," *Science Advances*, 10, eadl2013.

Koh, B. and F. Cui (2022), "An exploration of the relation between the visual attributes of thumbnails and the view-through of videos: The case of branded video content," *Decision Support Systems*, 160, 113820.

Lang, Annie (2000), "The Limited Capacity Model of Mediated Message Processing," *Journal of Communication*, 50 (1), 46–70.

Lee, J.E., S. Hur and B. Watkins (2018), "Visual communication of luxury fashion brands on social media: effects of visual complexity and brand familiarity," *Journal of Brand Management*, 25, 449–462.

Lennan, Christopher, Hao Nguyen and Dat Tran (2018), "Image quality assessment," https://github.com/idealo/image-quality-assessment.

Li, Xi, Mengze Shi and Xin (Shane) Wang (2019), "Video mining: Measuring visual information using automatic methods," *International Journal of Research in Marketing*, 36 (2), 216–231.

Liaukonyte, Jura, Thales Teixeira and Kenneth C. Wilbur (2015), "Television Advertising and Online Shopping," *Marketing Science*, 34 (3), 311–330.

Lindsay, Kate (2024), "Something went terribly wrong with online ads: What isn't an ad these days?" (accessed 2023-09-27), https://www.theatlantic.com/technology/archive/2024/02/online-ads-more-annoying/677576/.

Liu, Xuan, Savannah Wei Shi, Thales Teixeira and Michel Wedel (2018), "Video Content Marketing: The Making of Clips," *Journal of Marketing*, 82 (4), 86–101.

Loewenstein, Jeffrey, Rajagopal Raghunathan and Chip Heath (2011), "The Repetition-Break Plot Structure Makes Effective Television Advertisements," *Journal of Marketing*, 75 (5), 105–119.

Lowe, David G (2004), "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, 60, 91–110.

Lundberg, Scott and Su-In Lee (2017), "A unified approach to interpreting model predictions," https://arxiv.org/abs/1705.07874.

MacInnis, Deborah J., Ambar G. Rao and Allen M. Weiss (2002), "Assessing When Increased Media Weight of Real-World Advertisements Helps Sales," *Journal of Marketing Research*, 39 (4), 391–407.

McGranaghan, Matthew, Jura Liaukonyte and Kenneth C. Wilbur (2022), "How Viewer Tuning, Presence, and Attention Respond to Ad Content and Predict Brand Search Lift," *Marketing Science*, 41 (5), 873–895.

Menon, Satya and Barbara E. Kahn (1995), "The Impact of Context on Variety Seeking in Product Choices," *Journal of Consumer Research*, 22 (3), 285–295.

Miniukovich, Aliaksei and Maurizio Marchese (2020), "Relationship between visual complexity and aesthetics of webpages," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* New York, NY, USA Association for Computing Machinery https://doi.org/10.1145/3313831.3376602.

Morrison, Bruce John and Marvin J. Dainoff (1972), "Advertisement Complexity and Looking Time," *Journal of Marketing Research*, 9 (4), 396–400.

Mourchid, Youssef, Benjamin Renoust, Olivier Roupin, Lê Văn, Hocine Cherifi and Mohammed El Hassouni (2019), "Movienet: a movie multilayer network model using visual and textual semantic cues," *Applied Network Science*, 4 (1), 121 (accessed 2024-08-15), https://doi.org/10.1007/s41109-019-0226-0.

Nikolinakou, Angeliki and Karen Whitehill King (2018), "Viral video ads: Emotional triggers and social media virality," *Psychology & Marketing*, 35 (10), 715–726.

Olney, Thomas J., Morris B. Holbrook and Rajeev Batra (1991), "Consumer Responses to Advertising: The Effects of Ad Content, Emotions, and Attitude toward the Ad on Viewing Time," *Journal of Consumer Research*, 17 (4), 440–453.

Overgoor, Gijs, William Rand, Willemijn van Dolen and Masoud Mazloom (2022), "Simplicity is not key: Understanding firm-generated social media images and consumer liking," *International Journal of Research in Marketing*, 39 (3), 639–655.

Pauwels, Koen, Michael Peran, Zee Shah, German Schnaidt and Dauwe Vercamer (2023), "Sponsored brands video rings up clicks and sales in the short and long run," *Journal of Marketing Analytics*, 11 (3), 275–286.

Pieters, Rik, Michel Wedel and Rajeev Batra (2010), "The Stopping Power of Advertising: Measures and Effects of Visual Complexity," *Journal of Marketing*, 74 (5), 48–60.

Preacher, K. J. and A. F. Hayes (2008), "Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models," *Behavior Research Methods*, 40 (3), 879–891.

Rafieian, Omid and Hema Yoganarasimhan (2022), "Variety Effects in Mobile Advertising," *Journal of Marketing Research*, 59 (4), 718–738.

Reinecke, Katharina, Tom Yeh, Luke Miratrix, Rahmatri Mardiko, Yuechen Zhao, Jenny Liu and Krzysztof Z. Gajos (2013), "Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* New York, NY, USA Association for Computing Machinery https://doi.org/10.1145/2470654.2481281.

Schwenzow, Jasper, Jochen Hartmann, Amos Schikowsky and Mark Heitmann (2021), "Understanding videos at scale: How to extract insights for business research," *Journal of Business Research*, 123, 367–379.

Shani-Feinstein, Yael, Ellie J Kyung and Jacob Goldenberg (2022), "Moving, Fast or Slow: How Perceived Speed Influences Mental Representation and Decision Making," *Journal of Consumer Research*, 49 (3), 520–542.

Shehu, Edlira, Tammo H.A. Bijmolt and Michel Clement (2016), "Effects of Likeability Dynamics on Consumers' Intention to Share Online Video Advertisements," *Journal of Interactive Marketing*, 35 (1), 27–43.

Shin, Donghyuk, Shu He, Gene Moo Lee, Andrew B. Whinston, Suleyman Cetintas and Kuang-Chih Lee (2020), "Enhancing social media analysis with visual data analytics: A deep learning approach," in *SSRN* Amsterdam, The Netherlands, 1459–1492 https://papers.ssrn.com/sol3/papers.cfm?abstract$_id = number_here$.

Stayman, Douglas M. and Rajeev Batra (1991), "Encoding and Retrieval of Ad Affect in Memory," *Journal of Marketing Research*, 28 (2), 232–239.

Steenkamp, Jan-Benedict EM and Hans Baumgartner (1992), "The role of optimum stimulation level in exploratory consumer behavior," *Journal of Consumer Research*, 19 (3), 434–448 http://www.jstor.org/stable/2489400.

Stuppy, Anika, Jan R. Landwehr and A. Peter McGraw (2023), "The Art of Slowness: Slow Motion Enhances Consumer Evaluations by Increasing Processing Fluency," *Journal of Marketing Research*, .

Sweller, John (1988), "Cognitive load during problem solving: Effects on learning," *Cognitive Science*, 12 (2), 257–285.

Talebi, H. and P. Milanfar (2018*a*), "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, 27 (8), 3998–4011.

Talebi, Hossein and Peyman Milanfar (2017), "Introducing nima: Neural image assessment," https://research.google/blog/introducing-nima-neural-image-assessment/ Accessed: [insert date here].

Talebi, Hossein and Peyman Milanfar (2018*b*), "Nima: Neural image assessment," *IEEE transactions on image processing*, 27 (8), 3998–4011.

Teixeira, Thales, Michel Wedel and Rik Pieters (2012), "Emotion-Induced Engagement in Internet Video Advertisements," *Journal of Marketing Research*, 49 (2), 144–159.

Teixeira, Thales, Rosalind Picard and Rana el Kaliouby (2014), "Why, When, and How Much to Entertain Consumers in Advertisements? A Web-Based Facial Tracking Field Study," *Marketing Science*, 33 (6), 809–827.

Teixeira, Thales S., Michel Wedel and Rik Pieters (2010), "Moment-to-Moment Optimal Branding in TV Commercials: Preventing Avoidance by Pulsing," *Marketing Science*, 29 (5), 783–804.

Tellis, Gerard J., Deborah J. MacInnis, Seshadri Tirunillai and Yanwei Zhang (2019), "What Drives Virality (Sharing) of Online Digital Content? The Critical Role of Information, Emotion, and Brand Prominence," *Journal of Marketing*, 83 (4), 1–20.

Tong, X., Y. Chen, S. Zhou and S. Yang (2022), "How background visual complexity influences purchase intention in live streaming: The mediating role of emotion and the moderating role of gender," *Journal of Retailing and Consumer Services*, 67, 103031.

Toubia, Olivier, Jonah Berger and Jehoshua Eliashberg (2021), "How quantifying the shape of stories predicts their success," *Proceedings of the National Academy of Sciences*, 118 (26), e2011695118.

Villarroel Ordenes, Francisco, Dhruv Grewal, Stephan Ludwig, Ko De Ruyter, Dominik Mahr and Martin Wetzels (2019), "Cutting through Content Clutter: How Speech and Image Acts Drive Consumer Sharing of Social Media Brand Messages," *Journal of Consumer Research*, 45 (5), 988–1012.

Wang, Ao, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han and Guiguang Ding (2024), "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, .

Wang, Tian, Yang Chen, Hongqiang Lv, Jing Teng, Hichem Snoussi and Fei Tao (2021), "Online Detection of Action Start via Soft Computing for Smart City," *IEEE Transactions on Industrial Informatics*, 17 (1), 524–533 Conference Name: IEEE Transactions on Industrial Informatics (accessed 2024-08-15), https://ieeexplore.ieee.org/document/9099408/?arnumber=9099408.

Warriner, Amy Beth, Victor Kuperman and Marc Brysbaert (2013), "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior Research Methods*, 45, 1191–1207.

Welke, Dominik and Edward A. Vessel (2022), "Naturalistic viewing conditions can increase task engagement and aesthetic preference but have only minimal impact on EEG quality," *NeuroImage*, 256, 119218.

Wilms, Laura and Daniel Oberfeld (2018), "Color and emotion: effects of hue, saturation, and brightness," *Psychological Research*, 82, 896–914 https://doi.org/10.1007/s00426-017-0880-8.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest and Alexander M. Rush (2020), "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* Online Association for Computational Linguistics, 38–45 https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Wu, Yue, Yicong Zhou, George Saveriades, Sos Agaian, J. P. Noonan and P. Natarajan (2013), "Local shannon entropy measure with statistical tests for image randomness," *Information Sciences*, 222, 323–342.

Yang, Joonhyuk, Yingkang Xie, Lakshman Krishnamurthi and Purushottam Papatla (2022), "High-Energy Ad Content: A Large-Scale Investigation of TV Commercials," *Journal of Marketing Research*, 59 (4), 840–859.

Ye, Keren, Kyle Buettner and Adriana Kovashka (2018), "Story Understanding in Video Advertisements," arXiv:1807.11122 [cs] (accessed 2024-08-15), http://arxiv.org/abs/1807.11122.

Yoo, C. Y. (2007), "Implicit memory measures for web advertising effectiveness," *Journalism & Mass Communication Quarterly*, 84 (1), 7–23.

WEB APPENDIX

TABLE OF CONTENT

I: Related literature

Web Appendix A: OVERVIEW OF RELEVANT STUDIES ON TV AND ONLINE VIDEOS ADS

Author	**TV**	**Online**	**DV: Purchase**	**DV: Engagement**	**Visual variety**	**Scalable**	**Sample size**	**Datasets**	**Lab**	**Field**
This study	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	3,000+	2	**Yes**	**Yes**
Akpinar and Berger (2017)	No	Yes	Yes	Yes	No	No	240	1	Yes	Yes
Becker and Gijsenberg (2023)	Yes	No	Yes	No	No	No	247	1	No	Yes
Becker et al. (2019)	Yes	No	Yes	No	No	No	323	1	No	Yes
Becker et al. (2023)	Yes	No	No	No	No	No	1,315	1	Yes	Yes
Bijmolt et al. (1998)	Yes	No	No	No	No	No	1	1	Yes	No
Bruce et al. (2020)	Yes	No	No	No	No	No	325	1	No	Yes
Chandy et al. (2001)	Yes	No	No	No	No	No	72	1	No	Yes
Chandrasekaran et al. (2018)	Yes	No	No	No	No	No	293	1	No	Yes
Chang et al. (2023)	No	Yes	Yes	No	No[1]	Yes	12,600	2	Yes	Yes
Chen and Lee (2014)	No	Yes	No	Yes	No	No	83	1	Yes	No
Dall'Olio and Vakratsas (2023)	Yes	No	Yes	No	No	No	2,251	1	No	Yes
Deighton et al. (1989)	Yes	No	No	No	No	No	40	1	Yes	No
Draganska et al. (2014)	Yes	Yes	No	No	No	No	20	4	No	Yes
Eckler and Bolls (2011)	No	Yes	No	Yes	No	No	12	1	Yes	No
Edell and Burke (1987)	Yes	No	No	No	No	No	26	2	Yes	No
Fossen and Schweidel (2017)	Yes	No	No	No	No	No	9,103	1	No	Yes
Fossen and Schweidel (2019)	Yes	No	Yes	No	No	No	1,685	1	No	Yes
Guitart and Stremersch (2021)	Yes	No	Yes	No	No	No	2,000	1	No	Yes
Jia et al. (2020)	No	Yes	No	No	No	No	16	6	Yes	No
Liaukonyte et al. (2015)	Yes	No	Yes	No	No	No	1,224	1	No	Yes
Loewenstein et al. (2011)	Yes	Yes	Yes	No	No	No	967	2	Yes	Yes
McGranaghan et al. (2022)	Yes	No	No	No	No	Yes	6,650	1	No	Yes
MacInnis et al. (2002)	Yes	No	No	No	No	No	47	1	Yes	Yes
Nikolinakou and King (2018)	No	Yes	No	Yes	No	No	8	1	Yes	No
Olney et al. (1991)	Yes	No	No	No	No	No	146	1	No	yes
Pauwels et al. (2023)	No	Yes	Yes	No	No	Yes	25,364	1	No	Yes
Shehu et al. (2016)	No	Yes	No	Yes	No	No	120	1	Yes	No
Simmonds et al. (2020)	Yes	No	No	No	No	No	31	1	Yes	No
Southgate et al. (2010)	Yes	Yes	No	No	No	No	102	1	No	Yes
Stayman and Batra (1991)	Yes	No	No	No	No	No	4	2	Yes	No
Stuppy et al. (2023)	No	Yes	No	Yes	No	Partly	600+	7	Yes	Yes
Teixeira et al. (2010)	Yes	No	No	No	No	Partly	37	2	Yes	No
Teixeira et al. (2012)	No	Yes	No	No	No	Partly	28	1	Yes	No
Teixeira et al. (2014)	Yes	No	Yes	No	No	Partly	82	1	Yes	No
Tellis et al. (2019)	No	Yes	No	Yes	No	No	857	2	No	Yes
Yang et al. (2022)	Yes	No	No	No	No	Yes	27,000	2	No	Yes

1: Using a pixel-by-pixel measure on the frame level

II: Visual variety pipeline

Web Appendix B:Visual variety features and controls

Feature Extraction

The sequence of frames in each video is first divided into individual shots using OpenCV combined with `PySceneDetect's` ContentDetector, configured with a threshold of 27. This ensures that distinct scenes are accurately segmented based on changes in visual content. For each identified shot, three keyframes are selected: the first, middle, and last frames. To avoid the influence of abrupt transitions, the first and last frames are selected with a slight offset of two frames (approximately 0.07 seconds). This method captures the core visual elements of each shot while minimizing the risk of including distorted or transitional frames that could skew the analysis. To assess visual differences between keyframes at multiple levels of abstraction, four levels of content granularity are employed:

- **Pixel Level:** Pixel level variations are captured by calculating the Manhattan distance between corresponding pixels of two frames across the RGB color channels. Before calculation, the pixel values are normalized to the [0,1] range to ensure consistency.

- **Embedding Level:** High-level visual patterns and themes are analyzed by transforming each frame into a fixed-dimensional embedding vector using the pre-trained ViT-B/16 Vision Transformer (Dosovitskiy et al. 2020) from the HuggingFace transformers library (Wolf et al. 2020). The Vision Transformer (ViT) outputs an embedding of dimension 768 for each frame. The similarity between two frames is then quantified using cosine similarity, calculated with the `cosine_similarity` function from the `sklearn.metrics.pairwise` module.

- **Feature Level:** Local structural elements, such as edges and corners, are examined using the SIFT algorithm, implemented via OpenCV (`cv2`). Keypoints and descriptors are extracted from each frame, resulting in a variable number of descriptors per image, depending on the visual complexity of the frame. Matches between descriptors of two frames are identified using the `BFMatcher.knnMatch()` function from OpenCV, which

returns the k-best matches. The ratio test by Lowe (2004) is applied with a threshold of 0.75 to filter for /enquoteGood matches that are significantly similar.

- **Object Level:** Objects within frames are detected using the pre-trained YOLOv10 model (Wang et al. 2024). The model outputs bounding boxes, class labels, and confidence scores for detected objects. Post-processing and visualization of the detection results are handled using the `supervision` library. Objects are labelled across frames.

Compound Measure

The visual variety score is predicted using the DMLC XGBoost implementation (Chen and Guestrin 2016), a powerful gradient boosting framework. The model was trained and fine-tuned using a grid search approach with an 80/20 train/test split to identify the optimal hyperparameters. The final configuration `colsample_bytree=0.6`, `gamma=0.4`, `learning_rate=0.1`, `max_depth=4`, `min_child_weight=1`, `n_estimators=100`, and `subsample=0.6`, provided the best performance. The model's accuracy was evaluated using 10-fold cross-validation, resulting in an average Mean Squared Error (MSE) of 0.193 across all folds. This indicates that the model generalizes well to unseen data. To further interpret the influence of different features on the model's predictions, SHAP (SHapley Additive exPlanations) values were computed, providing insights into the contribution of each feature to the predicted visual variety scores (see Figure W1).

Controls

To account for external factors that might influence advertising effectiveness of the videos, several control variables were incorporated into the analysis. These controls ensure that we isolate the effect of visual variety on advertising effectiveness and our results are not confounded by other aspects of the video content.

Structural Features:

- **Duration**: The total length of each video was extracted using OpenCV.

Audio Features: Six audio-related features were controlled for to account for the potential influence of the audio track on advertising effectiveness:

- **Arousal and Valence**: These emotional dimensions of the audio were extracted using the pre-trained transformer model audeering/Wav2Vec2-Large-Robust from HuggingFace. This model is specifically trained to analyze emotional content in audio and speech, providing scores for arousal (intensity of emotion), dominance (control and influence), and valence (positivity or negativity of emotion).

- **Energy level and tempo**: Energy and tempo, which reflect the intensity and rhythm of the audio, were calculated following the approach of Yang et al. (2022) and the `librosa` package.

- **Frequency variation**: Frequency variation was calculated using the my-voice-analysis package (`https://shahabks.github.io/my-voice-analysis/`), to compute the average frequency (in Hz) and its variation. The latter is stored for each video as our frequency variation variable.

- **Transcript complexity**: Transcript complexity is calculated using the Flesch-Kincaid readability score (Kincaid et al. 1975). The score is calculated based on sentence length and word syllables using TextAnalyzer (`http://textanalyzer.org/lexica`).

Visual Features: Several additional visual features were controlled for to capture elements that might influence advertising effectiveness:

- **Unique faces**: The number of unique faces appearing in each video was calculated using the `DeepFace` package for face recognition and classification. Every fifth frame of the video was analyzed, with each detected face checked against a face database to determine whether it had appeared previously in the video. New faces were added to

the database, and the total number of unique faces was counted as a measure of visual variety related to human presence.

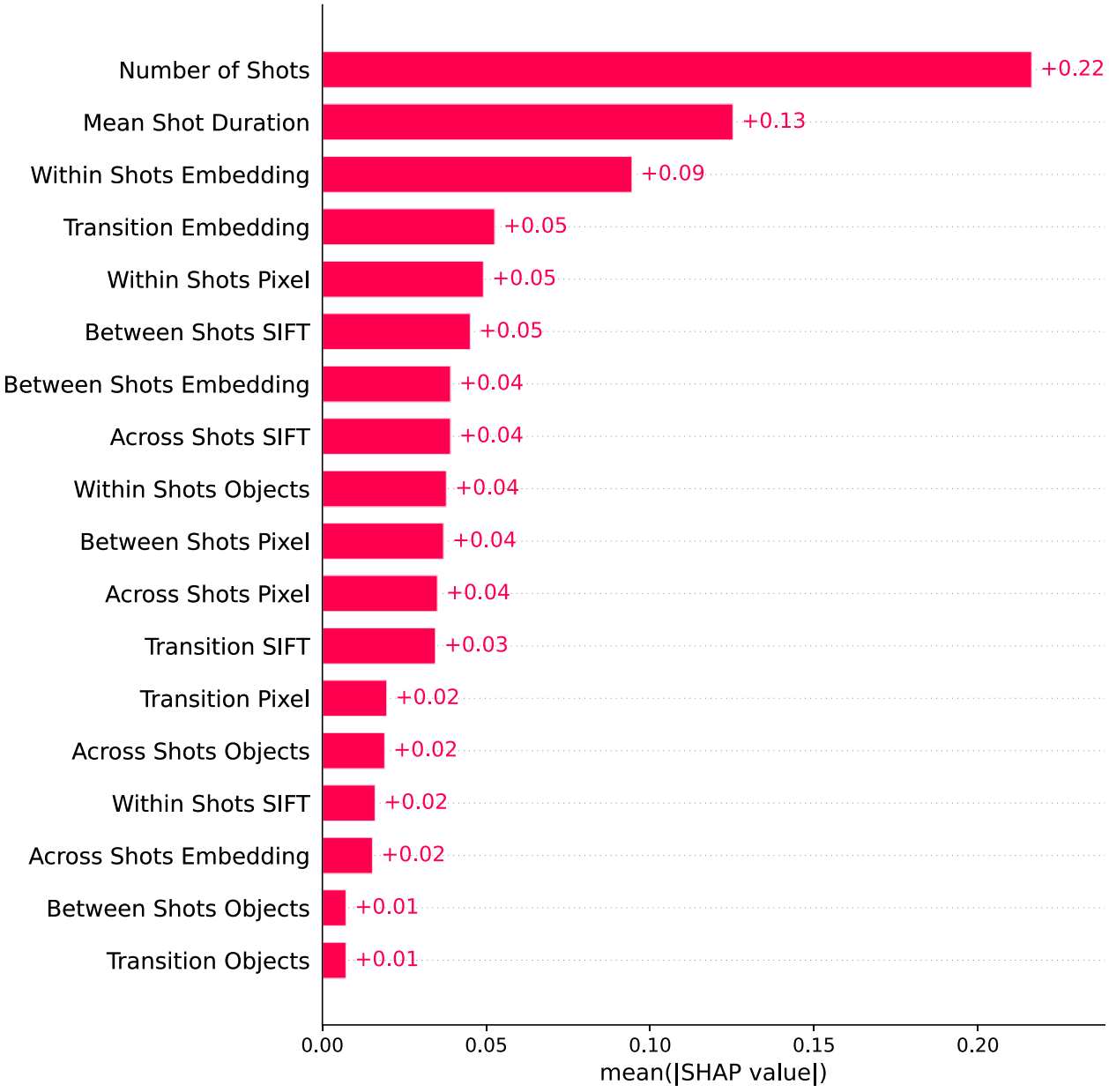
- **Face area**: This measure was derived by computing the bounding box areas for detected faces in each frame and dividing by the total frame area. The average was taken across all frames analyzed, providing a measure of the prominence of faces in the video.

- **Entropy** : Visual complexity was measured using Shannon entropy, calculated for all keyframes in the video and averaged across these keyframes. Entropy was computed using the `skimage.measure.shannon_entropy` function from the `scikit-image` library, which quantifies the randomness or disorder within an image. Higher entropy values indicate more complex and varied visual content.

- **Saturation and brightness**: These color-based features were calculated for all keyframes using OpenCV, then averaged across the keyframes for each video. Saturation represents the intensity of colors, while brightness indicates the lightness or darkness of the image. Both measures provide additional insight into the visual characteristics of the video.

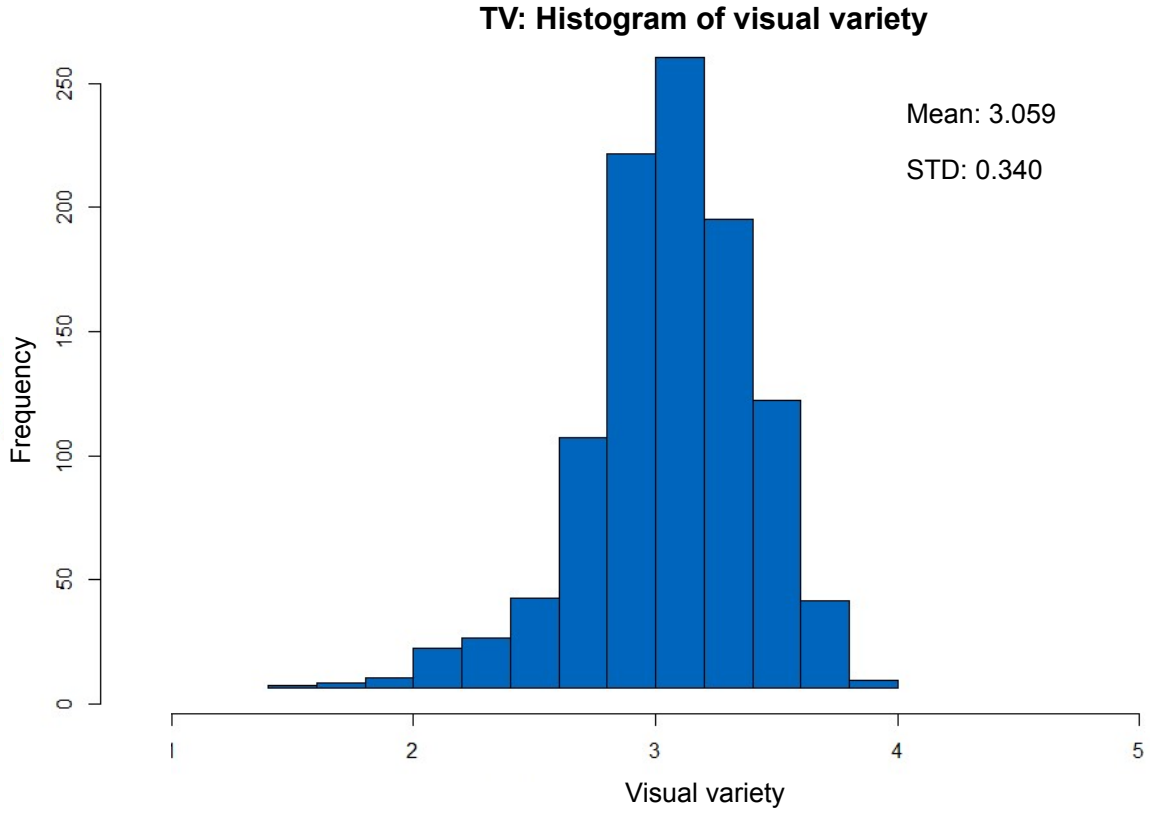
- **Image aesthetics**: The aesthetic quality of each video was assessed using the Neural Image Assessment (NIMA) model, developed by Google (Talebi and Milanfar 2018*b*). The NIMA model, implemented using the Idealo GitHub repository (Lennan et al. 2018), evaluates the aesthetic quality of images on a scale from 1 to 10. NIMA scores were calculated for all keyframes in the video and then averaged to provide a holistic measure of video aesthetics.

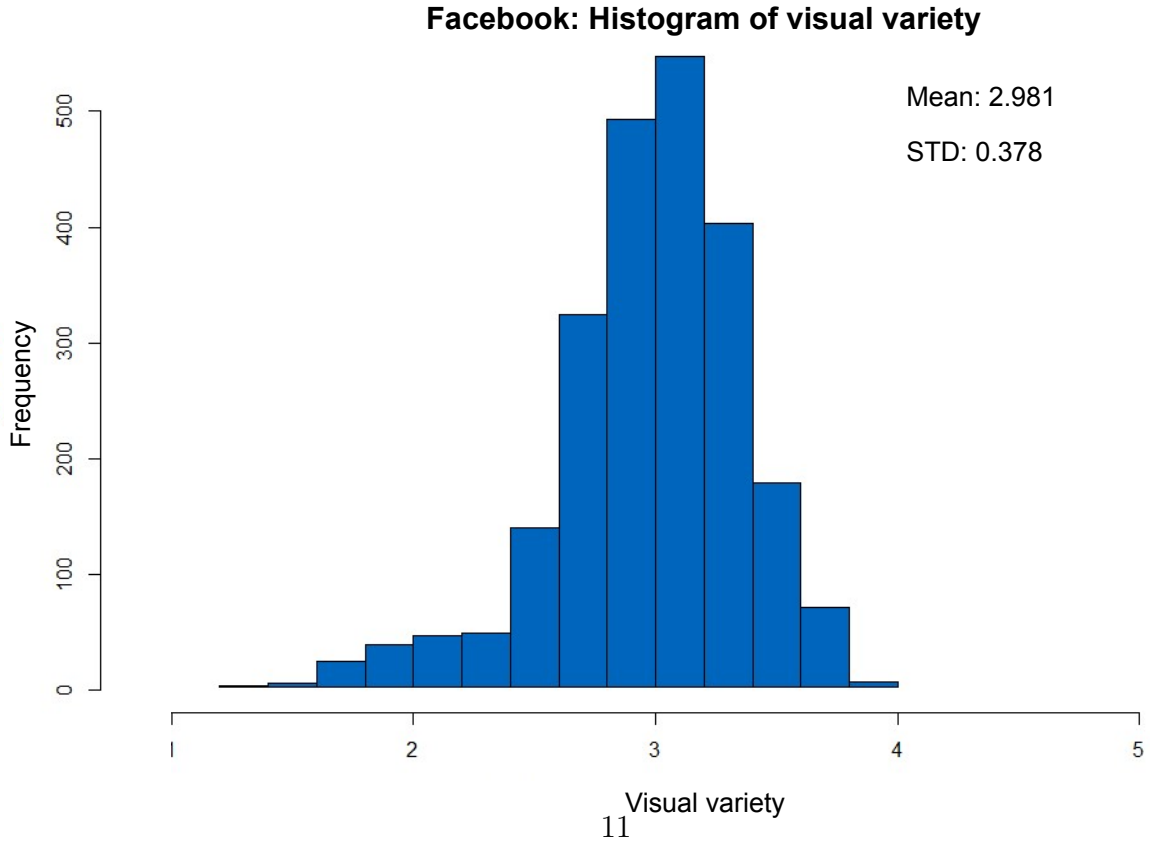
Figure W1: INFLUENCE OF EXTRACTED FEATURES ON VISUAL VARIETY



III: Dataset overview

Web Appendix C: DISTRIBUTION OF VISUAL VARIETY IN VIDEO ADS



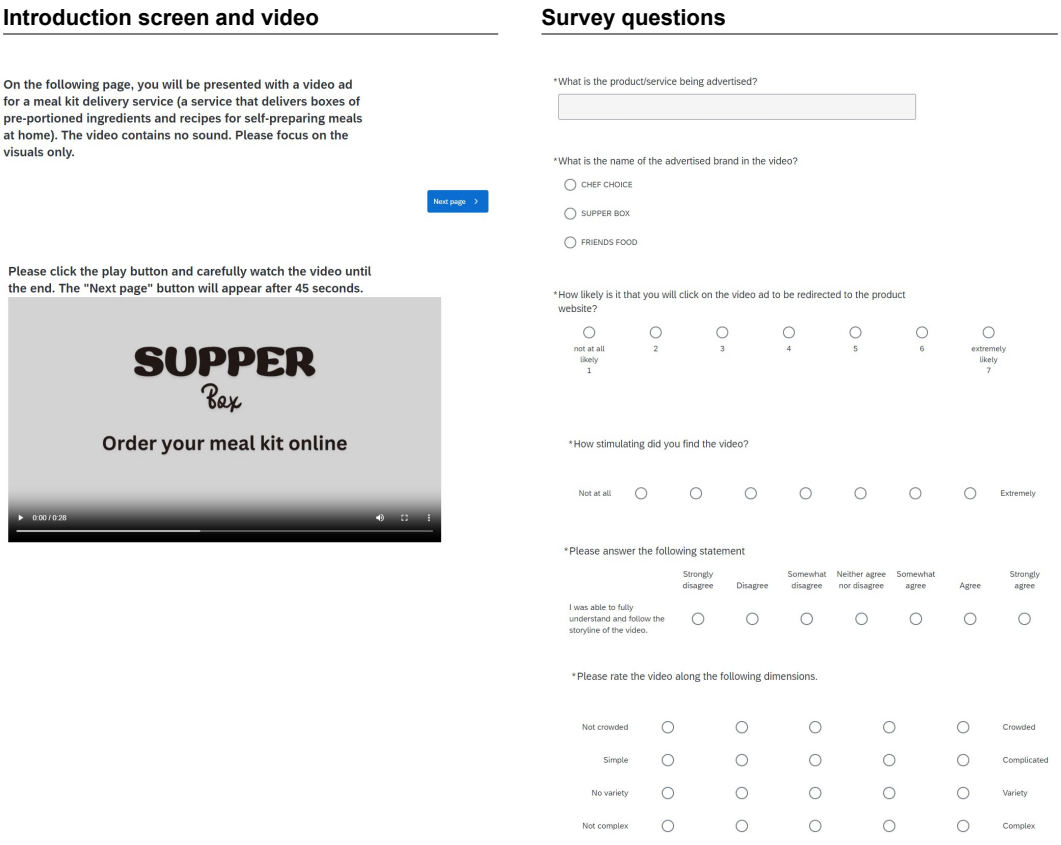


IV: Lab experiment

Web Appendix D: Experiment description

All three crafted videos ads promoted a fictitious brand (i.e., SUPPER BOX), which provides delivery of meal-kit boxes. The meal-kit boxes contain raw ingredients and the recipe to create a meal and are a service widely established in the USA, where we ran our survey. We collected in total 210 survey responses on Prolific based on an equal gender distribution (50% female, 50% male). We excluded 53 participants after an English screening test (−3) and multiple attention checks (−54), including brand and product category recall. Our final sample consists of 157 respondents (52.2% female, 47.8% male). The average time for completion of the entire survey was around 6 minutes. Figure W2 shows screenshots of our survey's setup for the lab experiment.

Figure W2: SCREENSHOTS OF SELECTED SCREENS SHOWN TO PARTICIPANTS



V: Technical recommendations for practitioners

Web Appendix E: TECHNICAL RECOMMENDATIONS TO OPTIMIZE VISUAL VARIETY

	Pre-production	**Post-production**
Pixel level	• Plan color and lighting schemes that change distinctively across scenes	• Enhance visual variety through color grading to ensure variation between and across shots
Embedding & feature level	• Choose visually distinct structural element, patterns and settings that differ in style	• Employ effects to alter patterns between shots
Object level	• Vary the number of objects across the scenes, change proportions and background elements	• Increase the number of distinct scenes to enhance object variety or modify object visibility across the video
Shot arrangement	• Strategically plan the narrative and scenes to allow for sequences where visually contrasting shots follow each other	• Arrange shots in a sequence that maximizes visual contrast between consecutive shots
Shot dynamics	• Plan a scenes to allow for a high variety of shot durations and angles	• Cut scenes into smaller shot increments and reduce duration

Legend

Low impact on visual variety

Medium impact on visual variety

High impact on visual variety

Additional references

Becker, Maren, Nico Wiegand and Werner J. Reinartz (2019), "Does It Pay to Be Real? Understanding Authenticity in TV Advertising," *Journal of Marketing*, 83 (1), 24–50.

Bijmolt, Tammo H. A., Wilma Claassen and Britta Brus (1998), "Children's Understanding of TV Advertising: Effects of Age, Gender, and Parental Influence," *Journal of Consumer Policy*, 21 (2), 171–194.

Bruce, Norris I., Maren Becker and Werner Reinartz (2020), "Communicating Brands in Television Advertising," *Journal of Marketing Research*, 57 (2), 236–256.

Chandy, Rajesh K., Gerard J. Tellis, Deborah J. Macinnis and Pattana Thaivanich (2001), "What to Say When: Advertising Appeals in Evolving Markets," *Journal of Marketing Research*, 38 (4), 399–414.

Chen, Tsai and Hsiang-Ming Lee (2014), "Why Do We Share? The Impact of Viral Videos Dramatized to Sell: How Microfilm Advertising Works," *Journal of Advertising Research*, 54 (3), 292–303.

Deighton, John, Daniel Romer and Josh McQueen (1989), "Using Drama to Persuade," *Journal of Consumer Research*, 16 (3), 335–343.

Eckler, Petya and Paul Bolls (2011), "Spreading the Virus," *Journal of Interactive Advertising*, 11 (2), 1–11.

Edell, Julie A. and Marian Chapman Burke (1987), "The Power of Feelings in Understanding Advertising Effects," *Journal of Consumer Research*, 14 (3), 421–433.

Nikolinakou, Angeliki and Karen Whitehill King (2018), "Viral video ads: Emotional triggers and social media virality," *Psychology & Marketing*, 35 (10), 715–726.

Simmonds, Lucy, Svetlana Bogomolova, Rachel Kennedy, Magda Nenycz-Thiel and Steven Bellman (2020), "A dual-process model of how incorporating audio-visual sensory cues in video advertising promotes active attention," *Psychology & Marketing*, 37 (8), 1057–1067.

Southgate, Duncan, Nikki Westoby and Graham Page (2010), "Creative determinants of viral video viewing," *International Journal of Advertising*, 29 (3), 349–368

C. Methods in Marketing: Leveraging Artificial Intelligence for Academic Information Dissemination

Authors Maximilian Witte
Keno Tetzlaff
Martin Reisenbichler
Mark Heitmann

Year 2025

Bibliographic Status Working paper *(2025 update with latest data)*
Accepted for EMAC 2024

Methods in Marketing: Leveraging Artificial Intelligence for Academic Information Dissemination

Maximilian Witte[1]
Keno Tetzlaff[2]
Martin Reisenbichler[3]
Mark Heitmann[4]

[2]Maximilian Witte is Doctoral Student at the University of Hamburg, Hamburg Business School, Moorweidenstrasse 18, 20148 Hamburg, Germany. maximilian.witte@uni-hamburg.de.

[2]Keno Tetzlaff is Doctoral Student at the University of Hamburg, Hamburg Business School, Moorweidenstrasse 18, 20148 Hamburg, Germany. keno.tetzlaff@uni-hamburg.de.

[3]Martin Reisenbichler is Assistant Professor at the University of Economics and Business in Vienna (WU), Welthandelsplatz 1, 1020 Vienna, Austria. martin.reisenbichler@wu.ac.at.

[4]Mark Heitmann is Professor of Marketing & Customer Insight at the University of Hamburg, Hamburg Business School, Moorweidenstrasse 18, 20148 Hamburg, Germany. mark.heitmann@uni-hamburg.de.

Methods in Marketing: Leveraging Artificial Intelligence for Academic Information Dissemination

Abstract

In the dynamically evolving field of marketing research, it is imperative that scholars follow the substantive and methodological developments. However, in the context of thousands of newly published articles each year, increasing diversity in methods, and further acceleration through the advent of artificial intelligence (AI), keeping track becomes a daunting task for marketing researchers. Some classical vehicles for information dissemination exist (e.g., editorials, benchmarks), but are typically not designed to cover developments in the entire marketing area and rely on error-prone manual work. We propose an AI-based approach for comprehensive and structured analysis of the entire marketing literature from 2010 to 2023. We demonstrate the increasing complexity in marketing (i.e., increase in both articles and methods per article), identify recent developments in the field, and strive to provide more structured access to article information.

Keywords: *text analysis, research methods, marketing domains*
Track: *Methods, Modelling & Marketing Analytics*

1 Introduction

In a landscape inundated with the publication of thousands of marketing articles per year (Bornmann and Mutz 2015), and with rapid transformations driven by emerging trends such as generative AI, big data, and machine learning that are gaining momentum (Kumar 2018), it becomes challenging for researchers to keep track of their dynamic field. However, a comprehensive understanding of their research domain is essential for scholars. Profound familiarity with previous research empowers academics to choose topics that matter and articulate their thoughts with precision, facilitating their active participation in academic discourse.

That is why the extraction and retrieval of information from previous research is a key to academic success, as exemplified by the many review articles that attempt to provide an overview and guidance on issues of importance within the academic marketing sphere (e.g., Grewal et al. 2020; Donthu et al. 2021; Shankar and Parsana 2022). However, while such means are helpful, extracting and providing information to researchers remains a challenging task for several reasons.

For example, the increased availability of novel data sources requires researchers to adapt new methods (Grewal et al. 2021). Identifying a suitable starting point and relevant resources might not be easy for some researchers (Berger et al. 2020) and, as Moorman et al. (2019) note, internal forces limit marketing researchers in the adoption of novel methods and stick to established topics, methods, and like-minded people instead.

Prior literature coined this exploration-exploitation trade-off (Sudhir 2016): exploration of novel topics and methods that oppose the exploitation and elaboration of known themes. While exploitation is a valid way to go, exploration is a prime route to follow when aiming to generate novel insights. However, current research often relies on repeating established and well-known methods and procedures (Inman et al. 2018), sometimes rendering the method obsolete by the time the article is published (Toubia 2022). For example, we observe these tendencies in the rather novel domain of text analysis. Transformers are objectively superior for sentiment classification for most applications. However, many recent marketing articles rely on dictionaries such as LIWC for sentiment classification (e.g., Woolley and Sharif 2021; Becker et al. 2022; Peng et al. 2022) and reference usage in a few older hallmark articles such as Berger and Milkman (2012) or Ludwig et al. (2013) that are up to 10 years old. Similarly, a latent moderation analysis Pieters et al. (2022) revealed that 95% of articles use the same method, although it is not optimal in most situations.

Collectively, this underscores the essential requirement for scholars to have a profound understanding of their research domain, remain attuned to the dynamic trends in market-

ing methodologies, and use algorithms skillfully to access pertinent knowledge. Within the context of this article, we present an innovative AI-based methodology capable of comprehensively analyzing the entirety of the marketing literature. This approach enables the extraction and identification of methodological and topical trends embedded within the literature. Our research holds tangible benefits for both researchers and practitioners. It facilitates easy access to novel methods and applications in marketing, thereby streamlining the process of staying abreast of advances in the field. Furthermore, our methodology contributes to strengthening the case for exploration over exploitation (Sudhir 2016), highlighting the importance of embracing new avenues and possibilities in marketing research and practice.

In the subsequent sections, we survey tools scholars currently rely on when turning to understanding their field including (i) journal editorials, (ii) method tutorial papers, and (iii) empirical method benchmarks. Then, we introduce our novel AI-based method to automatically extract knowledge from the marketing literature published from 2010 onward. Finally, we illustrate how themes and methods in marketing research develop over time, and how the marketing field generally evolves in terms of substantive and method contributions, exemplified on articles entailing LIWC and Topic Modeling applications.

2 Vehicles of Academic Information Dissemination

To provide the reader with some background on how scholars currently access information, we now turn to illustrate typical vehicles of scholarly information dissemination: editorials and commentaries (e.g., Yadav 2010; Moorman et al. 2019; Grewal et al. 2020), tutorial articles (e.g., Podsakoff et al. 2012; Hulland et al. 2018; Berger et al. 2020), and method benchmarks (e.g., Pieters et al. 2022; Shankar and Parsana 2022; Tetzlaff et al. 2023). Although each of these vehicles provides substantial value to a specific case, they are not designed to provide structured information for scholars on developments in the entire marketing field. That is why we propose a novel AI-based approach that allows for large-scale, automated analysis and grants researchers innovative angles on information dissemination.

First, editorials and commentaries analyze the development of the marketing field (e.g., Grewal et al. 2020, for JMR)). These articles are generic, cover broader developments, and attempt to distill future research directions by reflecting on the past (e.g., Moorman et al. 2019, for JM and marketing research at large).

Second, tutorial articles provide researchers with more specific information compared to editorials. For example, Berger et al. (2020) summarize applications of natural language processing (NLP) methods in marketing. Similarly, Hulland et al. (2018) collect best prac-

tices for survey research, while Podsakoff et al. (2012) inform about method bias in social sciences.

Third, with the advent of big data and AI in marketing, method benchmarks are an increasingly important tool for informing researchers about methods. Focused on specific tasks such as latent moderation analysis (Pieters et al. 2022), text analysis (Shankar and Parsana 2022), or image classification (Tetzlaff et al. 2024), they aim to provide guidance for the application of task-specific methods. For example, Pieters et al. (2022) compare six methods for Latent Moderation Analysis and Shankar and Parsana (2022) provide an empirical comparison of the NLP methods commonly used in marketing.

Each of the three tools mentioned has inherent design limitations. Editorials, for example, may lack comprehensive coverage due to topical interests or journal-specific focuses. Tutorials and benchmarks might exhibit bias by concentrating on specific subsections of marketing research, relying on keyword-based searches in publication databases, or unintentionally relying on citation networks (including like-minded articles, while excluding others).In summary, while these approaches remain viable, their inherent limitations prevent them from being optimal for large-scale, objective dissemination of research information.

3 Using AI to Inform Scholars on their Field

As emphasized in the preceding section, conventional avenues for accessing academic information on ongoing research and methodological trends remain valid, although not entirely ideal due to inherent shortcomings. Thus, we propose an innovative AI-based approach towards structured information dissemination from article full-texts. In this section, we delineate the framework of this approach that aims to rectify the limitations inherent in existing information channels.

3.1 Proposing an Automated Information Extraction Framework

Following previous literature on large-scale scientometric analysis (e.g., Grewal et al. 2020; Donthu et al. 2021), we extract rich information from a set of 20,000 marketing articles on the type of contribution, data sources and types, analytical methods, investigated constructs, and key findings. This structured information helps us to inform our analysis and thus enables marketing scholars to understand and explore developments in the field. To do so, we build an automatic extraction pipeline that analyzes the full-text manuscripts of all articles. This pipeline follows a five-step process: (1) define relevant information properties, (2) create datasets for the classification of relevant sentences in the article, (3) fine-tune a

sentence classification model to identify relevant text, (4) extract information using a Large Language Model (LLM), and (5) validate the extracted information using keyword search and manual inspection.

3.2 Extractive AI Model Tuning

Based on a small set of manually identified sentences, we create training data sets and fine-tune a few-shot text classification model (Tunstall et al. 2022) to detect potentially relevant sentences in the full text of the articles. This reduces noise from irrelevant context (e.g., literature review), which increases extraction precision and reduces computational cost (Liu et al. 2023). We use OpenAI's ChatGPT API to extract all information entities from the filtered sentences of the publications. Lastly, to minimize possible hallucinations of the language model, we validate all extracted categorical information through a keyword search in the full text of the publication and manual inspection.

4 Results

Next, we put our proposed AI approach to use and extract exemplary methodological information from thousands of marketing publications to inform scholars on the development of their field. More precisely, our sample contains 21,232 publications from journals in marketing research published between 2010 and 2023, published in all journals related to marketing, including all major journal rankings (FT50, UT Dallas, ABCD, Cnrs, EJL, ABS, Den, and Scimago) with a minimum rank of B.

Publications were accessed through Web of Science and their full-text manuscripts were manually downloaded by research assistants. We exclude articles without any empirical dataset from our analyses (i.e., editorials, tutorials, and commentaries).

Figure 1 shows the average number of analysis methods used per publication between 2010 and 2023. This number grew from 3.51 in 2010 to 4.22 in 2023, an increase of 20.36% that indicates a greater methodological complexity within individual articles. Probably, this is driven in part by the introduction of novel analysis methods (e.g., text classification) that feed into econometric models and are seldom used standalone. In addition to the increasing number of analysis methods used per publication, the increasing complexity for marketing researchers to navigate the field is highlighted by an increasing number of methods that are used each year in marketing publications (average growth rate of 9.61%).

Table 1 illustrates the developments in the use of data sources in marketing research. Although the use of survey and simulation data remains roughly equal between 2010 and 2023

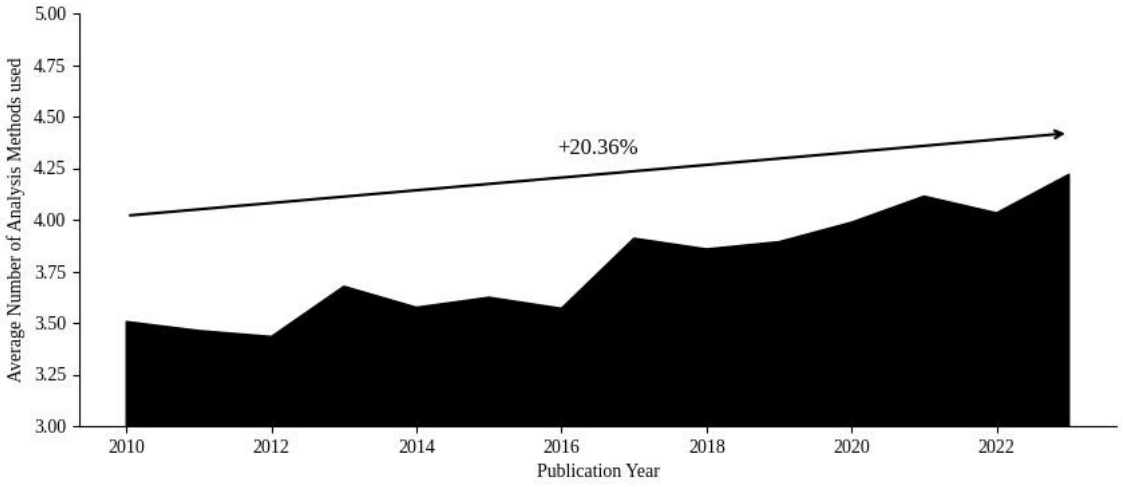


Figure 1: The average number of analysis methods used per publication between 2010 and 2023. The mean number of methods used increased by 20.36% between 2010 and 2023.

(both with a slight decline), the percentage of publications using observational data increased by 14.03% to 65.11% in 2023 and the use of experiment data increased by more than 40%. This reflects the transformation in marketing research that Kumar (2018) commented on in his recent editorial, where the growing popularity of digital platforms allows easier access to observational data and simplifies conducting field experiments in online environments.

Data Source (Year)	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Pct. Change
Uses Survey Data (%)	62.08	63.19	59.36	64.88	61.80	65.93	64.89	63.53	63.79	64.36	64.57	61.71	58.90	59.51	-4.14%
Uses Experiment Data (%)	31.64	33.42	36.45	33.89	37.95	36.49	33.83	33.11	34.70	34.45	35.27	35.39	37.32	44.56	40.83%
Uses Observational Data (%)	57.10	55.79	55.60	57.24	56.68	58.87	59.50	62.92	57.72	60.84	59.83	59.27	64.16	65.11	14.03%
Uses Simulation Data (%)	4.52	5.40	3.85	4.39	3.95	4.42	3.47	3.23	3.69	3.28	3.52	4.68	4.44	4.38	-3.01%

Table 1: The percentage of publications between the year of 2010 and 2023 that use specific sources of data.

Lastly, Figure 2 investigates how the use of specific analytical methods in marketing is related to the impact of an article. Traditional methods of analysis (e.g., regression models and analysis of variance) are still the most popular forms of analysis in marketing research.

Not surprisingly, machine learning (ML) methods in general and specific analytical methods from the wider machine learning domain, such as Linguistic Inquiry and Word Count, are increasingly used in marketing research publications (see left panel of Figure 2). The right panel of Figure 2 compares the impact of traditionally popular methods related to ML on an article's number of citations.

Publications that exploit traditional methods of analysis (e.g., regression, analysis of variance, factor analysis), accumulate an average of 53.83 citations during the first 10 years of publication, whereas a publication that leverages ML-based methods (e.g., NLP) will be

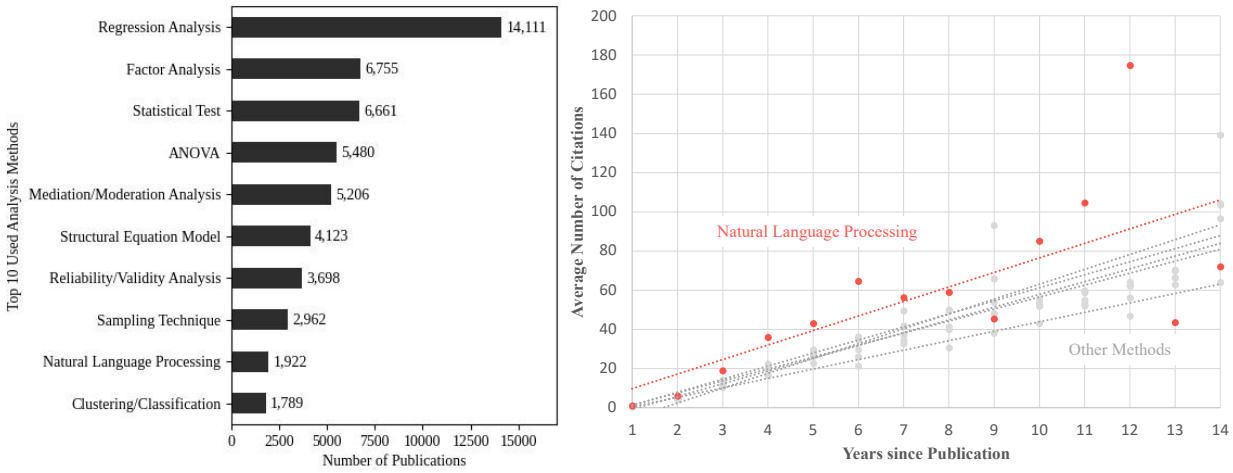


Figure 2: The left panel displays the top 10 most used methods of analysis in marketing research. The graph highlights that only a limited number of research methods are used in most publications. The right panel showcases the impact of method use on citation count for the top 5 methods and NLP. Emerging state-of-the-art methods in NLP influence a fast growth in citations.

cited 62.17 times on average (+30.22% compared to traditional analysis methods).

Researchers and practitioners interested in our methodological approach may find the BERD@NFDI web application (https://www.berd-nfdi.de/analytics-portal-new/), which is freely accessible, to be of great benefit. This application applies the processing techniques discussed in this paper with a somewhat different emphasis, enabling users to investigate the spread of ML methods in business and economics research fields. In our view, this tool is essential for aiding marketing researchers and practitioners in managing the extensive annual literature output, offering a search function for articles based on structured data extraction.

5 Discussion

Informing scholars on the development of themes and methods in their field of research is key to academic success. Although there are classic vehicles for information dissemination, such as editorials, tutorials, and method benchmarks, none of these are optimal in design when it comes to exploring novel directions, methods, and themes in a structured and systematic way.

To the best of our knowledge, we are the first paper to present a novel AI-based approach for the automatic extraction of information from thousands of marketing articles from 2010 onward. With the help of state-of-the-art, few-shot classification, and conversational AI, we extract six relevant information properties from the full-text of marketing research publications. Analyzing a sample over 21,000 marketing publications between 2010 and 2023, our

preliminary results indicate an increase in the diversity of analysis methods that are applied in marketing research over time. Moreover, we find that observational data is increasingly being used. Together, this suggests that marketing scholars put greater emphasis on the use of non-reactive (online) observations and quantitative methods, as they benefit from the realm and availability of big data through digitization and online platforms. Importantly, our results indicate a positive link between the explorative use of emerging methods for analysis (i.e., machine learning) and the impact of an article (i.e., citation count) that surpasses the impact of popular traditional methods (e.g., Structural Equation Modeling).

Our results highlight the applicability of AI for scholarly information extraction and insight generation at scale.These are of great value to researchers who want to gain a more profound understanding of fast-paced developments and trends in the marketing field. Through structured information and easier access to relevant articles, our results help scholars identify interesting topical areas alongside relevant methods for these. We hope this gives rise to a better alignment of the exploration-exploitation trade-off in research, where scholars can leverage past knowledge to explore and apply new themes and methods. With our innovative approach towards academic information dissemination, we resolve many shortcomings of existing attempts (e.g., editorials and benchmarks) and common dangers for the academic business such as falling into citation circles, constantly re-using and citing old methods, themes, and insights.

This analysis includes some limitations that call for future research. First, our data sample includes only publications published after 2010. Although 2010 seemed like a reasonable cut-off point for our data collection process, due to the recent shift of research driven by digitalization and machine learning-based research methods, the use of analytical methods in marketing research is influenced by the field's historical development. Future research could extend the analysis time frame to gather additional information. Lastly, our analysis is limited to only a few information properties extracted from the publications. Many other points of view, despite the focus on methods and data sources used in our analysis, could help researchers navigate the transforming field of marketing research.

References

Becker, Laura, Kristof Coussement, Marion Büttgen and Ellen Weber (2022), "Leadership in innovation communities: The impact of transformational leadership language on member participation," *Journal of Product Innovation Management*, 39 (3), 371–393.

Berger, Jonah, Ashlee Humphreys, Stephan Ludwig, Wendy W. Moe, Oded Netzer and David A. Schweidel (2020), "Uniting the Tribes: Using Text for Marketing Insight," *Journal of Marketing*, 84 (1), 1–25.

Berger, Jonah and Katherine L. Milkman (2012), "What Makes Online Content Viral?" *Journal of Marketing Research*, 49 (2), 192–205.

Bornmann, Lutz and Rüdiger Mutz (2015), "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references," *Journal of the Association for Information Science and Technology*, 66 (11), 2215–2222.

Donthu, Naveen, Werner Reinartz, Satish Kumar and Debidutta Pattnaik (2021), "A retrospective review of the first 35 years of the International Journal of Research in Marketing," *International Journal of Research in Marketing*, 38 (1), 232–269.

Grewal, Rajdeep, Sachin Gupta and Rebecca Hamilton (2020), "The Journal of Marketing Research Today: Spanning the Domains of Marketing Scholarship," *Journal of Marketing Research*, 57 (6), 985–998.

Grewal, Rajdeep, Sachin Gupta and Rebecca Hamilton (2021), "Marketing Insights from Multimedia Data: Text, Image, Audio, and Video," *Journal of Marketing Research*, 58 (6), 1025–1033.

Hulland, John, Hans Baumgartner and Keith Marion Smith (2018), "Marketing survey research best practices: evidence and recommendations from a review of JAMS articles," *Journal of the Academy of Marketing Science*, 46 (1), 92–108.

Inman, J Jeffrey, Margaret C Campbell, Amna Kirmani and Linda L Price (2018), "Our Vision for the Journal of Consumer Research: It's All about the Consumer," *Journal of Consumer Research*, 44 (5), 955–959.

Kumar, V. (2018), "Transformative Marketing: The Next 20 Years," *Journal of Marketing*, 82 (4), 1–12.

Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni and Percy Liang (2023), "Lost in the Middle: How Language Models Use Long Contexts," arXiv:2307.03172 [cs].

Ludwig, Stephan, Ko de Ruyter, Mike Friedman, Elisabeth C. Brüggen, Martin Wetzels and Gerard Pfann (2013), "More than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates," *Journal of Marketing*, 77 (1), 87–103.

Moorman, Christine, Harald J. van Heerde, C. Page Moreau and Robert W. Palmatier (2019), "Challenging the Boundaries of Marketing," *Journal of Marketing*, 83 (5), 1–4.

Peng, Ling, Geng Cui, Ziru Bao and Shuman Liu (2022), "Speaking the same language: the power of words in crowdfunding success and failure," *Marketing Letters*, 33 (2), 311–323.

Pieters, Constant, Rik Pieters and Aurélie Lemmens (2022), "Six Methods for Latent Moderation Analysis in Marketing Research: A Comparison and Guidelines," *Journal of Marketing Research*, 59 (5), 941–962.

Podsakoff, Philip M., Scott B. MacKenzie and Nathan P. Podsakoff (2012), "Sources of Method Bias in Social Science Research and Recommendations on How to Control It," *Annual Review of Psychology*, 63 (1), 539–569.

Shankar, Venkatesh and Sohil Parsana (2022), "An overview and empirical comparison of natural language processing (NLP) models and an introduction to and empirical application of autoencoder models in marketing," *Journal of the Academy of Marketing Science*, 50 (6), 1324–1350.

Sudhir, K. (2016), "Editorial—The Exploration-Exploitation Tradeoff and Efficiency in Knowledge Production," *Marketing Science*, 35 (1), 1–9.

Tetzlaff, Keno, Jochen Hartmann and Mark Heitmann (2023), "The Bigger Picture: A Comprehensive Review and Recommendations for Automated Image Classification in Marketing," *SSRN Electronic Journal*, .

Tetzlaff, Keno, Maximilian Witte, Jochen Hartmann and Mark Heitmann (2024), "The Language of Images: Performance of Image Classification Paradigms in Marketing," (accessed 2025-01-15), https://papers.ssrn.com/abstract=4224968.

Toubia, Olivier (2022), "Editorial: A New Chapter or a New Page for Marketing Science?" *Marketing Science*, 41 (1), 1–6.

Tunstall, Lewis, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat and Oren Pereg (2022), "Efficient Few-Shot Learning Without Prompts," arXiv:2209.11055 [cs].

Woolley, Kaitlin and Marissa A. Sharif (2021), "Incentives Increase Relative Positivity of Review Content and Enjoyment of Review Writing," *Journal of Marketing Research*, 58 (3), 539–558.

Yadav, Manjit S. (2010), "The Decline of Conceptual Articles and Implications for Knowledge Development," *Journal of Marketing*, 74 (1), 1–19.

IV. APPENDIX

1. SUMMARIES / ZUSAMMENFASSUNGEN

A. The Language of Images: Performance of Image Classification Paradigms in Marketing

Summary

Images say more than a thousand words. But can marketing leverage language modeling concepts to study image content? Marketing utilizes image classification for studying how advertising, social media, or e-commerce images relate to consumer perceptions and economic outcomes. To accomplish this, marketing publications apply variants of convolutional neural networks that analyze local image patterns by examining neighboring pixels. Recent transformer architectures, inspired by language modeling, study images more holistically by learning relationships across distant image parts. Even newer vision language models based on generative AI advances create detailed text interpretations of what they 'see' in images. We study the benefits of these advances based on 18 marketing-related datasets that cover what and who is visible, as well as how images are perceived. On average, language modeling concepts improve image classification accuracy by more than 10 percentage points. When training data is abundant, transformer architectures perform best and also most consistently across datasets. Performance of vision language models varies. Relative to alternatives, these models perform strongest with limited training data and for complex tasks focused on how images are perceived. Combining them with transformer-inspired architectures as a multi-paradigm ensemble achieves the best of both worlds, with the highest and most consistent performance across all tasks and datasets we study.

Zusammenfassung

Bilder sagen mehr als tausend Worte. Aber kann das Marketing Sprachmodellierungskonzepte nutzen, um Bildinhalte zu untersuchen? Das Marketing nutzt die Bildklassifizierung, um zu untersuchen, wie Bilder aus Werbung, sozialen Medien oder E-Commerce mit der Wahrnehmung der Verbraucher und den wirtschaftlichen Ergebnissen zusammenhängen. Zu diesem Zweck werden in Marketing-Publikationen Varianten von neuronalen Faltungsnetzwerken eingesetzt, die lokale Bildmuster durch die Untersuchung benachbarter Pixel analysieren. Neuere Transformer-Architekturen, die von der Sprachmodellierung inspiriert sind, untersuchen Bilder ganzheitlicher, indem sie Beziehungen zwischen entfernten Bildteilen lernen. Sogar neuere Bildsprachmodelle, die auf generativen KI-Fortschritten basieren, erstellen detaillierte Textinterpretationen dessen, was sie in Bildern „sehen“. Wir untersuchen die Vorteile dieser Fortschritte anhand von 18 marketingbezogenen Datensätzen, die erfassen, was und wer sichtbar ist und wie Bilder wahrgenommen werden. Im Durchschnitt verbessern die Sprachmodellierungskonzepte die Genauigkeit der Bildklassifizierung um mehr als 10 Prozentpunkte. Wenn viele Trainingsdaten vorhanden sind, erzielen Transformer-Architekturen die besten Ergebnisse und sind auch über die Datensätze hinweg am konsistentesten. Die Leistung von Bildsprachmodellen variiert. Im Vergleich zu den Alternativen schneiden diese Modelle bei begrenzten Trainingsdaten und bei komplexen Aufgaben, die sich auf die Wahrnehmung von Bildern konzentrieren, am besten ab. Kombiniert man sie mit Transformator-Architekturen als Multi-Paradigma-Ensemble, erhält man das Beste aus beiden Welten, mit der höchsten und konsistentesten Leistung über alle von uns untersuchten Aufgaben und Datensätze hinweg.

B. Valuable visual variety? How temporal dynamics of visual features in video ads affect ad effectiveness

Summary

In today's digital marketing landscape, video ads play a critical role in capturing consumer attention amid widespread content saturation. This paper investigates the impact of visual variety — defined as the temporal variability of a video's visual features — on advertising effectiveness. Utilizing machine learning, we develop a multidimensional visual variety measure and apply it to over 3,000 video ads across two media channels (TV, social media) and various product categories. Three studies reveal an inverted U-shaped relationship between visual variety and advertising effectiveness. A controlled lab experiment reveals that this relationship is mediated by both stimulation and comprehension. Whereas too little visual variety can lack stimulation, and too high visual variety can undermine comprehension, ads with medium visual variety strike an effective balance, being both simulating and comprehensible for consumers. We provide guidelines for advertisers on optimizing visual variety and introduce a web-based tool that calculates visual variety scores for videos.

Zusammenfassung

In der heutigen digitalen Marketinglandschaft spielen Videoanzeigen eine entscheidende Rolle, wenn es darum geht, die Aufmerksamkeit der Verbraucher inmitten der weit verbreiteten Inhaltssättigung zu gewinnen. In dieser Arbeit wird der Einfluss der visuellen Vielfalt - definiert als die zeitliche Variabilität der visuellen Merkmale eines Videos - auf die Werbewirksamkeit untersucht. Mithilfe von maschinellem Lernen entwickeln wir ein mehrdimensionales Maß für die visuelle Vielfalt und wenden es auf über 3.000 Videoanzeigen in zwei Medienkanälen (TV, soziale Medien) und verschiedenen Produktkategorien an. Drei Studien zeigen eine umgekehrt U-förmige Beziehung zwischen visueller Vielfalt und Werbewirksamkeit. Ein kontrolliertes Laborexperiment zeigt, dass diese Beziehung sowohl durch Stimulation als auch durch Verständnis vermittelt wird. Während eine zu geringe visuelle Vielfalt zu einem Mangel an Stimulation und eine zu große visuelle Vielfalt zu einer Beeinträchtigung des Verständnisses führen kann, stellen Anzeigen mit einer mittleren visuellen Vielfalt ein effektives Gleichgewicht dar, das für die Verbraucher sowohl simulierend als auch verständlich ist. Wir stellen Richtlinien für Werbetreibende zur Optimierung der visuellen Vielfalt zur Verfügung und stellen ein webbasiertes Tool vor, das die Bewertung der visuellen Vielfalt von Videos berechnet.

C. Methods in Marketing: Leveraging Artificial Intelligence for Academic Information Dissemination

Summary

In the dynamically evolving field of marketing research, it is imperative that scholars follow the substantive and methodological developments. However, in the context of thousands of newly published articles each year, increasing diversity in methods, and further acceleration through the advent of artificial intelligence (AI), keeping track becomes a daunting task for marketing researchers. Some classical vehicles for information dissemination exist (e.g., editorials, benchmarks), but are typically not designed to cover developments in the entire marketing area and rely on error-prone manual work. We propose an AI-based approach for comprehensive and structured analysis of the entire marketing literature from 2010 to 2023. We demonstrate the increasing complexity in marketing (i.e., increase in both articles and methods per article), identify recent developments in the field, and strive to provide more structured access to article information.

Zusammenfassung

In dem sich dynamisch entwickelnden Bereich der Marketingforschung ist es unerlässlich, dass Wissenschaftler den inhaltlichen und methodischen Entwicklungen folgen. Angesichts Tausender neu veröffentlichter Artikel jedes Jahr, zunehmender Vielfalt an Methoden und weiterer Beschleunigung durch den Aufstieg der künstlichen Intelligenz (KI) wird es jedoch zu einer Herausforderung für Marketingforscher, den Überblick zu behalten. Es gibt zwar einige klassische Mittel zur Informationsverbreitung (z. B. Leitartikel, Benchmarks), diese sind jedoch in der Regel nicht darauf ausgelegt, Entwicklungen im gesamten Marketingbereich abzudecken und beruhen auf fehleranfälliger manueller Arbeit. Wir schlagen einen KI-basierten Ansatz für eine umfassende und strukturierte Analyse der gesamten Marketingliteratur von 2010 bis 2023 vor. Wir zeigen die zunehmende Komplexität im Marketing auf (d. h. Anstieg sowohl der Artikelanzahl als auch der Methoden pro Artikel), identifizieren aktuelle Entwicklungen auf dem Gebiet und streben danach, einen strukturierteren Zugang zu Artikelinformationen zu bieten.

2. AFFIDAVIT

AFFIDAVIT

I hereby declare, Keno Tetzlaff, in lieu of an oath, that I have written the dissertation entitled

"Leveraging Machine Learning for Modelling Unstructured Data in Marketing"

autonomously - and if in cooperation with other scientists as described in the attached statement - according to § 6, subsections 4, 5, 7 of the prevailing doctoral regulations of the Faculty of Business Administration, and that I did not use any other aids than those I indicated herein. The parts taken literally or by sense from other works than mine are marked as such.

I assure that I did not take advantage of any commercial doctoral consultation nor was my work accepted or judged insufficient in an earlier doctoral procedure at home or abroad.

I assure that this printed copy with my original signature corresponds 100 % to the pdf file which I handed in for examination; and that all further prints of my thesis which might incur later shall correspond in full to this original printed thesis and the submitted pdf file.

_____ _____

Place, Date Signature