ARTICLE IN PRESS

International Journal of Research in Marketing xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Research in Marketing

journal homepage: www.elsevier.com/locate/ijresmar

Language of images: Classifying marketing images with transformers and vision language models

Maximilian Witte [a,*], Mark Heitmann [a,b], Jochen Hartmann [c], Keno Tetzlaff [a]

[a] University of Hamburg Business School, Moorweidenstrasse 18, 20148, Hamburg, Germany
[b] Nova School of Business and Economics, Universidade NOVA de Lisboa, Campus de Carcavelos, Portugal
[c] TUM School of Management, Technical University of Munich, Arcistrasse 21, 80333, Munich, Germany

ARTICLE INFO

Keywords:
Visual marketing
Automated image classification
Generative AI
Vision language models
Vision transformers
Convolutional neural networks

ABSTRACT

Visual communication is central to marketing. With the help of convolutional neural networks (CNNs) marketing has labeled large image datasets to understand visual impact. However, CNNs focus on local cues (e.g., smiles). They can miss marketing-relevant meanings shaped by context and configuration (e.g., joyful vs. sarcastic smiles). Recent advances like transformer-based vision models (TVMs) apply text-analytical concepts to image data. Vision language models (VLMs) such as GPT-5 or Phi-4 jointly represent images and text. These richer linguistic representations might succeed in classifications where CNNs fall short. Unlike CNNs, pretrained VLMs require no additional training, even for new image-related tasks. However, it remains unclear how accurate they classify marketing-relevant labels. Which of these paradigms and classification models should marketing rely on? Is any single model best suited for all applications? Drawing on prior marketing publications, we identify 18 datasets covering what and who appears in images, and how images are perceived. VLMs such as GPT-5 and Phi-4 achieve state-of-the-art accuracy across a wide range of image-related tasks without requiring task-specific fine-tuning. However, they should not be trusted blindly. They can result in unexpectedly high error rates for some tasks. A multi-paradigm ensemble of TVMs and VLMs can overcome these challenges. We conclude with recommendations when to test which models.

Introduction

Images attract attention and effectively communicate emotions, values, and product benefits. Consumers post images that contain brand content (Beichert et al., 2024) and marketers distribute a wealth of advertising images (e.g., Lee, 2021; Pieters et al., 2010; Rietveld et al., 2020). Marketers also crowd-source various design proposals or use generative AI to efficiently create many candidates for online advertising (Hartmann et al., 2025; Heitmann et al., 2025). Across these and many other applications, large-scale image classification supports marketing research by identifying the visual characteristics that drive impact and by enabling the selection of effective images from extensive content pools. This is achieved through automated assignment of relevant labels to entire images.

Convolutional neural networks (CNNs) have advanced marketing's ability to classify images and to understand and manage visual impact (see Dzyabura et al. 2022 for a broader image analytics overview). For example, Lee (2021) trains a CNN classifier on the

* Corresponding author.
E-mail addresses: maximilian.witte@uni-hamburg.de (M. Witte), mark.heitmann@uni-hamburg.de (M. Heitmann), jochen.hartmann@tum.de (J. Hartmann), keno.tetzlaff@uni-hamburg.de (K. Tetzlaff).

https://doi.org/10.1016/j.ijresmar.2026.01.001
Received 4 August 2023
Available online 11 January 2026

Please cite this article as: Maximilian Witte et al., *International Journal of Research in Marketing*, https://doi.org/10.1016/j.ijresmar.2026.01.001

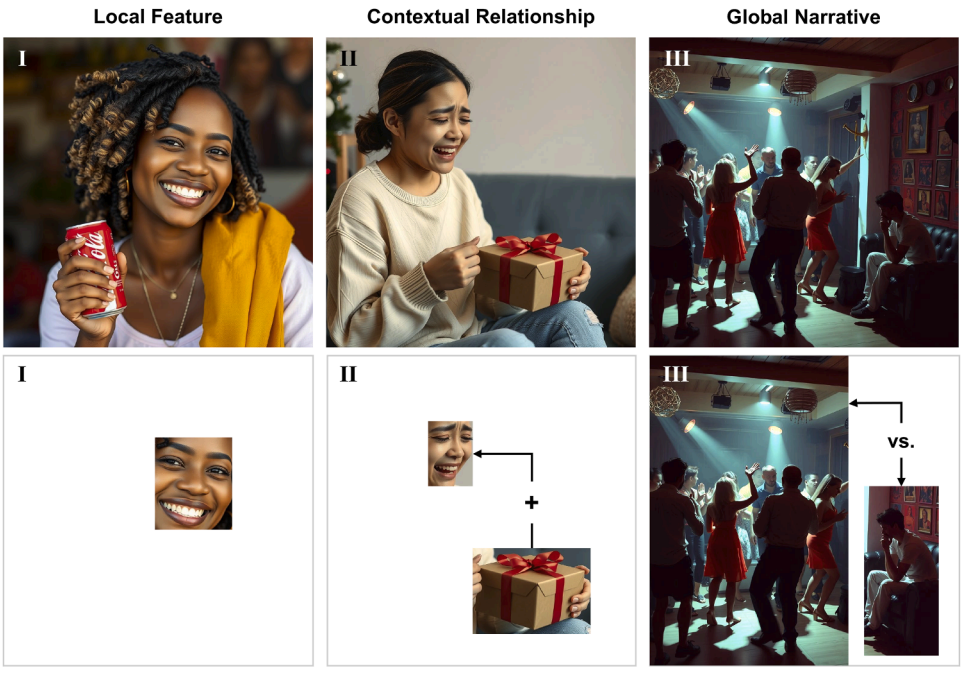

Fig. 1. Image classification scenarios with increasing difficulty. *Note:* Upper panel shows original images (FLUX-generated). The lower panel indicates key image regions and relationships needed for correct classification.

emotionality of brand images to understand whether emotionality drives shopping basket size. Other research classifies facial profile images to study how visible ethnicity relates to online responses (Gunarathne et al., 2022; Zhang et al., 2021).

While useful, CNNs have limitations in terms of accessibility and accuracy. They often need sophisticated training or fine-tuning, which requires expertise and computational resources. Among other things, it requires manually annotated training data and appropriate hyperparameter configurations. Even when done well, CNN classifications do not consistently match human judgments, limiting applications. Poor accuracy matters both practically and scientifically. For example, Dzyabura et al. (2023) use image classification to understand which images reduce online fashion return rates. In their research, 8.76% in classification accuracy translates into a 1% profit impact. Low accuracy also impedes the ability to detect econometric effects and can make scientific conclusions invalid (Frankel et al., 2022; Hartmann et al., 2019; Jaidka et al., 2020).

Recent computer vision advances promise improvements in accuracy and ease of application. This research evaluates alternative methods and paradigms from an applied perspective, presenting empirical evidence on accuracy and practical guidance for model choice in marketing.

To illustrate how the available paradigms differ, consider the three examples in Fig. 1. Marketing is often interested in detecting visual brand presence and in which contexts brand appear. For example, Hartmann et al. (2021) differentiate brand selfies (ego perspective) and consumer selfies (consumers looking into the camera) and classify social media imagery to understand which format is most desirable for marketing. Panel I shows an illustrative image. With CNN fine-tuning, these authors achieve high accuracy (predicting above 90% of human judgments correctly). CNNs work well for classifying *what* is visible. They are designed to identify strong signals based on informative groups of pixels (e.g., presence of face, hand, and product signals a consumer selfie). Similarly, key visual cues can suffice to detect *who* is visible in an image (e.g., in terms of gender, age, or ethnicity) (Gunarathne et al., 2022).

Next, consider panel II. Images have the power to evoke emotions about products, brands or more generally a depicted scene. *How* consumers perceive an image, often depends heavily on context and image configuration (Heitmann et al., 2025); which CNNs might not capture in full detail. For example, similar facial expressions can lead to a range of plausible interpretations–such as joy, distress, or embarrassment–depending on the context of the image, like receiving a gift, which helps make the emotion easier to interpret. Such relationships are more challenging to classify with CNNs that put little emphasis on relative location. Recently emerged transformer-based vision models (TVMs) include conceptual solutions. Rather than simply grouping pixels in an unordered fashion, TVMs follow techniques from text analysis, where relationships between textual elements matters and creates meaning. TVMs weight joint occurrence of image elements accordingly (Dosovitskiy et al., 2020). Conceptually, this capacity is useful to classify the facial expression in Panel II.

Panel III involves an entire visual story that is central to how the image is perceived. A simple count of the guests celebrating suggests overall positive emotions. However, the human eye notices the isolated single guest. If these more complex relations are not accounted for, automated classification is unlikely to match human perceptions. Vision language models (VLMs) like OpenAI's GPT-5 and Microsoft's Phi-4 (Abdin et al., 2024; OpenAI, 2025a), also known as multi-modal large language models, might better

capture such visual-emotional dynamics. These models are trained jointly on text and image data. By representing visual and linguistic information in joint multi-modal layers of abstractions VLMs relate visuals and rich textual descriptions more intricately than CNNs and TVMs (Bordes et al., 2024). With these capabilities, VLMs can create images based on detailed text prompts or conversely respond to targeted queries about what is visible on images. The latter is of primary interest to this investigation as it might allow better and easier image classification than TVMs and CNNs.

Although VLMs have not been evaluated for image classification in marketing, their capabilities, but also ease of use makes them appealing. Instead of connecting pixels with predefined labels, VLMs connect pixels with general-purpose concepts. These multimodal capabilities allow them to generalize to any task and linguistic conclusion without the need for fine-tuning, which makes any VLM application straightforward. Unlike conventional CNNs and TVMs, that both require adapting neural network weights to link image signals with novel labels, applying pretrained VLMs require neither AI training knowledge from users nor computational training resources. They also do not require manually annotated training images, which can be complex and costly to collect, in particular when many label categories exist or variance in subjective perceptions requires multiple annotations (Schamp et al., 2024). For these reasons, VLMs could enable marketing researchers to more efficiently explore many more questions than would be feasible with CNNs or TVMs. Furthermore, their inherent linguistic narrative might enable them to accurately code complex image patterns that TVMs and CNNs might miss.

However, VLMs can also hallucinate and create factually incorrect output that is unsupported by data. Their actual training material is seldom disclosed, so it is unclear which tasks are beyond their training and when they rely on out-of-scope reasoning (Luo et al., 2024; Yang et al., 2025). Furthermore, an unknown set of human-related queries can be restricted in these models to prevent misuse (e.g., human gender, age, ethnicity, beauty; Feuerriegel et al., 2024). How effective are VLMs compared to CNNs or TVMs in actual marketing-related applications? Which models should marketing rely on?

So far, no large scale comparison between VLMs and TVMs exists. The benefits and limitations compared to CNNs are not clear. A better understanding is important because marketing can be tempted to confuse impressive lighthouse performances with overall reliability (Hartmann et al., 2023). Comprehensively assessing all alternative CNNs, TVMs or VLMs including the associated hyperparameter optimization is computationally and analytically challenging to do for each application. Guidance is needed which alternatives are most promising to understand what warrants detailed assessments.

Drawing on more than 2200 experiments using nine classification methods across 18 datasets, we find that TVMs, such as ConvNeXt, outperform all major CNNs previously applied in marketing with improvements often exceeding 10% in classification accuracy. Since TVMs deliver consistently higher and practically meaningful accuracy across all examined marketing applications, CNNs have limited potential for future applications. VLMs show greater variability in performance across tasks. They can deliver high-quality predictions without requiring task-specific fine-tuning. However, they can also underperform dramatically, sometimes being up to 40% less accurate than TVMs. We find that combining VLMs and TVMs within a multi-paradigm ensemble leverages the strengths of both approaches with consistently high accuracy.

The remainder of this article is structured as follows. First, we review existing benchmarks and existing marketing applications and then discuss the conceptual differences between the image classification paradigms in more detail. Empirically, we compile datasets, tasks, and training data sample sizes that collectively reflect the range of marketing applications that we identified. For these typical marketing settings, we evaluate the performance of VLMs in comparison to widely used top-performing CNNs and TVMs. To better understand the reasons behind the observed performance differences, we extend our analysis by testing whether a multi-paradigm ensemble of VLMs and TVMs can compensate for the relative disadvantages of each approach. We conclude with practical recommendations to help marketing choose the right image classification paradigm and a step-by-step framework to apply and assess vision language models.

1. Prior research

1.1. Performance benchmarks

We are not aware of any comprehensive comparison of VLMs, TVMs, and CNNs. Encouraging benchmark studies outside of marketing suggest TVMs can outperform CNNs in terms of classification accuracy (e.g., Maurício et al., 2023). However, the datasets of these comparisons do not resemble the tasks and dataset characteristics that marketing is interested in. They train models on very large labeled training data, including hundreds of thousands of training examples limited to *what* is visible in an image (e.g., Deng et al. 2009). Marketing has different objectives. It is often faced with high effort of creating training data (Schamp et al., 2024). It often lacks the computational resources to run large-scale model training and is also interested in perceptual concepts. Despite the conceptual benefits, we lack evidence of whether and to what extent TVMs perform as well for typical marketing applications that include tasks around who is visible or how images are perceived. Since TVMs are often more complex models they might require more labeled training images than marketing can provide.

Due to their broad capabilities, prior assessments test VLMs on various tasks such as question answering, generic knowledge or producing working software code. For these tasks VLMs like GPT-5 and Phi-4 achieve top performance (Abdin et al., 2024). Similarly, in image-related tasks such as object counting, text recognition, or spatial relationships comprehension these models perform well (Abdin et al., 2024; Yao et al., 2025). To the best of our knowledge, a comprehensive image classification comparison with CNNs and TVMs does not exist. Such comparisons are particularly relevant for marketing, where research has primarily relied on CNNs. When examining VLM paradigms in isolation, recent evidence indicates that VLMs can represent both a strength and a limitation for

image classification. They help generalizing to unseen tasks, but can also produce hallucinations for standard tasks even ones that the model was originally trained on (Cooper et al., 2025).

Although we are not aware of an image classification comparison in marketing, various benchmarks on text classification have appeared (e.g., Alantari et al., 2021; Frankel et al., 2022; Jaidka et al., 2020; Shankar & Parsana, 2022). Although text classification is different from image classification, several insights might translate. For example, Hartmann et al. (2023) find that transformer-based text classification can achieve approximately 20% higher accuracy than conventional lexicon-based approaches. These are often used without further validation, yet can deviate substantially from human perception. They also find that major leaps in performance were due to entirely new paradigms (e.g, transformer architectures with transfer learning instead of traditional machine learning). The differences of the individual methods within each paradigm are smaller in comparison. Furthermore, Alantari et al. (2021) find that benchmark results on individual datasets do not generalize well without accounting for contextual factors, suggesting that meaningful comparisons in marketing research must be based on relevant applications and should not be restricted by specific datasets or tasks.

In this research, we will therefore investigate whether similar performance differences between paradigms exist in image classification and to which extent lighthouse illustrations of recent image classification methods generalize to various applied marketing tasks and datasets.

1.2. Image classification in marketing research

How does marketing research apply automated image classification? To obtain a comprehensive overview of peer-reviewed applications, we systematically analyze the full text of all articles published between 2012 and August 2025 in the marketing journals from all major journal rankings.[1] We start with 2012 since that is when the first relevant CNNs appeared (Krizhevsky et al., 2012). Based on nearly 27,000 articles that we assess in a semi-automated process, we identify a total of 49 peer-reviewed marketing publications that apply automated image classification (see Web Appendix A for a full overview). Overall, the publications span five distinct research domains, each investigating who, what, or how images influence marketing outcomes. This includes custom classification tasks (e.g., Liu et al., 2020) and extensions of image-related tasks to novel categories (e.g., Chuah & Yu, 2021) such as unknown brands or atypical visual stimuli.

A relevant group of articles applying image classification focuses on *advertising & promotional effectiveness*. This group leverages image classification to understand which visual attributes drive engagement. For instance, Hartmann et al. (2021) use the CNN VGG-16 to assign three custom labels to social media images. 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 and find that the inclusion of a face drives perceived deal quality of promotional images. Both studies use simple, objective annotations, reducing the need for independent ratings and allowing for efficient application to substantial sample sizes.

Second, 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 humanoid robots relate to social media sentiment by training a classification method on a small training sample of 223 images. While emotion classification uses established categories from computer vision, this application extends to novel visual stimuli (humanoid robots) not typically found in standard emotion datasets. Others study the impact of expected viewer judgments and how images are perceived. For instance, Troncoso and Luo (2023) investigate perceived job fit by designing a custom image-related task using freelancer profile images, which they analyze with multiple CNNs (Inception, ResNet, VGG). They find that better subjective fit of personal images drives hiring outcomes.

A third group of research centered around *brand management* also studies how images are perceived. 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. 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. In this application, only a few actual car designs are available for training a prediction model. Conversely, many ratings per design are needed to compute meaningful averages, resulting in limited training data of only contains 203 images. 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 is visible. 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 training dataset of 9,000 images to fine-tune a CNN 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.

These examples illustrate the breadth of insights made possible with automated image classification. Research includes different types of image-related tasks (what is visible, who is visible, how the image is perceived), it varies in the amount of annotated training data (ranging from a few hundred to many thousand observations) and in the number of relevant classes and the classification method (training models from scratch or fine-tuning CNNs like ResNet, VGG, Inception). Marketing builds on established computer vision

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

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., Chuah & Yu, 2021; Peng et al., 2020; 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 & Luo, 2022) • *Who* is on the image (e.g., Bharadwaj et al., 2022; Zhang et al., 2021) • *How* is the image perceived (e.g., Lee, 2021; Liu et al., 2020)
Dataset Characteristics	• *Number of training images:* Ranging from 203 to > 500,000 (Burnap et al., 2023; Zhang et al., 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 (Hartmann et al., 2021; Liu et al., 2020) • *Evaluation:* e.g., train-validation-test split, 5-fold cross validation (Hartmann et al., 2021; Liu et al., 2020) • *Metrics:* e.g., accuracy, AUC-ROC (Burnap et al., 2023; Troncoso & Luo, 2023; Zhang et al., 2021)

Note: CTR = click-through rate, CNN = convolutional neural network, AUC-ROC = area under the curve of the receiver operating characteristic, analysis based on 49 publications in marketing-related journals (see Web Appendix A).

applications (e.g., emotion recognition, product categorization), but also develops entirely novel tasks (e.g., brand selfie classification) or extends existing tasks to new classes or atypical visual examples (e.g., unknown brands or uncommon visual brand cues). Training typically includes 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. While published studies differ in various aspects, one important pattern emerges: all prior publications applied some form of model training; predominantly fine-tuning a pretrained CNN or training an entire model from scratch. All of the applications involved manual data labeling and model training, representing a significant time and resource investment.

It remains unclear whether language-inspired TVMs and VLMs can reduce the required effort for manual annotations and model training and how they would perform compared to the CNNs from previous applications. We will follow these types of applications, datasets, and tasks to investigate this question. Next, we will discuss these alternative image classification paradigms in more detail.

2. Paradigms of image classification

Fig. 2 presents the three main paradigms in image classification and contrasts their methodological foundations. Unlike traditional marketing models with few interpretable parameters (e.g., price elasticities), image analysis requires millions or even billions of parameters to process raw pixel values across color channels. This scale makes direct parameter estimation infeasible, necessitating transfer learning approaches instead.

Within this context, VLMs have the deepest ties between language and images due to their joint training on both image and text data. TVMs are built on ideas from language modeling like the self-attention mechanism in the transformer architecture (see Hartmann et al., 2023, for performance in text analysis), but do not integrate actual language (text) into model training. Novel classes or types of data (e.g., humanoid robots instead of human faces) continue to require fine-tuning to extend the base model capabilities. This means selected layers of these models need to be retrained based on manually annotated training data. Pretrained VLMs, on the other hand, can provide inferences without task-specific fine-tuning.

Convolutional neural networks (CNNs)

CNNs are typically used by either training a model from scratch (Hu et al., 2019) or more recently by fine-tuning pretrained models (e.g., Hartmann et al., 2021). Pretrained CNNs, such as VGG19 or ResNet152, are typically trained on large-scale image datasets such as ImageNet (Deng et al., 2009). Entire open-source neural network models can be downloaded and adapted to any custom task.

CNNs identify image objects based on informative groups of pixels (e.g., associated with eyes or a mouth to identify presence of a human face). They accomplish this by connecting neural network layers of increasing abstraction. At the lowest level of the image data itself, CNNs scan small image sections (e.g., a 3×3 kernel as indicated in Fig. 2) (Semih Kayhan & Van Gemert, 2020). The same scanning process is applied across the entire image, so CNNs can identify patterns no matter where they appear. Repeating the same process across sliding kernels is also parameter-efficient. With fewer parameters to train less training data is needed for fine-tuning models to applied marketing.

Because of this architecture, location information, distance, and possible relationships between local cues carries little information in CNNs (Semih Kayhan & Van Gemert, 2020). This is less relevant for many standard tasks. For example, hotel room marketing might be interested in understanding which rooms to highlight on booking sites such as booking.com. To investigate this questions, images

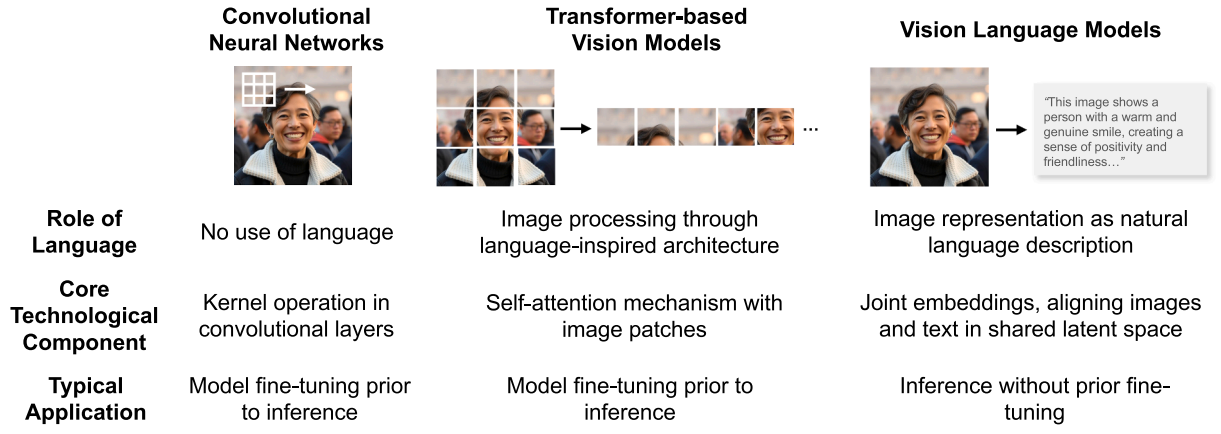

Fig. 2. Overview of image classification paradigms.

need to be classified into visible room types like bathrooms or living and sleeping rooms. For this task it is less relevant where exactly objects like sinks or beds appear or how objects relate to each other. Their presence alone suffices for room classification. Other more nuanced emotional reactions to the same images often hinge on intricate relationships among image content. CNN fine-tuning for such perceptual categories is likely less effective.

This limitation in modeling image relationships persists because fine-tuning adjusts parameter values but not the model architecture. It assumes that low-level features such as edges or simple shapes remain relevant across tasks, and that task-specific meanings can be captured by adapting higher, more abstract layers. Keeping most of the model fixed allows adaptation with relatively little data–often hundreds to a few thousand labeled examples–and reduces the risk of overfitting (poor performance beyond the training data).

Although simpler than training a model from scratch, fine-tuning CNNs still requires time and effort of adapting a neural network, choosing an architecture (e.g., which CNN to use), tuning hyperparameters, and often engaging in trial-and-error to find a workable setup. It also does not address the lack of attention to contextual relationships.

Transformer-based vision models (TVMs)

TVMs incorporate language modeling principles to pixel-based image processing. In language modeling, Vaswani et al. (2017) introduced the transformer architecture, which uses a self-attention mechanism to identify the most relevant relationships between words within a sequence. For example, to understand the sentence "I threw the ball to her", the model must connect the words describing who performed the action ("I") and who received it ("her"). Unlike CNNs that analyze information through local receptive fields, TVMs process relationships between elements through self-attention. They learn which parts of an input sequence to emphasize or to ignore, enabling them to capture long-range dependencies.

The vision transformer model (ViT) (Dosovitskiy et al., 2020) 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). The ViT can flexibly learn long-range dependencies between distant image parts (e.g., smile and gift in Fig. 1). However, fewer assumptions than CNNs translate into larger training data requirements to ensure they perform well (Dosovitskiy et al., 2020). Intermediate TVMs like ConvNeXt build on CNNs, but introduce larger kernels, organizes them in sequences, and add residual connections that help take location into account (Liu et al., 2022), which has proven effective in applications outside of marketing (Maurício et al., 2023). Due to its architecture, ConvNeXt relies more strongly on nearby pixels than ViT and might miss global relationships, but requires fewer training data than ViT.

Conceptually, TVMs might be more powerful than CNNs for complex image classification tasks that require understanding relationships between distant objects or identifying patterns across the entire scene. For example, a person wearing expensive jewelry can signal glamour, a baby wearing similar jewelry can signal playfulness and fun. A marketer wishing to track brand personality based on consumer image posts, needs to capture such differences to draw valid inferences about visible brand personality. On the other hand, the added complexity of TVMs might introduce unwanted errors, should location carry little meaning.

Vision language models (VLMs)

The ability of VLMs to produce detailed textual responses including image descriptions has recently attracted much attention (Bordes et al., 2024). Since VLMs are trained by encoding images together with associated text descriptions into a shared embedding space (Abdin et al., 2024), they are able to represent images through rich semantic representations.

CNNs and TVMs are pretrained on large datasets of many images but relatively few object classes (e.g., 1M labeled images with 1000 object classes in the visual recognition challenge from ImageNet). VLMs capture many more possible classes based on the vast

amount of available text data that includes abstract concepts, objects, and the relationships between them (Bordes et al., 2024). Due to their linguistic capabilities, these models can infer which textual summary fits best to a combination of image content. Irrespective of the set of training categories in databases like ImageNet, any classification category or any classification task can be semantically inferred without task-specific training or fine-tuning. Such zero-shot classification can be done for any novel task. Consequently, VLMs require no data annotation, fine-tuning knowledge, or associated computational resources.

On the other hand, VLMs can produce factually incorrect output (hallucinations) that arise from architectural characteristics of the model (e.g., Shang et al., 2024) and from biases in the pre-training data (e.g., Chen et al., 2024; Huang et al., 2025; Shu et al., 2025). A common issue is that pre-training on large-scale image-text pairs can bias the model toward concepts that frequently co-occur even when these concepts are not visible in an actual image (e.g., the VLM "sees" a computer monitor in an image containing only mouse and keyboard; Chen et al., 2024).

VLM hallucinations are not limited to, but occur most often for, images and tasks outside the training data of the base model (Luo et al., 2024; Yang et al., 2025). Some out-of-scope cases can be anticipated based on their novel marketing nature (e.g., brand personality categories that are developed for a specific brand or specific industry), but most are unknown to marketing. Furthermore VLMs operate and classify within the bounds permitted by their developers. Since VLMs can handle any conceivable classification, developers must implement restrictions to prevent misuse (e.g., profiling on sensitive attributes, claims about people, image based decision-making). These restrictions can undermine accuracy even when the model produces output (e.g., OpenAI, 2025b). Similar to the training data limitations, marketing can anticipate issues for some evident applications (e.g., classifying ethnicity of consumers), but the full scope is not known for each marketing application.

Fine-tuning VLMs might appear a reasonable solution to such hallucination problems. While conceptually appealing, VLM architecture makes fine-tuning VLMs less effective than fine-tuning CNNs and TVMs (Hao et al., 2025; Zhai et al., 2023). Typical fine-tuning datasets in marketing contain only a few hundred to a few thousand labeled examples (see Table 1). This represents a tiny fraction of the data seen during VLM pre-training. As a result, fine-tuning can lead either to negligible changes in model behavior or to severe overfitting, depending on the choice of hyperparameters (Hao et al., 2025; Zhai et al., 2023). Instead of fine-tuning, all parameters are kept constant and custom tasks are approached by directly prompting the model (Rathje et al., 2024).

For these reasons, VLMs are both promising and conceptually constrained. Their actual utility for image classification in marketing remains an empirical question that we study in this research. Numerous VLMs exist to date (e.g., GPT-5, Phi-4, Llama-3.2-Vision). Most widely used VLMs share similar architectural designs and pre-training strategies (e.g., Abdin et al., 2024; MetaAI, 2024; OpenAI, 2025b). While commercial models are trivial to apply, they require users to upload data into cloud services with associated data security concerns. They are also costly to apply for complex tasks with many novel labels or large datasets. As their architectures change, exact replication cannot be guaranteed. We therefore examine a prominent commercial model (GPT-5) and a prominent open-source alternative (Phi-4) that we deploy locally to avoid data security concerns, minimize cost scaling, and ensure replicability (Abdin et al., 2024; OpenAI, 2025b).

3. Empirical analysis of image classification methods

Datasets and methodology

To systematically evaluate how the three 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 (e.g., Hartmann et al., 2025; Wang et al., 2020) and complement these by public image datasets available on Kaggle that mirror typical marketing classification tasks.

Table 2 provides an overview of all datasets in our empirical evaluation. The total compilation of all datasets contains representative examples of each research question and image-related tasks related to what is visible, who is visible, and how image are perceived, allowing us to assess the stability of results across applications. For an illustrative example of each dataset, see Web Appendix B.

3.1. Methods

We compare representative methods from the three key paradigms of image classification. Specifically, we include all CNNs previously used in peer-reviewed marketing research, i.e., InceptionV3 (Szegedy et al., 2015), ResNet152 (He et al., 2015), VGG19 (Simonyan & 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 C for a more detailed description of all methods).

We include the TVMs ViT (i.e., ViT; Dosovitskiy et al., 2020) and ConvNeXt (Liu et al., 2022) in versions that are pretrained on ImageNet to make them comparable to CNNs pretrained on the same data, following common practice in marketing.

We prompt both VLMs (GPT-5 and Phi-4) in the same way to assign the most appropriate class from the set of all possible labels. We use the following prompt template “Assign the best fitting class to describe the image by choosing one of the following classes: [all class names with examples]”, that we adapt slightly when needed (see Web Appendix D for a detailed description). Because VLMs operate solely through prompting, we apply both zero-shot and few-shot in-context learning strategies. For each image, we first obtain two independent predictions, one without example guidance and one with example guidance. We then integrate these two predictions and query the model once more to produce the final classification.

Table 2
Overview of all datasets in the empirical evaluation.

Dataset	Research domain	Image-related task	Number of images	Classes	Image dimensions	Source
AIDA Funnel Car	Advertising & Promotional Effectiveness	How (low, medium, or high AIDA)	559	3	341 × 283	Heitmann et al. (2025)
Brand Logo	Brand Management	What (e.g., electronics, foods, sports)	158,652	9	482 × 396	Wang et al. (2020)
Brand Selfie	Consumer Insights & 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 & Behavior	How (e.g., anger, joy, surpise)	8,350	6	722 × 555	Panda et al. (2018)
Face Sentiment	Consumer Insights & Behavior	How (savory, unsavory)	12,420	2	295 × 350	Good guys-Bad Guys (n.d.)
Fashion Products	Product Design & Management	What (e.g., bags, shoes, watches)	44,441	25	60 × 80	Fashion Product Images (n.d.)
Generated Aesthetics	Advertising & Promotional Effectiveness	How (low, medium, high)	10, 320	3	512 × 512	Hartmann and Exner (2024)
Generated Ethnicity	Consumer Insights & Behavior	Who (e.g., asian, black, white)	10,001	4	256 × 256	Generated.photos (n.d.)
Generated Gender	Consumer Insights & Behavior	Who (female, male)	10,001	2	256 × 256	Generated.photos (n.d.)
Image Quality	Advertising & Promotional Effectiveness	How (e.g., low, medium, very high)	10,073	5	512 × 384	Hosu et al. (2020)
Image Sentiment	Consumer Insights & Behavior	How (e.g., highly positive, neutral, negative)	12,550	5	465 × 400	Image Sentiment Polarity (n.d)
Luxury Brands	Brand Management	What (e.g., Chanel, Gucci, Versace)	2,184	7	150 × 150	Brand Styles (n.d)
Outdoor Scenery	Media Management	What (e.g., buildings, forest, street)	17,034	6	150 × 150	Intel Scene Classification (n.d)
Public Places	Media Management	What (e.g., bathroom, conference room, kitchen)	9,456	21	700 × 700	Bylinskii et al. (2015)
Stock-photo Category	Media Management	What (e.g., arts & culture, business, nature)	8,000	16	200 × 221	Unsplash Images (n.d.)
Store Item Colour	Product Design & Management	What (e.g., black, red, yellow)	6,239	12	224 × 224	Store Items Color (n.d.)
Unbiased Emotion	Consumer Insights & Behavior	How (e.g., anger, fear, love)	3,045	6	837 × 659	Panda et al. (2018)

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

3.2. Training

Since collecting training data can be costly, in particular when multiple annotations are needed, we investigate the role of the training data size. Note, training data only matters for CNNs and TVMs as VLMs are used without task-specific training. 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), reflecting typical training data sizes in marketing.

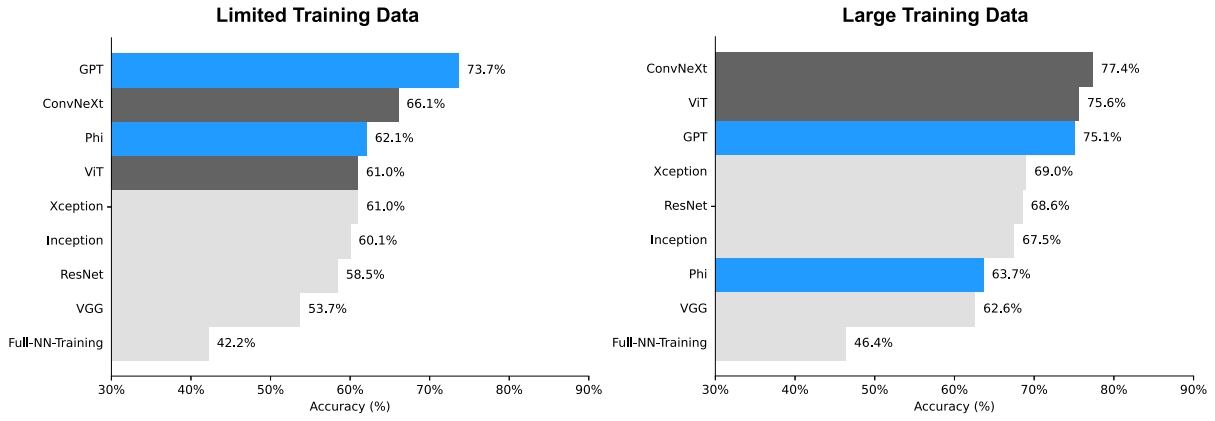

Fig. 3. Average accuracy across all datasets across methods. *Note:* Performances of CNNs are displayed in light gray, TVMs in dark gray, and VLMs in blue.

To enable a fair comparison across methods, we construct a unified hold-out set that is shared across all training conditions. The size of this hold-out set is determined by the dataset with the fewest available images, namely, the AIDA Funnel Car dataset. This results in a fixed hold-out set of 359 test examples per dataset. We perform hyperparameter optimization for each method, dataset, and training data size combination, because the optimal settings may vary across paradigms and tasks.

We evaluate 9 methods (5 CNNs, 2 TVMs, 2 VLMs) under large and limited training data conditions, including 9 hyperparameter optimization runs across 18 data sets. This results in 2275 total experiments (7 fine-tunable methods x 18 datasets × 2 data availability conditions × 9 hyperparameter optimization runs + 2 methods without task-specific fine-tuning x 18 datasets × 2 data availability conditions).[2] We measure accuracy as the primary performance metric because it is the most widely reported metric in marketing publications.

3.3. Accuracy of image classification paradigms

Average performance across datasets

We begin by computing the global average accuracy across all datasets and tasks to establish an expected benchmark accuracy for comparison. To ensure such comparisons are meaningful, we distinguish between large-data and limited-data scenarios, as the amount of training data required for fine-tuning is a key differentiating factor between methods (see Fig. 3).

VLMs require no fine-tuning, training data collection or manual data annotation. Can they match or even exceed the performance of fine-tuned CNNs or TVMs? Obviously the amount of available training data plays a role. When training data is limited, average accuracy of GPT is about 74%. This is better than TVMs like ConvNeXt (66%) and ViT (61%), which make 8% and 13% more mistakes when assigning labels.

These differences diminish, when more training data is available. With large training data, ConvNeXt (77.4%) and ViT (75.6%) perform similar to GPT (75.1%).[3] The open-source VLM Phi does not reach GPT performance on average with a sizeable difference of about 12% accuracy. Since GPT is a commercial model, large-scale data labels can be more costly. On average researchers must therefore make trade-offs between application costs and accuracy when choosing a VLM.

Among CNNs, Xception is consistently best performing at 61% (limited training data) and 69% (large training data). However, all of the CNNs do not reach the performance of TVMs and VLMs. For limited training data, fine-tuned CNNs are between 13% (Xception) and 20% (VGG) less accurate than the top performing benchmark (GPT). For large training data, this difference between ranges between 8% (Xception) and 14% (VGG) compared to ConvNeXt (the top benchmark). For both large and small training data, ConvNeXt and GPT perform consistently better in meaningful and practically relevant ways suggesting CNNs have limited benefits on average.

This is interesting because ConvNeXt is a more complex model and contains more parameters compared to CNNs. However additional complexity, does not seem to inhibit relative performance even when training data is limited to a few hundred observations. ViT, in contrast, is more sensitive to training data size. ViT labels 13% fewer images correctly than GPT and is about as accurate as Phi for limited training data (61.0%). For large training data, it performs almost as good as ConvNeXt (75.6% vs. 77.4%).

These average observations suggest that VLMs can compete with fine-tuned models. When training data collection is costly or researchers are faced with other data collection limitations, VLMs represent low-cost readily available alternatives that researchers

[2] The AIDA funnel car dataset contains fewer than 1000 images and is excluded from the large training data condition, reducing the total by 65 experiments.

[3] GPT and Phi performance differs by training data availability because the large training data excludes the AIDA funnel car dataset.

should take advantage of. They change the rules of the game as researchers can explore many different research questions with ease without collecting training data and the complexities of fine-tuning neural network models.

However, VLMs can hallucinate and might not work for all tasks and datasets equally well. For these reasons fine-tuning still plays a relevant role. This is best done with TVMs and among them with ConvNeXt. On average ConvNext is consistently the better choice than all conventional CNNs (including Xception, ResNet, Inception and VGG). For both TVMs and CNNs, adding training data pays off with average gains in accuracy between 8% and 12%. These observations suggest, marketing researchers should strive for at least 1000 manually annotated training images, whenever feasible.

Training a full neural network is conceptually inferior as the richness of images requires more image labels than marketing can often collect. We observe false classifications for the majority of observations (accuracy of 42.2% and 46.4% for limited and large training data).

Collectively, these average findings suggest that adopting language-related models like TVMs or VLMs has meaningful and practically relevant benefits in expected accuracy over the CNNs that have been used in marketing. Note, however, that these average accuracy values have high variance, which might mask differences in relative performance across individual tasks and datasets.

This is particularly relevant for few-shot applications of VLMs, where base model training data, prohibited tasks, and capabilities are opaque to applied marketing. Strengths and weaknesses of CNNs and TVMs may likewise vary by classification task, leading to possible differences in relative accuracy across applications. Methods that outperform on one dataset may underperform on another, making it essential to examine the stability of relative performance. We therefore investigate whether average values provide reasonable expectations by inspecting individual datasets next.

Performance on dataset level

Fig. 4 shows the attained accuracies across all datasets, separated into limited-data and large-data settings. The various intersections between the performance lines demonstrate that relative performance depends on the image classification tasks and datasets. This means researchers cannot make meaningful inferences about accuracy from individual dataset observations. The fact that one model performes well for a particular academic task does not imply similar performance in other applications.

For the two VLMs, GPT (solid blue) and Phi (dotted blue), in particular, it is evident that researchers must make decisions based on the data at hand. In most cases, GPT with no fine-tuning frequently surpasses the best fine-tuned models, even when large quantities of training data are available (e.g., for expressed emotions or brand logo classification). For most applications, the open-source model Phi performs 5-10 percentage points less accurately than GPT. Interestingly, it exhibits many of the same strengths and weaknesses as GPT, suggesting that the VLM architecture itself lends well to certain task types.

However, larger differences exist for some applications where the training data likely differs. For example, GPT was likely exposed to more brand visuals than Phi so brand logos and other iconic luxury brand assets are better detected with GPT. This results in strong differences with Phi falsely assigning about 60% of images to the wrong luxury brands and GPT identifying almost 80% correctly.

Conversely, Phi can also perform better, e.g., when inferring human sentiment judgments. Since models are constantly improved and new ones appear, such differences clarify that it is essential to compare multiple VLMs to ensure adequate performance. Given applications require little setup and fine-tuning, new VLMs appear and model improvements can alter relative performance, there are few reasons not to benchmark multiple VLMs for each application.

More importantly, these models should not be trusted blindly despite their impressive capabilities. Novel typologies (e.g., ratings of image quality) may not align with their training data. Labeling more than 50% of images incorrectly is clearly problematic for any empirical investigation.

Note that both VLMs are inaccurate on the generated gender and ethnicity datasets, likely due to ethical safeguard restrictions. In an extreme case, GPT refuses to provide class predictions for more than 90 percent of the test data for the ethnicity dataset, making accuracy estimation infeasible. For this reason, we omit GPT accuracy for that dataset. In other cases, these models do produce labels but often reach incorrect conclusions. This implies that model restrictions are not necessarily transparent in marketing applications.

For these reasons, it is essential to always evaluate VLM accuracy using manually annotated marketing images. Since VLMs can be applied without manual annotations, marketers may be tempted to skip validation to avoid the effort of manual annotation. However, as demonstrated here, such shortcuts risk erroneous conclusions associated with serious misinterpretations of visuals.

Manually annotating images is not only useful to validate VLM performance, it also allows researchers to fine-tune TVMs. Whenever VLMs perform poorly, TVMs (black) are most accurate. With sufficient training data and for most tasks, TVMs classify 70% to over 90% of images in accordance with human judgment. TVMs, particularly ConvNeXt, are clearly the most robust. Exceptions occur primarily for subjective evaluations–such as perceived image quality, conveyed emotion, or aesthetic appeal–where human disagreement makes reliable modeling difficult. Nonetheless, across all tasks we find that either a VLM or a TVM correctly classifies 60% or more of images when ample training data are available.

When fine-tuning models it is important to consider the homogeneity of visual information. Seemingly trivial content like brand logos can be challenging to train to a TVM, since logos appear across diverse contexts (e.g., advertising, products, sponsorships), often in distorted or stylized forms (e.g., on T-shirts), or in different versions (e.g., Adidas stripes vs. Trefoil). It can require more annotated training data than is readily available.

In terms of consistency, CNNs (grey) display similar substantial variability in accuracy as VLMs, albeit on a much lower level. While they occasionally approach TVM performance, they usually lag far behind. For example, ResNet's accuracy fluctuates by more than 30 percentage points across datasets and tasks, regardless of data size. In addition to their overall poor performance, this volatility complicates reliable performance expectations. Choosing a CNN would require careful model selection and extensive hyperparameter

Fig. 4. Method performance per dataset. *Note:* Number of classes is provided in parentheses to the right of each dataset name. Blue solid line represents GPT, blue dotted line Phi. The black solid line stands for ConvNeXt and the black dotted line for ViT.

tuning to identify the most effective configuration for a specific application, an effort that rarely yields accuracy comparable to top benchmarks.

These observations underscore that lighthouse leaderboard results based on isolated applications are not representative of the broader set of marketing-related classification tasks, even when tasks might initially appear straightforward. Careful evaluation of accuracy on independent hold-out data is therefore essential, not only for relative comparisons to identify the best method but also to ensure that conclusions are valid. This is particularly important for VLMs that often do not provide confidence scores and do not require any labeled training data to make inferences.

While the relative performance of individual methods varies widely, the various image tasks represented in Fig. 4 can be grouped according to the core marketing tasks of what is visible, who is visible, and how images are perceived. Because certain architectures may align more closely with specific types of tasks, we next analyze performance as a function of the classification task group.

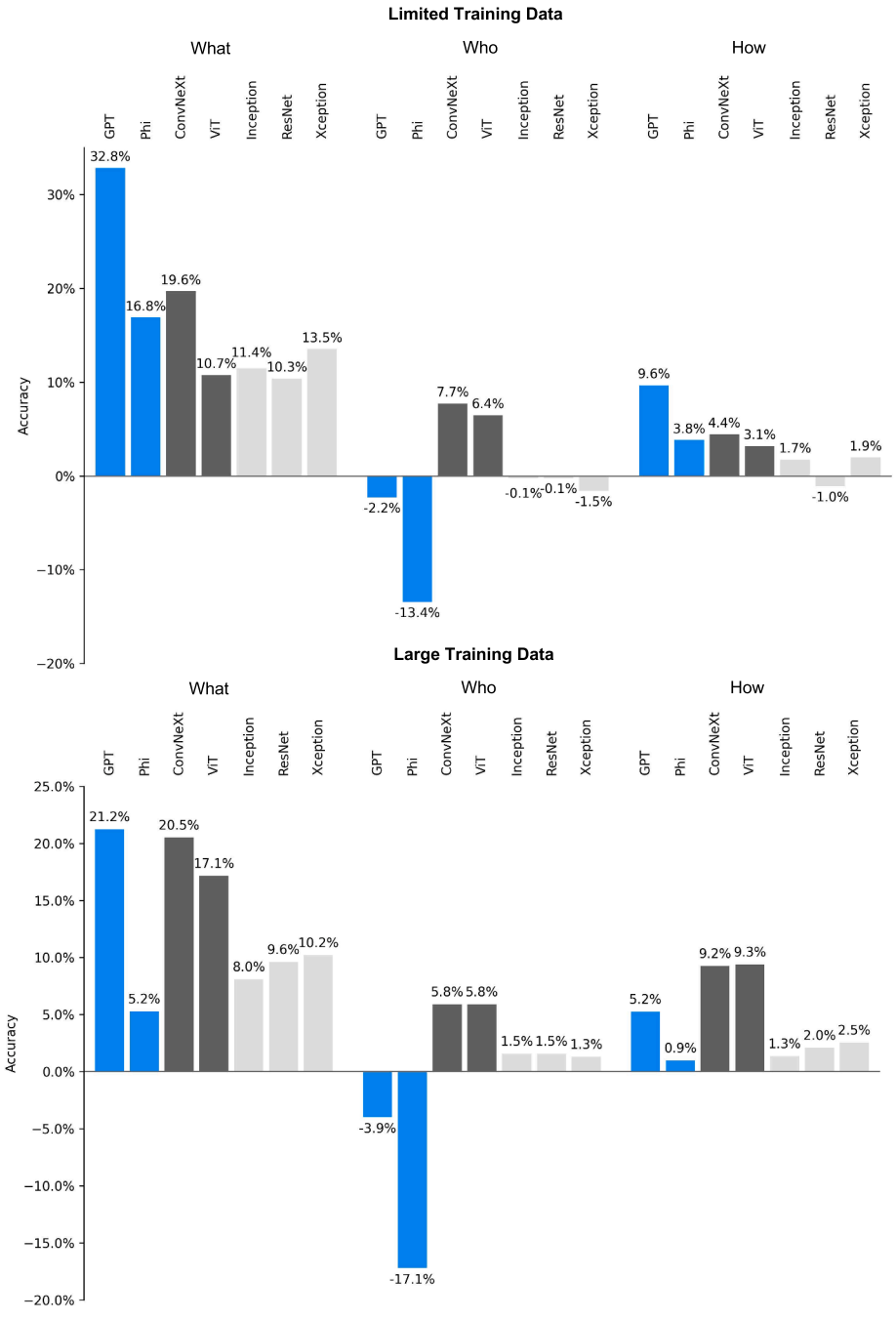

Fig. 5. Accuracy depending on image-related task. *Notes:* Model performance compared to VGG model (baseline).

Performance across image-related task types

Theoretically, image context plays a stronger role for classifications concerning how images are perceived than for those addressing who or what is visible. Consequently, we would expect relatively higher performance from TVMs and VLMs on how-related tasks. Fig. 5 compares all methods to VGG, which serves as the weakest overall benchmark. We exclude models trained from scratch, as it combines the highest training effort with the lowest performance across all applications (see Fig. 4), i.e., is irrelevant for practical group-level conclusions.

When training data is limited, GPT indeed outperforms all alternatives on how-related tasks. Phi performs comparably to TVMs, GPT achieves an outperformance of 9.6%, Phi 3.8%, and ConvNeXt 4.4% compared to VGG. However, this pattern reverses when larger

training datasets are available. With sufficient data, both ConvNeXt and ViT provide similarly robust classifications, outperforming VGG by 9.2% and 9.3%, compared to GPT's 5.2%. In line with expectations, all CNNs perform clearly worse on how-related tasks. Apparently, the contextual understanding of TVMs, combined with fine-tuning for the specific task, enables them to surpass VLMs once adequate training data are provided.

Contrary to expectations, VLMs demonstrate their strongest advantages for limited training data and what-tasks (GPT outperforming VGG by 32.8%, while ConvNeXt provides only 19.6% outperformance). When larger training datasets are available, fine-tuning ConvNeXt results in performance comparable to GPT (20.5% for ConvNeXt and 21.2% for GPT), whereas conventional CNNs fail to achieve competitive accuracy even for relatively objective what-tasks.

These findings suggest that VLMs can perform on par with top-performing TVMs for what and how tasks, despite requiring neither fine-tuning nor additional training data. Since much of marketing research focuses on these task types, the ability of VLMs to handle them efficiently provides researchers with new flexibility to explore questions that would otherwise demand extensive manual labeling and model optimization.

In contrast, VLMs perform poorly on the who-tasks that marketing research has examined thus far. Results from datasets related to ethnicity or gender detection indicate that this underperformance likely stems from ethically motivated restrictions designed to prevent potential misuse. TVMs and CNNs, on the other hand, can be fine-tuned for any image-labeling task because their architectures do not restrict any form of data (e.g., sensitive classes). In these contexts, TVMs are over 20% more accurate for tasks around ethnicity, gender, or other personal characteristics than VLMs.

Overall, these observations reinforce our earlier conclusion that language-inspired approaches offer substantial benefits for image classification in marketing research. Although VLM performance varies considerably across datasets, much of this variability arises from restricted who-tasks. When VLMs underperform, TVMs consistently achieve the highest classification accuracies, particularly when large volumes of labeled data are available. No task category shows promising relative CNN performance, providing no empirical justification for their continued application in marketing image analysis.

Examining task patterns and conceptual distinctions reveal that VLMs and TVMs are suited to different types of image classification challenges. Differences may emerge at the dataset level, as observed so far, but they may also stem from dataset composition and variation in performance for individual images. Up to this point, we have assumed that researchers would select a single model per dataset. However, the observed complementarity between paradigms suggests untapped potential in using multiple models concurrently. Given that TVMs and VLMs have proven most promising for image classification in marketing, the following section investigates their combination in a multi-paradigm ensemble to assess whether such integration can further enhance performance.

3.4. Accuracy of image classification ensemble methods

To examine in which ways VLMs and TVMs might complement each other, we implement an ensemble of the best performing VLM and TVM (GPT and ConvNeXt, respectively). This adds 315 additional experiments to our empirical analysis, increasing the total to 2590 experiments. We combine the class-probability outputs of both models through their respective confidence distributions (i.e., soft voting; Kumar & Jain, 2020). Both models independently predict probabilities across all classes, and the ensemble head learns to weight these predictions according to their relative confidence. When models disagree, the ensemble learns how to associate confidence scores with true labels and adjusts the weights accordingly to identify the most likely classification.

GPT's strong probabilistic calibration and semantic reasoning make it particularly suited for confidence-based integration, while ConvNeXt provides stable, well-calibrated visual probabilities, together forming a coherent foundation for this confidence-related approach. We employ a simple ensemble head consisting of a single neural network layer with 256 neurons (see Web Appendix D for details on the implementation).

For limited training data, the ensemble model combining GPT and ConvNeXt delivers consistently strong performance across tasks. It frequently matches or exceeds the accuracy of the best individual models, providing an improvement of up to 4 percentage points (see right panel of Fig. 6). The left panel of Fig. 6 shows that only few models achieve higher accuracy than the ensemble, indicating low performance variance and high robustness relative to all singular models.

With large training data, the ensemble achieves the highest average accuracy on what (84.3%) and how (71.9%) tasks, and nearly matches ConvNeXt on who tasks (91.8% vs. 92.1%). Importantly, it provides the most consistent top performance: on average, fewer than one model outperforms the ensemble across, what, who, and how tasks (0.4, 0.5, and 0.7 models respectively). In contrast, ConvNeXt, the second most consistent model, is outperformed by 1.9, 0, and 2.8 individual methods on average across these task types.

With limited training data, the ensemble remains competitive but VLMs narrow the gap. GPT achieves 81.2% on what tasks versus 80.7% for the ensemble, while the ensemble leads on how tasks (64.4% vs. 62.0% for GPT). Notably, individual VLMs show high variance: Phi is outperformed by 8.0 models on who tasks, while the ensemble is surpassed by only 0.5 to 1.6 models on average. This stability makes the ensemble a practical solution when maximizing accuracy matters, though the added implementation effort may not be justified when VLMs alone perform adequately (see Web Appendix E for dataset-level results).

These findings suggest that TVMs and VLMs not only perform well on distinct types of visual data but also capture complementary representational features. In particular, when individual models fall short of achieving high accuracy, the ensemble provides additional capacity for performance gains. Notably, these improvements were realized using a comparatively simple input generation procedure and ensemble head architecture. Beyond the scope of this comparison, more sophisticated integration strategies, such as a transformer-based fusion or learned cross-modal attention, may further enhance overall performance.

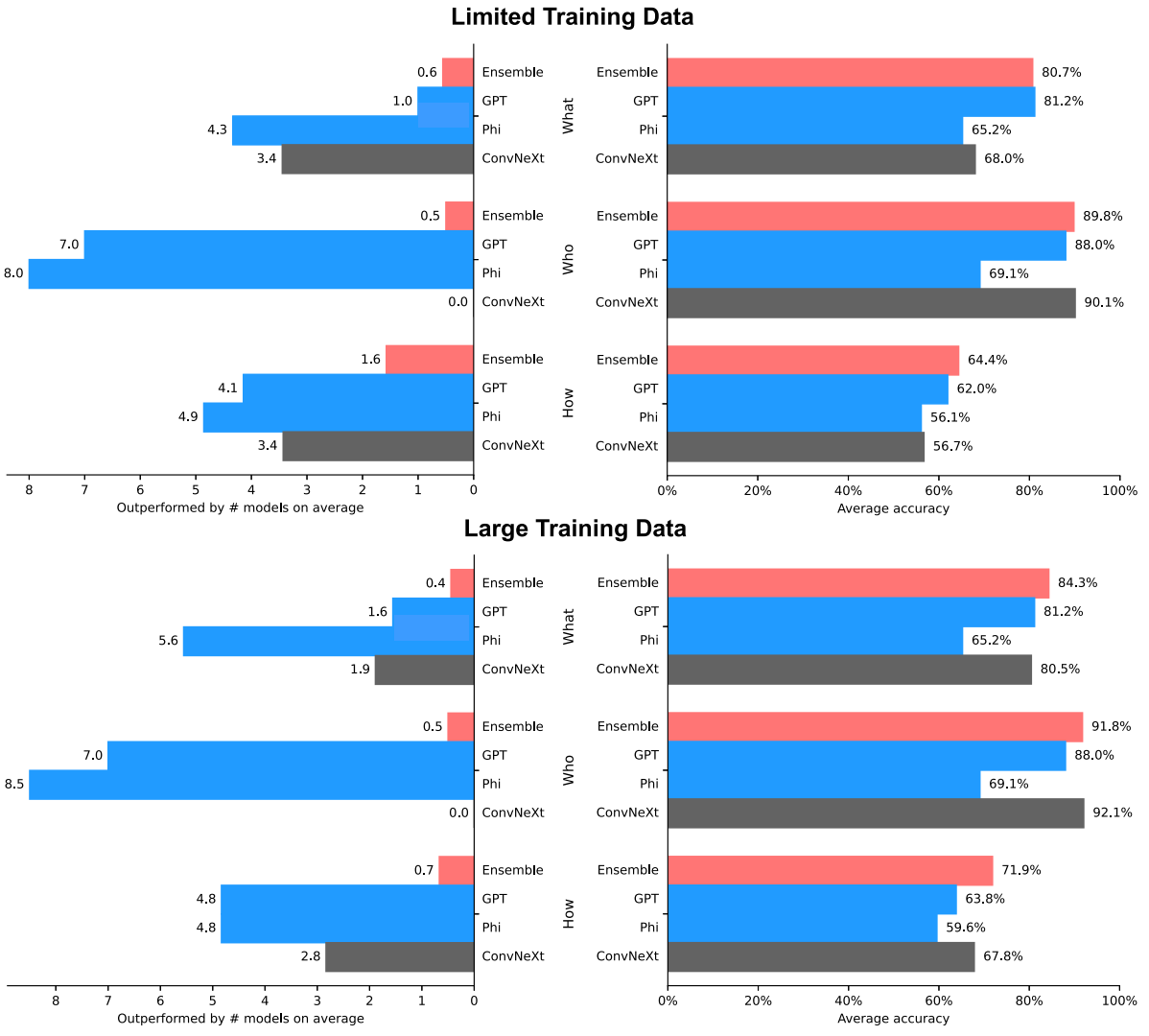

Fig. 6. Ensemble performance across all datasets. *Note:* GPT and Phi performance differs by training size because the large training data excludes the AIDA funnel car dataset.

4. General discussion

Summary

Visual content is rapidly proliferating and increasing in importance for marketing research and practice. Computer vision drives 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.

So far, existing marketing research has predominantly relied on fine-tuned CNNs based on custom training data. In contrast, the emerging paradigm of VLMs enables users to apply automated image classification without the need for task-specific training, requiring minimal effort and knowledge. However, given the recency of this novel image classification paradigm, its usefulness compared to alternative is not clear. Based on an overview of current practices in automated image classification within marketing research, this study conducted over 2500 experiments on the relative performance of CNNs, TVMs, and VLMs on 18 distinct datasets that reflect typical marketing applications.

Our empirical analysis yields several findings to guide method selection in marketing research. VLMs without task-specific training can achieve performance comparable to fine-tuned methods, thereby making advanced image analysis more accessible to researchers with limited machine learning expertise. Lowering the need for training will likely accelerate the adoption of these models (Dekimpe & Hanssens, 2000) and enable researchers to explore many more questions with ease. However, while VLMs provide high accuracies

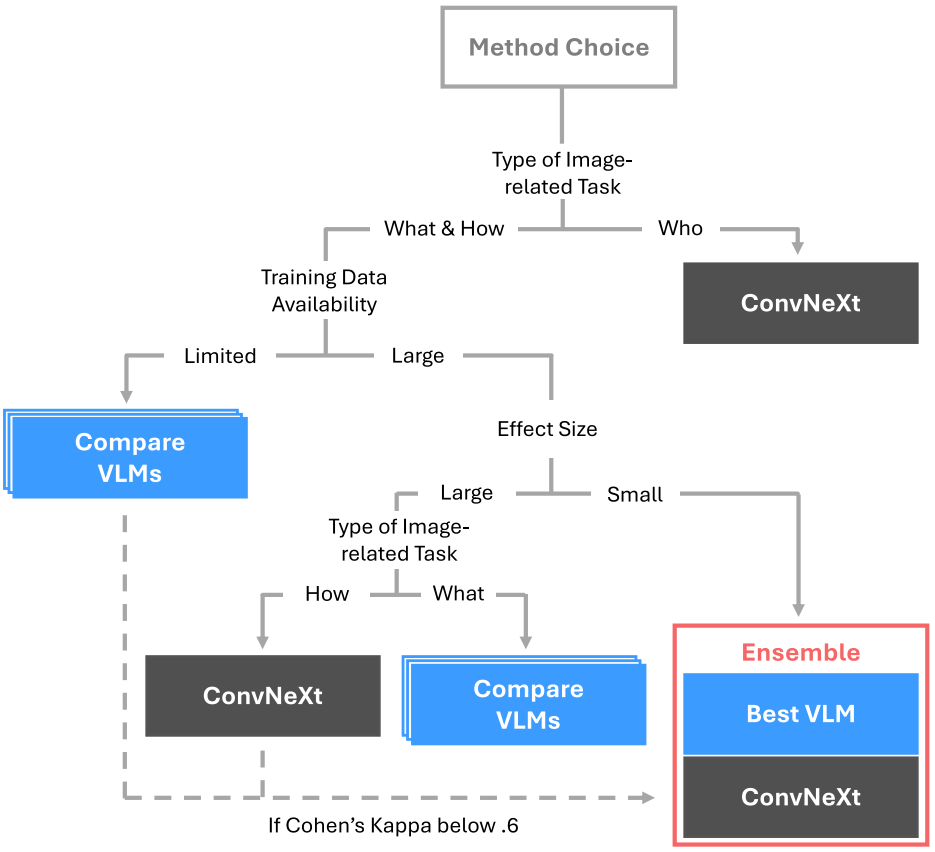

Fig. 7. Decision framework for method choice.

for many image-related tasks their performance varies substantially. It is therefore essential to evaluate VLMs on a hold-out dataset. In several applications task-specific training with TVMs performed much better. However, across all scenarios, CNNs underperform relative to newer paradigms, making them less relevant for future marketing applications.

Choosing an image classification method

Choosing the right image classification method is crucial because poor classification accuracy can bias econometric estimates and lead to suboptimal business decisions (Frankel et al., 2022; Hartmann et al., 2019; Jaidka et al., 2020). On the other hand, assessments are costly and our review suggest very few researchers in marketing compare alternative methods. To understand which methods to focus on, Fig. 7 provides a decision framework that is based on our empirical findings.

Irrespective of the available training data, we found no application where VLMs would outperform the best performing TVM, ConvNeXt, on Who-related tasks. This might be driven by related restricted or novel tasks and visuals that standard VLMs are not trained on. For these Who-related marketing tasks, researchers will still need to fine-tune a model and ConvNeXt works consistently best.

The pattern is different for What and How tasks. Marketers often need to collect multiple ratings per image to compute a meaningful perceptual average. This limits the amount of training images that researchers can collect ratings on. We found that marketing research often relies on limited training data with several hundred training images. For these applications VLMs performed better than any fine-tuned model. While we studied two prominent VLMs, many others exist. Testing VLMs for any individual application is easy because no fine-tuning is involved. According to Fig. 4 relative performance varies, so multiple VLMs should be assessed.

A key challenge is determining when VLM performance justifies skipping fine-tuning. Accuracy alone provides limited guidance, as no accepted threshold exists for scientific applications that generalizes across applications. We recommend using Cohen's Kappa, an established metric for inter-rater agreement (e.g., Huang & Rust, 2024; Leung et al., 2022; Schamp et al., 2023), treating the AI model as an additional annotator. Kappa complements accuracy by accounting for chance agreement based on class distributions. Following Landis and Koch (1977), we recommend using VLMs when Kappa exceeds 0.6, indicating substantial agreement with human labels. When Kappa falls between 0.4 and 0.6, researchers can cautiously interpret initial findings but should examine confusion matrices to assess whether errors are systematic. When no VLM reaches this threshold, implementing the multi-paradigm ensemble is recommended.

Table 3
Step-by-step guide: evaluating vision language model performance.

Step	Description
1	**Define classification task:**
	• Specify type of image-related task (what/who/how classification)
	• Identify target classes (e.g., happy, sad, neutral)
2	**Prepare evaluation dataset:**
	• Create hold-out sample of 100-250 manually labeled images
	• Ensure sufficient observations of minority class for imbalanced datasets
	• Have multiple annotators label ambiguous cases to establish ground truth
3	**Choose models and design prompt:**
	• Identify relevant VLMs for evaluation (e.g., OpenAI GPT, Microsoft Phi, Meta Llama)
	• Create clear and specific instructions
	• Consider providing context or examples (e.g., few-shot prompting)
	• Test reasoning structures (e.g., chain-of-thought and role-based prompting)
4	**Run inference and calculate performance metrics:**
	• Run inference on entire hold-out dataset after estimating cost and runtime based on a sample
	• Compute accuracy (correct predictions / total predictions) and Cohen's Kappa
	• For imbalanced datasets, calculate precision, recall, F1-score, and ROC-AUC
	• Identify classes with poor performance through confusion matrix
5	**Assess marketing viability:**
	• Compare performance against predetermined threshold
	• Evaluate business impact of errors
	• Consider cost-benefit given zero training requirements
6	**Document edge cases:**
	• Identify systematic failure patterns
	• Note content restriction refusals
	• Record unexpected or nonsensical outputs
7	**Make implementation decision:**
	• Deploy if performance meets requirements
	• Use fine-tuned method if inadequate performance
	• Document rational for method selection in research

When more training data is available fine-tuning has more benefits. For small effect sizes and low statistical power, small improvements in image classification performance matter. In this research, we found that VLMs and TVMs have complementary capabilities. Even a simple soft-voting ensemble with a single classification head resulted in systematically better performance for many applications. However, the benefits are limited to a few percentage points and the additional effort is high. Whenever small effect sizes make it worthwhile, we recommend identifying the most suitable VLM, then combining it with ConvNeXt in a multi-paradigm ensemble, and using the best performing classification from the ensemble or its two components.

When smaller differences in effect sizes matter less, VLMs on their own provide reliable results for most What tasks. Similar to limited data settings, it is advisable to compare multiple VLMs. Different tasks and datasets are better suited to different VLMs, as the models are trained on distinct data sources and operate under varying restrictions.

However, VLMs proved less reliable for How tasks. This outcome is intuitive because standard VLMs are not typically trained on data that align with the specific perceptual questions or target populations that marketing research aims to capture. For such applications, fine-tuning ConvNeXt offers a more robust solution and generally yields higher performance than VLMs alone. Nevertheless, even a fine-tuned ConvNeXt does not always reach a Kappa value above 0.6. When model performance remains below this threshold, adopting the multi-paradigm ensemble is recommended.

Based on Fig. 7 marketing researchers can narrow down the solution space for image classification to the methods most likely to provide the best effort-accuracy trade-off. However, regardless of the chosen method, performance evaluation remains essential, as actual performance is difficult to predict and can vary far more than leaderboard results outside of marketing would suggest.

4.1. VLM-Based image classification in marketing

Fine-tuning CNNs and TVMs follows similar principles. VLMs on the other hand are different in execution. Table 3 presents a structured, step-by-step framework for VLM application.

The evaluation process begins with clearly defining the classification objective. Since VLM performance depends on the image task, researchers should specify the type of image-related task (e.g., object, person, or perception identification) and define target classes before implementation.

Although VLMs can be appied without training data, manually labeling a test dataset is essential. Even simple tasks, such as detecting a well-known brand, can prove surprisingly challenging for VLMs. Since training data and possible restrictions are not clear to researchers, they can't anticipate such challenges. Automated classifications can detect valuable nuances in images that humans might miss (e.g., on x-ray images for medical applications). The marketing applications we found are all about what humans can see. When human annotators disagree, automated classifications cannot surpass that upper bound. Assessing human-level performance is

therefore essential for properly contextualizing model performance in this setting. They should rather set an acceptable performance threshold (e.g., Cohen's Kappa of 0.6) given statistical power requirements and guided by benchmarks and human-level performance.

To assess performance, researchers should construct an evaluation dataset by having human raters annotate images to ensure reliable ground truth data. Ambiguous or subjective cases should be labeled by multiple raters to enhance robustness and reduce labeling bias (Schamp et al., 2024). When appropriate, labels should be sourced from individuals representing the target population (e.g., consumers). If possible class distributions should be balanced to prevent skewed or misleading performance estimates.

After preparing the evaluation dataset, researchers should select suitable VLMs based on available computational resources, expertise, and budget. Open-source model can be executed locally, but this can require high computational resources and GPU memory. On the other hand, it is the cheapest way to annotate very large datasets and the model can be saved to facilitate later replication. Cloud solutions are easier to apply, but models can change without notice and budget requirements scale with the number of images. For VLMs prompt phrasing is important and iteratively refined to clearly express the classification task and, when suitable, include examples (few-shot prompting).

Before running a full-scale evaluation on test data, researchers should estimate computational cost and runtime on a small subset to ensure scalability. This not as clear for VLMs as it is for CNNs or TVMs and depends on cloud service pricing and required computational resources. The primary metric for classification performance that is conventionally used in marketing is accuracy, defined as the proportion of correct predictions. However, this often ignores possible class imbalance. If one class far exceeds all others, a model that assigns all images to that class achieves high accuracy without adding useful information. We therefore recommend, to study precision, recall, F1-score, and ROC-AUC in particular when classes are imbalanced. A confusion matrix further supports diagnostic analysis by revealing systematic misclassifications. These observations provide a basis for refining prompts, improving datasets, or selecting alternative models.

Once quantitative results are obtained, they should be interpreted in light of practical constraints and intended use cases. Researchers should evaluate performance and assess the real-world impact of misclassifications. We found several datasets and tasks where VLMs did not match the accuracy of ConvNeXt or ViT. This includes cases with small differences and high performance of all models. Whenever performance suffices for the task at hand, it is not meaningful to start costly data collection or fine-tuning to gain extra percentage points. Beyond aggregate metrics, documenting model limitations is essential for transparency and replicability. Researchers should record systematic errors, such as recurring misinterpretations or refusals due to content restrictions, and note unexpected or nonsensical outputs.

The final step involves deciding on the model's deployment strategy. If the VLM meets or exceeds defined performance requirements, it can be implemented directly. Conversely, if performance remains inadequate, fine-tuned models and ensembles between VLMs and TVMs should be explored (see Fig. 7). This requires choosing hyperparameters. We found that the commonly adjusted parameters of learning rate, batch size, and number of training epochs have strong impact. The best set can be identified using grid search, which is also what we have done in this research. In cases where exhaustive grid search is not computationally feasible, we recommend employing automated hyperparameter optimization techniques, such as those proposed by Akiba et al. (2019). The rationale for model selection (including empirical results, prompt design choices, identified limitations, and hyperparameters) should be clearly documented to support reproducibility and transparency in future research.

To support accessible and reproducible workflows, we provide cost-efficient and user-friendly python scripts for implementing the VLMs and our multi-paradigm ensemble (available at https://github.com/marketing-and-customer-insight/Language_of_Images_Classifying_Marketing_Images).

Limitations & future research

Despite the breadth of datasets and methods included in this investigation, it cannot cover all conceivable image classifications tasks. We caution that other (novel) tasks can result in different levels of accuracy. Similarly, we only included two VLMs (GPT-5 and Phi-4) in the main analysis. Since we compare multiple datasets and tasks, costs and computational resources were a limiting factor for us. However, other VLMs exist and the performance and multi-modal reasoning capabilities will continue to advance (e.g., OpenAI, 2025c). While overall performance might improve, relative limitations in terms of restricted tasks in Who domains or less relevant training data for novel How tasks and/or specific target groups will likely remain.

Computing resources might appear a limitation of applied image classification. Note, however, that none of the implementations of this research require computational resources beyond what has been already applied in marketing. Individual fine-tuned paradigms also do not differ in relevant ways. Only the ensemble methods require training two models, which increases the computational effort.[4] 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 computational times a minor concern for most practical and scientific applications.

During our empirical analysis, we implemented the VLMs using standard prompts and methodologies to make results comparable. We found a relatively small impact of alternative prompting strategies on the tasks we studied but the role of prompting is task dependent and should be further explored for individual applications. Combining VLMs with TVMs in an ensemble is also promising. We

[4] To reduce the time required for training and inference, researchers can use parallel computations using multiple GPUs (see Dehghani et al. 2023, Sreedhar et al. 2024 for information on applying parallel computations).

have experimented with more sophisticated language modeling and text embeddings to provide richer information to the ensemble, which has not improved performance. However, there is potential in other ensemble architectures, combining multiple VLMs and testing alternative prompting strategies.

Our empirical comparison focuses on image classification. 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, e.g., brand logos (Hartmann et al., 2021). 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 results of this research and the rapid pace of development in VLMs suggests it is meaningful to assess them for object detection as well (e.g., prompting "What can you see on this image? Please provide the coordinates of the bounding boxes including the focal objects.")

Looking further ahead, future research can also explore VLMs 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 undoubtedly continue to emerge, many of the insights from our analysis are likely to remain relevant: (1) VLMs can achieve high-quality predictions without task-specific fine-tuning. (2) For who-related classification tasks or when large training datasets are available, ConvNeXt generally outperforms VLMs and provides a robust solution. (3) When effect sizes are small, maximizing accuracy becomes critical; in such cases, researchers should compare multiple VLMs and implement an ensemble combining TVMs and VLMs. (4) Conventional CNNs have become less relevant for marketing as modern language-based architectures offer superior performance.

Our findings demonstrate that VLMs have fundamentally changed image classification in marketing research. These developments make advanced image analysis more accessible to marketing researchers than ever and can improve understanding of how visual communication affects consumer behavior and market outcomes. We hope our work encourages marketing research to rapidly adopt these novel approaches for analyzing large-scale image data.

CRediT authorship contribution statement

Maximilian Witte: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Data curation, Conceptualization, Investigation; **Mark Heitmann:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization; **Jochen Hartmann:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Conceptualization; **Keno Tetzlaff:** Writing – original draft, Methodology, Data curation, Conceptualization.

Data availability

All code associated with this publication is available at https://github.com/marketing-and-customer-insight/Language_of_Images_Classifying_Marketing_Images. All data associated with this publication is available at https://openbigdata.org/resource/datasets-for-image-classification-in-marketing/.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – NFDI 27/1-2026, project number 460037581.

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.ijresmar.2026.01.001.

References

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., Lee, J. R., Lee, Y. T., Li, Y., Liu, W., Mendes, C. C. T., Nguyen, A., Price, E., Rosa, G., Saarikivi, O. et al. (2024). Phi-4Technical Report. https://doi.org/10.48550/arXiv.2412.08905

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. https://arxiv.org/abs/1907.10902

Alantari, H. J., Currim, I. S., Deng, Y., & Singh, S. (2021). An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. *International Journal of Research in Marketing*. https://doi.org/10.1016/j.ijresmar.2021.10.011

Beichert, M., Bayerl, A., Goldenberg, J., & Lanz, A. (2024). Revenue generation through influencer marketing. *Journal of Marketing*, *88*(4), 40–63. Number: 4. https://doi.org/10.1177/00222429231217471

Bharadwaj, N., Ballings, M., Naik, P. A., Moore, M., & Arat, M. M. (2022). A new livestream retail analytics framework to assess the sales impact of emotional displays. *Journal of Marketing*, *86*(1), 27–47. https://doi.org/10.1177/00222429211013042

Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., Ibrahim, M., Hall, M., Xiong, Y., Lebensold, J., Ross, C., Jayakumar, S., Guo, C., Bouchacourt, D., Al-Tahan, H. et al. (2024). An Introduction to Vision-Language Modeling. https://doi.org/10.48550/arXiv.2405.17247

Brand Styles (n.d). Ph.d. thesis. https://www.kaggle.com/datasets/olgabelitskaya/style-color-images.

Burnap, A., Hauser, J. R., & Timoshenko, A., et al. (2023). Product aesthetic design: A machine learning augmentation. *Marketing Science*, *42*(6), 1029–1056. https://doi.org/10.1287/mksc.2022.1429

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research*, *116*(B), 165–178. https://doi.org/10.1016/j.visres.2015.03.005

Cascio Rizzo, G. L., Berger, J. A., & Zhou, M., et al. (2024). How Hand Movement Shapes Communication's Impact. https://papers.ssrn.com/abstract=4962927

Chang, H. H., Mukherjee, A., & Chattopadhyay, A. (2023). More voices persuade: The attentional benefits of voice numerosity. *Journal of Marketing Research*, *60*(4), 687–706. https://doi.org/10.1177/00222437221134115

Chen, X., Ma, Z., Zhang, X., Xu, S., Qian, S., Yang, J., Fouhey, D. F., & Chai, J. (2024). Multi-Object Hallucination in Vision-Language Models. https://doi.org/10.48550/arXiv.2407.06192

Chollet, F. (2017). Xception: Deep Learning With Depthwise Separable Convolutions. http://arxiv.org/abs/1610.02357.

Chuah, S. H.-W., & Yu, J. (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

Cooper, A., Kato, K., Shih, C.-H., Yamane, H., Vinken, K., Takemoto, K., Sunagawa, T., Yeh, H.-W., Yamanaka, J., Mason, I., & Boix, X. (2025). Rethinking VLMs and LLMs for image classification. *Scientific Reports*, *15*(1), 19692. Publisher: Nature Publishing Group. https://doi.org/10.1038/s41598-025-04384-8

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A., Caron, M., Geirhos, R., Alabdulmohsin, I., Jenatton, R., Beyer, L., Tschannen, M., Arnab, A., Wang, X., Riquelme, C., Minderer, M., Puigcerver, J., Evci, U. et al., et al. (2023). Scaling Vision Transformers to 22Billion Parameters. https://doi.org/10.48550/arXiv.2302.05442

Dekimpe, M. G., & Hanssens, D. M. (2000). Time-series models in marketing: Past, present and future. *International Journal of Research in Marketing*, *17*, 183–193. https://doi.org/https://doi.org/10.1016/S0167-8116(00)00014-8

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://arxiv.org/abs/2010.11929.

Dzyabura, D., El Kihal, S., Hauser, J. R., & Ibragimov, M. (2023). Leveraging the power of images in managing product return rates. *Marketing Science*, *42*(6), 1125–1142. https://doi.org/10.1287/mksc.2023.1451

Dzyabura, D., El Kihal, S., & Peres, R. (2022). Image analytics in marketing. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of market research* (pp. 665–692). Cham: Springer International Publishing.

E-commerce Products (n.d). Ph.d. thesis. https://www.kaggle.com/datasets/kennethrithvik/shopee.

Fashion Product Images (n.d). Ph.d. thesis. https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-small

Feng, Y., Chen, H., & Kong, Q., et al. (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, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, *66*(1), 111–126. https://doi.org/10.1007/s12599-023-00834-7

Frankel, R., Jennings, J., & Lee, J. (2022). Disclosure sentiment: Machine learning vs. dictionary methods. *Management Science*, *68*(7), 5514–5532. https://doi.org/10.1287/mnsc.2021.4156

Generated.photos (n.d). Ph.d. thesis. https://generated.photos

Good guys-Bad Guys (n.d). Ph.d. thesis. https://www.kaggle.com/gpiosenka/good-guysbad-guys-image-data-set.

Grewal, R., Gupta, S., & Hamilton, R. (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

Gunarathne, P., Rui, H., & Seidmann, A. (2022). Racial bias in customer service: Evidence from twitter. *Information Systems Research*, *33*(1), 43–54. https://doi.org/10.1287/isre.2021.1058

Hao, Y., Pan, X., Zhang, H., Ye, C., Pan, R., & Zhang, T. (2025). Understanding Overadaptation in Supervised Fine-Tuning: The Role of Ensemble Methods. https://doi.org/10.48550/arXiv.2506.01901

Hartmann, J., & Exner, Y. (2024). GenImageNet. (Accessed July 14, 2024) https://osf.io/8ctjy/.

Hartmann, J., Exner, Y., & Domdey, S. (2025). The power of generative marketing: Can generative AI create superhuman visual marketing content? *International Journal of Research in Marketing*, *42*(1), 13–31. https://doi.org/10.1016/j.ijresmar.2024.09.002

Hartmann, J., Heitmann, M., Schamp, C., & Netzer, O. (2021). The power of brand selfies. *Journal of Marketing Research*, *58*(6), 1159–1177. Number: 6. https://doi.org/10.1177/00222437211037258

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, *40*(1), 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, *36*(1), 20–38. https://doi.org/10.1016/j.ijresmar.2018.09.009

He, J., Li, B., & Wang, X. S. (2023). Image features and demand in the sharing economy: A study of airbnb. *International Journal of Research in Marketing*, *40*(4), 760–780. https://doi.org/10.1016/j.ijresmar.2023.04.001

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. http://arxiv.org/abs/1512.03385

Heitmann, M., Jansen, T. P. J., Reisenbichler, M., & Schweidel, D. A., et al. (2025). EXPRESS: Picture perfect: Engaging customers with visual generative AI. *Journal of Marketing*, . Publisher: SAGE Publications Inc. https://doi.org/10.1177/00222429251356993

Hosu, V., Lin, H., Sziranyi, T., & Saupe, D. (2020). KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment. https://doi.org/10.1109/TIP.2020.2967829

Hu, M. M., Dang, C. I., & Chintagunta, P. K. (2019). Search and learning at a daily deals website. *Marketing Science*, *38*(4), 609–642. https://doi.org/10.1287/mksc.2019.1156

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1–55. https://doi.org/10.1145/3703155

Huang, M.-H., & Rust, R. T. (2024). The caring machine: Feeling AI for customer care. *Journal of Marketing*, *88*(5), 1–23. Publisher: SAGE Publications Inc. https://doi.org/10.1177/00222429231224748

Image Sentiment Polarity (n.d). Ph.d. thesis. Publication Title: data.world https://data.world/crowdflower/image-sentiment-polarity.

Intel Scene Classification (a). Ph.d. thesis. https://www.kaggle.com/puneet6060/intel-image-classification

Jaidka, K., Giorgi, S., Schwartz, H. A., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (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

Klostermann, J., Plumeyer, A., Böger, D., & Decker, R. (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

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*. https://papers.nips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kumar, A., & Jain, M. (2020). Ensemble Learning for AI Developers: Learn Bagging, Stacking, and Boosting Methods with Use Cases. Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4842-5940-5

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. Publisher: International Biometric Society https://doi.org/10.2307/2529310

Lee, J. K. (2021). Emotional expressions and brand status. *Journal of Marketing Research*, *58*(6), 1178–1196. https://doi.org/10.1177/00222437211037340

Leung, F. F., Gu, F. F., Li, Y., Zhang, J. Z., & Palmatier, R. W. (2022). Influencer marketing effectiveness. *Journal of Marketing*, *86*(6), 93–115. Number: 6. https://doi.org/10.1177/00222429221102889

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., & Li, C. (2024). LLaVA-OneVision: Easy Visual Task Transfer. https://doi.org/10.48550/arXiv.2408.03326

Li, X., Shi, M., & Wang, X. S. (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

Liu, L., Dzyabura, D., & Mizik, N., et al. (2020). Visual listening in: Extracting brand image portrayed on social media. *Marketing Science*, *39*(4), 669–686. https://doi.org/10.1287/mksc.2020.1226

Liu, X., Shi, S. W., Teixeira, T., & Wedel, M. (2018). Video content marketing: The making of clips. *Journal of Marketing*, *82*(4), 86–101. https://doi.org/10.1509/jm.16.0048

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. http://arxiv.org/abs/2201.03545.

Luo, R., Li, Y., Chen, L., He, W., Lin, T.-E., Liu, Z., Zhang, L., Song, Z., Xia, X., Liu, T., Yang, M., & Hui, B. (2024). DEEM: Diffusion Models Serve as the Eyes of Large Language Models for Image Perception. https://doi.org/10.48550/arXiv.2405.15232

Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, *13*(9), 5521. https://doi.org/10.3390/app13095521

MetaAI (2024). Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Nanne, A. J., Antheunis, M. L., van der Lee, C. G., Postma, E. O., Wubben, S., & van Noort, G. (2020). The use of computer vision to analyze brand-related user generated image content. *Journal of Interactive Marketing*, *50*(1), 156–167. https://doi.org/10.1016/j.intmar.2019.09.003

OpenAI (2025a). GPT-5 prompting guide. https://cookbook.openai.com/examples/gpt-5/gpt-5_prompting_guide#optimizing-intelligence-and-instruction-following.

OpenAI (2025b). Introducing 4o Image Generation. Accessed May 30, 2025 https://openai.com/index/introducing-4o-image-generation/.

OpenAI (2025c). Thinking with images. Accessed May 02, 2025 https://openai.com/index/thinking-with-images/.

Overgoor, G., Rand, W., van Dolen, W., & Mazloom, M. (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.

Panda, R., Zhang, J., Li, H., Lee, J.-Y., Lu, X., & Roy-Chowdhury, A. K. (2018). Contemplating visual emotions: Understanding and overcoming dataset bias. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer vision - ECCV 2018* (pp. 594–612). Cham: Springer International Publishing (*vol. 11206*). https://doi.org/10.1007/978-3-030-01216-8_36

Peng, L., Cui, G., Chung, Y., & Zheng, W. (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, R., Wedel, M., & Batra, R. (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

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT Is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121. Publisher: Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.2308950121

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 779–788). https://doi.org/10.1109/CVPR.2016.91

Rietveld, R., Van Dolen, W., Mazloom, M., & Worring, M. (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, C., Hartmann, J., & Herhausen, D. (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

Schamp, C., Heitmann, M., Bijmolt, T. H. A., & Katzenstein, R. (2023). The effectiveness of cause-related marketing: a meta-analysis on consumer responses. *Journal of Marketing Research*, *60*(1), 189–215. https://doi.org/10.1177/00222437221109782

Schwenzow, J., Hartmann, J., Schikowsky, A., & Heitmann, M. (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

Semih Kayhan, O., & Van Gemert, J. C. (2020). On translation invariance in CNNs: convolutional layers can exploit absolute spatial location. In *2020IEEE/CVF Conference on computer vision and pattern recognition (CVPR)* (pp. 14262–14273). Seattle, WA, USA: IEEE. https://doi.org/10.1109/CVPR42600.2020.01428

Shang, Y., Zeng, X., Zhu, Y., Yang, X., Fang, Z., Zhang, J., Chen, J., Liu, Z., & Tian, Y. (2024). From Pixels to Tokens: Revisiting Object Hallucinations in Large Vision-Language Models. https://doi.org/10.48550/arXiv.2410.06795

Shankar, V., & Parsana, S. (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. https://doi.org/10.1007/s11747-022-00840-3

Shu, Y., Lin, H., Liu, Y., Zhang, Y., Zeng, G., Li, Y., Zhou, Y., Lim, S.-N., Yang, H., & Sebe, N. (2025). When Semantics Mislead Vision: Mitigating Large Multimodal Models Hallucinations in Scene Text Spotting and Understanding. https://doi.org/10.48550/arXiv.2506.05551

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doi.org/10.48550/arXiv.1409.1556

Sreedhar, K., Clemons, J., Venkatesan, R., Keckler, S. W., & Horowitz, M. (2024). Vision Transformer Computation and Resilience for Dynamic Inference. https://doi.org/10.48550/arXiv.2212.02687

Store Items Color (n.d). Ph.D. thesis. https://www.kaggle.com/datasets/imoore/6000-store-items-images-classified-by-color

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision. Accessed October 14, 2023 http://arxiv.org/abs/1512.00567.

Troncoso, I., & Luo, L. (2023). Look the part? the role of profile pictures in online labor markets. *Marketing Science*, *42*(6), 1080–1100. https://doi.org/10.1287/mksc.2022.1425

Unsplash Images (n.d). Ph.d. thesis. https://www.kaggle.com/datasets/prathameshbhalekar/images-scraped-from-unsplash

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

Wang, J., Min, W., Hou, S., Ma, S., Zheng, Y., & Jiang, S. (2020). LogoDet-3K: A Large-Scale Image Dataset for Logo Detection. http://arxiv.org/abs/2008.05359.

Yang, Y., Patel, A., Deitke, M., Gupta, T., Weihs, L., Head, A., Yatskar, M., Callison-Burch, C., Krishna, R., Kembhavi, A., & Clark, C. (2025). Scaling Text-Rich Image Understanding via Code-Guided Synthetic Multimodal Data Generation. https://doi.org/10.48550/arXiv.2502.14846

Yao, R., Zhang, B., Huang, J., Long, X., Zhang, Y., Zou, T., Wu, Y., Su, S., Xu, Y., Zeng, W., Yang, Z., Li, G., Zhang, S., Li, Z., Chen, Y., Xiong, S., Xu, P., Zhang, J., Zhou, B., Clifton, D., & Gool, L. V. (2025). LENS: Multi-level Evaluation of Multimodal Reasoning with Large Language Models. https://doi.org/10.48550/arXiv.2505.15616

Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y. J., & Ma, Y. (2023). Investigating the Catastrophic Forgetting in Multimodal Large Language Models. https://doi.org/10.48550/arXiv.2309.10313

Zhang, M., & Luo, L. (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, S., Friedman, E. M. S., Srinivasan, K., Dhar, R., & Zhang, X. (2024). Serving with a smile on airbnb: Analyzing the economic returns and behavioral underpinnings of the host's smile. *Journal of Consumer Research*, *51*(6), ucae049. https://doi.org/10.1093/jcr/ucae049

Zhang, S., Mehta, N., Singh, P. V., & Srinivasan, K. (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

Zhang, S., Xu, K., & Srinivasan, K. (2023). Frontiers: Unmasking social compliance behavior during the pandemic. *Marketing Science*, *42*(3), 440–450. https://doi.org/10.1287/mksc.2022.1419