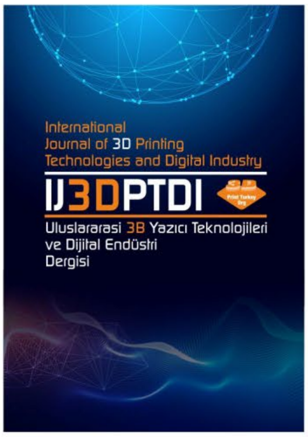

International
Journal of 3D Printing
Technologies and Digital Industry
IJ3DPTDI
Uluslararası 3B Yazıcı Teknolojileri
ve Dijital Endüstri
Dergisi

ULUSLARARASI 3B YAZICI TEKNOLOJİLERİ
VE DİJİTAL ENDÜSTRİ DERGİSİ

INTERNATIONAL JOURNAL OF 3D PRINTING
TECHNOLOGIES AND DIGITAL INDUSTRY

ISSN:2602-3350 (Online)

URL: https://dergipark.org.tr/ij3dptdi

PERFORMANCE COMPARISON OF VISION-LANGUAGE MODELS IN IMAGE CLASSIFICATION

Yazarlar (Authors): Dogukan Ozeren*, Mehmet Erkan Yuksel, Asım Sinan Yuksel

Bu makaleye şu şekilde atıfta bulunabilirsiniz (To cite to this article): *Ozeren D., Yuksel M. E., Yuksel A. S., "Performance Comparison of Vision-Language Models In Image Classification" Int. J. of 3D Printing Tech. Dig. Ind., 9(2): 247-262, (2025).*

DOI: 10.46519/ij3dptdi.1693015

Araştırma Makale/ Research Article

Erişim Linki: (To link to this article): https://dergipark.org.tr/en/pub/ij3dptdi/archive

PERFORMANCE COMPARISON OF VISION-LANGUAGE MODELS IN IMAGE CLASSIFICATION

Dogukan Ozeren[a]*, Mehmet Erkan Yuksel[a], Asım Sinan Yuksel[b]

[a]Burdur Mehmet Akif Ersoy University, Engineering and Architecture Faculty, Computer Engineering Department, TÜRKİYE
[b]Süleyman Demirel University, Engineering and Natural Sciences Faculty, Computer Engineering Department, TÜRKİYE

* *Corresponding Author:* erkanyuksel@mehmetakif.edu.tr

(**Received:** 06.05.25; **Revised:** 20.06.25; **Accepted:** 28.06.25)

ABSTRACT

Vision-Language Models (VLMs) have introduced a new paradigm shift in image classification by integrating visual and textual modalities. While these models have demonstrated strong performance on multimodal tasks, their effectiveness in purely visual classification remains underexplored. This study presents a comprehensive, metric-driven comparative analysis of eight state-of-the-art VLMs—GPT-4o-latest, GPT-4o-mini, Gemini-flash-1.5-8b, LLaMA-3.2-90B-vision-instruct, Grok-2-vision-1212, Qwen2.5-vl-7b-instruct, Claude-3.5-sonnet, and Pixtral-large-2411—across four datasets: CIFAR-10, ImageNet, COCO, and the domain-specific New Plant Diseases dataset. Model performance was evaluated using accuracy, precision, recall, F1-score, and robustness under zero-shot and few-shot settings. Quantitative results indicate that GPT-4o-latest consistently achieves the highest performance on typical benchmarks (accuracy: 0.91, F1-score: 0.91 on CIFAR-10), substantially surpassing lightweight models such as Pixtral-large-2411 (accuracy: 0.13, F1-score: 0.13). Near-perfect results on ImageNet and COCO likely reflect pre-training overlap, whereas notable performance degradation on the New Plant Diseases dataset underscores domain adaptation challenges. Our findings emphasize the need for robust, parameter-efficient, and domain-adaptive fine-tuning strategies to advance VLMs in real-world image classification.

Keywords: Vision-Language Models, Image Classification, Multimodal Learning, Zero-Shot Classification, Few-Shot Learning, Model Generalization.

1. INTRODUCTION

The field of computer vision has advanced rapidly, driven by breakthroughs in deep learning, resulting in outstanding achievements in tasks such as object detection, segmentation, and classification. Traditionally, Convolutional Neural Networks (CNNs) have dominated image classification, using hierarchical feature extraction to achieve high accuracy across diverse datasets [1-4]. In recent years, Vision Transformers (ViTs) have emerged, utilizing self-attention mechanism to capture long-range dependencies and enhance robustness to complex visual patterns [5]. These architectures have established new benchmarks on various datasets such as CIFAR-10 [6], ImageNet [7], MNIST [8], CelebA [9], and COCO [10], solidifying their position as the standard for image classification tasks.

The emergence of VLMs marks a significant paradigm shift by integrating both visual and textual modalities, thereby offering a novel approach to image understanding. Trained on large-scale datasets comprising image-text pairs, these models utilize cross-modal learning to generate enriched semantic representations. Unlike traditional vision-only architectures, VLMs augment visual information with linguistic context, enabling enhanced reasoning in multimodal tasks such as image captioning, visual question answering (VQA), and scene understanding. However, their effectiveness in pure image classification—where explicit textual context is absent—remains an open research question [11-14].

Despite their potential, VLMs encounter several challenges when applied to image classification tasks. Unlike CNNs and ViTs, which are optimized for extracting discriminative features from visual inputs, VLMs often rely on multimodal embeddings that may not be fully utilized in vision-only tasks. Typically, VLMs exhibit higher computational and memory requirements, resulting in increased inference latency compared to deep learning models, raising concerns about their efficiency for high-performance classification tasks. In addition, their reliance on pretraining corpora comprising image-text pairs poses risks of biases, domain dependencies, and reduced generalization when applied to vision-only tasks. Therefore, rigorous comparative analyses of VLMs and unimodal vision models are crucial to understand their advantages and limitations in image classification.

Traditional image classification models operate exclusively within the visual domain, extracting features from pixel-level data to identify patterns, textures, and object structures. CNN-based architectures, such as VGG [15], Inception [16], ResNet [17], DenseNet [18-20], and EfficientNet [21], have demonstrated remarkable success in large-scale image classification due to their use of local receptive fields, parameter sharing, and deep hierarchical structures. ViTs have further advanced the field by leveraging the self-attention mechanism that allow models to capture long-range dependencies and improve feature learning across entire images, achieving significant performance on diverse benchmark datasets and proving their robustness in various real-world applications [22-33]. In contrast, VLMs use a fundamentally different method by integrating both visual and textual inputs to learn multimodal representations [34-37]. Prominent examples include GPT-4V (OpenAI), Gemini 1.5 (Google DeepMind), LLaVA-Next (Meta), Claude 3 (Anthropic), and Qwen-VL (Alibaba Cloud), all of which have achieved notable success in multimodal tasks such as image captioning, VQA, and cross-modal retrieval.

Despite their advantages in semantic reasoning, the application of VLMs to classification tasks that lack explicit textual context remains a critical area of research. Unlike tasks that require joint vision-language understanding, image classification relies primarily on intrinsic visual characteristics, such as color, shape, texture, and spatial relationships, raising the critical questions about whether VLMs can outperform (or even match) established single-modality models without fully leveraging their linguistic capabilities. While VLMs present advantages such as zero-shot classification, transfer learning, and improved generalization, they also pose notable challenges. Their reliance on large-scale multimodal pretraining corpora increases the risk of domain biases, limiting their effectiveness in exclusively visual tasks. Additionally, VLMs require substantial computational resources, making them less efficient and scalable compared to traditional CNNs and ViTs for high-throughput image classification scenarios. Consequently, evaluating their performance on standard and domain-specific classification benchmarks is crucial to understanding the feasibility and limitations of VLMs in vision-centric applications.

In this study, we address the following research questions:

- How do VLMs perform in terms of classification accuracy, precision, recall, and F1-score across diverse datasets?
- Can VLMs generalize effectively to visual domains without textual context, or do they exhibit limitations in such settings?
- What computational trade-offs arise when using VLMs for large-scale classification tasks?
- How robust are these models to data variations, including domain shifts and input noise?

To answer these questions, we conduct extensive evaluations on eight state-of-the-art VLMs—GPT-4o-latest, GPT-4o-mini, Gemini-flash-1.5-8b, LLaMA-3.2-90B-vision-instruct, Grok-2-vision-1212, Qwen2.5-vl-7b-instruct, Claude-3.5-sonnet, and Pixtral-large-2411—across four benchmark datasets: CIFAR-10, ImageNet, COCO, and New Plant Diseases [38] (as a domain-specific dataset). Our comparative analysis focuses on performance metrics such as accuracy, precision, recall, F1-score, robustness, and computational efficiency in zero-shot and few-shot classification settings.

In summary, this study presents a comprehensive and quantitative benchmarking analysis of various VLMs across multiple standard and domain-specific image classification datasets. The key contributions of this research are as follows:

- *Comprehensive Benchmarking of VLMs:* We systematically evaluate eight state-of-the-art VLMs across four diverse datasets under both zero-shot and few-shot settings. This extensive analysis offers valuable insights into the generalization capabilities of VLMs across domains with varying levels of complexity.
- *Novel Analysis of Prompting Strategies:* We investigate the impact of zero-shot and few-shot prompting strategies on VLM performance, providing a detailed understanding of how prompt engineering shapes classification outcomes across different contexts.
- *Domain-Specific Dataset Evaluation:* We use the New Plant Diseases dataset to assess VLM performance on fine-grained, domain-specific classification tasks, addressing an area that remains largely unexplored in the existing literature.

2. RELATED WORK

Traditional Deep Learning Approaches: Over the past decade, image classification has experienced substantial advancement, driven predominantly by advances in deep learning. CNNs have become the cornerstone of modern computer vision, demonstrating remarkable performance in tasks such as object detection, segmentation, and classification. Their strength lies in their capacity to learn hierarchical representations of visual data, effectively capturing both low-level features, such as edges and textures, as well as high-level semantic information [39-41]. However, their inherently unimodal architecture limits their ability to incorporate external information, such as textual cues, thereby constraining their effectiveness in tasks that require contextual reasoning. Successive architectures, including VGG, Inception, ResNet, DenseNet, and EfficientNet consolidated CNNs as the dominant method for image classification.

Emergence of Multimodal Models: To overcome the limitations of unimodal models, multimodal learning approaches have gained traction. CLIP introduced contrastive learning on large-scale image-text datasets, enabling zero-shot generalization [32]. ALIGN further scaled this paradigm, improving robustness and cross-domain transfer ability via vast, noisy data [12]. These models demonstrate the potential of large-scale multimodal pretraining to generalize across diverse vision tasks, including classification, detection, and style transfer, without requiring task-specific supervision [22-37].

The transformative impact of transformer architectures in natural language processing (NLP) has catalyzed their widespread adoption in vision tasks. Pioneering models such as VisualBERT [23], LXMERT [24], and ViLT [30] have substantially advanced unified visual-linguistic modeling by effectively integrating multimodal data, thereby achieving state-of-the-art performance across a range of multimodal task. However, these models are typically computationally intensive and demonstrate limitations in scenarios where textual information is limited or absent, as in traditional image classification tasks.

The introduction of the Vision Transformer (ViT) [28], which encodes images as sequences of fixed-size patches, marked a paradigm shift by enabling the efficient modeling of long-range dependencies within visual data. This approach effectively challenged the long-standing dominance of CNNs in image classification. This transformer-based approach inspired subsequent developments in multimodal learning. Models like ViLT, VisualBERT, and LXMERT integrate vision and language employing unified transformer architectures, fusing modalities via cross-attention and joint token processing. They achieved competitive results across tasks such as image classification, visual question answering, and image captioning. However, these models pose substantial challenges in terms of scalability and efficiency. Their high number of parameters and reliance on large training corpora require significant computational resources. Methods such as knowledge distillation, pruning, lightweight transformer design, and efficient fine-tuning have been proposed to reduce the computational load and improve model scalability [42-49].

While multimodal models such as CLIP and ALIGN have achieved remarkable performance on a range of benchmark datasets (e.g., ImageNet, COCO), their applicability to specialized domains warrants further examination. In practical scenarios—such as medical diagnosis or agricultural disease detection—visual distinctions are often subtle and may not be represented in generic benchmark datasets. Multimodal models tailored to specific domains, when fine-tuned on specialized data, can surpass the performance of general-purpose models like CLIP. This underscores the critical role of domain adaptation in ensuring robust and accurate outcomes in specialized contexts [50–53]. Another key challenge is model robustness. Although multimodal models exhibit high accuracy under standard testing conditions, they are frequently vulnerable to adversarial perturbations, distributional shifts, and noisy inputs. Such brittleness significantly constrains their suitability for safety-critical applications. To address these challenges, recent research has focused on robust training methodologies that incorporate adversarial data augmentation, uncertainty quantification, and distribution-aware loss optimization [54–56].

As VLMs continue to evolve, several promising research directions have emerged that aim to enhance their effectiveness, efficiency, and robustness, particularly in the context of complex multimodal tasks such as image classification. These directions reflect the field's growing demand for models that are not only powerful but also adaptable, scalable, and resilient in real-world deployments.

- **Efficiency Optimization:** Reducing the resource demands of transformer-based multimodal architectures pose substantial barriers to their practical use, particularly in latency-sensitive or resource-constrained environments. To mitigate these limitations, recent efforts have focused on techniques such as sparse attention, quantization, adapter-based fine-tuning, efficient pretraining, knowledge distillation, and lightweight transformer architecture design.

- **Domain Adaptation and Transfer Learning:** Pretrained VLMs often struggle with distribution shifts in domain-specific applications. Emerging methods, including adapter-based modular tuning, prompt-based adaptation, multi-stage domain-specific pretraining, seek to adapt these models to specialized domains like medical imaging, remote sensing, and agriculture, where data is typically limited and imbalanced. Moreover, the integration of auxiliary supervision signals, such as domain ontologies or metadata, can further refine the model's representations to align with the statistical and conceptual structure of the target domain.

- **Robustness and Reliability:** Despite high accuracy on benchmark datasets, many VLMs remain vulnerable to input noise and adversarial manipulation. Researchers have proposed incorporating adversarial training, uncertainty modeling, and robustness certification frameworks to enhance model stability under real-world conditions.

These directions point toward a future where VLMs are not only accurate, but also efficient, generalizable, and trustworthy—traits necessary for their successful integration into specialized real-world applications. In summary, the integration of visual and linguistic modalities has demonstrated considerable promise in image classification. While models like CLIP and ALIGN have set benchmarks in zero-shot generalization, challenges related to robustness, efficiency, and domain-specific adaptation remain. Addressing these limitations is critical for realizing the full potential of VLMs in both research and industry.

3. MATERIAL AND METHOD

We outline the experimental design to evaluate the performance of VLMs in image classification. The methodology includes model selection, dataset preparation, preprocessing protocols, prompt strategy (zero-shot and few-shot), and system configuration to ensure fair comparisons across models.

3.1. Model Selection

Eight VLMs were selected based on three criteria: (1) demonstrated performance in existing benchmarks, (2) capability to process multimodal inputs with prompt-based classification, and (3) availability through open-source implementations or public APIs. The analysis encompasses both closed (proprietary) and open-source models, spanning the spectrum from large-scale semantic reasoners to lightweight, deployable systems. The use of closed models (e.g., GPT-4o-latest, Claude 3.5) is acknowledged as a limitation to reproducibility and interpretability, as internal architectures and processing protocols are not transparent. Table 1 list the models we evaluated. These models represent diverse design philosophies and serve as proxies for evaluating trade-offs between performance, scalability, and computational efficiency.

Table 1. Examined VLMs and their architectural characteristics.

Model Name	Developer	Architecture Type
GPT-4o-latest	OpenAI	Transformer architecture with specific enhancements for multimodal capabilities (handling both text and images), autoregressive generation, and potentially Sparse Mixture of Experts for efficiency.
GPT-4o-mini	OpenAI	Transformer-based autoregressive model designed to be smaller and more computationally efficient than its larger counterpart, GPT-4.
Gemini-flash-1.5	Google	Transformer-based architecture that integrates multimodal capabilities for both text and image processing.
LLaMA-3.2-90B-vision-instruct	Meta	Multimodal transformer-based model that combines the LLaMA architecture with advanced vision processing capabilities. It benefits from large-scale pretraining with 90 billion parameters, cross-modal learning, and potential optimizations like sparse attention or Mixture of Experts (MoE).
Grok-2-vision-1212	xAI	Multimodal VLM based on the transformer architecture.
Qwen-2-VL-7B-instruct	Alibaba	Transformer-based multimodal model that integrates both text and image inputs.
Claude-3.5-sonnet	Anthropic	Transformer-based language model focused on creative text generation tasks.
Pixtral-large-2411	Mistral	Sophisticated transformer-based architecture, combining a large multimodal decoder with a dedicated vision encoder and an extensive context window

Table 1. Examined VLMs and their architectural characteristics (cont.).

Model Name	Features	Characteristics
GPT-4o-latest	- Multimodal capabilities: Text, image, audio, video. - Advanced image captioning and interpretation. - Supports real-time speech interaction and multimedia processing.	Designed for versatile real-time content generation and interaction across various formats, while reducing hallucinations.
GPT-4o-mini	- Smaller, more efficient variant of GPT-4. - Cost-effective with reduced memory and computation requirements. - Excellent for text and image tasks. - Efficient vision-text alignment.	Focuses on reducing computational overhead while maintaining strong performance across NLP and VLM tasks.
Gemini-flash-1.5	- Optimized for speed and quality. - Integrates advanced multimodal reasoning. - Supports large-scale image captioning, processing, and QA tasks. - Faster inference, better image-text coherence.	Designed to increase processing speed and reduce latency, providing a robust solution for VL tasks.
LLaMA-3.2-90B-vision-instruct	- 90B parameters with a VL instruction-following focus. - Utilizes cross-attention layers for effective image processing and captioning. - Large-scale model for fine-grained classification. - Capable of processing both text and image data, making it suitable for a wide range of vision-language tasks such as image captioning, VQA, and instruction-based visual tasks.	Highly capable of handling multimodal tasks such as VQA and document processing.
Grok-2-vision-1212	- High-resolution image processing. - Designed for daptive image classification and fine-grained visual understanding. - Suitable for large-scale deployment. - High-speed multimodal processing. - Optimized for real-time tasks.	Focuses on vision-based processing, particularly suited for large and complex visual datasets.
Qwen-2-VL-7B-instruct	- Specialized in multimodal tasks involving images and text. - Capable of resolving dynamic image resolutions. - Optimized for instruction-following in multimodal tasks. - Low-computation multimodal learning.	Features dynamic resolution processing to enable scalable VLM deployment.

Claude-3.5-sonnet	- Fast and affordable version of Claude. - Uses dynamic token generation and advanced contextualization for multimodal content. - Safety-focused and interpretable. - Strong context-aware reasoning in vision.	Aims for high safety levels and reduced hallucinations, designed for robust and safe conversational AI.
Pixtral-large-2411	- High-performing model designed for visual reasoning tasks. - Optimized for image captioning and visual content understanding. - Large-scale image processing. - Specializes in high-resolution image classification.	Specializes in integrating visual understanding with text-based queries, enhancing interactive visual tasks.

3.2. Dataset Preparation

To conduct a comprehensive evaluation of VLMs across different visual recognition scenarios, we utilized four publicly available datasets: CIFAR-10, ImageNet, COCO, and New Plant Diseases dataset. These datasets were selected based on their diversity in image resolution, domain complexity, multimodal richness, and relevance to both general-purpose and domain-specific classification tasks. By employing these datasets under zero-shot and few-shot configurations, we were able to systematically investigate how VLMs leverage pretrained multimodal knowledge to perform classification in both standard and specialized domains without extensive retraining. Table 2 provides a summary of the datasets we used.

Table 2. Benchmark datasets.

Dataset	Domain	Number of Classes	Number of Samples	Characteristics
CIFAR-10	General object recognition	10	60K	Low-resolution, simple images, basic objects
ImageNet	Large-scale object classification	1000	1.2M	High-resolution, complex real-world images
COCO	Multimodal scene understanding	80	124K	Complex scenes, multiple objects per image, caption annotations
New Plant Diseases	Agricultural disease classifciation	38	55K	Fine-grained domain-specific classification, subtle visual differences

The selection of these datasets was carefully made to cover different aspects of classification and multimodal reasoning:

- **CIFAR-10:** It is a low-resolution dataset used for benchmarking and rapid prototyping. It enables evaluation of performance on low-resolution, small-sized images, providing insights into the robustness of VLMs under constrained visual input conditions.

- **ImageNet:** It is a large-scale database of annotated images used for image classification, object detection, and object localization. It is the de facto standard for large-scale object recognition and offers a benchmark for evaluating VLMs' generalization ability on high-resolution natural images.

-

- **COCO:** It is a large-scale image recognition dataset for object detection, segmentation, and captioning tasks. It introduces multimodal complexity through images with multiple objects and rich textual descriptions, allowing us to assess VLMs' capacity for reasoning in complex visual scenes.

- **New Plant Diseases:** It is a domain-specific agricultural dataset designed to address the challenges of fine-grained image classification. It focuses on distinguishing subtle visual differences between healthy and diseased plant specimens, presenting a rigorous test of the domain adaptation capabilities of VLMs. By requiring models to detect nuanced patterns across closely related classes, this dataset serves as a valuable benchmark for evaluating the robustness and generalization performance of VLMs in specialized, real-world scenarios beyond traditional image classification domains.

This diverse selection enables an extensive analysis of VLMs across different visual domains and levels of task difficulty, ensuring that both generalist and specialist scenarios are rigorously evaluated.

3.3. Preprocessing

To ensure methodological consistency and preserve the integrity of performance comparisons across models, we adopt a model-specific preprocessing strategy tailored to the architectural and training specifications of each model under evaluation. The selected models represent a diverse range of architectures and visual tokenization mechanisms. Therefore, standardized preprocessing is neither feasible nor methodologically sound.

- **Image Format and Quality Control:** All images are first converted to high-quality RGB format (PNG or high-resolution JPEG) to maintain fidelity. The sRGB color space is enforced across the dataset to ensure consistency with the visual encoders' training environments.
- **Normalization:** For closed-source models (GPT-4o, Gemini, Claude, Grok), normalization was handled internally by the API or runtime environment. No external pixel scaling or transformation is applied. For open-source models (LLaMA-3.2, Qwen2.5-vl, Pixtral), normalization was applied using model-specific routines.
- **Prompt Engineering:** Prompt design plays a critical role in determining VLM performance, particularly in few-shot settings. For zero-shot classification, prompts consisted of simple class lists or concise descriptions. In few-shot scenarios, the prompts incorporated either 5 or 10 class-labeled exemplars following a consistent template. Although contrastive learning is frequently cited as the foundation for prompt design in CLIP-like settings, in this study's few-shot configuration, we enhanced in-context prompting by providing explicit class examples without introducing additional contrastive objectives beyond those inherent to the original models. All text prompts were tokenized using each model's native tokenizer. For open-source models, we employed the official implementations available through Hugging Face or GitHub repositories. For proprietary models, prompts were formatted as plain text and submitted alongside the corresponding image via the respective API. Prompt templates were rigorously standardized across models to ensure consistent semantic intent in both few-shot and zero-shot evaluations.
- **Inference Protocols:** For open-source models, inference was performed using either the default configurations or the recommended settings specified in the official repositories (e.g., greedy decoding). For closed, API-based models, default parameter values (e.g., temperature and top-k) were adopted unless the official documentation explicitly required alternative specifications. It is important to acknowledge that, due to proprietary restrictions, access to or modification of the full set of inference parameters was not always possible for these closed APIs. Runtime latency was monitored in a qualitative manner. The practical aspects of all inference protocols were evaluated to the extent permitted by the transparency limitations of each API.

3.4. Experimental Setup

Input-Output Structure and Classification Flow: For all datasets, each VLM takes as input an image accompanied by a text prompt designed to guide the model in performing the classification task. The output generated by the model is a predicted class label, selected from the set of true class labels corresponding to the respective dataset (CIFAR-10, COCO, ImageNet, or New Plant Diseases). The true class labels serve as the ground truth for the evaluation. Fig. 1 provides an overview of the general processing pipeline employed in the classification experiments. The predicted label is compared against the ground truth to evaluate performance metrics.

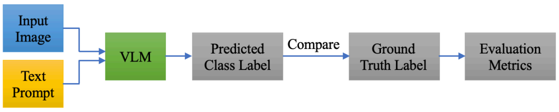
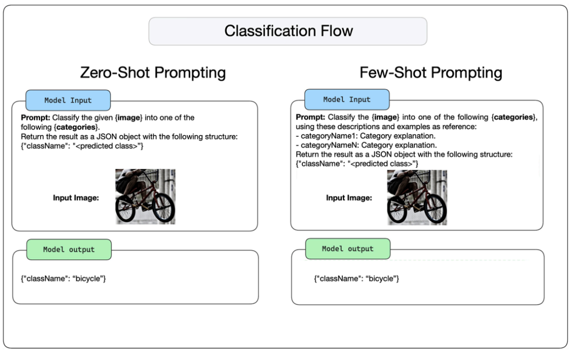


Figure 1. General classification pipeline for VLMs.

Fig. 2 illustrates the input-output structure of two prompting strategies applied in image classification tasks. In zero-shot prompting (left), the model receives an input image and a prompt specifying a set of categories, returning a predicted class in JSON format. In few-shot prompting (right), the prompt additionally includes category descriptions and example pairs to guide the model's prediction. Both strategies produce outputs in a consistent structured format, enabling direct comparison of performance across different prompting conditions.



Prompt: Classify the given {**image**} into one of the following {**categories**}.
Return the result as a JSON object with the following structure:
{"className": "<predicted class>"}

{"className": "bicycle"}

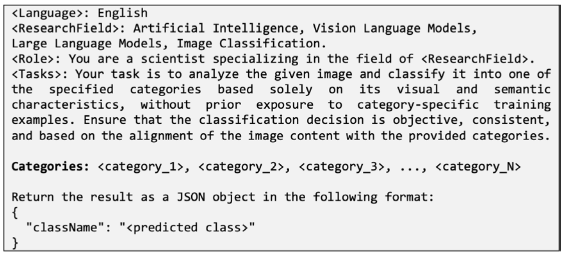
Prompt: Classify the {**image**} into one of the following {**categories**}, using these descriptions and examples as reference:
- categoryName1: Category explanation.
- categoryNameN: Category explanation.
Return the result as a JSON object with the following structure:
{"className": "<predicted class>"}

{"className": "bicycle"}

Figure 2. Classification flow for zero-shot and few-shot prompting in VLM.

To evaluate the classification capabilities of VLMs, we designed two experimental scenarios: zero-shot classification and few-shot classification.

Zero-Shot Classification: Models perform classification without access to any task-specific labeled examples. They leverage their pretrained visual and linguistic representations to infer class labels based on textual prompts. Each model receives a set of descriptive prompts corresponding to candidate classes and selects the most semantically relevant label based on its internal reasoning. Fig. 3 shows a zero-shot prompt template.

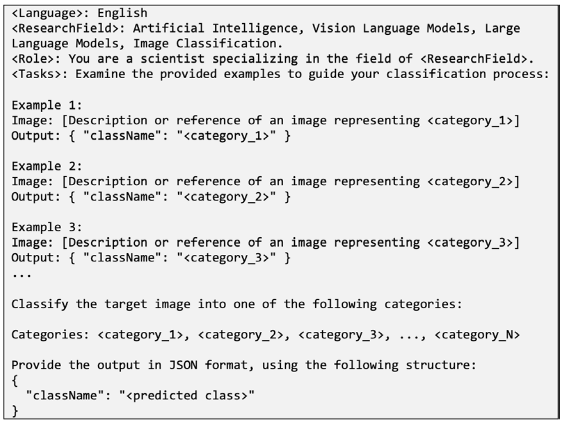


```
<Language>: English
<ResearchField>: Artificial Intelligence, Vision Language Models,
Large Language Models, Image Classification.
<Role>: You are a scientist specializing in the field of <ResearchField>.
<Tasks>: Your task is to analyze the given image and classify it into one of
the specified categories based solely on its visual and semantic
characteristics, without prior exposure to category-specific training
examples. Ensure that the classification decision is objective, consistent,
and based on the alignment of the image content with the provided categories.

Categories: <category_1>, <category_2>, <category_3>, ..., <category_N>

Return the result as a JSON object in the following format:
{
  "className": "<predicted class>"
}
```

Figure 3. Zero-shot prompt template.

Few-Shot Classification: To evaluate the models' adaptability to low-resource settings, we used labeled exemplars. Fig. 4 shows a few-shot prompt template.

- 5-shot: Each class is represented by five labeled examples.
- 10-shot: Each class is represented by ten labeled examples.



```
<Language>: English
<ResearchField>: Artificial Intelligence, Vision Language Models, Large
Language Models, Image Classification.
<Role>: You are a scientist specializing in the field of <ResearchField>.
<Tasks>: Examine the provided examples to guide your classification process:

Example 1:
Image: [Description or reference of an image representing <category_1>]
Output: { "className": "<category_1>" }

Example 2:
Image: [Description or reference of an image representing <category_2>]
Output: { "className": "<category_2>" }

Example 3:
Image: [Description or reference of an image representing <category_3>]
Output: { "className": "<category_3>" }
...

Classify the target image into one of the following categories:

Categories: <category_1>, <category_2>, <category_3>, ..., <category_N>

Provide the output in JSON format, using the following structure:
{
  "className": "<predicted class>"
}
```

Figure 4. Few-shot prompt template.

These exemplars are incorporated into the input prompt in a structured in-context learning format. When supported by the model, contrastive learning techniques are employed to enhance inter-class discrimination. This setup is designed to evaluate each model's capacity to generalize from limited supervision, thereby simulating real-world scenarios in which labeled data is scarce or expensive to obtain.

3.5. Hardware Configuration

Experiments were conducted on a high-performance computing platform equipped with 64 GB of RAM, a 1 TB SSD, a 16-core CPU, a 40-core GPU, and a 16-core Neural Engine. The software environment comprised PyTorch 2.1, TensorFlow 2.10, and the Hugging Face Transformers library, ensuring full compatibility with open-source model architectures. Proprietary models were accessed through their respective public APIs.

3.6. Evaluation Metrics

Model performance was evaluated using standard classification metrics:

- *Accuracy:* Proportion of correctly classified images.
- *Precision:* Proportion of true positives among predicted positives.
- *Recall:* Proportion of true positives among actual positives.
- *F1 score:* Harmonic mean of precision and recall, providing a balanced metric under class imbalance.

These metrics were calculated for each model-dataset pair and macro-averaged across all classes for comparative evaluation.

4. FINDINGS

We present a comparative analysis of eight state-of-the-art VLMs for image classification: Qwen2.5-vl-7b-instruct, Gemini-flash-1.5-8b, Grok-2-vision-1212, Pixtral-large-2411, GPT-4o-latest, GPT-4o-mini, Claude-3.5-sonnet, and LLaMA-3.2-90B-vision-instruct. These models were evaluated on four benchmark datasets—CIFAR-10, ImageNet, COCO, and New Plant Diseases— under both zero-shot and few-shot settings. Performance was assessed based on accuracy, precision, recall, and F1 score. Quantitative results are shown in Tables 3-6.

Results on CIFAR-10 Dataset: Table 3 reveals substantial variability in performance across the evaluated models. GPT-4o-latest achieved the highest scores, with an accuracy and F1-score of 0.91, followed closely by GPT-4o-mini, which attained an F1-score of 0.89. Gemini-flash-1.5-8b and LLaMA-3.2-90B-vision-instruct also demonstrated competitive results, with F1-scores of 0.80 and 0.78, respectively. In contrast, Pixtral-large-2411 exhibited markedly poor performance, with both accuracy and F1-score at 0.13. These results indicate that while large-scale VLMs effectively exploit advanced semantic reasoning capabilities, lightweight models may compromise classification performance in favor of computational efficiency.

Table 3. Results on CIFAR-10.

Model	Accuracy		Precision		Recall		F1 Score	
	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot
Qwen2.5-vl-7b-instruct	0.76	0.66	0.82	0.84	0.76	0.66	0.79	0.74
Gemini-flash-1.5-8b	0.70	0.77	0.87	0.83	0.70	0.77	0.78	0.80
Grok-2-vision-1212	0.65	0.68	0.73	0.74	0.65	0.68	0.69	0.71
Pixtral-large-2411	0.11	0.13	0.17	0.13	0.11	0.13	0.13	0.13
Claude-3.5-sonet	0.57	0.56	0.63	0.59	0.57	0.56	0.60	0.58
GPT-4o-latest	0.92	0.91	0.92	0.91	0.92	0.91	0.92	0.91
GPT-4o-mini	0.90	0.87	0.91	0.90	0.90	0.87	0.90	0.89
LLaMA-3.2-90B-vision-instruct	0.75	0.75	0.81	0.81	0.75	0.75	0.78	0.78

Results on ImageNet Dataset: Table 4 shows that most models achieved near-perfect performance, with accuracies and F1-scores of 0.99 or 1.00. An exception was LLaMA-3.2-90B-vision-instruct, which exhibited substantially lower performance, with an accuracy of 0.57 and an F1-score of 0.64. The consistently high scores across models suggest significant overlap between the ImageNet dataset and the models' pretraining corpora, which may artificially inflate their measured capabilities on this benchmark. These findings highlight the need for caution when interpreting such results as evidence of true generalization.

Table 4. Results on ImageNet.

Model	Accuracy		Precision		Recall		F1 Score	
	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot
Qwen2.5-vl-7b-instruct	0.99	0.98	0.99	0.99	0.99	0.98	0.99	0.99
Gemini-flash-1.5-8b	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Grok-2-vision-1212	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Pixtral-large-2411	0.98	0.99	0.99	0.99	0.98	0.99	0.98	0.99
Claude-3.5-sonet	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
GPT-4o-latest	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GPT-4o-mini	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LLaMA-3.2-90B-vision-instruct	0.98	0.57	0.98	0.71	0.98	0.57	0.98	0.64

Results on COCO Dataset: As seen in Table 5, all models achieved near-perfect classification performance (F1-scores≈1.00). It is important to note, however, that COCO, primarily an object detection and captioning dataset, was adapted for classification by assigning a dominant label to each image. This methodological simplification likely reduced task complexity, which may explain the models' uniformly high scores.

Table 5. Results on COCO.

Model	Accuracy		Precision		Recall		F1 Score	
	Zero -shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot
Qwen2.5-vl-7b-instruct	1.00	0.99	1.00	1.00	1.00	0.99	1.00	1.00
Gemini-flash-1.5-8b	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Grok-2-vision-1212	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00
Pixtral-large-2411	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Claude-3.5-sonet	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GPT-4o-latest	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GPT-4o-mini	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00
LLaMA-3.2-90B-vision-instruct	0.97	1.00	0.98	1.00	0.97	1.00	0.97	1.00

Results on New Plant Diseases Dataset: Table 6 reveals considerable performance degradation on the domain-specific New Plant Diseases dataset. GPT-4o-latest again demonstrated the strongest performance (accuracy: 0.64, F1-score: 0.66). In contrast, Qwen2.5-vl-7B-instruct and Pixtral-large-2411 performed poorly, with F1-scores of 0.20 and 0.29, respectively. Several models, such as Gemini-flash-1.5-8b and LLaMA-3.2-90B-vision-instruct, exhibited high precision but markedly low recall, indicating a conservative classification bias that favors precision at the cost of missing positive cases. This behavior is particularly concerning in high-stakes domains such as agriculture and healthcare.

Table 6. Results on New Plant Diseases.

Model	Accuracy		Precision		Recall		F1 Score	
	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot
Qwen2.5-vl-7b-instruct	0.28	0.20	0.39	0.20	0.28	0.20	0.33	0.20
Gemini-flash-1.5-8b	0.57	0.50	0.61	0.64	0.57	0.50	0.59	0.57
Grok-2-vision-1212	0.49	0.49	0.61	0.54	0.49	0.49	0.54	0.52
Pixtral-large-2411	0.39	0.26	0.40	0.31	0.39	0.26	0.39	0.29
Claude-3.5-sonet	0.46	0.46	0.50	0.47	0.46	0.46	0.48	0.47
GPT-4o-latest	0.62	0.64	0.66	0.68	0.62	0.64	0.64	0.66
GPT-4o-mini	0.49	0.46	0.52	0.55	0.49	0.46	0.50	0.50
LLaMA-3.2-90B-vision-instruct	0.24	0.45	0.42	0.64	0.24	0.45	0.30	0.53

4.1. Key Observations

Model Scale vs. Efficiency: Large-scale models (e.g., GPT-4o-latest) demonstrated superior performance across general and specialized datasets but incurred significant computational costs. Lightweight models (e.g., GPT-4o-mini, Qwen2.5-vl-7b-instruct) offered a more favorable balance between efficiency and moderate classification accuracy.

- **Dataset Bias:** The exceptionally high scores on ImageNet and COCO indicate possible overlaps with pretraining datasets, warranting further evaluations on out-of-distribution (OOD) benchmarks.

- **Domain Adaptation Challenges:** The substantial performance drop on New Plant Diseases dataset underscores the importance of domain adaptation (and the potential of prompt engineering or fine-tuning), especially in specialized domains where fine-grained visual details are critical.

- **Precision-Recall Imbalance:** Models like LLaMA-3.2-90B-vision-instruct exhibited high precision but poor recall on domain-specific datasets, an imbalance that could lead to critical misclassifications in sensitive applications.

- **Overall Best Performer:** GPT-4o-latest consistently achieved the highest or near-highest scores across all datasets and settings, demonstrating superior visual-text alignment, generalization, and robustness, especially in zero-shot and few-shot settings. This reinforces the effectiveness of large-scale multimodal training in zero-shot and few-shot settings.

5. DISCUSSION

We provided a comprehensive analysis of the observed results, discussing the strengths and limitations of each model across four critical dimensions: classification performance, generalization capability, computational efficiency, and robustness to data variations.

- **Classification Performance:** Large-scale VLMs such as GPT-4o-latest, LLaMA-3.2-90B-vision-instruct, and Claude-3.5-sonnet consistently exhibited strong classification performance, particularly under zero-shot settings. Their robust semantic alignment between visual inputs and textual prompts contributed significantly to high accuracy and F1-scores. GPT-4o-latest, in particular, demonstrated exceptional adaptability across both general-purpose (CIFAR-10, ImageNet) and domain-specific (New Plant Diseases) datasets, outperforming other models in most settings. This highlights the advantage of large-scale, multimodal pretraining for cross-domain image classification tasks. Conversely, models such as Pixtral-large-2411 consistently underperformed across benchmarks. Despite architectural strengths for high-resolution visual reasoning, its poor results in standard classification tasks suggest that architectural specialization alone does not guarantee competitive general-purpose performance. In few-shot settings (5-shot and 10-shot), most models demonstrated performance improvements, suggesting that even a limited number of labeled examples can significantly enhance the classification abilities of VLMs. Nonetheless, lightweight models like GPT-4o-mini and Qwen2.5-vl-7b-instruct maintained a trade-off between moderate classification accuracy and operational efficiency.

- **Generalization Across Domains:** While nearly all models achieved near-perfect scores on ImageNet and COCO, these results must be interpreted cautiously. The likelihood of pretraining data overlap introduces a confounding factor that limits the interpretation of these scores as indicators of true generalization. In contrast, the performance on the New Plant Diseases dataset revealed critical weaknesses. Several models—including Grok-2-vision-1212 and Claude-3.5-sonnet—exhibited sharp performance drops, highlighting limited generalization when faced with specialized domains characterized by fine-grained visual distinctions and domain-specific patterns. GPT-4o-latest demonstrated the highest cross-domain robustness, maintaining relatively strong performance even under domain shifts. This suggests that extensive multimodal pretraining with diverse datasets can, to some extent, confer improved transferability. However, even the best-performing models displayed vulnerabilities, emphasizing the ongoing need for task-specific calibration, domain-adaptive fine-tuning, and improved prompt design strategies to ensure reliable performance across varied real-world applications.

- **Computational Efficiency and Scalability:** A clear trade-off emerged between model scale and computational efficiency. Large models like GPT-4o-latest and LLaMA-3.2-90B-vision-instruct, while achieving superior performance, impose significant computational burdens, potentially limiting their deployment in resource-constrained environments. In contrast, lightweight models such as GPT-4o-mini and Qwen2.5-vl-7b-instruct offered lower computational costs with only a moderate reduction in classification performance. These models represent practical alternatives for applications requiring real-time inference or deployment on edge devices. The findings reinforce the importance of model compression, parameter-efficient tuning, and adaptive architectures in future research to balance accuracy with scalability and resource demands.

- **Robustness to Data Variations:** Robustness testing indicated that models with extensive multimodal pretraining—particularly GPT-4o-latest—demonstrated greater stability under noisy or perturbed inputs. These models maintained consistent performance across minor adversarial attacks and distributional shifts. However, several lightweight and specialized models (e.g., Pixtral-large-2411, Grok-2-vision-1212) exhibited higher sensitivity to such variations, leading to degraded performance. The observed precision-recall imbalances, particularly on the New Plant Diseases dataset, further highlight the fragility of some models under domain-specific and imbalanced class distributions. High precision coupled with low recall indicates conservative decision thresholds, which, while minimizing false positives, increase the risk of critical misclassifications—an unacceptable trade-off in high-risk applications like healthcare diagnostics or agricultural monitoring. Robustness, therefore, remains a critical research frontier. Future work should incorporate adversarial training, uncertainty modeling, and formal robustness certification into the VLM development pipeline.

5.1. Implications for Future Research

The findings of this study point to several key directions for advancing VLM-based image classification:

- Enhanced Multimodal Architectures: Future models must better balance visual and textual processing, ensuring that unimodal tasks like image classification are not disadvantaged by excessive reliance on language inputs.
- Parameter-Efficient Fine-Tuning: Methods such as Low-Rank Adaptation (LoRA), prompt tuning, and adapter modules offer promising avenues for adapting large VLMs to domain-specific tasks without prohibitive computational overheads.
- Robustness Optimization: Addressing sensitivity to distributional shifts, adversarial perturbations, and input noise is paramount. Techniques such as distributionally robust optimization and adversarial data augmentation should be incorporated into training regimes.
- Domain Adaptation Strategies: Tailoring VLMs for specialized domains will require sophisticated fine-tuning techniques that minimize catastrophic forgetting while enhancing domain-specific feature extraction.

Collectively, these research directions aim to build VLMs that are not only accurate and generalizable but also efficient, scalable, and resilient, thereby unlocking their full potential for real-world deployment.

6. CONCLUSION

This study presents a comprehensive comparative evaluation of eight state-of-the-art VLMs across diverse image classification benchmarks and data regimes (zero-shot, few-shot). The findings reveal that significant trade-offs exist between accuracy, cross-domain generalization, and computational efficiency. Large-scale VLMs demonstrate strong performance, underscoring the benefits of large-scale multimodal pretraining for robust semantic understanding and cross-domain adaptability. However, reliability concerns remain due to inconsistencies on specialized datasets and susceptibility to distribution shifts. Near-perfect results on datasets like ImageNet and COCO raise concerns about overlap with pretraining corpora, emphasizing the need for rigorous out-of-distribution (OOD) evaluations to accurately assess model generalization. Performance gaps on domain-specific datasets (e.g., New Plant Diseases) further exposed limitations in domain adaptability, while precision-recall imbalances in several models highlighted reliability concerns for critical applications.

On the other hand, Lightweight VLMs offer improved efficiency but lagged in accuracy and robustness, reinforcing the persistent trade-off between model scale and operational practicality. Notably, VLMs could not consistently outperform unimodal vision models, suggesting that multimodal integration alone is insufficient for all classification tasks. To bridge the gaps we identified in our experiments, future research should focus on enhancing robustness to distributional shift via adversarial training and uncertainty modeling, employing advanced domain adaptation and parameter-efficient fine-tuning strategies, improving computational scalability, and explicitly evaluating models on OOD benchmarks for realistic deployment scenarios. Transparent reporting and open protocols are crucial for reproducibility, particularly given the limitations of closed-source models. These directions are critical for the evolution of VLMs into reliable, efficient, and generalizable systems suitable for real-world deployment across diverse domains.

ACKNOWLEDGEMENTS

This research was supported by The Scientific and Technological Research Council of Türkiye (TÜBİTAK) through the 1005-National New Ideas and Products Research Support Program (Project No: 124E769). The study was conducted within the framework of this project, aligned with its technological scope and objectives. This work is based on the Master's thesis titled *"Performance Comparison of Vision-Language Models in Image Classification"* completed at the Department of Computer Engineering, Institute of Science and Technology, Burdur Mehmet Akif Ersoy University, Türkiye.

FINANCIAL DISCLOSURE

The authors report no financial support for the research, authorship or publication of this study.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ETHICS STATEMENT

No ethical approval was required for this study.

REFERENCES

1. LeCun, Y., Bengio, Y., and Hinton, G. "Deep learning", Nature, Vol. 521, Issue 7553, Pages 436-444, 2015.

2. Yao, G., Lei, T., and Zhong, J. "A review of convolutional-neural-network-based action recognition", Pattern Recognition Letters, Vol. 118, Pages 14-22, 2019.

3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... and Bengio, Y. "Generative adversarial networks", Communications of the ACM, Vol. 63, Issue 11, Pages 139-144, 2020.

4. Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., … and Farhan, L. "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions", Journal of Big Data, Vol. 8, Issue 53, 2021.

5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... and Polosukhin, I. "Attention is all you need", arXiv preprint arXiv: 1706.03762v7.

6. Krizhevsky, A., and Hinton, G. "Learning multiple layers of features from tiny images", https://www.cs.toronto.edu/~kriz/cifar.html, 2009.

7. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L. "Imagenet: A large-scale hierarchical image database", IEEE Conference on Computer Vision and Pattern Recognition, 248-255, 2009.

8. LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. "Gradient-based learning applied to document recognition", Proceedings of the IEEE, Vol. 86, Issue 11, Pages 2278-2324, 1998.

9. Liu, Z., Luo, P., Wang, X., and Tang, X. "Deep learning face attributes in the wild," IEEE International Conference on Computer Vision, Pages 3730-3738, 2015.

10. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... and Dollar, P. "Microsoft COCO: Common objects in context", arXiv preprint arXiv: 1405.0312v3, 2014.

11. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., … and Sutskever, I. "Learning transferable visual models from natural language supervision", arXiv preprint arXiv: 2103.00020, 2021.

12. Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., ... and Duerig, T. "Scaling up visual and vision-language representation learning with noisy text supervision", arXiv preprint arXiv: 2102.05918v2, 2021.

13. Lai, Z., Saveris, V., Chen, C., Chen, H. Y., Zhang, H., … and Yang Y. "Revisit large-scale image-caption data in pre-training multimodal foundation models", arXiv preprint arXiv: 2410.02740v1, 2024.

14. Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... and Simonyan, K. "Flamingo: A visual language model for few-shot learning", arXiv preprint arXiv: 2204.14198v2, 2022.

15. Simonyan, K. and Zisserman, A. "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv: 1409.1556v6, 2015.

16. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... and Rabinovich, A. "Going deeper with convolutions", arXiv preprint arXiv: 1409.4842v1 2015.

17. He, K., Zhang, X., Ren, S., and Sun, J. "Deep residual learning for image recognition", arXiv preprint arXiv: 1512.03385v1, 2016.

18. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. “Densely connected convolutional networks”, arXiv preprint arXiv: 1608.06993v5, 2017.

19. Wang, Z., Lu, Y., Li, W., Wang, S., Wang, X., and Chen, X. “Single image super-resolution with attention-based densely connected modüle”, Neurocomputing, Vol. 453, Pages, 876-884, 2021.

20. Yildiz, E., Yuksel, M. E., and Sevgen, S. “A single-image GAN model using self-attention mechanism and DenseNets”, Neurocomputing, Vol. 596, Issue, 127873, 2024.

21. Tan, M. and Le, Q. V. “EfficientNet: Rethinking model scaling for convolutional neural networks”, arXiv preprint arXiv: 1905.11946v5, 2019.

22. Lu, J., Batra, D., Parikh, D., and Lee, S. “ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks”, arXiv preprint arXiv: 1908.02265v1, 2019.

23. Li, L. H., Yatskar, M., Yin, D., Hsieh, C. J., and Chang, K. W. “VisualBERT: A simple and performant baseline for vision and language”, arXiv preprint arXiv: 1908.03557v1, 2019

24. Tan, H. and Bansal, M. “LXMERT: Learning cross-modality encoder representations from transformers”, arXiv preprint arXiv: 1908.07490v3, 2019.

25. Chen, Y. C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., and Liu, J. “UNITER: Universal image-text representation learning”, arXiv preprint arXiv: 1909.11740v3, 2020.

26. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., … and Gao, J. “Oscar: Object-semantics aligned pre-training for vision-language tasks”, arXiv preprint arXiv: 2004.06165v5, 2020.

27. Li, J., Selvaraju, R. R., Gotmare, A. D., … and Hoi, S. “Align before fuse: Vision and language representation learning with momentum distillation”, arXiv preprint arXiv: 2107.07651v2, 2021.

28. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... and Houlsby, N. “An image is worth 16x16 words: Transformers for image recognition at scale”, arXiv preprint arXiv: 2010.11929v2, 2021.

29. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., ... and Guo, B. “Swin Transformer: Hierarchical vision transformer using shifted windows”, arXiv preprint arXiv: 2103.14030v2, 2021.

30. Kim, W., Son, B., and Kim, I. “ViLT: Vision-and-Language Transformer without Convolution or Region Supervision”, arXiv preprint arXiv: 2102.03334v2, 2021.

31. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. “Training data-efficient image transformers & distillation through attention”, arXiv preprint arXiv: 2012.12877v2, 2021.

32. Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., …, and Keutzer, K., “How much can CLIP benefit vision-and-language tasks?”, arXiv preprint arXiv: 2107.06383, 2021.

33. Li, J., Li, D., Xie, S., and Li, F. F. “BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation”, arXiv preprint arXiv: 2201.12086v2, 2022.

34. Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., ... and Cucchiara, R. “The revolution of multimodal large language models: a survey”, arXiv preprint arXiv: 2402.12451v2, 2024.

35. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., … and Ji, R. “MME: A comprehensive evaluation benchmark for multimodal large language models”, arXiv preprint arXiv: 2306.13394v4, 2024.

36. Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., … and Chen, K. “MultiModal-GPT: A vision and language model for dialogue with humans”, arXiv preprint arXiv: 2305.04790v3, 2023.

37. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., … and Zhou, J. "Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond." arXiv preprint arXiv: 2308.12966v3, 2023.

38. Mohanty, S. P., Hughes, D. P., and Salathé, M. "Using deep learning for image-based plant disease detection", Frontiers In Plant Science, Vol. 7, Issue 215232, 2016.

39. Dosovitskiy, A. and Brox, T., "Discriminative unsupervised feature learning with exemplar convolutional neural networks", arXiv preprint arXiv: 1406.6909v2, 2015.

40. Chollet, F. "Xception: Deep Learning with Depthwise Separable Convolutions", arXiv preprint arXiv: 1610.02357v3, 2017.

41. Zeiler, M. D. and Fergus, R. "Visualizing and understanding convolutional networks", arXiv preprint arXiv: 1311.2901v3, 2014.

42. Chen, F., Zhang, D., Han, M., Chen, X., … and Xu, B. "VLP: A survey on vision-language pre-training", arXiv preprint arXiv: 2202.09061v4, 2022.

43. Long, S., Cao, F., Han, S. C., and Yang, H. "Vision-and-language pretrained models: A survey", arXiv preprint arXiv: 2204.07356v5, 2022.

44. Gao, P., Geng, S., Zhang, R., Ma, T., … and Qiao, Y. "CLIP-adapter: Better vision-language models with feature adapters", arXiv preprint arXiv: 2110.04544v2, 2025.

45. Gou, J., Yu, B., Maybank, S. J., and Tao, D. "Knowledge distillation: A survey", International Journal of Computer Vision, vol. 129, Pages. 1789-1819, 2021.

46. Sanh, V., Debut, L., Chaumond, J., and Wolf, T. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", arXiv preprint arXiv: 1910.01108v4, 2020.

47. Sanh, V., Wolf, T., and Ruder, S. "Movement pruning: Adaptive sparsity by fine-tuning", arXiv preprint arXiv: 2005.07683v2, 2020.

48. Zhou, K., Yang, J., Loy, C. C., and Liu, Z. "Conditional prompt learning for vision-language models", arXiv preprint arXiv: 2203.05557v2, 2022.

49. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., … and Wu, Y. "CoCa: Contrastive captioners are image-text foundation models", arXiv preprint arXiv: 2205.01917v2, 2022.

50. Liu, M., Li, B., and Yu, Y. "Fully fine-tuned CLIP models are efficient few-shot learners", arXiv preprint arXiv: 2407.04003v1, 2024.

51. Wang, S., Wang, J., Wang, G., Zhang, B., Zhou, K., and Wei, H., "Open-vocabulary calibration for fine-tuned CLIP", arXiv preprint arXiv: 2402.04655v4, 2024.

52. Chen, J., Yang, D., Jiang, Y., Li, M., … and Zhang, L. "Efficiency in focus: LayerNorm as a catalyst for fine-tuning medical visual language pre-trained models", arXiv preprint arXiv: 2404.16385v1, 2024.

53. Duan, Z., Cheng, H., Xu, D., Wu, X., Zhang, X., … and Xie, Z. "CityLLaVA: Efficient fine-tuning for VLMs in city scenario", arXiv preprint arXiv: 2405.03194v1, 2024.

54. Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N. M. M., and Lin, M. "On evaluating adversarial robustness of large vision-language models", arXiv preprint arXiv: 2305.16934v2, 2023.

55. Zhou, W., Bai, S., Mandic, D. P., Zhao, Q., and Chen, B. "Revisiting the adversarial robustness of vision language models: a multimodal perspective. arXiv preprint arXiv: 2404.19287, 2024.

56. Li, L., Guan, H., Qiu, J., and Spratling, M. "One prompt word is enough to boost adversarial robustness for pre-trained vision-language models", arXiv preprint arXiv: 403.01849v1, 2024.