Online ISSN: 3104-9338
Print ISSN: 3104-932X

Investigation of Partial Image Classification Methods

Ziwen Dong[1], Ijazul Haq[2], Shan Huang[1], Jin Y. Du[2*]

[1] Guangdong Janus Biotechnology Co., Ltd.

[2] Guangdong CAS Angels Biotechnology Co., Ltd.

Received: April 7, 2026

Revised: April 14, 2026

Accepted: April 15, 2026

Published online: April 23, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 5 (May 2026)

* Corresponding Author: Jin Y. Du (jinyduphd@gmail.com)

Abstract. Recognizing an object based on only partial information is a common task that humans perform every day. In this study, we explore how accurate several computer algorithms, including traditional methods and LVMs (Large Vision Models), perform at image classification using a novel dataset comprised of 10 different animal classes. The traditional methods we use are Resnet and Transformer, while the LVMs are GPT-4, Claude, Gemini, LLaVa, Qwen, and CLIP (Contrastive Language-Image Pre-training). The dataset consists of 16K manually cropped images, providing a unique challenge in assessing the models' ability to recognize images based on incomplete information. The results indicate significant variations in model performance. Swin Transformer achieves the best accuracy, outperforming even humans. On the other hand, LVMs under zero-shot underperform humans; but benefit from few-shot preparation.

Keywords: *Image Classification; Large Vision Models; Computer Vision; GPT; Generative AI; Machine Learning*

1. Introduction

Large Language Models (LLMs), such as GPT-3 [1], PaLM [2], LlaMA [3] and Vicuna [4], are well known for making advancements in the field of NLP (Natural Language Processing) using extensive pretraining and vast network architectures. Recently, GPT-4 [5] extended these advancements to multimodal data, spurring the rapid development of Large Multimodal Models (LMMs). These multimodal models harness the knowledge from LLMs to effectively align visual features with textual data. LMMs have been used in a variety of tasks, ranging from object detection to complex scene understanding [6]. However, the robustness of these models in dealing with partial or incomplete visual data remains relatively underexplored. This gap is

particularly evident in the realm of image classification, where the integrity of the visual input is typically assumed to be complete.

Partial image classification presents a unique set of challenges and opportunities. It involves the identification and classification of objects from images where only fragments are visible. This scenario is common in real-world applications such as surveillance, where objects of interest are often occluded or only partially visible due to various obstructions. Additionally, partial image classification can enhance the capability of autonomous systems, such as drones, robots, or self-driving cars, which must operate effectively in dynamic environments with incomplete information.

The objective of this study is to evaluate the performance of vision models on a novel dataset specifically designed for this purpose. The dataset comprises 16K instances of partial animal images, carefully crafted to simulate a range of occlusions and partial visibility scenarios. The models investigated in this study include: GPT-4 [5], Claude [7], Gemini [8], LLaVa [9], Qwen [10], CLIP [11], Resnet [12], and Swin-Transformer [13]. Furthermore, we compared the models' performance with that of humans. To test human performance, we developed a specialized crowdsourcing platform and asked volunteer annotators to identify the correct animal category from the partial images presented to them.

Preliminary investigations into this dataset have revealed substantial variations in the performance of different LVMs, indicating that the ability to effectively handle partial images may be an important discriminator of model capability. This paper seeks to understand these variations in depth, exploring how different architectures and training strategies affect the efficacy of LVMs in partial image classification.

Figure 1. Images of partially visible animals.

This study may find applications in the fields of computer vision, robotics, and self-driving cars, among others. For instance, robots are now capable enough to identify and recognize objects or animals in images where a sufficient portion of the image is visible; however, they

may struggle in cases such as shown in Figure 1 (left), where only part of the animal (cat) is visible. Similarly, self-driving cars can easily identify animals when they are clearly visible, but in cases such as Figure 1 (right), where only a portion of the animal's body is visible, the car may overlook the animal.

The remainder of this paper is organized as follows: Section 2 provides background information and reviews existing literature on image classification, especially with LVMs. Section 3 describes our dataset collection and preparation. Section 4 introduces the methods used. Section 5 discusses how we prompt LVMs. Section 6 details the hyperparameter settings and testing details. Section 7 presents the testing results and the conclusions. Finally, discussions for future work are given in Section 8.

2. Background and Literature Review

Computer vision is the branch of AI concerned with making decisions based on input images or videos. This paper is concerned with one of the central problems of computer vision: image classification. To address this problem, researchers have proposed various models and techniques such as deep learning convolutional neural networks (CNNs) [21]. Within the realm of CNNs, some renowned models include VGGNet [29], Resnet [12], GoogLeNet [30] and MobileNets [31]. The emergence of the Vision Transformer [32] and Swin Transformer [13] model architectures has brought about a new transformation in the field of computer vision, largely due to their innovative application of self-attention mechanisms. These mechanisms allow the models to capture complex relationships and dependencies within images, leading to improved performance compared to traditional CNNs [32]. Building upon this foundation, PMANet [38] further explores the potential of attention mechanisms to address the challenges of skin disease classification. More recently, the application of large AI models has become the trend. Models such as GPT-4 [5], which use more than one trillion parameters, have shown the potential to handle a wide and complex range of tasks.

2.1. Large Vision Models for Image Classification

Li et al. [14] conduct a systematic study on the reliability of Large Vision-Language Models in image understanding tasks, highlighting their tendency to produce semantically inconsistent outputs when interpreting visual content. To address this issue, they propose a polling-based query strategy that reformulates model evaluation as a binary decision problem, thereby improving the robustness of LVM predictions. Liu et al. [15] comprehensively evaluate the performance of LVLMs in Optical Character Recognition (OCR). Xu et al. [16] introduce

LVLM-Ehub, a comprehensive benchmarking framework for evaluating LVLMs, providing a dual evaluation approach combining quantitative capability assessments and an online arena platform for more dynamic, user-involved testing. MM-Vet, a benchmark developed by Yu et al. [17], evaluates LMMs on complex tasks that integrate multiple Vision-Language (VL) capabilities including recognition, OCR, knowledge, language generation, spatial awareness, and mathematics. MM-Vet encompasses 16 tasks derived from these capabilities, offering a nuanced framework to assess how well models perform across varied and integrated tasks. Li et al. [18] design a benchmark and toolkit, ELEVATER (Evaluation of Language-augmented Visual Task-level Transfer), to evaluate the performance of language-augmented visual models, addressing the need for standardized testing methods to identify the challenges associated with assessing these models' transferability across diverse datasets and tasks. It features 20 image classification datasets and 35 object detection datasets, each enhanced with external knowledge to test models under zero-shot, few-shot, and full model fine-tuning scenarios. Yin et al. [19] introduce the Language-Assisted Multi-Modal (LAMM) framework, an open-source endeavor for evaluating and enhancing MLLMs (Multi-model Large Language Models), which focuses on the integration of language and visual data, creating an ecosystem that supports the development of AI agents capable of complex, real-world tasks.

2.2. Partial Image Classification

Making classification decisions based on partial images poses a unique and significant challenge in the field of computer vision. Unlike typical image classifications, where the completeness and clarity of images are generally assumed, partial image classification requires advanced models to maintain high accuracy even when critical visual information is missing. This scenario tests the models' capability to leverage contextual clues and extract meaningful insights from incomplete data to make informed predictions. Such capabilities are particularly crucial in applications like surveillance systems and autonomous navigation, where understanding partially visible objects can make a substantial difference. Several studies have investigated the complexities of image classification under conditions of partial visibility or object occlusion. For instance, subspace decomposition-based methods attempt to estimate missing deep features from occluded images to improve overall classification robustness [39]. Additionally, deep feature augmentation strategies further address this pressing issue by enriching feature representations during the training phase, thereby alleviating performance degradation that is often caused by partial observation [40]. Jeong, Lee, and Son [37] conduct an insightful study specifically focused on classification based on partial images, but they

concentrate on vehicles as the primary subject matter. Furthermore, recent benchmark studies systematically evaluate the robustness of modern deep learning models under varying degrees of object occlusion, revealing that even state-of-the-art architectures experience significant performance drops when faced with incomplete visual information [41]. This highlights the ongoing need for innovative approaches in the domain of partial image classification.

2.3. Datasets/Benchmarks for Image Classification

Benchmark datasets play a crucial role in the field of image classification, serving as essential tools for evaluating and comparing the performance of different algorithms. These datasets provide standardized samples that enable researchers and practitioners to measure the effectiveness of their approaches objectively. As illustrated in Table 1, several notable benchmark datasets are commonly utilized in image classification tasks, each offering unique characteristics and challenges that contribute to advancing the development and refinement of classification algorithms.

Table 1. Benchmark datasets for image classification.

Dataset	Number of Instances	Number of Categories	Application	Reference
ImageNet -1K [26]	≈1.4million	1000	Image classification, Object detection, Image segmentation, etc.	https://www.image-net.org/
CIFAR-10 [33]	60000	10	Image classification	https://www.cs.toronto.edu/~kriz/cifar.html
MNIST [34]	70000	10	Image classification, Handwritten digit recognition	http://yann.lecun.com/exdb/mnist/
COCO [35]	330000	80	Image classification, Object detection, Image segmentation, etc.	https://cocodataset.org/
PASCAL-VOC [36]	11530	20	Image classification, Object detection, Image segmentation, etc.	http://host.robots.ox.ac.uk/pascal/VOC/

3. Dataset Development

This section details how we collected and prepared the dataset used for this paper.

Collection Stage. For this experiment, we manually constructed an artificial dataset by gathering 4,000 full-body images across 10 distinct animal classes (bear, cat, cow, dog, donkey, elephant, horse, lion, sheep, and tiger), each comprising 400 images. These images were sourced from multiple websites using a combination of methods, including web searches with

specific keywords and accessing public image datasets.

Preprocessing Stage. During the preprocessing stage, each image was carefully manually cropped to isolate four partial images, as shown in Figure.2, each representing a distinct part of the animal—head, torso, tail, and leg. The cropping rectangles was chosen to fully encompass the target part and not much more. No fixed aspect ratio was imposed, as the contour of each body part varied across images. The background was not removed.

The result was a dataset comprising 16,000 partial images, organized into four subsets corresponding to each body part.

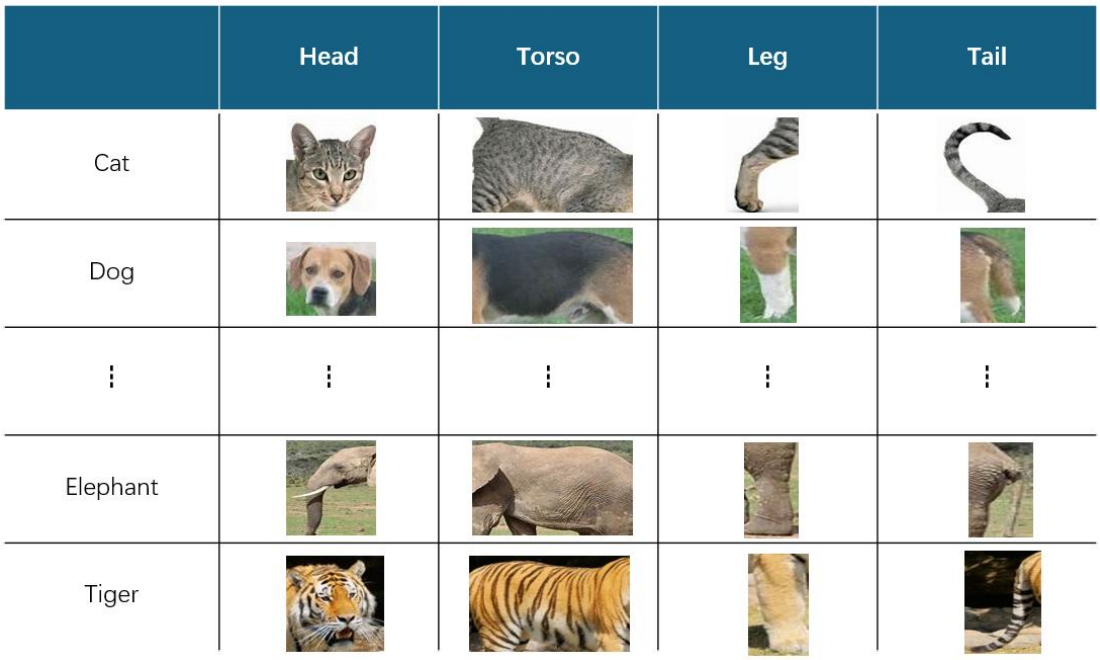

Figure 2. Images of partially visible animals.

4. Methods

This section details the algorithms/methods used in this paper.

4.1. Resnet

Convolutional Neural Networks (CNNs) are a cornerstone in the advancement of deep learning techniques for computer vision. These networks excel in automatically extracting features from images by using convolutional layers in a neural network, leading to breakthrough performances in a wide array of visual tasks, from image classification [21] to object localization [22] and semantic segmentation [23].

Residual Network (ResNet) [12] is a widely recognized and influential CNN architecture that introduced the concept of residual units specifically to tackle the vanishing and exploding

gradient problems frequently encountered during the training of deep neural networks. By incorporating skip connections, ResNet significantly mitigates these gradient-related issues, thereby facilitating the successful training of much deeper networks than was previously feasible. This innovative architectural design not only stabilizes the training process of deep networks but also leads to marked improvements in model performance across various tasks. For the purposes of our experiment, we have chosen to utilize ResNet-18, which serves as a suitable balance between complexity and efficiency while still delivering robust results.

4.2. Swin Transformer

Swin Transformer [13] is a vision modeling approach that incorporates the self-attention mechanism [24] into a deep neural network. This mechanism allows a model to compute relationships between each element in a sequence and all other elements. The architecture of Swin Transformer leverages a hierarchical feature map and a technique known as shifted windowing to synergistically fuse local and global modeling capabilities. This integration not only achieves superior performance on multiple visual tasks but also offers a novel perspective for the design of visual models. We test three Swin-Transformer models: Tiny, Base and Large. They differ in their architectural complexity, particularly in their embedding dimensions, depths and attention heads.

4.3. Large Vision Models

LVMs are a significant leap in the development of AI applied to computer vision. These models, developed using deep learning techniques, leverage vast amounts of visual data to learn rich, complex representations of images. This capability enables them to achieve superior performance on a variety of visual tasks, ranging from basic image classification to more complex applications like scene reconstruction and semantic segmentation. LVMs are transfer-based models, characterized by their large number of parameters, enabling them to capture intricate patterns that are indiscernible to simpler models.

4.3.1. GPT-4

GPT-4 [5] is widely recognized as the most well-known large language model (LLM) available today. In this study, we test its offshoot large vision model (LVM): GPT-4 Vision. Developed by OpenAI and founded on the sophisticated Transformer architecture [24], this advanced model is capable of processing information across diverse modalities, including natural language text and images. By effectively combining its capabilities in both domains, GPT-4 Vision can thoroughly analyze the content and characteristics of images, enabling

precise classification, recognition, and interpretation of visual data in various contexts. This multifaceted approach enhances its utility in applications requiring multimodal understanding.

4.3.2. Gemini

Gemini [8] is a sophisticated family of AI models developed by Google, designed to process an extensive range of modalities, including natural language text, audio recordings, videos, and even programming code. This versatile model family is available in three distinct sizes: Ultra, Pro, and Nano [25], with each variant tailored for different computational scales—from handling highly complex tasks on data centers to optimizing applications that run efficiently on mobile devices. According to the findings presented in [25], Gemini Ultra sets a new standard by advancing the state of the art on 30 out of 32 established benchmarks. For our study, we have chosen to utilize the Gemini-Pro-Vision model, which strikes a balance between performance and computational efficiency, making it well-suited for our specific requirements.

4.3.3. Claude

Claude, a family of transformer-based language models developed by the American AI startup Anthropic [7], represents a significant advancement in natural language processing. Currently on version 3, Claude is available in three scales: Opus, Sonnet and Haiku. Claude's capabilities include advanced reasoning, vision analysis, code generation and multilingual processing. We use Sonnet.

4.3.4. Qwen-VL

Qwen-VL is a series of large-scale vision-language models developed by the Chinese company Alibaba. Starting from a text-only language model, it was trained on a dataset of image-caption-box tuples. Qwen-VL set new records for generalist models of comparable scale across a range of vision benchmarks including image captioning, question answering and visual grounding [10]. We use Qwen-VL-Max.

4.3.5. LLaVA

LLaVA (Large Language and Vision Assistant) is a publicly available large multi-modal model developed by Haotian Liu from the University of Wisconsin-Madison and Chunyuan Li from Microsoft Research through instruction tuning of GPT-4. It effectively combines a vision encoder with a large language model (LLM). The developers conducted extensive testing of LLaVA on two primary tasks: describing images with text, evaluated using GPT-4, and answering multiple-choice questions from Science QA; results are reported in [9]. For our study, we utilize LLaVA version 1.6 to leverage its capabilities in multimodal understanding.

4.3.6. CLIP

CLIP (Contrastive Language-Image Pre-training) is an advanced neural network developed by OpenAI that effectively combines the fields of natural language processing and computer vision. This innovative model is specifically designed for tasks related to image classification, harnessing a diverse dataset consisting of a wide variety of images paired with corresponding natural language captions. One of CLIP's standout features is its high degree of flexibility, which allows users to select specific categories for image classification from which the model can make choices. Notably, CLIP is intended for general use and does not support further training; therefore, our experiment focuses solely on zero-shot results. Radford, Kim et al. [11] conducted comprehensive testing of CLIP across more than 30 existing computer vision datasets, demonstrating that in some instances, its accuracy is comparable to that of models specifically trained on subsets of these datasets. For our analysis, we evaluate four distinct versions of CLIP that were downloaded from the Hugging Face website: vit-base-patch32, vit-large-patch14, plip, and metaclip-b32-400m, each selected for their unique characteristics and capabilities.

5. Prompting Large Vision Models

In the field of computer vision and multimodal AI, prompting techniques are employed to optimize the performance of pre-trained models on specific tasks by guiding the model towards what specifically is desired for output. As the development of large vision models has progressed, the importance of prompt engineering has become increasingly evident. A model’s performance may be sensitive to the specific wording of the prompt, in ways that are highly non-obvious to humans. For text-only tasks, the most advanced models have advanced to the point that the best prompts can often be obtained by simply asking the model to write the prompt itself. However, for multimodal tasks, designing suitable prompts is still very much an active area of research, involving a back-and-forth between the user and the model.

In our study, we manually designed our prompts, using a little bit of trial and error to see which prompts worked best. We aimed to enhance the performance of LVMs, thus narrowing the gap between pre-trained capabilities and specialized task performance. A prompt that works best for one model may not be best for a different model; thus, the prompts we settled on differed slightly between models.

5.1. Prompt Engineering and Experimental Design

In this research, we use a structured prompt engineering strategy for LVMs. The prompts we

used for each of the LVMs are shown in Appendixes A and B. But all designs are based on the following principles:

5.1.1. Structure of the Prompt

To ensure the accuracy of model classification and facilitate automated post-processing, the core prompt used in our experiments is composed of the following components:

Task Orientation and Role-Setting: The model is explicitly instructed to act as an "image classifier" and is informed that the task is part of an academic research study aimed at evaluating its performance. This helps to activate the model's knowledge representations relevant to this specific domain.

Label Set Constraint: The classification scope is strictly defined as a fixed set of labels: [bear, cat, cow, dog, donkey, elephant, horse, lion, sheep, tiger]. By clearly defining this label set, we mitigate the risk of the model outputting irrelevant categories.

Response Formatting: The Mandatory Response Formatting component is designed to enforce output consistency and robustness through two key constraints. First, the model's output is strictly limited to the category name, devoid of any explanatory text or punctuation. Second, the model is prohibited from issuing refusal responses, such as "the image is blurry", compelling it to make its best guess even in situations of uncertainty.

5.1.2. Robustness and Consistency

To accommodate the demands of large-scale dataset evaluation and ensure the reliability of experimental results, the following mechanisms were implemented:

First, the LVM temperature parameter was set to 0.01 to maximize consistency and reproducibility of results.

Second, a robust error-handling process was set up to address potential anomalies during large-scale API calls, such as network fluctuations or content filtering interceptions. If the intended classification process fails, then "unknow" is outputted and we move on to the next image.

5.2. 3-shot setting

To further enhance the performance of Large Vision Models on specific recognition tasks, this study uses a few-shot learning strategy when available. The core of this strategy lies in constructing a multi-turn dialogue structure that includes a System Role, a User Role, and an Assistant Role, thereby providing the model with task guidance and a paradigmatic reference.

During the inference process, the functions of each role are as follows.

System Role: Responsible for establishing task rules. Defines the model's identity as an image classifier and the label constraints it must follow.

User Role: During the 3-shot phase, inputs 3 example images. During the testing phase, inputs the test image.

Assistant Role: During the 3-shot phase, inputs the labels of the 3 images.

Note that LVMs typically have policies constraining usage. These constraints may include restrictions on inputs and outputs. Due to these policies, we encountered instances where models refused to provide answers or offered invalid responses; we counted these as incorrect answers when calculating accuracy. The number of such refusals, and common invalid responses, are detailed in Appendix C.

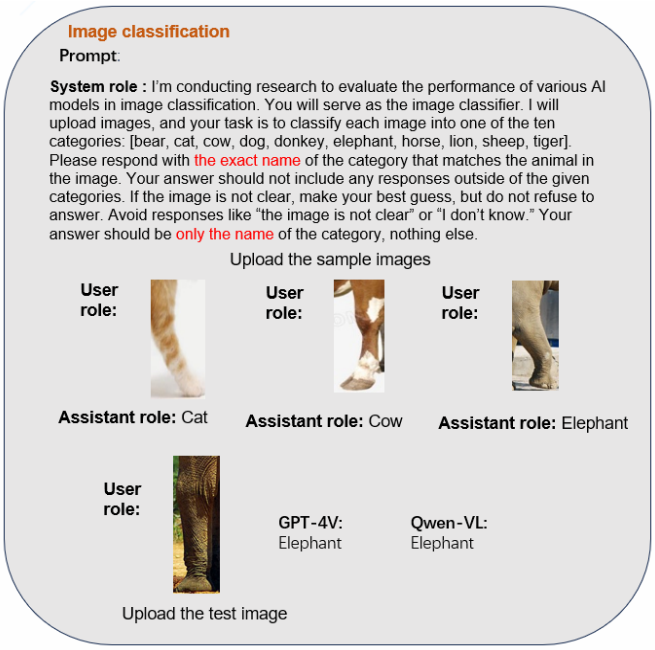

Figure 3. 3-shot experiment with GPT-4 and Qwen-VL.

6. Experimental Setup

This section details our experiments.

The dataset was partitioned into a 7:3 ratio, to form the training and testing sets respectively. That is, the training set consisted of 11200 images, 2800 from each part, and the testing set consisted of 4800 images, 1200 from each part.

6.1. Human Annotation

Comparison with human judgment is common practice for evaluating computer models,

especially the large AI models developed recently [20]. By comparing the accuracy of human annotators with that of the models, we can better understand the strengths and limitations of these models in partial classification tasks. This comparison also provides insights into the potential areas where AI models need further improvement to match or surpass human performance.

We developed a specialized online crowdsourcing platform. The platform was designed to ensure a user-friendly interface, facilitating efficient and accurate responses from the participants. First, volunteers were recruited through a mobile messaging system. Then they were required to register on the platform, to ensure the validity and integrity of the responses. They were informed about the nature of the task and provided with a brief tutorial on how to use the platform. Images from the test set of the dataset were presented one by one to the annotators. Each annotator was shown a series of partial animal images, with each image accompanied by a list of the 10 possible categories. The annotators were required to select the category they believed the partial image belonged to. Volunteers were free to stop whenever they chose. Each annotator was allowed a maximum of 200 annotations to avoid placing excessive weight to any one annotator. This task was considered finished once all the testing images had been annotated exactly once. Figure 4 is a screenshot of the platform.

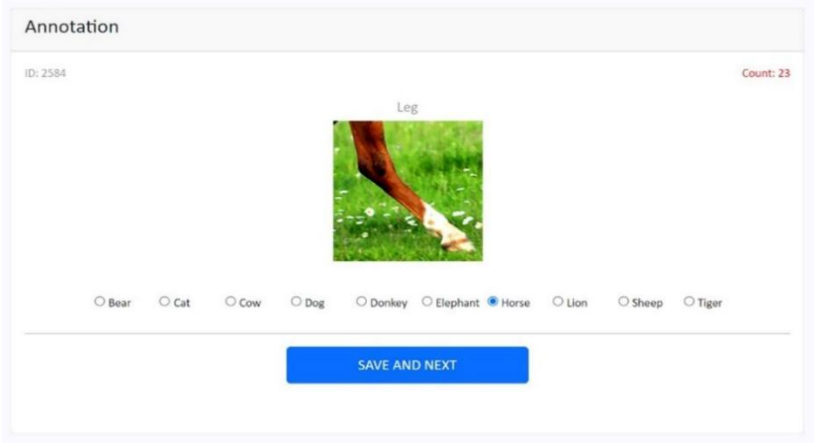

Figure 4. Screenshot of the platform.

The results of human annotation are discussed in detail in Appendix D.

6.2. Model Preprocessing and Hyperparameters

The ResNet and Swin Transformer models that we use are both pre-trained on the ImageNet dataset. For these two models, we resized the input images to 224x224 pixels and normalized using the mean and standard deviation values for the ImageNet dataset, which are [0.485, 0.456, 0.406] and [0.229, 0.224, 0.225] respectively. The ResNet-18 architecture is trained using the

Stochastic Gradient Descent (SGD) optimizer [27], initialized with a learning rate of 0.01 and supplemented with weight decay of 1e-4. The training process utilizes a batch size of 64 and spans 30 epochs. The Swin Transformer models, encompassing the Tiny, Base and Large configurations, are trained using the Adaptive Moment Estimation with Weight Decay (AdamW) optimizer [28] with an initial learning rate of 0.0001 and a weight decay rate of 1e-6. All of these models undergo training with a batch size of 16, covering a total of 30 epochs.

The CLIP models we use are zero-shot models that align image features with text features, eliminating the need for parameter setting or prompts. The CLIP models are merely informed of the available classes.

LVMs are governed by one primary hyperparameter: temperature (T), which is set to 0.01 across all models to minimize output randomness.

To evaluate the Resnet and Swin Transformer models, the full training set was used. On the other hand, the LVMs and CLIP models are intended to be used with little or no further training by the user. For the zero-shot experiments, we of course did not use the training set. For the few-shot experiments, we randomly chose 3 examples from the training set to provide to the models as sample correct classifications. We then tested each example from the testing set as usual. Note that Gemini, GPT-4, Claude and Qwen support such few-shot training; but the other LVMs do not.

7. Results and Conclusions

To ensure a fair comparison of performance across different methods in this task, we categorize the methods into two groups for independent evaluation: traditional deep learning methods and LVMs. We include the results of Human Annotators as a reference standard in both comparison categories.

Traditional deep learning methods (Table 2) are trained on only our dataset, reflecting their learning capability on the target task. Comparing them with the human annotators for this task allows us to assess the potential of these methods to approach or surpass human-level performance after learning from the task-specific data.

Table 2. Performance Comparison of Deep Learning Methods with Human Annotators.

Method	Head only	Torso only	Tail only	Leg only
Human Annotators	0.9348	0.7289	0.5708	0.5263
Swin-Transformer (tiny)	0.9120	0.8320	0.5808	0.4930
Swin-Transformer (base)	0.9420	0.8760	0.6920	0.5370
Swin-Transformer (large)	0.9720	0.9160	0.7660	0.6590

LVMs (Table 3) are pre-trained on a large and diverse corpus. We then evaluate them under

zero-shot or few-shot settings for this task, testing their generalization ability on unseen data. Comparing their results with human annotators helps gauge the "out-of-the-box" performance of such general-purpose models with little or no task-specific training.

Table 3. Performance Comparison of LVMs with Human Annotators.

Method	Head only	Torso only	Tail only	Leg only
Human Annotators	0.9348	0.7289	0.5708	0.5263
Vit-base-patch32	0.8841	0.6267	0.3267	0.3084
Vit-large-patch14	0.9320	0.7575	0.4375	0.4220
Plip	0.6491	0.2050	0.3566	0.1908
Metaclip-b32	0.8750	0.5775	0.3366	0.2675
LLaVA-1.6	0.8083	0.5134	0.2316	0.2461
Gemini-pro-vision (0-	0.9248	0.7705	0.4183	0.3891
shot) (3-shot)	0.9550	0.8767	0.5200	0.4600
GPT4-vision (0-shot)	0.9216	0.6883	0.4358	0.3908
(3-shot)	0.9192	0.7208	0.4567	0.4175
Claude3-sonnet (0-shot)	0.7750	0.4670	0.2760	0.2640
(3-shot)	0.8767	0.6483	0.3383	0.3214
Qwen-vl-max (0-shot)	0.9107	0.8110	0.4758	0.5208
(3-shot)	0.9367	0.8925	0.5575	0.6067

As expected, both human annotators and all methods are least accurate when less informative partial images (e.g., a leg or a tail) are presented.

Based on our results, we draw the following conclusions:

(1) Among all evaluated methods, Swin Transformer (Large) achieves the best overall performance, outperforming both human annotators and all LVMs across all body parts. In particular, it reaches an accuracy of 0.9720 on head images and 0.7660 on tail images, substantially exceeding human performance (0.9348 and 0.5708, respectively).

(2) Among the LVMs, Qwen-vl-max emerges as the best-performing LVM. Under the 3-shot setting, it achieves 0.9367 accuracy on head images, 0.8925 on torso images, 0.5575 on tail images, and 0.6067 on leg images, outperforming human annotators on torso and leg classification. However, it slightly underperforms human annotators on tail classification (0.5575 vs. 0.5708).

(3) Except for Qwen-vl-max, the remaining evaluated large vision models (LVMs) consistently underperform compared to human annotators when it comes to classifying legs and tails, particularly in a zero-shot setting. While human annotators achieve impressive accuracies exceeding 0.52 on leg images and 0.57 on tail images, these LVMs continually fall below these established benchmarks. This discrepancy highlights the challenges faced by current models in accurately recognizing and classifying these specific features, underscoring the superior performance of human judgment in this context. The findings suggest a need for further

refinement and development of LVMs to enhance their capability in such classification tasks.

(4) 3-shot prompting provides a general performance improvement for the evaluated LVMs. Supplying only three examples leads to accuracy gains in most cases, with notable improvements for challenging parts such as legs and tails. For instance, Qwen-vl-max improves from below human accuracy in the zero-shot setting to above human accuracy in the 3-shot setting for leg classification (0.6067 vs. 0.5263).

8. Discussion

This study investigated the capability of large vision models (LVMs) and traditional models to recognize objects under conditions of partial visibility, reflecting real-world challenges like occlusion. By assessing performance across different body parts with varying informativeness, we evaluated the performance gap between these models under zero-shot and three-shot settings for LVMs. This analysis highlights how each model adapts to incomplete visual information.

Several limitations must be acknowledged when interpreting our results. First, the LVMs evaluated in this study were subject to usage policies and safety constraints. We observed instances where models refused to respond or produced invalid outputs, especially under zero-shot settings. These responses were treated as incorrect predictions, potentially underestimating the visual recognition capability of these models. More, while LVM accuracy improved in the 3-shot setting, these results remained sensitive to specific prompt engineering. Factors such as the choice of examples, their sequence, and wording can influence outcomes, meaning the reported performance reflects one prompting setup rather than the absolute ceiling of LVM potential.

Overall, traditional transformer-based architectures, such as the Swin-Transformer, establish a strong performance ceiling, consistently outperforming both humans and large vision models (LVMs) in partial-object recognition tasks. Current LVMs, while flexible and capable of zero-shot inference, still fall short of specialized models when faced with incomplete visual information. Their performance improves significantly under few-shot conditions, indicating they benefit from even small amounts of task-specific adaptation. While LVMs hold considerable promise, their ability to reliably interpret partial images remains limited compared to dedicated vision transformers. Future task-aligned optimization will be essential for them to match or exceed the robustness of specialized models. For tasks involving partially visible objects—such as wildlife monitoring, medical imaging, or surveillance—specialized vision transformers currently provide the most reliable performance available.

Appendix A: Prompts for LVMs (0-shot)

Table 4. Prompts for LVMs(0-shot).

GPT4-vision	Gemini-pro-vision
I'm conducting a study to evaluate the performance of different AI models in image classification for educational and research purposes. You will play the role of an image classifier. I will provide images that showcase parts of an animal, and your task is to identify the animal from a predetermined list of categories. The list consists of 10 categories: 1.dog 2. cat 3. Tiger 4. cow 5. donkey 6. elephant 7. bear8.sheep 9. horse10. lion Please respond with only the exact name of the category that matches the animal in the image. Choose your answer strictly from the provided list, and refrain from including any responses outside of these categories. If the image is not clear you can make your best guess, but avoid responses like "unknown" or "I'm sorry".	I'm conducting a study to evaluate the performance of different AI models in image classification for educational and research purposes. You will play the role of an image classifier. I will provide images that showcase parts of an animal, and your task is to identify the animal from a predetermined list of categories. The list consists of 10 categories: 1. dog 2. cat 3. tiger 4. cow 5. sonkey 6. Elephant 7. bear 8. sheep 9. horse 10. lion Please respond with only the exact name of the category that matches the animal in the image. Choose your answer strictly from the provided list, and refrain from including any responses outside of these categories. If the image is not clear you can make your best guess, but avoid responses like "unknown" or "I'm sorry". For example: User: <upload image> Assistant: cat User: <upload image> Assistant: dog User: <upload image> Assistant: horse etc.

Table 5. Prompts for LVMs(0-shot).

Claude3-sonnet	Qwen-vl-max	LLaVA-1.6
I'm conducting a study to evaluate the performance of different AI models in image classification for educational and research purposes. You will play the role of an image classifier. I will provide images that showcase parts of an animal, and your task is to identify the animal from a predetermined list of categories. The list consists of 10 possible categories： 1. dog 2. cat 3. tiger 4. cow 5. donkey 6. elephant 7. bear 8. sheep 9. horse 10. lion Your answer must be one and only one of the possible categories. You just need to output the category, nothing else. If the image is not clear you can make your best guess, but avoid responses like "Based on the provided image" or "furry".	I'm conducting a study to evaluate the performance of different AI models in image classification for educational and research purposes. You will play the role of an image classifier. I will provide images that showcase parts of an animal, and your task is to identify the animal from a predetermined list of categories. The list consists of 10 categories： 1. dog 2. cat 3. tiger 4. cow 5. donkey 6. elephant 7. bear 8. sheep 9. horse 10. lion Please respond with only the exact name of the category that matches the animal in the image. Choose your answer strictly from the provided list, and refrain from including any responses outside of these categories. If the image is not clear you can make your best guess, but avoid responses like "Based on the image provided" or "I understand".	I'm conducting a study to evaluate the performance of different AI models in image classification for educational and research purposes. You will play the role of an image classifier. I will provide images that showcase parts of an animal, and your task is to identify the animal from a predetermined list of categories. The list consists of 10 categories: 1.dog 2. cat 3. tiger 4. cow 5. donkey 6. elephant 7. bear 8. sheep 9. horse 10. lion Please respond with only the exact name of the category that matches the animal in the image. Choose your answer strictly from the provided list, and refrain from including any responses outside of these categories. If the image is not clear you can make your best guess, but avoid responses like "not clear".

Appendix B: Prompts for LVMs (3-shot)

Table 6. Prompts for LVMs(3-shot).

Gemini-pro-vision	Gemini-pro-vision	Qwen-vl-max	Qwen-vl-max
"I'm conducting research to evaluate the performance of various AI models in image classification tasks. You will serve as the image classifier for this study. I will upload images, and your task is to classify each image into one of the ten specified categories: [bear, cat, cow, dog, donkey, elephant, horse, lion, sheep, tiger]. Please respond with the exact name of the category that best matches the animal depicted in the image. Your answer should not include any responses outside of the provided categories. If the image is unclear or difficult to interpret, make your best guess without refusing to answer. Avoid phrases like 'the image is not clear' or 'I don't know.' Your response should consist solely of the name of the category, nothing more. Thank you for your cooperation." After establishing these detailed instructions, we proceeded to pass three examples to the model. For each example, we uploaded the image as the 'user' and returned the true category of the image as the 'model."	"I'm conducting research to evaluate the performance of various AI models in image classification tasks. You will serve as the image classifier for this study. I will upload images, and your task is to classify each image into one of the ten specified categories: [bear, cat, cow, dog, donkey, elephant, horse, lion, sheep, tiger]. Please respond with the exact name of the category that best matches the animal depicted in the image. Your answer should not include any responses outside of the provided categories. If the image is unclear or difficult to interpret, make your best guess without refusing to answer. Avoid phrases like 'the image is not clear' or 'I don't know.' Your response should consist solely of the name of the category, nothing more. Thank you for your cooperation." After establishing these detailed instructions, we proceeded to pass three examples to the model. For each example, we uploaded the image as the 'user' and returned the true category of the image as the ' assistant.'	"I'm conducting research to evaluate the performance of various AI models in image classification tasks. You will serve as the image classifier for this study. I will upload images, and your task is to classify each image into one of the ten specified categories: [bear, cat, cow, dog, donkey, elephant, horse, lion, sheep, tiger]. Please respond with the exact name of the category that best matches the animal depicted in the image. Your answer should not include any responses outside of the provided categories. If the image is unclear or difficult to interpret, make your best guess without refusing to answer. Avoid phrases like 'the image is not clear' or 'I don't know.' Your response should consist solely of the name of the category, nothing more. Thank you for your cooperation." After establishing these detailed instructions, we proceeded to pass three examples to the model. For each example, we uploaded the image as the 'user' and returned the true category of the image as the ' assistant.'	"I'm conducting research to evaluate the performance of various AI models in image classification tasks. You will serve as the image classifier for this study. I will upload images, and your task is to classify each image into one of the ten specified categories: [bear, cat, cow, dog, donkey, elephant, horse, lion, sheep, tiger]. Please respond with the exact name of the category that best matches the animal depicted in the image. Your answer should not include any responses outside of the provided categories. If the image is unclear or difficult to interpret, make your best guess without refusing to answer. Avoid phrases like 'the image is not clear' or 'I don't know.' Your response should consist solely of the name of the category, nothing more. Thank you for your cooperation." After establishing these detailed instructions, we proceeded to pass three examples to the model. For each example, we uploaded the image as the 'user' and returned the true category of the image as the ' assistant.'

Appendix C: Invalid Responses from LVMs

Table 7. Number of invalid responses from 0-shot.

Model	Test Images	Valid Responses	Invalid Responses	No Response	Total Invalid (No Response + Invalid Response)
GPT4	4800	4550	26	224	250
Gemini		4472	201	127	328
Claude		4630	82	88	170
Qwen		4606	184	10	194
LLaVa		4399	324	77	401

Table 8. Number of invalid responses from 3-shot.

Model	Test Images	Valid Responses	Invalid Responses	No Response	Total Invalid (No Response + Invalid Response)
GPT4	4800	4513	182	105	287
Gemini		4783	0	17	17
Claude		4726	70	4	74
Qwen		4636	52	112	164

Table 9. Examples of invalid responses and their frequencies from 0-shot.

GPT4		Gemini	
Cheetah	3	I'm sorry…	114
Giraffe	8	It's not possible to identify	3
Zebra	8	The image is too blurry to…	6
(No animal stated)	2	Wolf	4
Goat	4	Snake	10

Table 10. Examples of invalid responses and their frequencies from 0-shot.

Claude		Qwen	
(No animal stated)	65	(Image not recognized)	83
Giraffe	2	Zebra	7
Rabbit	2	Fox	2
Snake	2	Goat	3
Zebra	3	Ok.	75

Table11. Examples of invalid responses and their frequencies from 0-shot.

LLava	
The image provided is not clear enough to…	246
Not clear	21

Table 12. Examples of invalid responses and their frequencies from 3-shot.

GPT4	
The image is not clear…	118
(None of the ten categories)	20
Zebra	6
Buffalo	3

Table 13. Examples of invalid responses and their frequencies from 3-shot.

Claude		Qwen	
Goat	27	Bull	7
I apologize…	11	Leopard	2
Pig	6	Zebra	16
Zebra	5	Snake	7
Rhino	2	Mushroom	2
Leopard	2	Deer	2

Appendix D: LVM Costs

Open-source algorithms are freely available for public use, incurring no additional costs beyond the operating expenses associated with the user's computing resources. However, it is important to note that the following four large language models (LVMs) are not open source and do impose charges for usage; these fees are typically proportional to the volume of input and output processed. As of July 11, 2024, we have compiled a list of their pricing structures, which reflects the varying costs associated with their deployment in practical applications. This information will provide potential users with valuable insights into the financial implications of utilizing these proprietary models.

Table 14. Prices for Closed Source LVMs

Model	Input Price (per 1000 tokens)	Output Price (per 1000 tokens)	Source
GPT-4-vision	$ 0.06	$ 0.12	https://azure.microsoft.com/zh-cn/pricing/details/cognitive-services/openai-service/
Qwen-vl-max	$ 0.003	$ 0.003	https://help.aliyun.com/document_detail/2712568.html?spm=a2c4g.2712587.0.0.7dd53809r05POG
Claude-sonnet	$ 0.003	$0.015	https://docs.anthropic.com/zh-CN/docs/models-overview
Gemini-pro-vision	$0.0035	$0.0105	https://ai.google.dev/pricing?hl=zh-cn

Appendix E: Human Annotation Results

Unlike machines, the accuracy of human annotators is influenced by a variety of factors, including motivation, attention, and fatigue. As a result, we anticipate that human annotation will exhibit the highest degree of variability in performance when compared to all existing machine methods. This inherent variability makes our reported accuracy particularly susceptible to questioning and scrutiny. In this section, we aim to provide a comprehensive analysis along with additional information that will help assess and better understand this variability, ultimately shedding light on the complexities involved in human annotation

processes.

A total of 83 annotators participated in the study. Among these annotators, the highest accuracy achieved was an impressive 90.48%, while the lowest recorded accuracy was significantly lower at 40.00%. Although this represents a substantial range of performance, it cannot be interpreted at face value due to several critical factors. Specifically, the difficulty of the images presented varied considerably, with heads being the easiest to classify and legs and tails proving to be much more challenging. Additionally, the number of annotations attempted by each annotator also varied widely, ranging from a minimum of just 6 to a maximum of 200. These factors contribute to the nuanced understanding of accuracy levels among human annotators and highlight the importance of contextualizing these results.

To investigate how accuracy may vary depending on the specific group of human annotators selected, we randomly divided the 83 annotators into four groups consisting of either 20 or 21 members each. We then calculated the accuracies for each individual group to assess this variability in human performance. Importantly, the assignment of annotators into these groups was conducted independently of the total number of annotations each group was responsible for, ensuring a fair evaluation of their performance. The results of this analysis are presented below, providing valuable insights into the impact of group selection on annotation accuracy.

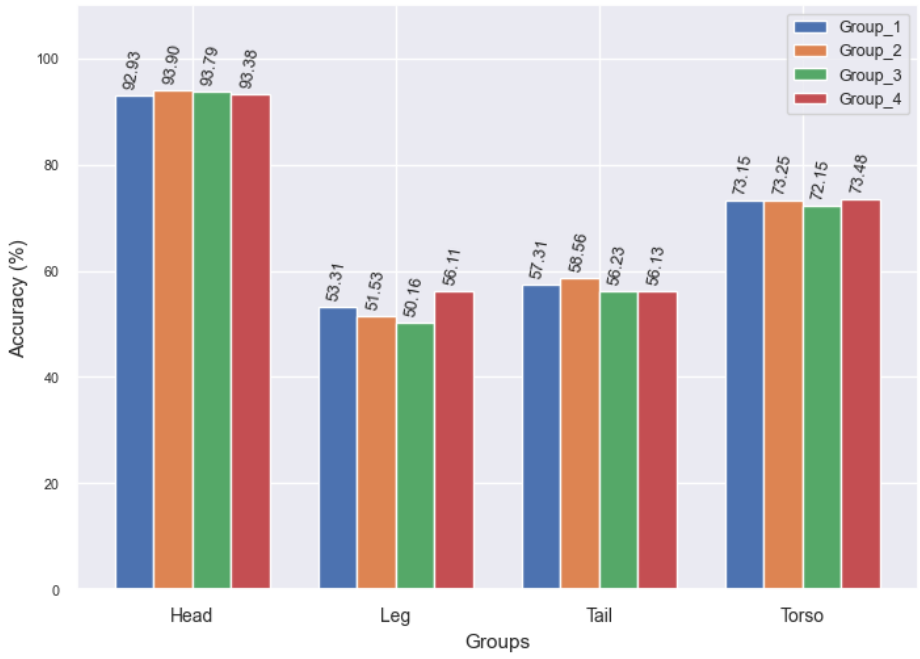

Figure 6. Variation in Annotator Accuracy across Groups

To test how accuracy may depend on which group of human annotators is selected, we randomly assigned the 83 annotators into 4 groups of 20 or 21 and calculated the accuracies for

each group. This being a test of variation in human performance, the assignment into groups was done regardless of how many or how few total annotations each group was responsible for. The results are shown below.

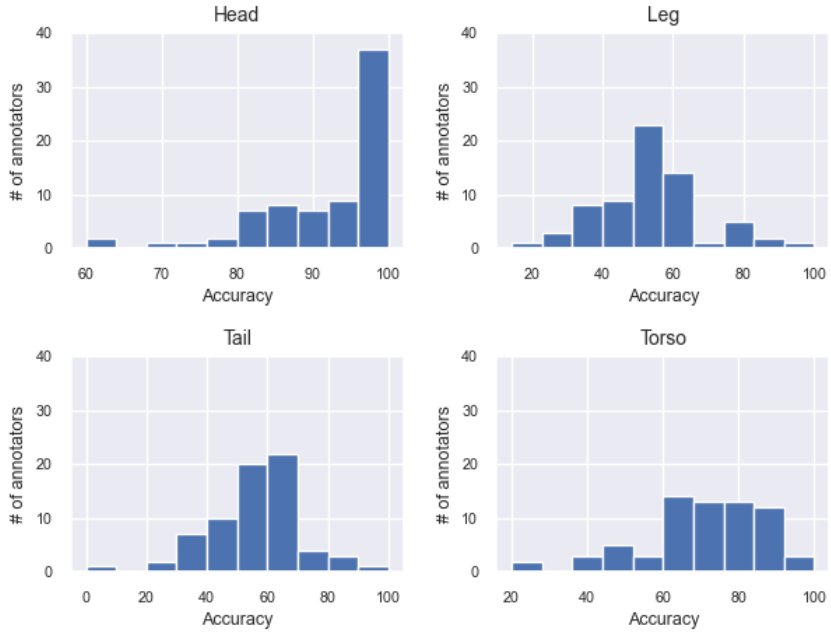

Figure 7. Variation in Annotator Accuracy across Individuals

Declarations

Availability of supporting data

The datasets used and / or analyzed during the current study are available from the first author on reasonable request.

Competing interests

The authors declare no conflict of interest.

Funding

Not applicable.

Authors' contributions

Dong, Ziwen wrote the necessary computer code, ran the algorithms and managed the results, and wrote most of the initial draft of this manuscript. Ijazul, Haq created and managed the human annotation platform, supervised the LVM prompt-writing, and wrote part of the initial draft of this manuscript. Huang, Shan helped check Dong, Ziwen's computer code. Du, Jin supervised this project and edited this manuscript.

Acknowledgements

The authors thank Du, Ruxu for the idea and motivation behind this project.

References

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. A*dvances in neural information processing systems*, 33, 1877-1901.

[2] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2023). Palm: Scaling language modeling with pathways. *Journal of machine learning research,* 24(240), 1-113.

[3] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:*2302.13971.

[4] Team, V. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. *Vicuna: An open-source chatbot impressing gpt-4 with,* 90.

[5] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). *Gpt-4 technical report. arXiv preprint arXiv*:2303.08774.

[6] Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., ... & Cucchiara, R. (2024). The revolution of multimodal large language models: A survey. *Findings of the association for computational linguistics: ACL 2024*, 13590-13618.

[7] Sparkman, M., & Witt, A. (2025). Claude AI and literature reviews: An experiment in utility and ethical use. *Library Trends*, 73(3), 355-380.

[8] Pichai, S., & Hassabis, D. (2023). Introducing Gemini: Our largest and most capable AI model. *Google. https://blog.google/technology/ai/google-gemini-ai/*

[9] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in neural information processing systems*, 36, 34892-34916.

[10] Team, Q. (2023). Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv*:2308.12966.

[11] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *In International conference on machine learning* (pp. 8748-8763).

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[13] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *In Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012-10022).

[14] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J. R. (2023). Evaluating object hallucination in large vision-language models. *In Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 292-305).

[15] Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., ... & Bai, X. (2024). Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), 220102.

[16] Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., ... & Luo, P. (2024). Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3), 1877-1893.

[17] Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., ... & Wang, L. (2023). Mm-vet: Evaluating large Multimodal models for integrated capabilities. *arXiv preprint arXiv*:2308.02490.

[18] Li, C., Liu, H., Li, L., Zhang, P., Aneja, J., Yang, J., ... & Gao, J. (2022). Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35, 9287-9301.

[19] Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., ... & Ouyang, W. (2023). Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36, 26650-26685.

[20] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science?. *Computational Linguistics*, 50(1), 237-291.

[21] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

[22] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv*:1312.6229.

[23] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

[24] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30.*

[25] Team, G., Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., ... & Blanco, L. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv*:2312.11805.

[26] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. *In 2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee.

[27] Bottou, L. (2010, September). Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers (pp. 177-186). *Heidelberg: Physica-Verlag HD*.

[28] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv*:1412.6980.

[29] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556.

[30] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

[31] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv*:1704.04861.

[32] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv*:2010.11929.

[33] Krizhevsky, A. , & Hinton, G. (2009). Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4).

[34] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (2002). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

[35] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). *Cham: Springer International Publishing.*

[36] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2), 303-338.

[37] Jeong, J. , Lee, J. Y. , & Son, Y. (2018). A study of partial image classification of vehicles using finger gestures. *International Journal of Grid and Distributed Computing*, 11, 111-122.

[38] Zhao, G., Zhang, C., Wang, X., Lin, B., & Yan, F. (2024). PMANet: Progressive multi-stage attention networks for skin disease classification. *Image and Vision Computing*, 149, 105166.

[39] Cen, F., & Wang, G. (2019). Boosting occluded image classification via subspace decomposition-based estimation of deep features. *IEEE transactions on cybernetics*, 50(7), 3409-3422.

[40] Cen, F., Zhao, X., Li, W., & Wang, G. (2021). Deep feature augmentation for occluded image classification. *Pattern Recognition*, 111, 107737.

[41] Kassaw, K., Luzi, F., Collins, L. M., & Malof, J. M. (2025). Are deep learning models robust to partial object occlusion in visual recognition tasks?. *Pattern Recognition*, 112215.