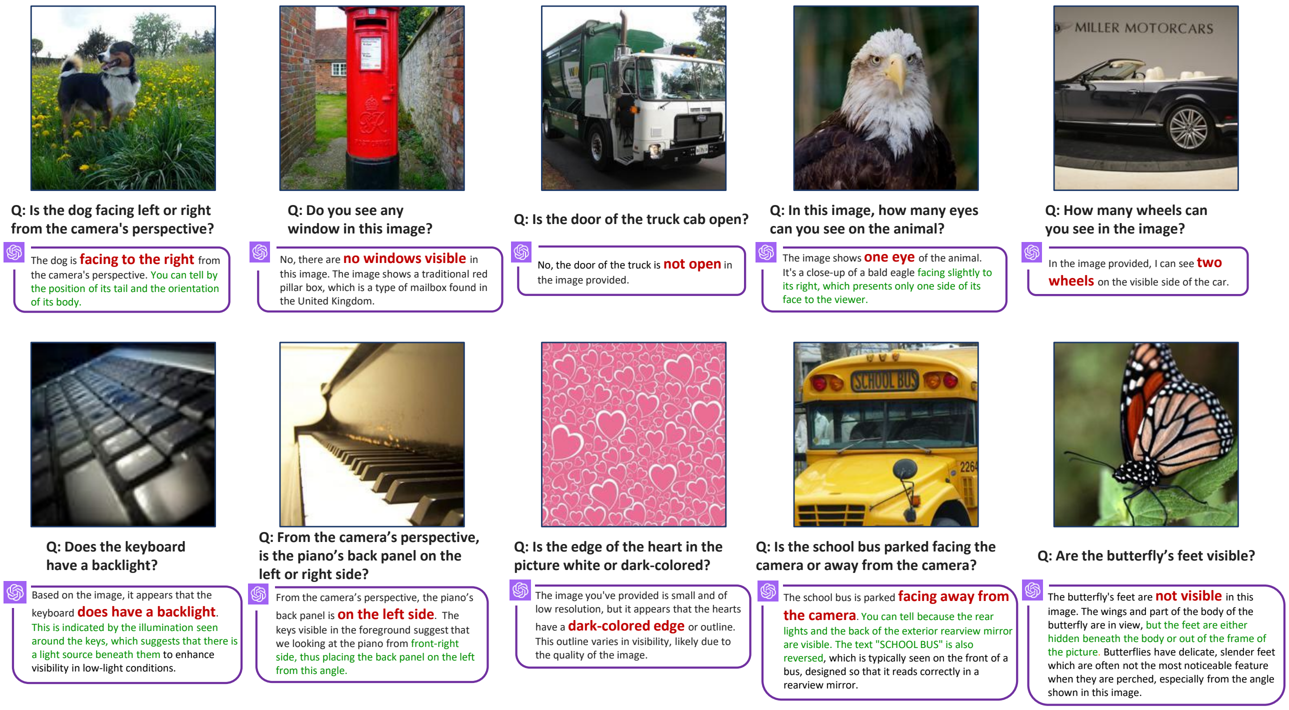
# Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

Shengbang Tong[1] Zhuang Liu[2] Yuexiang Zhai[3]
Yi Ma[3] Yann LeCun[1] Saining Xie[1]

[1]New York University [2]FAIR, Meta [3]UC Berkeley



Q: Is the dog facing left or right from the camera's perspective?

The dog is **facing to the right** from the camera's perspective. You can tell by the position of its tail and the orientation of its body.

Q: Do you see any window in this image?

No, there are **no windows visible** in this image. The image shows a traditional red pillar box, which is a type of mailbox found in the United Kingdom.

Q: Is the door of the truck cab open?

No, the door of the truck is **not open** in the image provided.

Q: In this image, how many eyes can you see on the animal?

The image shows **one eye** of the animal. It's a close-up of a bald eagle facing slightly to its right, which presents only one side of its face to the viewer.

Q: How many wheels can you see in the image?

In the image provided, I can see **two wheels** on the visible side of the car.

Q: Does the keyboard have a backlight?

Based on the image, it appears that the keyboard **does have a backlight**. This is indicated by the illumination seen around the keys, which suggests that there is a light source beneath them to enhance visibility in low-light conditions.

Q: From the camera's perspective, is the piano's back panel on the left or right side?

From the camera's perspective, the piano's back panel is **on the left side**. The keys visible in the foreground suggest that we looking at the piano from front-right side, thus placing the back panel on the left from this angle.

Q: Is the edge of the heart in the picture white or dark-colored?

The image you've provided is small and of low resolution, but it appears that the hearts have a **dark-colored edge** or outline. This outline varies in visibility, likely due to the quality of the image.

Q: Is the school bus parked facing the camera or away from the camera?

The school bus is parked **facing away from the camera**. You can tell because the rear lights and the back of the exterior rearview mirror are visible. The text "SCHOOL BUS" is also reversed, which is typically seen on the front of a bus, designed so that it reads correctly in a rearview mirror.

Q: Are the butterfly's feet visible?

The butterfly's feet are **not visible** in this image. The wings and part of the body of the butterfly are in view, but the feet are either hidden beneath the body or out of the frame of the picture. Butterflies have delicate, slender feet which are often not the most noticeable feature when they are perched, especially from the angle shown in this image.

Figure 1. Instances are systematically identified where the visual question answering (VQA) capabilities of GPT-4V [41] fall short (`Date accessed: Nov 04, 2023`). Our research highlights scenarios in which advanced systems like GPT-4V struggle with seemingly simple questions due to inaccurate visual grounding. Text in **red** signifies an incorrect response, while text in **green** represents hallucinated explanations for the incorrect answer. All the images referenced are sourced from ImageNet-1K and LAION-Aesthetic datasets.

## Abstract

*Is vision good enough for language? Recent advancements in multimodal models primarily stem from the powerful reasoning abilities of large language models (LLMs). However, the visual component typically depends only on the instance-level contrastive language-image pre-training (CLIP). Our research reveals that the visual capabilities in recent MultiModal LLMs (MLLMs) still exhibit systematic shortcomings. To understand the roots of these errors, we explore the gap between the visual embedding space of CLIP and vision-only self-supervised learning. We identify "CLIP-blind pairs" – images that CLIP perceives as similar despite their clear visual differences. With these pairs, we construct the Multimodal Visual Patterns (MMVP) benchmark. MMVP exposes areas where state-of-the-art systems, including GPT-4V, struggle with straightforward questions across nine basic visual patterns, often providing incorrect answers and hallucinated explanations. We further evaluate various CLIP-based vision-and-language models and found a notable correlation between visual patterns that challenge CLIP models and those problematic for multimodal LLMs. As an initial effort to address these issues, we propose a Mixture of Features (MoF) approach, demonstrating that integrating vision self-supervised learning features with MLLMs can significantly enhance their visual grounding capabilities. Together, our research suggests visual representation learning remains an open challenge, and accurate visual grounding is crucial for future successful multimodal systems.*

# 1. Introduction

Multimodal Large Language Models (MLLMs) [8, 13, 31, 40] have been rapidly developing in recent times. MLLMs integrate images into large language models (LLMs) and leverage the powerful abilities of LLMs [41, 59, 69], showcasing remarkable proficiency in tasks such as image understanding, visual question answering, and instruction following. In particular, the recently released GPT-4V(ision) [40] has pushed performance to an unprecedented level [41, 63].

Beneath the advancements of these models, we find there exists a notable weakness: they still exhibit visual shortcomings, some of which are surprisingly elementary and evident (see Figure 1). We ask: *Where do these problems originate? Is it a deficiency in visual modality, language understanding, or their alignment?* In this work, we suggest that these shortcomings observed in MLLMs might stem from a problem related to the **visual representations**.

At their core, most MLLMs [8, 31, 71] are built on *pretrained* vision [43, 54] and language [59, 68, 69] models. These models are connected using various types of adapters [2, 26, 31] to integrate the different modalities. A natural hypothesis is that any limitation in the pretrained vision models can cascade into the downstream MLLMs that adopt them. Studies have explored a similar issue for language. For example, Tong et al. [57], Yuksekgonul et al. [65] demonstrate that failure patterns in the pretrained text encoder [43, 44] will lead to downstream failures in text-guided generative models [22, 46].

On the vision side, most open-source MLLMs [2, 26, 31] adopt the pretrained Contrastive Language-Image Pre-Training (CLIP) model [43] as the visual encoder. We begin by identifying failure examples that CLIP struggles to encode properly (Section 2). Inspired by Tong et al. [57], we exploit the *erroneous agreements* in the embedding space. If two visually different images are encoded similarly by CLIP, then at least one of the images is likely ambiguously encoded. We call such a pair of images a *CLIP-blind* pair. To measure the visual similarity between images, we use a vision-only self-supervised encoder such as DINOv2 [42]. In this context, *CLIP-blind* pairs are images with similar CLIP embeddings but different DINOv2 embeddings.

We discover that these CLIP-blind pairs indeed lead to errors in downstream MLLMs. With these pairs, We introduce the **M**ulti**M**odal **V**isual **P**atterns (MMVP) benchmark. This benchmark is specifically designed to inquire about differences in CLIP-blind pairs and evaluate the visual abilities of *state-of-the-art* MLLMs with straightforward questions. We evaluate a variety of open-source [8, 30, 31, 71] and closed-source models [13, 41] including GPT-4V [40], and conduct a user study to measure human performance. The results show that MLLM models struggle with straightforward visual questions. Most of these models perform below the level of random guessing, with GPT-4V being the exception. Yet, even GPT-4V exhibits a considerable disparity in performance – exceeding 50% – compared to human performance.

Having identified a large number of individual failure instances in MLLMs, we continue to study the systematic visual patterns in MMVP which CLIP models struggle (Section 3). We summarize nine prevalent patterns of the CLIP-blind pairs in MMVP, such as “orientation”, “counting”, and “viewpoint”, which pose significant challenges for the CLIP vision encoder. Notice that there has been significant and ongoing progress in scaling up both training data and model size for CLIP [10, 43, 54, 62, 66]. We categorize examples from MMVP into visual patterns to systematically assess whether scaling alone can mitigate these challenges. Our findings suggest that 7 out of the 9 identified visual patterns cannot be resolved by any large-scale CLIP-based models, indicating that model/data scaling alone is not sufficient. Moreover, we identify a strong correlation between the visual patterns that challenge CLIP models and the performance of MLLMs. If CLIP struggles with a particular visual pattern, such as “orientation”, MLLMs will likely also fall short. This shows that the CLIP vision encoders could become a bottleneck in such systems.

Finally, we take a step towards improving the visual grounding of MLLMs. Since the visual shortcomings of MLLMs stem from their reliance on the CLIP model, we investigate the impact of integrating vision-centric representations into MLLMs (Section 4). Specifically, we explore ways to incorporate a vision-only self-supervised model, such as DINOv2 [42], to enhance the visual grounding capabilities of MLLMs. We refer to these techniques as Mixture-of-Features (MoF). First, we linearly mix CLIP and DINOv2 features in different ratios, which we refer to as Additive-MoF (A-MoF). This process reveals that DINOv2 features are more effective in visual grounding, though they come at the cost of diminished instruction-following ability. To address this, we introduce Interleaved-MoF (I-MoF) that spatially mixes visual tokens from both CLIP and DINOv2 models. We find that this practice significantly enhances visual grounding while maintaining the instruction-following capabilities.

# 2. The Multimodal Visual Patterns (MMVP) Benchmark

Currently, the majority of open-source MLLMs [8, 31, 71] use the *off-the-shelf* CLIP vision encoders to process images. In this section, we begin by identifying CLIP-blind pairs in the CLIP model (Section 2.1). Subsequently, we construct the Multimodal Visual Patterns-MLLM (MMVP-MLLM) benchmark using these CLIP-blind pairs (Section 2.2). We evaluate SOTA MLLMs including GPT-4V on the benchmark (Section 2.3) and find that all the tested models struggle with simple questions on visual details. A

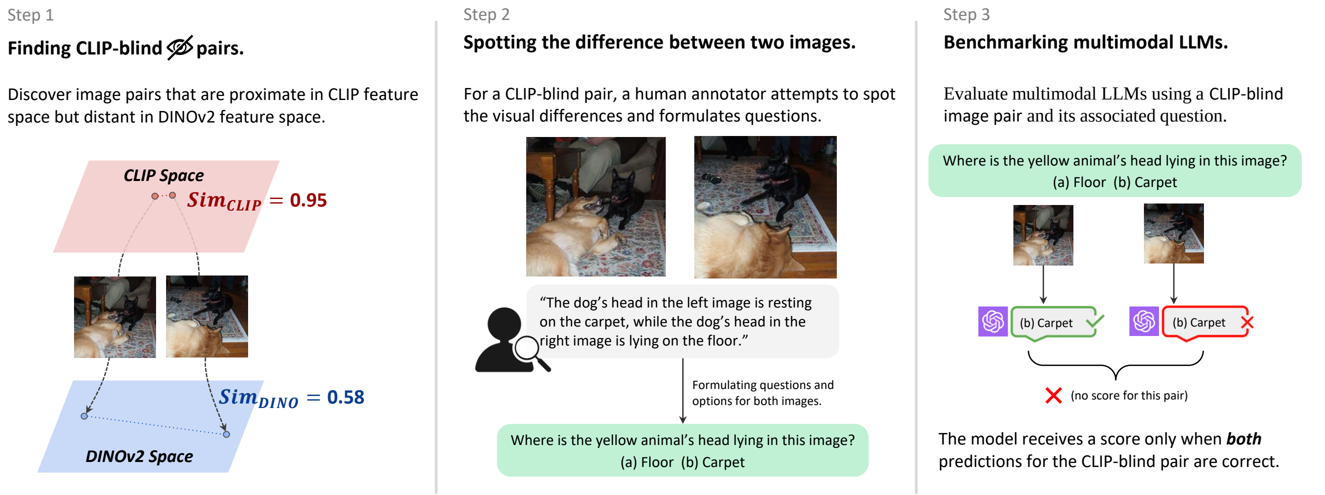

Step 1
Finding CLIP-blind pairs.
Discover image pairs that are proximate in CLIP feature space but distant in DINOv2 feature space.
CLIP Space
$Sim_{CLIP}$ = 0.95
$Sim_{DINO}$ = 0.58
DINOv2 Space
Step 2
Spotting the difference between two images.
For a CLIP-blind pair, a human annotator attempts to spot the visual differences and formulates questions.
"The dog's head in the left image is resting on the carpet, while the dog's head in the right image is lying on the floor."
Formulating questions and options for both images.
Where is the yellow animal's head lying in this image? (a) Floor (b) Carpet
Step 3
Benchmarking multimodal LLMs.
Evaluate multimodal LLMs using a CLIP-blind image pair and its associated question.
Where is the yellow animal's head lying in this image? (a) Floor (b) Carpet
(b) Carpet ✓
(b) Carpet ✗
✗ (no score for this pair)
The model receives a score only when *both* predictions for the CLIP-blind pair are correct.

Figure 2. Constructing MMVP benchmark via CLIP-blind pairs. **Left:** We start with finding CLIP-blind pairs that have similar CLIP embedding but different DINOv2 embedding. **Center:** We manually inspect the differences between pair-wise images and formulate questions based on the differences in the images. **Right:** We ask MLLMs the question alongside the CLIP-blind pair. The model receives a score only when both questions for the CLIP-blind pair are answered correctly.

visualization of this process is provided in Figure 2.

### 2.1. Finding CLIP-blind Pairs

It is challenging to directly find instances (images) that the CLIP vision encoder struggles to encode "properly". To circumvent this issue, we extend the idea proposed in Tong et al. [57] to automatically find blind pairs in vision models. The underlying principle is simple: if two images, despite having stark visual differences, are encoded similarly by the CLIP vision encoder, then one of them is likely encoded ambiguously (See Figure 2 left for example). To measure the visual difference between two images, we examine the images' representations within a reference model: a vision-only self-supervised model trained without any language guidance, e.g., DINOv2 [42]. These models are shown to capture more visual details and information [42, 53].

We take the corpus datasets, ImageNet [47] and LAION-Aesthetics [48], to collect these CLIP-blind pairs.

For each pair, we compute its CLIP embeddings using CLIP-ViT-L-14 [9, 43] model and their DINOv2 embeddings using DINOv2-ViT-L-14 [9, 42] model. We return pairs such that the cosine similarity exceeds 0.95 for CLIP embeddings and less than 0.6 for DINOv2 embeddings.

### 2.2. Designing Benchmark from CLIP-blind Pairs

We introduce the Multimodal Visual Patterns (MMVP) benchmark, and a Visual Question Answering (VQA) benchmark. Utilizing the collected CLIP-blind pairs, we carefully design 150 pairs with 300 questions. For each CLIP-blind pair of images, we manually pinpoint the visual details that the CLIP vision encoder overlooks (see the middle of Figure 2) and craft questions that probe these visual details, for example "Is the dog facing left or right?" (See the right of Figure 2 and more examples in Figure 3). The primary goal is to determine whether MLLM models would fail when posed with these seemingly basic questions and overlook critical visual details. Hence, the questions are intentionally straightforward and unambiguous.

### 2.3. Benchmark Results

We assess the questions on *SOTA* open-source models (LLaVA-1.5 [31], InstructBLIP [8], Mini-GPT4 [71]) and closed-source models (GPT-4V [40], Gemini [14], Bard [13]) We leave details of how we access the model in Appendix B.1. In our evaluation, each question is queried independently, eliminating any biases from chat histories. We also evaluate human performance through a user study where users are presented with 300 questions in a randomized sequence. For any given pair of images, we consider a pair of images to be correctly answered if both the questions associated with the pair are answered accurately.

**Human study confirms questions are straightforward.** As shown in Figure 4, human participants accurately answer an average of 95.7% of the questions. This high accuracy rate underscores the ease of the questions. More details can be found in Appendix B.4.

**Current MLLMs struggle with visual details.** As shown in Figure 4, there is a significant performance gap

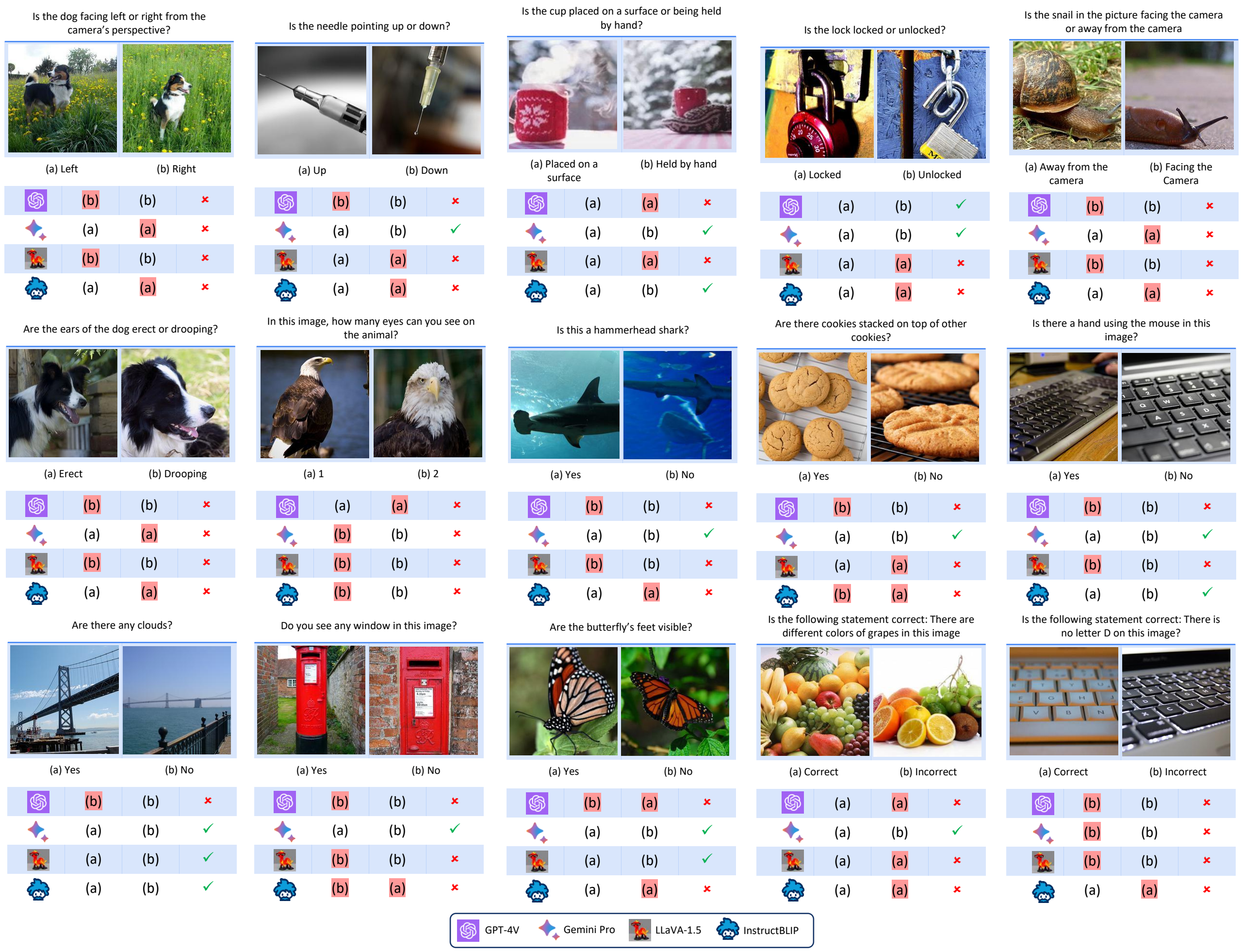

Is the dog facing left or right from the camera's perspective?
(a) Left
(b) Right
Is the needle pointing up or down?
(a) Up
(b) Down
Is the cup placed on a surface or being held by hand?
(a) Placed on a surface
(b) Held by hand
Is the lock locked or unlocked?
(a) Locked
(b) Unlocked
Is the snail in the picture facing the camera or away from the camera
(a) Away from the camera
(b) Facing the Camera
Are the ears of the dog erect or drooping?
(a) Erect
(b) Drooping
In this image, how many eyes can you see on the animal?
(a) 1
(b) 2
Is this a hammerhead shark?
(a) Yes
(b) No
Are there cookies stacked on top of other cookies?
(a) Yes
(b) No
Is there a hand using the mouse in this image?
(a) Yes
(b) No
Are there any clouds?
(a) Yes
(b) No
Do you see any window in this image?
(a) Yes
(b) No
Are the butterfly's feet visible?
(a) Yes
(b) No
Is the following statement correct: There are different colors of grapes in this image
(a) Correct
(b) Incorrect
Is the following statement correct: There is no letter D on this image?
(a) Correct
(b) Incorrect
GPT-4V
Gemini Pro
LLaVA-1.5
InstructBLIP

Figure 3. **Examples of Questions in the MMVP benchmark.** Incorrect answers are shaded in red. A model is considered correct only if it answers both questions in a pair correctly. Both leading closed-source models (GPT-4V, Gemini) and open-source models (LLaVA-1.5, InstructBLIP) fail these simple visual questions. (See Appendix B.2 for all the questions in MMVP benchmark.)

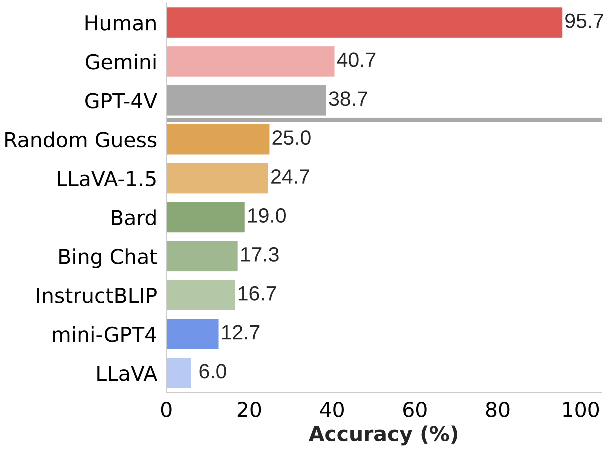


Figure 4. **Benchmark results of current *SOTA* MLLM models and humans.** We evaluate benchmark questions for current *SOTA* MLLM models and human performances through user studies.

between human and MLLM models, despite the latter often demonstrating impressive results [6, 27]. Models except GPT-4V and Gemini, scored below random guess level (25%). Most advanced GPT-4V and Gemini also face challenges in addressing basic visual grounding questions. Figures 1 and 3 provide examples of errors made by models. The outcomes suggest that irrespective of model size or training data, struggle with visual details.

We have also conducted an ablation study, such as swapping options and changing notations in the question formulation (see Appendix B.3 for more details), to further confirm that this poor performance stems from visual incapability, not hallucination in the language models.

## 3. Systematic Failures in CLIP

In the previous section, we identify CLIP-blind pairs and use them to find failures in MLLMs. Here, we delve deeper into these pairs to investigate (i) systematic visual patterns

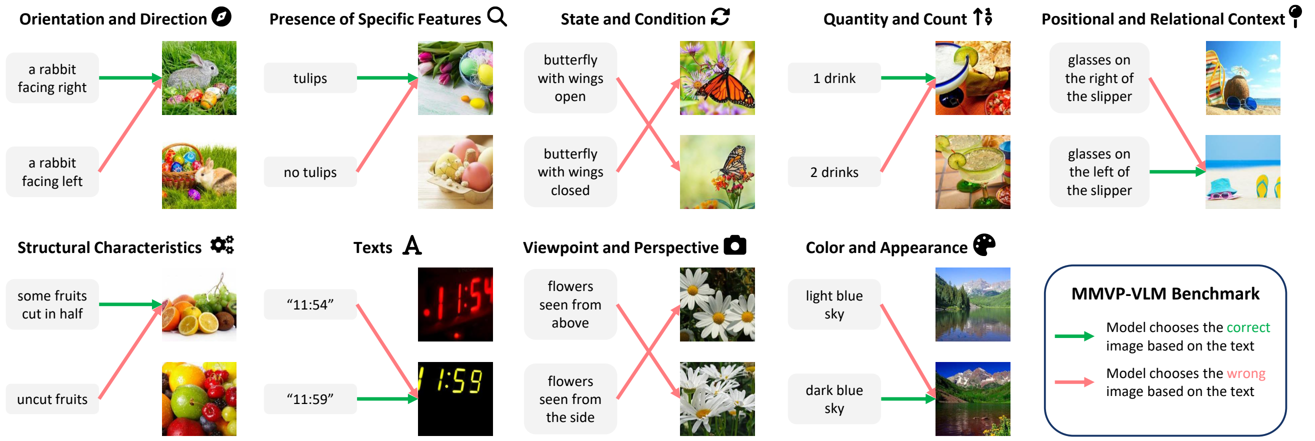

Orientation and Direction
a rabbit facing right
a rabbit facing left
Presence of Specific Features
tulips
no tulips
State and Condition
butterfly with wings open
butterfly with wings closed
Quantity and Count
1 drink
2 drinks
Positional and Relational Context
glasses on the right of the slipper
glasses on the left of the slipper
Structural Characteristics
some fruits cut in half
uncut fruits
Texts
"11:54"
"11:59"
Viewpoint and Perspective
flowers seen from above
flowers seen from the side
Color and Appearance
light blue sky
dark blue sky
MMVP-VLM Benchmark
Model chooses the correct image based on the text
Model chooses the wrong image based on the text

Figure 5. **Examples from MMVP-VLM**. MMVP-VLM consists of image pairs across nine visual patterns. The examples in the figure are from EVA01 ViT-g-14 model [54], one of the largest CLIP models that also fails to choose the right image given the text description.

emerged from CLIP-blind pairs (Section 3.1), (ii) whether these visual patterns pose challenges for CLIP-based models with massive scaling up (Section 3.2), and (iii) the correlation between failure patterns in CLIP models and those in MLLMs (Section 3.3).

### 3.1. Visual Patterns in CLIP-blind Pairs

Having identified the CLIP-blind pairs, we summarize systematic visual patterns that the CLIP vision encoders might consistently misinterpret. It is too abstract to directly capture systematic visual patterns in the CLIP-blind pairs. Therefore, we turn to the questions and options from the MMVP benchmark. With these questions, we transform abstract visual patterns in images into clearer, language-based descriptors that are easier to categorize.

In this work, we use GPT-4 [41] to categorize general patterns by prompting it with the following:

> **User**
> I am analyzing an image embedding model. Can you go through the questions and options, trying to figure out some general patterns that the embedding model struggles with? Please focus on the visual features and generalize patterns that are important to vision models
> [MMVP Questions and Options]

We identify 9 visual patterns:

- Orientation and Direction
- Presence of Specific Features
- State and Condition
- Quantity and Count
- Positional and Relational Context
- Color and Appearance
- Structural and Physical Characteristics
- Text
- Viewpoint and Perspective

These visual patterns suggest that CLIP vision encoders overly focus on high-level semantic understanding, overlooking intricate details of the visual world. Full descriptions of the visual patterns can be found in Appendix D.

### 3.2. The MMVP-VLM Benchmark

CLIP-based models have developed rapidly since the introduction in the first paper [43]. We want to test whether these visual patterns still impose challenges to the more recent CLIP models [10, 54, 62, 66], which significantly scale up in terms of training data and model size. In doing so, we introduce a new benchmark: MMVP-VLM to systematically study if CLIP models handle this visual pattern well.

We distill a subset of questions from the MMVP benchmark into simpler language descriptions and categorize them into visual patterns. To maintain a balanced number of questions for each visual pattern, we add a few questions, if needed, to ensure that each visual pattern is represented by 15 text-image pairs. Examples of pairs are shown in Figure 5. A pair is deemed correctly answered if the model can accurately match both image-text combinations.

We evaluate MMVP-VLM on a variety of CLIP models [10, 43, 54, 62, 66]. These models vary in aspects like size, training data, and methodology. As evidenced in Table 1, increasing network size and training data only aids in identifying two visual patterns – "color and appearance" and "state and condition". The rest of the visual patterns continue to challenge all CLIP-based models. We also find that the ImageNet-1k zero-shot accuracy is not a definitive indicator of a model's performance regarding visual patterns. This underscores the necessity for additional evaluation metrics, such as MMVP-VLM, to accurately assess the model's capabilities in areas beyond image classification.

### 3.3. How CLIP's Errors Affect MLLMs

After analyzing the visual patterns that CLIP models struggle with, we pose the following question: Is there a correla-

|  | Image Size | Params (M) | IN-1k ZeroShot | 🧭 | 🔍 | 🔄 | ⇅ | 📍 | 🎨 | ⚙ | A | 📷 | MMVP Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| OpenAI ViT-L-14 [43] | $224^2$ | 427.6 | 75.5 | 13.3 | 13.3 | 20.0 | 20.0 | 13.3 | 53.3 | 20.0 | 6.7 | 13.3 | 19.3 |
| OpenAI ViT-L-14 [43] | $336^2$ | 427.9 | 76.6 | 0.0 | 20.0 | 40.0 | 20.0 | 6.7 | 20.0 | 33.3 | 6.7 | 33.3 | 20.0 |
| SigLIP ViT-SO-14 [66] | $224^2$ | 877.4 | 82.0 | 26.7 | 20.0 | 53.3 | 40.0 | 20.0 | 66.7 | 40.0 | 20.0 | 53.3 | 37.8 |
| SigLIP ViT-SO-14 [66] | $384^2$ | 878.0 | 83.1 | 20.0 | 26.7 | 60.0 | 33.3 | 13.3 | 66.7 | 33.3 | 26.7 | 53.3 | 37.0 |
| DFN ViT-H-14 [10] | $224^2$ | 986.1 | 83.4 | 20.0 | 26.7 | 73.3 | 26.7 | 26.7 | 66.7 | 46.7 | 13.3 | 53.3 | 39.3 |
| DFN ViT-H-14 [10] | $378^2$ | 986.7 | 84.4 | 13.3 | 20.0 | 53.3 | 33.3 | 26.7 | 66.7 | 40.0 | 20.0 | 40.0 | 34.8 |
| MetaCLIP ViT-L-14 [62] | $224^2$ | 427.6 | 79.2 | 13.3 | 6.7 | 66.7 | 6.7 | 33.3 | 46.7 | 20.0 | 6.7 | 13.3 | 23.7 |
| MetaCLIP ViT-H-14 [62] | $224^2$ | 986.1 | 80.6 | 6.7 | 13.3 | 60.0 | 13.3 | 6.7 | 53.3 | 26.7 | 13.3 | 33.3 | 25.2 |
| EVA01 ViT-g-14 [54] | $224^2$ | 1136.4 | 78.5 | 6.7 | 26.7 | 40.0 | 6.7 | 13.3 | 66.7 | 13.3 | 13.3 | 20.0 | 23.0 |
| EVA02 ViT-bigE-14+ [54] | $224^2$ | 5044.9 | 82.0 | 13.3 | 20.0 | 66.7 | 26.7 | 26.7 | 66.7 | 26.7 | 20.0 | 33.3 | 33.3 |

Table 1. Performance of various CLIP based models on different visual patterns in MMVP-VLM benchmark. Models scaled up in resolution show minimal improvement, whereas a slight advantage is observed when scaling up the network. For each visual pattern, ImageNet-1k Zero-shot accuracy and MMVP average, we use light gray to highlight the best performance. For most of the visual patterns, all CLIP-based methods show struggle, as evident from the scores. We use symbols for visual patterns due to space limit: 🧭: Orientation and Direction, 🔍: Presence of Specific Features, 🔄: State and Condition, ⇅: Quantity and Count, 📍: Positional and Relational Context, 🎨: Color and Appearance, ⚙: Structural and Physical Characteristics, A: Texts, 📷: Viewpoint and Perspective.

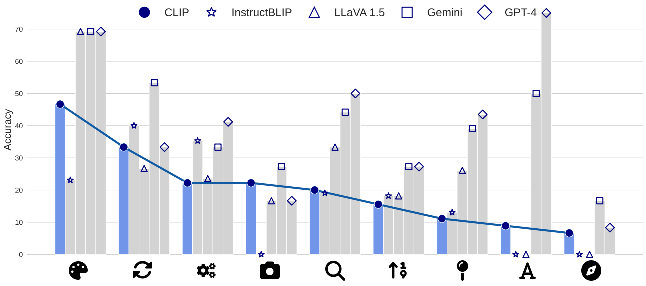


Figure 6. **CLIP and MLLM's performance on visual patterns.** If CLIP performs poorly on a visual pattern such as " 🧭 orientation", MLLMs also underperform on the visual pattern.

tion between the underperformance of CLIP and MLLMs' visual incapability? To explore this, we categorize questions from MMVP into these visual patterns summarized and calculate each MLLM's performance on these patterns.

In Figure 6, we plot CLIP's performance and MLLMs' performance for each visual pattern. When the CLIP vision encoder underperforms on a certain visual pattern, the MLLM tends to exhibit similar shortcomings. Open-source models such as LLaVA 1.5 [30] and InstructBLIP [8] that explicitly use the CLIP vision encoder display a strong correlation in performance.

Further, we calculate the Pearson Correlation Coefficient between the CLIP model and MLLM's performance on each visual pattern. Results show that LLaVA 1.5 and InstructBLIP all possess a coefficient score greater than 0.7. This high score indicates a strong correlation that weaknesses in visual pattern recognition in the CLIP model are transferred to MLLMs. More details on the Pearson Correlation Coefficient can be found in Appendix C.

## 4. Mixture-of-Features (MoF) for MLLM

Based on our exploration in earlier sections, a natural question arises: *If open-sourced MLLM's visual shortcomings come from the CLIP vision encoder, how do we build a more competent visual encoder?* In this section, we take initial steps to answer the question by studying Mixture-of-Features (MoF). We start with additive MoF that mixes CLIP features and vision-only SSL model features. Results show that each encoder presents unique advantages and limitations when employed as the pretrained model in MLLM (Section 4.2). We subsequently propose Interleaved MoF that integrates the features from both CLIP and SSL into MLLM to enhance visual grounding without compromising the model's ability to follow instructions (Section 4.3).

### 4.1. Experiment Setting

We adopt LLaVA [30, 31] as the framework to study visual encoders in MLLM. LLaVA uses a pretrained CLIP encoder and trains an adapter to align visual tokens with language tokens in the LLM. (See left side of Figure 7). We use DINOv2 [42] as the vision-only SSL model in our work because it is currently the most scalable vision-only model. Our exploration includes the use of two visual encoders: CLIP-ViT-L-14 [43] and DINOV2-ViT-L-14 [42]. To ensure consistent and fair comparisons, we train and finetune our model with the same experiment setting in LLaVA. We include the additional experimental details in Appendix A.

### 4.2. Additive MoF

We add a pretrained DINOv2 encoder into MLLM and mix the CLIP pretrained encoder with it. We use a coefficient $\alpha$ to control the portion of CLIP features and $1 - \alpha$ to control the amount of DINOv2 features and *linearly* add them

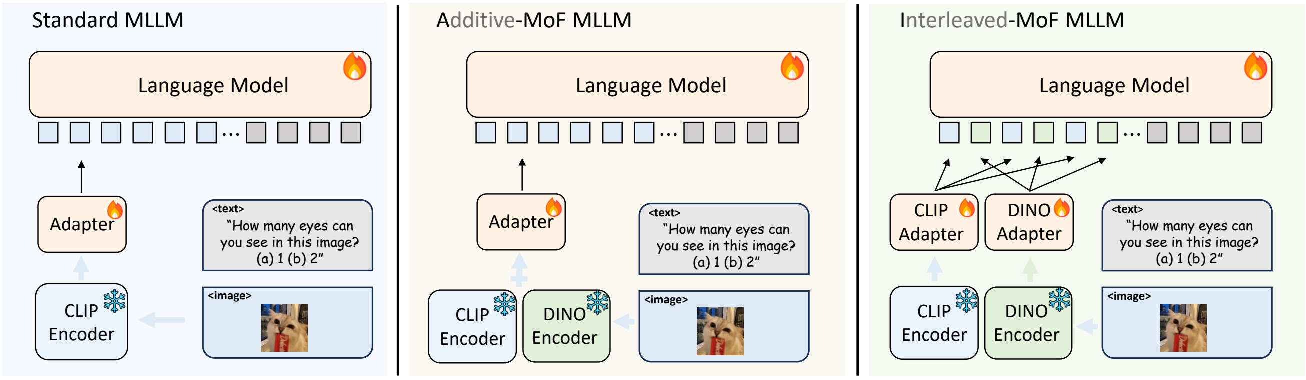


Figure 7. **Different Mixture-of-Feature (MoF) Strategies in MLLM.** *Left*: Standard MLLM that uses CLIP as *off-the-shelf* pretrained vision encoder; *Middle*: Additive-MoF (A-MoF) MLLM: Linearly mixing CLIP and DINOv2 features before the adapter; *Right*: Interleaved-MoF (I-MoF MLLM) Spatially interleaving CLIP visual tokens and DINOv2 visual tokens after the adapter.

together (See middle part of Figure 7 for visualization).

We evaluate the model's visual grounding ability by the MMVP proposed earlier in Section 2 and the model's instruction-following capability by LLaVA benchmark introduced in Liu et al. [31]. Initially, we conduct five experiments where we linearly transition from using 100% CLIP features to 100% DINOv2 features. In these tests, the DINOv2 feature proportions are set at $\{0.00, 0.25, 0.50, 0.75, 1.00\}$. To further verify the observed trends, we introduce two additional experiments with DINOv2 proportions of $\{0.625, 0.875\}$. Our findings, presented in Table 2, reveal two insights:

1. As the proportion of DINOv2 features increases, MLLM exhibits a decline in its instruction-following capability. Notably, there is a sharp decrease when the DINOv2 proportion reaches 87.5%.
2. A higher proportion of DINOv2 features enhances the model's visual grounding capability, but this advantage diminishes when the DINOv2 proportion surpasses 0.75, at which point instruction-following is notably impaired.

Hence, if we were to add DINOv2 features or completely replace CLIP with DINOv2, it would result in a trade-off between visual grounding and instruction-following. A higher proportion of DINOv2 features improves the model's visual perception at the expense of its ability to follow linguistic instructions, while CLIP features enhance language comprehension but reduce visual grounding.

## 4.3. Interleaved MoF

We propose interleaved MoF to leverage advantages from both CLIP and DINOv2 embeddings to enhance image representation. An image concurrently passes into CLIP and DINOv2 encoders, and the resulting embeddings are individually processed by adapters. We take the processed features from CLIP and DINOv2 and interleave them while maintaining their original spatial order. We then feed the interleaved features to LLM (See right part of Figure 7).

| method | SSL ratio | MMVP | LLaVA |
|---|---|---|---|
| LLaVA | 0.0 | 5.5 | **81.8** |
| LLaVA + A-MoF | 0.25 | 7.9 (+2.4) | 79.4 (-2.4) |
| | 0.5 | 12.0 (+6.5) | 78.6 (-3.2) |
| | 0.625 | 15.0 (+9.5) | 76.4 (-5.4) |
| | 0.75 | **18.7** (+13.2) | 75.8 (-6.0) |
| | 0.875 | 16.5 (+11.0) | 69.3 (-12.5) |
| | 1.0 | 13.4 (+7.9) | 68.5 (-13.3) |

Table 2. **Empirical Results of Additive MoF.** We use DINOv2 as the image SSL model in our work. With more DINOv2 features added, there is an improvement in visual grounding, while a decline in instruction following ability.

| method | res | #tokens | MMVP | LLaVA | POPE |
|---|---|---|---|---|---|
| LLaVA | $224^2$ | 256 | 5.5 | 81.8 | 50.0 |
| LLaVA | $336^2$ | 576 | 6.0 | 81.4 | 50.1 |
| LLaVA + I-MoF | $224^2$ | 512 | 16.7 (+10.7) | 82.8 | 51.0 |
| LLaVA$^{1.5}$ | $336^2$ | 576 | 24.7 | 84.7 | 85.9 |
| LLaVA$^{1.5}$ + I-MoF | $224^2$ | 512 | 28.0 (+3.3) | 82.7 | 86.3 |

Table 3. **Empirical Results of Interleaved MoF.** Interleaved MoF improves visual grounding while maintaining same level of instruction following ability.

We summarize the results in Table 3. Under the LLaVA setting, interleave MoF significantly enhances visual grounding, with a 10.7% increase observed in MMVP, without compromising the model's ability to follow instructions. This experiment is replicated with the LLaVA-1.5 setting and under various image resolution settings, yielding similar enhancements in performance. We also evaluate on POPE [27] which is designed to test hallucination in visual grounding. Interleaved-MoF also shows consistent improvement against the original LLaVA models. Merely increasing the image resolution, and consequently, the number of tokens does not boost visual grounding capabilities. Instead, it is the interleaving of MoF between vision-only SSL models and VLM models that leads to improved performance in visual grounding tasks. We conduct more experiments using MAE or MoCoV3 as vision-only SSL models in I-MoF and show similar improvements in visual grounding tasks in Appenfix E.1. We also evaluated Interleaved MoF on additional benchmarks such as MM-Bench [32] and GQA [21], finding that Interleaved MoF achieves similar performance on these benchmarks. Please refer to Appendix E.2 for more results on these benchmarks.

## 5. Related Works

**Multimodal LLMs.** We study the limitations of Multimodal LLMs [8, 13, 30, 31, 40] and explore possible ways to improve these models. Multimodal LLMs build from pretrained Large Language Models [3, 41, 58, 59, 69] and CLIP vision encoder [43, 54]. These systems then use an adapter, such as MLPs [30, 31], Q-Former [8, 26], and gated attention [2, 25], to integrate the pretrained CLIP vision encoder into LLMs. More recently, instructBLIP [8], LLaVA-1.5 [30] highlight the importance of high-quality training data. Yet, there is a scarcity of research focusing on the impact of visual encoders, which is an important gap our work aims to address through a systematic study.

**Evaluating Multimodal LLMs.** MMVP assesses MLLMs using a set of simple yet critical Visual Question Answering (VQA) questions constructed from CLIP-blind pairs. Previous benchmarks such as TextVQA [52], VQAv2 [15], and GQA [21] have centered on traditional VQA queries. Recently, there are works like MM-Vet [64], POPE [27], and MM-Bench [32] designed to specifically evaluate multimodal LLMs including hallucination, reasoning, and robustness. The previous benchmarks and evaluations have shown that Multimodal LLMs can suffer from hallucination [28, 29], catastrophic forgetting [67] and lack of robustness [11]. In taking a step back to the fundamentals, our work uncovers that even the most advanced multimodal LLMs, such as GPT-4V [40], Gemini [14], Bard [30], and LLaVA-1.5 [30], are not immune to stumbling over elementary visual questions. We also identified part of the problem as being the incapable visual encoder.

**Visual Encoders.** MMVP-VLM provides a detailed analysis of the visual capabilities of various CLIP variants [43, 54, 62, 66]. These models mostly follow the method proposed in Radford et al. [43] that uses contrastive loss to train on large volumes of image-text pairs. They differ in training data [62], training recipes [54], and objective functions [66]. Nonetheless, our studies show that all of these CLIP variants struggle with simple visual patterns such as “orientation”, “count”, “presence of specific features”, *etc*. Another line of research focuses on vision-only self-supervised learning (SSL). This category includes contrastive SSL [5, 7, 16, 17] and mask-based SSL [4, 18, 70]. SLIP [39] explores the synergy between CLIP and contrastive SSL, but focusing primarily on standard classification tasks. In fact, a common practice to evaluate the quality of these vision models is through linear probing or fine-tuning on ImageNet [45, 47]. Although current evaluation methods provide a basic level of assessment on representation quality, our findings indicate a growing detachment from the needs of recent use cases. As demonstrated in the MoF experiments in Section 4, the CLIP vision model and the vision-only SSL models learn complementary features. However, the linear probing accuracy on ImageNet alone provides a limited understanding of feature utility in MLLMs. This observation suggests the need for more diverse evaluations [61] in visual representation learning, to better align with current and emerging applications.

**Ambiguities in Embedding Models.** Our work exploits CLIP-blind pairs within the CLIP vision embedding space to generate examples of failures in CLIP models and subsequently MLLMs. This concept has ties to previous research focused on documenting failure modes in text embedding models [12, 36, 55]. More recently, Thrush et al. [56], Yuksekgonul et al. [65] and Hsieh et al. [19] study the binding problems CLIP faces in processing text queries, noting that CLIP models treat text input as a bag of words. Tong et al. [57] examines the implications for downstream text-guided generative models. Tschannen et al. [60] suggests image captioners as promising alternatives to CLIP for improving attribute binding. Our work focuses on the visual patterns.

## 6. Discussion

Circling back to the very first question we ask: is vision good enough for language? Perhaps not yet, as our study shows that vision models might become a bottleneck in multimodal systems. MLLMs fail in simple questions because their pre-trained CLIP vision encoders overlook crucial visual details in images, and systematically fail to sort important visual patterns. Yet, CLIP-type models remain the most scalable and widely used vision models today. Contrary to the popular belief that data and model scaling is a panacea, our research demonstrates that scaling alone does not rectify the inherent deficiencies in CLIP models.

Our study reveals that popular visual representation learning models – vision-and-language models and vision-only self-supervised learning models – excel in different aspects. The distinction in their capabilities go beyond conventional benchmarks such as linear probing or zero-shot accuracy on ImageNet. Although a carefully designed Mixture-of-Features approach could alleviate visual limitations and utilize the strengths of these two learning paradigms, it is necessary to develop new evaluation metrics to facilitate the development of new visual representation learning algorithms. We hope our work can motivate further innovation in vision models.

**Acknowledgements.** We thank Penghao Wu, Muzi Tao, Erik Jones, Michael Psenka, Daniel Yeh, Druv Pai, Chen Sun for helpful discussions and feedback. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise. This research is also supported by Intel, Google TRC program, the Google Cloud Research Credits program with the award GCP19980904, and an Amazon Research Award Fall 2023. The authors thank hyperbolic labs for supporting part of the experiments. All experiments and data processing were performed at NYU.

## References

[1] ShareGPT, 2023.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeruIPS*, 2022.

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

[4] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

[5] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. 2022.

[6] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICML*, 2021.

[10] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

[11] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[12] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *NAACL*, 2019.

[13] Google. Bard, 2023.

[14] Google. Gemini, 2023.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 2020.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[19] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *NeurIPS*, 2023.

[20] Jennifer Hu and Roger Levy. Prompt-based methods may underestimate large language models' linguistic generalizations. In *EMNLP*, 2023.

[21] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

[22] Heewoo Jun and Alex Nichol. Shap-E: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

[23] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[25] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*, 2023.

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[27] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[28] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you

see? an image-context reasoning benchmark challenging for GPT-4V (ision), LLaVA-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023.

[29] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

[30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023.

[32] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2017.

[34] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

[35] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.

[36] Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *NAACL*, 2019.

[37] Microsoft. newbing, 2023.

[38] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual question answering by reading text in images. In *ICDAR*, 2019.

[39] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pretraining. In *ECCV*, 2022.

[40] OpenAI. GPT-4V(ision) System Card, 2023.

[41] OpenAI. Gpt-4 technical report, 2023.

[42] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[44] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

[45] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *NeurIPS*, 2021.

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[48] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

[49] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *ECCV*, 2022.

[50] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[51] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, 2020.

[52] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, 2019.

[53] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of MAE pre-pretraining for billion-scale pretraining. In *ICCV*, 2023.

[54] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[55] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *ACL*, 2019.

[56] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022.

[57] Shengbang Tong, Erik Jones, and Jacob Steinhardt. Mass-producing failures of multimodal systems with language models. In *NeurIPS*, 2023.

[58] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[59] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLaMA 2: Open foundation and fine-tuned chat models. 2023.

[60] Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. *NeurIPS*, 2023.

[61] Kirill Vishniakov, Zhiqiang Shen, and Zhuang Liu. Convnet vs transformer, supervised vs clip: Beyond imagenet accuracy, 2024.

[62] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. *arXiv preprint arXiv:2309.16671*, 2023.

[63] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The Dawn of LMMs: Preliminary Explorations with GPT-4V (ision). *arXiv preprint arXiv:2309.17421*, 2023.

[64] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[65] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2022.

[66] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

[67] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

[68] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[69] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena, 2023.

[70] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *ICLR*, 2021.

[71] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.