Verb Mirage: Unveiling and Assessing Verb Concept Hallucinations in Multimodal Large Language Models

Zehao Wang[1], Xinpeng Liu[1,2], Yudonglin Zhang[1], Xiaoqian Wu[1], Zhou Fang[1], Yifan Fang[1], Junfu Pu[3], Cewu Lu[1,2*], Yong-Lu Li[1,2*]

[1]Shanghai Jiao Tong University
[2]Shanghai Innovation Institute
[3]ARC Lab, Tencent PCG
{davidwang200099, rightoverthere, enlighten, joefang, despr0, lucewu, yonglu_li}@sjtu.edu.cn,
xinpengliu0907@gmail.com, jevinpu@tencent.com

Abstract

Multimodal Large Language Models (MLLMs) have garnered significant attention recently and demonstrate outstanding capabilities in various tasks such as OCR, VQA, captioning, *etc*. However, hallucination remains a persistent issue. While numerous methods have been proposed to mitigate hallucinations, achieving notable improvements, these methods primarily focus on mitigating hallucinations related to **object/noun concepts**. Verb concepts, which are crucial for understanding human actions, have been largely overlooked. In this paper, to the best of our knowledge, we are the **first** to investigate the **verb hallucination** phenomenon of MLLMs from various perspectives. Our findings reveal that most state-of-the-art MLLMs suffer from severe verb hallucination. To assess the effectiveness of existing mitigation methods for object concept hallucination in relation to verb hallucination, we evaluated these methods and found that they do not effectively address verb hallucination. To address this issue, we propose a baseline method based on fine-tuning with rich verb knowledge, achieving decent superiority. The experiment results demonstrate that our method significantly reduces hallucinations related to verbs.

Code —
https://github.com/davidwang200099/Verb_Mirage

Introduction

Multimodal Large Language Models (MLLMs) (Zhu et al. 2024; Liu et al. 2023; Chen et al. 2024b; Bai et al. 2023; Li et al. 2024c) have drawn much attention in both the research and industry communities. Armed with high-quality data, a large number of parameters, and efficient instruction-following fine-tuning, they achieve significant success in various tasks, including OCR, VQA, and image captioning, demonstrating a strong generalization ability.

However, MLLMs' performance improvement could be hindered by hallucination. Typically, hallucination (Ji et al. 2023; Liu et al. 2024c) means the output of MLLM contains content against facts, irrelevant or nonsensical given context, such as prompt or multimodal input. To test MLLM hallucination in different tasks, many benchmarks (Liu et al. 2024d; Cai et al. 2024; Fu et al. 2025) have been made, allowing people to assess MLLMs' abilities in various aspects. To mitigate MLLM hallucination, many methods (Huang et al. 2024; Leng et al. 2024; Sun et al. 2024; Yin et al. 2024; Zhou et al. 2023) have been proposed, successfully relieving hallucination to a large extent.

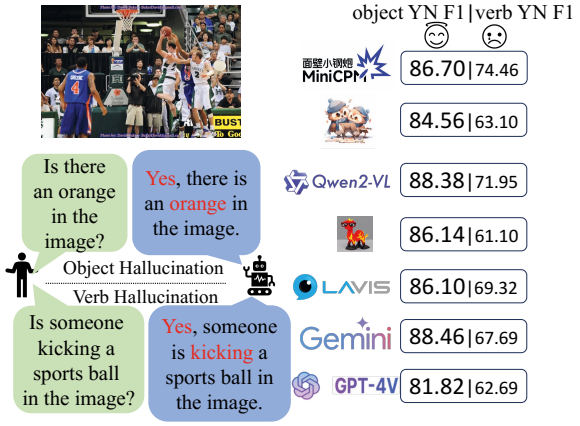


Figure 1: Besides the well-discussed object hallucination, in this paper, we unveil the severe **verb hallucination** of state-of-the-art MLLMs with our designed benchmarks. All models show low object hallucination (on POPE) but severe verb hallucination. Gemini-1.5-Flash and GPT-4-Turbo are tested with 100 randomly sampled questions.

However, existing benchmarks and methods mainly target hallucination about **objects/noun-related** concepts. **Action/verb-related** concepts, which are crucial to understanding human events, are overlooked.

To this end, we propose to dig into the verb hallucination problem. We build the **first** verb-hallucination-oriented benchmark, which is based on existing datasets (Chao et al. 2015; Sigurdsson et al. 2018) without the need for extra manual annotations. As MLLMs are a cooperation of vision and language modalities, we probe MLLM verb hallucination given both different visual inputs and language inputs, covering different query conditions, different imaging conditions, and different semantic conditions. Extensive experiments show that all MLLMs perform poorly on many as-

*Corresponding authors.

pects and thus show severe verb hallucination.

Moreover, we test existing low-cost hallucination mitigation methods on widely used MLLMs and show that they all fail in mitigating verb hallucination. To somewhat relieve verb hallucination, we propose a baseline method based on parameter-efficient fine-tuning with verb structure knowledge. Experiments show that our method successfully relieves verb hallucination, but its performance is still far from satisfactory. Finally, we explore the reason for verb hallucination and discuss possible future solutions.

In conclusion, our contributions are:

1. To our knowledge, we point out and analyze MLLM's verb hallucination for the first time and probe this phenomenon under different conditions.
2. We study the influence of existing training-free and finetuning-based methods on MLLM verb hallucination. We find that fine-tuning is still the most promising way for verb hallucination mitigation.
3. We probe model behavior from the perspective of vision-language interaction and token uncertainty, and study how well MLLMs learn verbs.

Related Work

MLLM Benchmarks. Before the emergence of MLLMs, great efforts have been made to build datasets on tasks like image captioning (Chen et al. 2015; Sharma et al. 2018; Young et al. 2014), VQA (Goyal et al. 2017; Hudson and Manning 2019; Marino et al. 2019), OCR (Singh et al. 2019), action recognition and detection (Gu et al. 2018; Li et al. 2019, 2020a,b,c), *etc*. However, they mainly assess domain-specific models. To fully evaluate MLLMs, more benchmarks have been proposed (Liu et al. 2024d; Ying et al. 2024; Yue et al. 2024; Li et al. 2024a; Tong et al. 2024; Li et al. 2025) to test different aspects and subtasks. Benchmarks are also proposed to conduct detailed assessments on specific aspects, like BenchLMM (Cai et al. 2024) for robustness against image styles and MMSpuBench (Ye et al. 2024b) for robustness against spurious correlations.

MLLM Hallucination. Among the emerging benchmarks, hallucination has become a focus. Typically, hallucination means that the contents generated by models are untruthful, against facts, or nonsensical (Ji et al. 2023; Liu et al. 2024c; Zhang et al. 2025b). Many benchmarks on hallucination have been proposed (Guan et al. 2024; Kaul et al. 2024; Wang et al. 2024b; Chen et al. 2024a; Zhong et al. 2024). Among different types of hallucination, object hallucination (Rohrbach et al. 2018) is deeper explored. Binary questions are used to probe hallucination about a certain class of objects in POPE (Li et al. 2023). CHAIR score (Rohrbach et al. 2018) is used to measure object hallucination in image captioning. In most previous studies, only object concept hallucinations were covered by identifying whether MLLMs refer to objects nonexisting in the image or incorrect attributes of objects. Though some other phenomena are studied, such as event hallucination (Jiang et al. 2024), relationship hallucination (Wu et al. 2024), *etc.*, verb-related concepts, which are crucial to understanding human actions, were still neglected. Event is a complex combination of objects, verbs, adjectives, adverbs, *etc.*, and Hal-eval (Jiang et al. 2024) tests event hallucination by introducing nonexisting objects and can be bypassed by mitigating object hallucination. Verbs are a kind of relation, but visual relationship also covers *spatial relations, ownership, subject-place relations, attributes*, *etc.*, which account for the majority. In this paper, we focus on the understanding of **humans existing in the image**, *i.e.*, **a human-centric problem**.

Hallucination Mitigation. Researchers have revealed the reasons for hallucination from many different aspects and proposed hallucination mitigation methods. Over-attention to summarizing tokens (Huang et al. 2024) and language prior (Leng et al. 2024) are recognized to be correlated with object hallucination. To mitigate the bias or errors in training data, researchers proposed mitigating bias (Liu et al. 2024a; Hu et al. 2023; Gunjal, Yin, and Bas 2024) in the dataset or enriching the annotation (Zhai et al. 2023). Moreover, some suggest post-processing at inference time by adjusting decoding strategy (Leng et al. 2024; Chen et al. 2024c) or correcting the output of MLLMs with the help of expert models (Zhou et al. 2023; Yin et al. 2024).

Probing on MLLM IO Conditions

As shown in Fig. 2, we probe verb hallucination from various perspectives: MLLM behavior given different question formats, image qualities, verb semantics, viewpoints, *etc*.

We select HICO (Chao et al. 2015) and CharadesEgo (Sigurdsson et al. 2018) as the main datasets for probing verb hallucination. HICO contains 47K images with dense annotations. It includes 600 action classes formed by 80 object classes and 117 verb classes. WordNet (Miller 1995) was used to handle synonym and hierarchy problem. It has rich verb labels and is thus suitable for evaluating MLLM verb understanding with minor manual adaptation. Meanwhile, CharadesEgo contains 7K videos of daily indoor activities. In each scenario, the same actor is recorded with both an egocentric and an exocentric camera. It contains 157 action classes, each formed by a verb and an object.

We test several open-sourced and close-sourced MLLMs including InstructBLIP-7B (Dai et al. 2023), LLaVA-V1.5-7B (Liu et al. 2024b), mPLUG-Owl2 (Ye et al. 2024a), Qwen-VL-Chat (Bai et al. 2023), MiniCPM-Llama3-V2.5 (Yao et al. 2024), Qwen2-VL-7B-Instruct (Wang et al. 2024a), Molmo-7B-D (Deitke et al. 2025), GPT-4 (Achiam et al. 2023), Gemini-1.5 (Team et al. 2023), *etc*. They have different ranks on leaderboards and show outstanding results on benchmarks targeting object concepts.

Probing on Query Conditions

Question Formats Next, we probe the relation between verb hallucination and QA format. MLLMs with low hallucinations should give hallucination-free answers, given different question formats. Thus, we evaluate verb hallucination using different question formats, including Multiple Choice (MC) questions with only one correct answer each and Yes-or-No (YN) questions. Here, we do not introduce

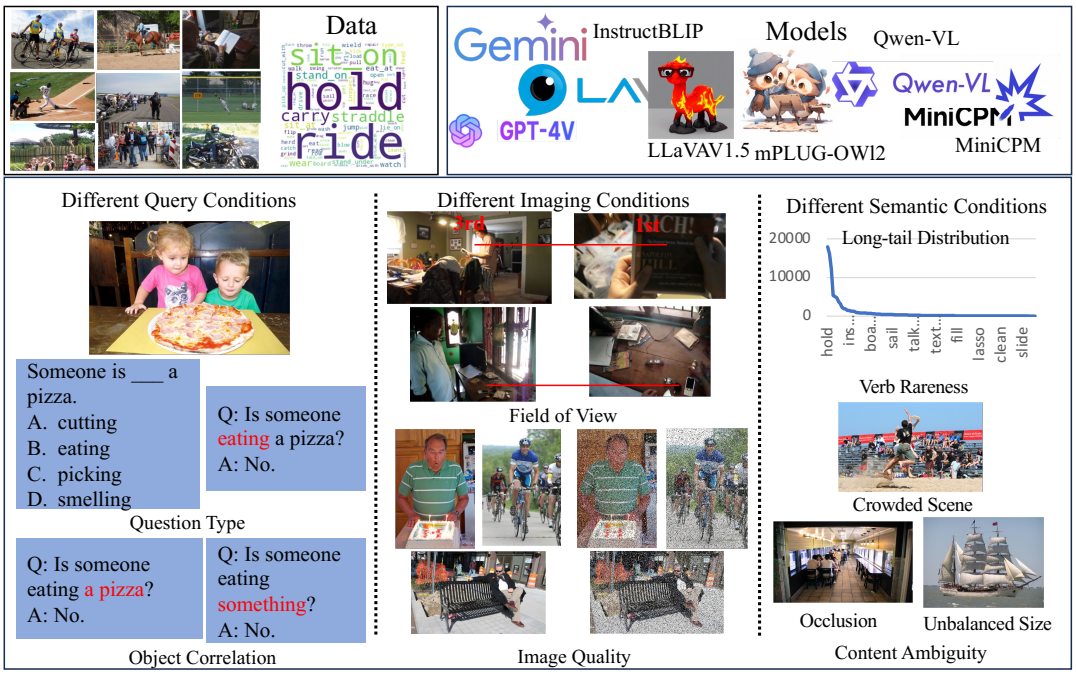


Figure 2: We probe MLLM verb hallucination from various perspectives, *eg*., question formats, the existence of object correlation, different fields of view, image qualities, verb semantics, and image semantics.

free-form image captioning and blank-filling because these two forms require rule-based post-processing on verbs and may lead to severe misclassification errors.

We aim to verify none other than verb hallucination, so we do our best to omit relevance to object hallucination. For each *verb-object* tuple in the labels, we form questions or options by altering the verb and leaving the object unchanged. For YN questions, when building negative samples, we randomly choose verbs that are plausible for the objects but not carried out in the image. For example, if an image contains a person holding a cup, we may ask MLLM, "Is someone holding a cup? Is someone washing a cup?" We regard accuracy, precision, recall, and F1 score as the metrics for hallucination. Similarly, when building MC questions, for a sample image, we randomly choose a verb presented in the image and three verbs possibly performable upon objects but not presented. We introduce circular evaluation (Liu et al. 2024d) and regard accuracy as a metric for MLLM verb hallucination. Then, the relevance of object hallucination can be minimized. However, we must point out that as a substantial proportion of verbs are transitive verbs, the influence of objects can not be completely omitted. The construction of benchmarks is detailed in Suppl. Sec. 1.

Object Correlation Sometimes we focus on human interaction with a certain class of object, but sometimes our focus on verbs may be object class agnostic. Specifically, we may wonder "Is someone holding a cup in the image?" However, sometimes we may also want to know "Is someone eating something in the image?" Therefore, we test MLLM verb understanding, given reference to objects and not. Among these two conditions, we believe questions without object correlation (*ie*, "Is someone eating something?") have less relevance to object hallucination. Still, questions with object correlation are also very practical in daily use.

Analysis The results are shown in Tab. 1, giving us rich clues about MLLM verb hallucination.

Heavy reliance on objects. MLLMs show drastic performance degradation on MC questions without reference to objects. Detailed statistics on YN questions based on object classes referred to in the questions also reveal that MLLM verb understanding relies heavily on object reference. We analyze some commonly used datasets (Sharma et al. 2018; Changpinyo et al. 2021; Lin et al. 2014; Ordonez, Kulkarni, and Berg 2011; Krishna et al. 2017; Hudson and Manning 2019; Goyal et al. 2017) for MLLM pretraining and investigate the number of nouns and verbs in the datasets in Fig. 4(a). We can see that the number of nouns is 4-10 times the number of verbs. One reason for this unbalanced ratio of nouns and verbs is that datasets on action understanding have not attracted enough attention from the MLLM community. Another reason is that there are many more nouns than verbs in English. Research on shortcut learning (Geirhos et al. 2020) also sheds light on the reason for MLLMs' overreliance on nouns.

Inability to refuse. All MLLMs have high recall but low precision, meaning that MLLMs tend to give *Yes* whether a certain verb is presented in the image or not. Binary questions require MLLMs to have a deep understanding of verb concepts in images, which are more difficult to answer than MC questions, but vitally important in daily use.

Similar Bias. We ensemble the answers given by three outstanding models and show results in Tab. 3. There is no

Model	YN w/ obj acc	YN w/ obj prec	YN w/ obj recall	YN w/o obj acc	YN w/o obj prec	YN w/o obj recall	MC w/ obj acc	MC w/o obj acc
Molmo-7B-D	59.16	44.91	96.63	46.13	38.39	99.19	**60.64**	56.78
Qwen2-VL-7B	75.51	58.37	93.75	74.69	57.43	95.87	**71.47**	65.31
MiniCPM	80.91	66.83	85.41	79.14	63.33	90.33	**66.39**	60.77
Qwen-VL-Chat	78.06	62.37	87.02	79.24	65.09	82.68	**55.95**	54.57
mPLUG-Owl-2	62.94	47.38	95.99	62.61	47.25	94.94	**63.91**	62.60
LLaVA V1.5	52.16	40.99	97.35	59.16	45.10	98.06	**57.37**	51.00
InstructBLIP	72.53	55.79	86.77	73.82	57.25	87.79	**13.48**	6.25

Table 1: Results on YN and MC questions w/ and w/o object reference. Red: high recall. Blue: low precision. **Bold**: higher MC acc w/ object reference than w/o object referece.

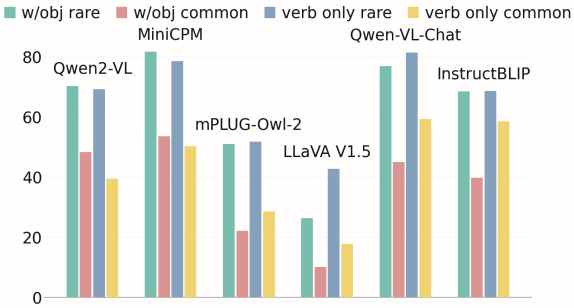


Figure 3: Comparison of YN questions with correct answer *No* on rare and common subsets.

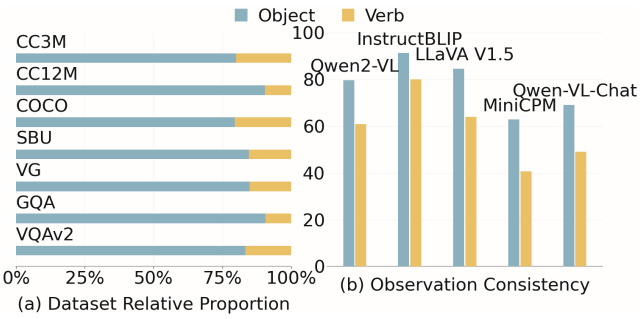


Figure 4: Comparison between objects and verbs.

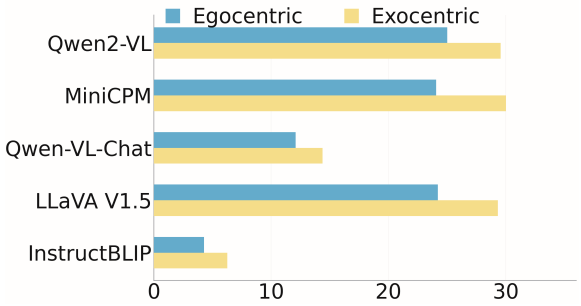


Figure 5: Performance comparison on egocentric and exocentric verb understanding (question type: MCQ).

substantial improvement over the individual models, showing that the models share similar biases.

Probing on Imaging Conditions

High-Quality and Low-Quality Images Previous research (Li et al. 2022; Leng et al. 2024) has revealed that visually distorted images can hinder both humans and models from recognizing the contents in images well. However, the relation between visual distortion and verb hallucination is unexplored. Do MLLMs hallucinate in the same way when given high-quality images and visually distorted images? Is verb understanding more sensitive to visual distortion than object understanding for MLLMs? Here, we add pepper-salt noise as a visual distortion to images, each pixel of which is affected at a probability of 75%.

We evaluate MLLM verb understanding with both high-quality and visually-distorted images and report performance and error consistency following (Geirhos et al. 2021) in Tab. 2. All tested MLLMs show obvious performance degradation. Error consistency in the form of Cohen's Kappa (Cohen 1960) measures MLLM consistency of answers given different visual conditions and provides a guideline for MLLM performance improvement. We can see that some MLLMs with low ranks do not have bad error consistency, but MLLMs with higher ranks show low error consistency. This means that their verb hallucination can be easily induced by visual distortion.

As a control experiment, besides verb understanding, we also built a test set for MLLMs' object understanding with the same set of images. The observation consistency in terms of Cohen's Kappa is reported in Fig. 4(b). From the result, we can see that MLLM shows much higher inconsistency in verb understanding than object understanding, which means that visual distortion does more harm to verb understanding than object understanding. We attribute this result to the sparsity of verb semantics in pixel space.

In conclusion, **visual distortion affects both object understanding and verb understanding of MLLMs, but the effect on verb understanding is greater.**

Egocentric and Exocentric Images Recently, MLLMs have shown outstanding capabilities in recognition and reasoning, and a growing body of research has delved into leveraging MLLMs for various tasks in robotics, where egocentric images are widely used. Therefore, a study of MLLM's understanding of verbs in egocentric images holds significant importance. To evaluate MLLMs' understanding of verbs in egocentric images and the performance gap between egocentric and exocentric images, we build a small test set using Charades-Ego and conduct experiments on different MLLMs.

Given the same scenario recorded with exocentric and egocentric cameras, we also probe MLLMs' understanding with MC and YN questions. The results are shown in Fig. 5 and Tab. 4. There is a substantial gap between MLLMs' understanding of exocentric and egocentric images.

In conclusion, **MLLMs can not understand verb concepts in egocentric images as well as exocentric images. MLLMs are better at exocentric verb understanding.**

	YN verb only					MC verb only		
	w/o Pepper Salt		w/ Pepper Salt		YN Err. Cons.	w/o Pepper Salt	w/ Pepper Salt	MC Err. Cons.
	YN acc	YN F1	YN acc	YN F1		MC acc	MC acc	
MiniCPM-Llama3-V2.5	79.14	74.46	67.40	57.23	26.12	60.77	40.50	37.20
Qwen-VL-Chat	79.24	72.84	66.64	61.88	38.47	54.57	33.98	43.38
LLaVA V1.5	59.16	61.79	51.29	57.35	73.85	51.00	49.97	68.37
InstructBLIP	73.82	69.30	71.04	67.02	74.16	6.25	6.34	82.33

Table 2: Performance comparison for images w/ and w/o pepper-salt noise. Underline: higher performance w/o peppser-salt noise. Red/Blue: good/bad error consistency (Err. Cons.).

	YN w/ obj			YN verb only		
	acc	prec	recall	acc	prec	recall
MiniCPM	80.91	66.83	85.41	79.14	63.33	90.33
Qwen-VL-Chat	78.06	62.37	87.02	79.24	65.09	82.68
InstructBLIP	72.53	55.79	86.77	73.83	57.26	87.80
Ensemble	75.59	58.74	91.17	79.81	64.48	89.12

Table 3: Ensembled accuracy/precision/recall on YN question w/wo object reference.

View	Egocentric			Exocentric		
Model	acc	prec	recall	acc	prec	recall
Qwen2-VL-7B	60.10	60.31	59.12	63.01	60.98	72.25
MiniCPM-Llama3-V2.5	59.42	62.60	46.78	62.00	67.44	46.39
Qwen-VL-Chat	57.06	61.87	36.79	60.62	64.94	46.19

Table 4: Comparison on egocentric and exocentric verb understanding (question type: YN).

Probing on Semantic Conditions

Rare and Common Verbs Verbs follow a long-tailed distribution in action datasets because of all sorts of difficulties in the process of dataset collection. However, understanding *rare* and *common* verbs is equally important in real-world applications. We hypothesize that MLLMs tend to hallucinate more on rare verbs and try to prove it on existing datasets.

We divide the negative samples into two subsets: the rare set and the common set. In the rare set, the verb in question lies in the *tail* of HICO verb distribution, while in the common set, the verb in question lies in the *head*. Specifically, for YN questions with an object reference, the rare set contains all HOIs whose annotations make up less than 20% among HOIs relevant to the same object class. For YN questions without object reference, the common set contains all questions containing `hold`, `ride`, `sit_on`, `straddle`, and `carry`, making up 50% of the verb annotations in the HICO dataset. From Fig. 3, we can see that MLLMs tend to refuse existent rare verbs but accept nonexistent common verbs in images. This phenomenon reveals that the long-tailed distribution of verb annotations limits MLLM verb understanding. How to understand rare verbs remains a problem, and there is a large room for action, data collection, and curation. In conclusion, **MLLMs can not understand rare verbs as well as common verbs, *i.e.*, long-tail affects a lot**.

Image Content Ambiguity Ambiguity always exists in real-world scenarios. Understanding verbs in crowded or heavily occluded scenarios is important in many fields such as surveillance, social robotics, and visual reasoning. To assess MLLMs' verb understanding given images with ambiguous content, we select images from HICO containing content ambiguities, form an ambiguous subset, and compare them with images with less content ambiguity. Specifically, our ambiguous subset contains many contributing factors to ambiguity:

- **Imbalanced human-object relative size.** Imbalanced human-object relative sizes can add difficulties to MLLM's perception of humans and objects. The existence of verbs relies heavily on the accurate perception of humans and objects in images. Potential failures in perception bring great difficulties to the recognition of verbs.
- **Crowded scene.** A highly complicated scene structure can distract MLLMs. To judge the existence of a verb, MLLMs must analyze all humans and objects and draw a comprehensive conclusion. The large number of humans and objects puts heavy burdens on MLLMs.
- **Occlusion.** We can recognize humans and objects according to parts of them and analyze their relationship, even with prominent occlusion. Thus, visual reasoning under occluded scenarios is important in fields like action understanding and robot manipulation.

From Fig. 6, we have the following finding: **MLLMs show performance degradation when images contain content ambiguity. How to understand verbs given ambiguous content is an open problem to be solved.**

Probing on Model Behaviors

In this section, we take LLaVA V1.5 as an example and study the relationship between its behavior and verb hallucination. Our study finds that although verb hallucination shares some commonality with object hallucination, they are fundamentally different.

Vision-Language Interaction

Key Image Area Attention. First, we hypothesize that models pay less attention to key information in images and text when they give hallucinated answers. From Fig. 7(a), we can see that there is an obvious distinction of distribution between hallucinated attention and non-hallucinated attention,

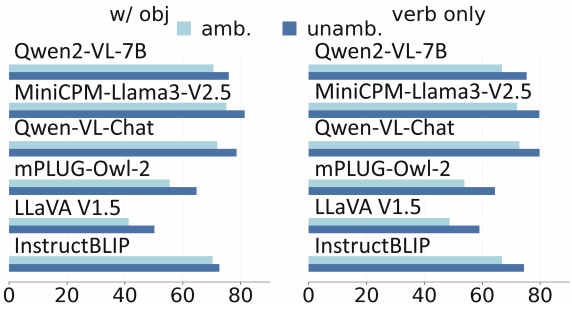
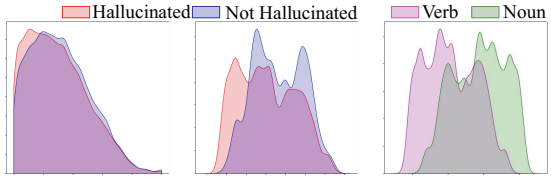


Figure 6: MLLM accuracy on ambiguous (Amb.) and unambiguous (Unamb.) subsets.



(a) Key Image Area Attn. (b) Verb Token Attn. (c) Verb vs. Noun Attn.

Figure 7: Attention for YN questions with object references.

showing **a strong correlation between inadequate attention to key areas and hallucination**.

Visual Token Attention. The Visual Token Attention is defined as

$$\underset{j}{\text{mean}} \frac{\sum_{i \in V} \alpha_{ij}}{\sum_{i \in V} \alpha_{ij} + \sum_{k \in T} \alpha_{kj}}, \quad (1)$$

where α_{ij} represents the attention weight assigned to token i at head j in the last transformer layer, V denotes the set of visual tokens, and T denotes the set of textual tokens. We visualize it in Fig. 8 and we find:

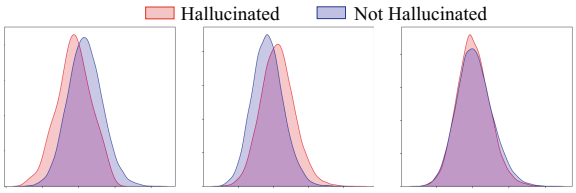
1. For questions with correct answer *No* hallucinated models tend to give more attention to visual tokens.
2. For MC questions, there is no obvious difference in vision token attention between hallucinated models and non-hallucinated models.

Therefore, **unlike what has been found on object hallucination, more attention to visual tokens does not exclude verb hallucination.**

Token Uncertainty

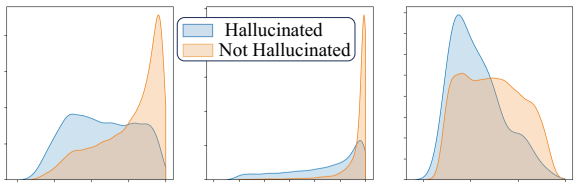
Token Uncertainty and Hallucination We dig into the uncertainty (Zhou et al. 2023) of MLLMs via visualizing the distribution of probabilities of predicted tokens of the widely-used open-sourced LLaVA V1.5. From Fig. 9, we can see that there is a substantial difference between correct answers and hallucinated answers: hallucinated tokens are mostly given with low probability, which means that the model is confused about the answers. For Yes-or-No questions, we visualize questions to which the model gives answers *Yes* and *No* separately. The observations are:

1. LLaVA V1.5 tends to answer *Yes* with high confidence, but *No* with relatively lower confidence.



(a) positive samples (b) negative samples (c) MC questions

Figure 8: Probability distribution of visual token attention.



(a) Option Tokens (b) "Yes" Tokens (c) "No" Tokens

Figure 9: Distribution of token uncertainty.

2. Non-hallucinated answers are given with higher confidence than hallucinated ones, regardless of *Yes* and *No*.
3. For MC questions, answers given by LLaVA V1.5 with higher confidence are very likely to be correct.

In conclusion, MLLM verb hallucination has something in common with MLLM object hallucination. **Uncertainty is strongly related to MLLM verb hallucination.**

mAP vs. Acc A natural question is, does low accuracy mean poor verb understanding? For example, in Tab. 1, LLaVA V1.5 shows an accuracy of 52.16 on YN questions with object references. To answer this question, we compare its performance with a strong baseline: HICO-finetuned CLIP (Radford et al. 2021) with outstanding results (Li et al. 2024b). To rule out the influence of objects, we test mAP under the "Known Object" setting (Chao et al. 2015). We use the probability of *Yes* tokens to calculate mAP. LLaVA V1.5 achieves an mAP of 68.41 while HICO-finetuned CLIP has an mAP of 60.45. The detailed AP gap of each HOI class is in Fig. 11. LLaVA V1.5 outperforms CLIP in most HOI classes. This means that LLaVA V1.5, although showing severe verb hallucination, does know well which verbs an image is more likely to contain. Therefore, one of the sources of verb hallucination is the **mis-calibration of tokens.**

Field of View and Uncertainty Previous research (Chen et al. 2024c; Zhang et al. 2025a) has revealed the use of field of view ensembling in hallucination mitigation. (Zhang et al. 2025a) has shown the influence on the change of token uncertainty. We also try to analyze the effect of field of view cropping on MLLM verb understanding. We select from HICO a set of images. For each image, there is only one HOI instance (to rule out the disturbance of multiple instances). The HOI instance is also small enough so that cropped images are substantially different from the original images. The token uncertainty on MC questions visualized in Fig. 10 shows that the uncertainty and accuracy do not change much: different from object, verb understanding is

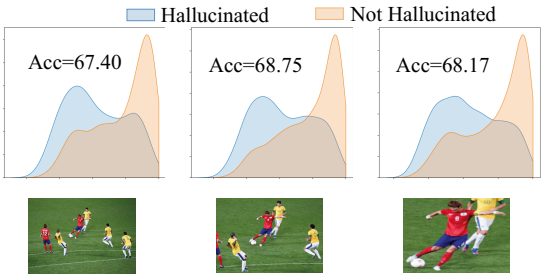


Figure 10: The influence of crop size on token uncertainty.

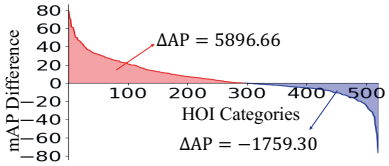


Figure 11: AP Comparison of LLaVA V1.5 and CLIP.

not substantially affected by the field of view: MLLMs **do not always know the verb concepts in the area**.

Hallucination Mitigation Methods

Training-Free Methods

OPERA (Huang et al. 2024), VCD (Leng et al. 2024), and Nullu (Yang et al. 2025) are outstanding training-free hallucination mitigation methods. They do not require fine-tuning and thus are low-cost and have more general applicability.

We present the results of OPERA in Tab. 5, showing the marginal or even negative effect of OPERA. Though OPERA tries to punish overreliance on summary tokens and force more attention on visual tokens, we have discussed in Sec. that more attention to visual tokens does not exclude hallucination. If the reward is given for attention to visual tokens, hallucination may be worsened.

VCD regards language prior as a hallucination-inducing factor, uses visual distortion to trigger it, and proposes contrastive decoding to mitigate hallucination. The results of VCD are in Tab. 5. VCD shows inconsistent effects on three models and thus fails on our benchmarks. To dig deeper, we compute Qwen-VL's KL divergence between the original token distribution $p_\theta(y_t|v, x, y_{<t})$ and contrasted token distribution $p_{vcd}(y_t|v, v', x, y_{<t}))$ of **20K** samples, and find that the KL divergence of **18.6K** samples is **0**, meaning that MLLMs **rely heavily on language prior to understand verbs, and such prior can not be easily omitted**.

Nullu identifies a low-rank hallucination subspace from truthful and hallucinated responses, projects model weights into its null space to suppress false priors, and reduces hallucination without high cost. Its negative results in Tab. 5 show that **MLLM layers have not formed reliable distinguishment between truthful and hallucinated verbs**.

Influence of Fine-tuning

REVERIE (Zhang et al. 2024), Haloquest (Wang et al. 2024b), and Octopus (Suo et al. 2025) are outstanding fine-tuning methods, and show decent results on object-centric benchmarks. Octopus train models to adopt different contrastive decoding methods in different cases, while the other two methods fine-tune models with meticulously designed training sets. However, these training sets do not contain much verb knowledge, and we show their limitation in Tab. 5. To mitigate verb hallucination to some extent without sacrificing MLLM's ability in other perspectives, we explore the effect of datasets with rich verb knowledge. We try to advance the MLLM with Pangea (Li et al. 2024b), which organizes existing heterogeneous action datasets in a unified way. It builds a mapping from action labels to abstract verb semantics. 290 frequent verb nodes in VerbNet (Schuler 2005) are selected and a one-to-290 mapping is built. It gives us a whole picture of diverse verbs and carries the structure knowledge of verb relationships. We select 60K samples from Pangea according to the proportion of the source dataset and build an instruction-tuning dataset. Details are in Suppl. Sec. 4. It contains **280 out of 290** nodes in Pangea P2S mapping and covers a wide range of verb semantics. Following common practice (Zhang et al. 2024; Wang et al. 2024b), we fine-tune LLaVA V1.5 with LoRA. The results are shown in Tab. 5. Although Pangea only contains **rough** action labels, it proves more helpful to verb hallucination mitigation than previous training sets. The performance of the original/fine-tuned model on MMMU (Yue et al. 2024): 32.2/33.0, MathVista (Lu et al. 2024): 23.6/24.5. The result is not seriously impaired. Our method may also be integrated into MoE in case of interference. In the future, mining more action data according to the structured verb semantics and fine-tuning MLLMs on them can be a promising way to mitigate verb hallucination.

	YN w/ obj		YN w/o obj		MC w/ obj	MC w/o obj
	acc	F1	acc	F1	acc	acc
LLaVA V1.5	52.16	57.69	59.16	61.79	57.37	51.00
OPERA	42.46	53.69	54.45	59.23	57.28	51.13
VCD	52.38	58.04	57.86	60.85	54.26	48.94
REVERIE	40.67	52.66	41.9	53.03	37.88	41.32
Haloquest	70.57	64.89	72.73	**63.00**	55.20	47.45
Nullu	51.99	57.90	59.22	61.78	55.98	53.17
Octopus	52.12	57.73	53.52	58.71	46.59	47.55
Ours	**78.48**	**68.13**	**77.37**	61.61	**61.73**	**60.79**

Table 5: Comparison on existing methods.

Conclusion

In this paper, to our best knowledge, we first reveal MLLM verb hallucination and build a benchmark to probe it from various perspectives. Our experiment reveals that MLLMs suffer from severe verb hallucination in many ways, and existing training-free hallucination mitigation methods fail. Fine-tuning is still the most promising way. However, how to fine-tune existing models efficiently is still a problem to be explored. Experiments show that MLLM verb hallucination is quite different from object hallucination. Moreover, whether there are effective training-free verb hallucination mitigation methods is still an open question.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62306175, the Shanghai Committee of Science and Technology, China (Grant No. 24511103200), the National Key Research and Development Project of China (No. 2022ZD0160102), and Shanghai Artificial Intelligence Laboratory, XPLORER PRIZE grants.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; et al. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Cai, R.; Song, Z.; Guan, D.; Chen, Z.; Luo, X.; Yi, C.; and Kot, A. 2024. BenchLMM: Benchmarking cross-style visual capability of large multimodal models. In *ECCV*.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Chao, Y.-W.; Wang, Z.; He, Y.; Wang, J.; and Deng, J. 2015. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Chen, X.; Ma, Z.; Zhang, X.; Xu, S.; Qian, S.; Yang, J.; Fouhey, D. F.; and Chai, J. 2024a. Multi-Object Hallucination in Vision-Language Models. In *NeurIPS*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*.

Chen, Z.; Zhao, Z.; Luo, H.; Yao, H.; Li, B.; and Zhou, J. 2024c. HALC: Object Hallucination Reduction via Adaptive Focal-Contrast Decoding. In *ICML*.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1): 37–46.

Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *NeurIPS*.

Deitke, M.; Clark, C.; Lee, S.; Tripathi, R.; Yang, Y.; Park, J. S.; Salehi, M.; et al. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *CVPR*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; Wu, Y.; and Ji, R. 2025. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. In *NeurIPS*.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Geirhos, R.; Narayanappa, K.; Mitzkus, B.; Thieringer, T.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2021. Partial success in closing the gap between human and machine vision. In *NeurIPS*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 6904–6913.

Gu, C.; Sun, C.; Ross, D. A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. 2024. HallusionBench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*.

Gunjal, A.; Yin, J.; and Bas, E. 2024. Detecting and preventing hallucinations in large vision language models. In *AAAI*.

Hu, H.; Zhang, J.; Zhao, M.; and Sun, Z. 2023. CIEM: Contrastive Instruction Evaluation Method for Better Instruction Tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12): 1–38.

Jiang, C.; Jia, H.; Dong, M.; Ye, W.; Xu, H.; Yan, M.; et al. 2024. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In *ACM MM*.

Kaul, P.; Li, Z.; Yang, H.; Dukler, Y.; Swaminathan, A.; Taylor, C.; and Soatto, S. 2024. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *CVPR*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123: 32–73.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*.

Li, B.; Ge, Y.; Ge, Y.; Wang, G.; Wang, R.; Zhang, R.; and Shan, Y. 2024a. SEED-Bench: Benchmarking Multimodal Large Language Models. In *CVPR*.

Li, H.; Li, N.; Chen, Y.; Zhu, J.; Guo, Q.; Lu, C.; and Li, Y.-L. 2025. The Labyrinth of Links: Navigating the Associative Maze of Multi-modal LLMs. In *ICLR*.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *EMNLP*.

Li, Y.-L.; Liu, X.; Lu, H.; Wang, S.; Liu, J.; Li, J.; and Lu, C. 2020a. Detailed 2D-3D Joint Representation for Human-Object Interaction. In *CVPR*.

Li, Y.-L.; Liu, X.; Wu, X.; Li, Y.; and Lu, C. 2020b. Hoi analysis: Integrating and decomposing human-object interaction. *NeurIPS*.

Li, Y.-L.; Liu, X.; Wu, X.; Li, Y.; Qiu, Z.; Xu, L.; Xu, Y.; Fang, H.-S.; and Lu, C. 2022. Hake: a knowledge engine foundation for human activity understanding. *IEEE T-PAMI*, 45(7): 8494–8506.

Li, Y.-L.; Wu, X.; Liu, X.; Wang, Z.; Dou, Y.; Ji, Y.; Zhang, J.; Li, Y.; Lu, X.; Tan, J.; and Lu, C. 2024b. From Isolated Islands to Pangea: Unifying Semantic Space for Human Action Understanding. In *CVPR*.

Li, Y.-L.; Xu, L.; Liu, X.; Huang, X.; Xu, Y.; Wang, S.; Fang, H.-S.; Ma, Z.; Chen, M.; and Lu, C. 2020c. PaStaNet: Toward Human Activity Knowledge Engine. In *CVPR*.

Li, Y.-L.; Zhou, S.; Huang, X.; Xu, L.; Ma, Z.; Fang, H.-S.; Wang, Y.; and Lu, C. 2019. Transferable Interactiveness Knowledge for Human-Object Interaction Detection. In *CVPR*.

Li, Z.; Yang, B.; Liu, Q.; Ma, Z.; Zhang, S.; Yang, J.; Sun, Y.; Liu, Y.; and Bai, X. 2024c. Monkey: Image resolution and text label are important things for large multi-modal models. In *CVPR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2024a. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. In *ICLR*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024b. Improved baselines with visual instruction tuning. In *CVPR*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *NeurIPS*.

Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024c. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2024d. Mmbench: Is your multi-modal model an all-around player? In *ECCV*.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; et al. 2024. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. In *ICLR*.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.

Ordonez, V.; Kulkarni, G.; and Berg, T. 2011. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object Hallucination in Image Captioning. In *EMNLP*.

Schuler, K. K. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Sigurdsson, G. A.; Gupta, A.; Schmid, C.; Farhadi, A.; and Alahari, K. 2018. Actor and observer: Joint modeling of first and third-person videos. In *CVPR*.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *CVPR*.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; et al. 2024. Aligning large multimodal models with factually augmented rlhf. In *ACL Findings*.

Suo, W.; Zhang, L.; Sun, M.; Wu, L. Y.; Wang, P.; and Zhang, Y. 2025. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *CVPR*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tong, S.; Liu, Z.; Zhai, Y.; Ma, Y.; LeCun, Y.; and Xie, S. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; et al. 2024a. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Z.; Bingham, G.; Yu, A.; Le, Q.; Luong, T.; and Ghiasi, G. 2024b. HaloQuest: A Visual Hallucination Dataset for Advancing Multimodal Reasoning. In *ECCV*.

Wu, M.; Ji, J.; Huang, O.; Li, J.; Wu, Y.; Sun, X.; and Ji, R. 2024. Evaluating and Analyzing Relationship Hallucinations in Large Vision-Language Models. In *ICML*.

Yang, L.; Zheng, Z.; Chen, B.; Zhao, Z.; Lin, C.; and Shen, C. 2025. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *CVPR*.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024a. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*.

Ye, W.; Zheng, G.; Ma, Y.; Cao, X.; Lai, B.; Rehg, J. M.; and Zhang, A. 2024b. MM-SpuBench: Towards Better Understanding of Spurious Biases in Multimodal LLMs. In *NeurIPS 2024 Workshop on RBFM*.

Yin, S.; Fu, C.; Zhao, S.; Xu, T.; Wang, H.; Sui, D.; Shen, Y.; Li, K.; Sun, X.; and Chen, E. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Sci. China Inf. Sci.*, 67(12).

Ying, K.; Meng, F.; Wang, J.; Li, Z.; Lin, H.; et al. 2024. MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI. In *ICML*.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2: 67–78.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*.

Zhai, B.; Yang, S.; Xu, C.; Shen, S.; Keutzer, K.; and Li, M. 2023. Halle-switch: Controlling object hallucination in large vision language models. *arXiv preprint arXiv:2310.01779*.

Zhang, J.; Khayatkhoei, M.; Chhikara, P.; and Ilievski, F. 2025a. MLLMs Know Where to Look: Training-free Perception of Small Visual Details with Multimodal LLMs. In *ICLR*.

Zhang, J.; Wang, T.; Zhang, H.; Lu, P.; and Zheng, F. 2024. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *ECCV*.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; et al. 2025b. Siren's song in the AI ocean: a survey on hallucination in large language models. *Computational Linguistics*.

Zhong, W.; Feng, X.; Zhao, L.; Li, Q.; Huang, L.; Gu, Y.; Ma, W.; Xu, Y.; and Qin, B. 2024. Investigating and Mitigating the Multimodal Hallucination Snowballing in Large Vision-Language Models. In *ACL*.

Zhou, Y.; Cui, C.; Yoon, J.; Zhang, L.; Deng, Z.; Finn, C.; Bansal, M.; and Yao, H. 2023. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. In *ICLR*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *ICLR*.