🔍 See Different, Think Better: Visual Variations Mitigating Hallucinations in LVLMs

Anonymous Author(s)
Submission Id: 1833

Abstract

Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities in visual understanding and multimodal reasoning. However, LVLMs frequently exhibit hallucination phenomena, manifesting as the generated textual responses that demonstrate inconsistencies with the provided visual content. Existing hallucination mitigation methods are predominantly text-centric, the challenges of visual-semantic alignment significantly limit their effectiveness, especially when confronted with fine-grained visual understanding scenarios. To this end, this paper presents **ViHallu**, a **Vi**sion-Centric **Hallu**cination mitigation framework that enhances visual-semantic alignment through **Visual Variation Image Generation** and **Visual Instruction Construction**. ViHallu introduces ***visual variation images*** with controllable visual alterations while maintaining the overall image structure. These images, combined with carefully constructed visual instructions, enable LVLMs to better understand fine-grained visual content through fine-tuning, allowing models to more precisely capture the correspondence between visual content and text, thereby enhancing visual-semantic alignment. Extensive experiments on multiple benchmarks show that ViHallu effectively enhances models' fine-grained visual understanding while significantly reducing hallucination tendencies. Furthermore, we release ViHallu-Instruction, a visual instruction dataset specifically designed for hallucination mitigation and visual-semantic alignment. Code and dataset will be released after acceptance.

CCS Concepts

• **Computing methodologies** → **Computer vision**; **Natural language processing**.

Keywords

Large Vision-Language Model, Hallucination, Visual-Semantic Alignment

ACM Reference Format:
Anonymous Author(s). 2025. 🔍 See Different, Think Better: Visual Variations Mitigating Hallucinations in LVLMs. In *Proceedings of Proceedings of the 33th ACM International Conference on Multimedia (MM '25)*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

MM '25, Dublin, Ireland

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

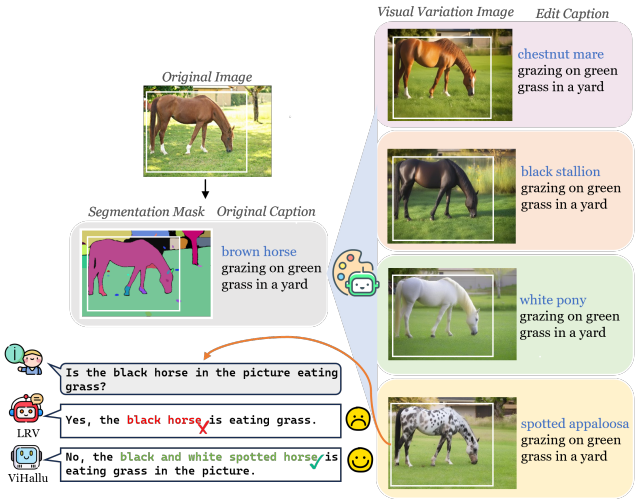

Figure 1: By inputting segmentation masks along with guiding caption into the controllable T2I model, the target visual variation image can be generated, maintaining structural similarity with the original image while exhibiting controlled local alterations.

Recently, Large Vision-Language Models (LVLMs) [14, 20, 21] have shown remarkable potential in tasks such as image description [14] and visual question answering [6, 21]. However, due to weak vision-text alignment, these models tend to generate hallucinations [16, 22, 45], manifesting as incorrectly identifying nonexistent objects, misattributing object properties, or misrepresenting spatial relationships. These issues not only diminish model reliability but also significantly constrain the practical application of LVLMs in critical domains such as medical diagnosis and autonomous driving.

The visual hallucination in LVLMs stems primarily from visual-semantic misalignment, where the model's textual outputs fail to accurately correspond to the visual elements present in the images. This cross-modal misalignment becomes particularly pronounced when models attempt to process fine-grained visual distinctions. For instance, models frequently struggle to differentiate between semantically similar scenes like *"a speckled black and white bird observing a chestnut brown horse eating grass in a meadow"* versus *"a chestnut brown bird eating grass while watching a speckled black and white horse standing in a meadow"*, despite the stark visual contrast these scenes represent. Such fundamental visual-semantic misalignments severely impair the visual understanding capability of LVLMs.

Previous research has primarily adopted text-centric approaches to mitigate hallucination in LVLMs, including improving the quality

of the dataset through noise and error elimination [41] and the incorporation of negative textual examples [12, 18, 44]. LRV-Instruction [18] proposed enhancing the robustness of LVLMs against hallucinations by incorporating additional negative instructions. HalluciDoctor [41] focused on reducing hallucination impacts by detecting and eliminating hallucinated text content in visual instruction data. However, these text-centric strategies demonstrate limited efficacy in addressing fine-grained visual distinctions and do not sufficiently enhance the model's visual comprehension capabilities. As illustrated in Figure 1, when presented with a black-and-white spotted appaloosa horse, LRV misinterprets the prompt to indicate that the image contains a black horse. Consequently, developing vision-centric approaches to enhance visual-semantic alignment may be promising for addressing these visual hallucination.

To strengthen visual-semantic alignment in LVLMs, we introduce ***visual variation samples***, which exhibit fine-grained visual differences (such as scene elements, object categories, and attributes, etc.), as shown in the right side of Figure 1. **By training models to recognize these fine-grained visual distinctions, we enable models to more precisely capture the correspondence between visual content and text, thereby enhancing visual-semantic alignment**. However, such visual variation samples rarely occur in natural datasets. This raises two critical questions: How can we effectively generate high-quality visual variation samples? And how can we leverage these samples to enhance the visual-semantic alignment of LVLMs?

Motivated by recent advances in Text-to-Image (T2I) generation [8, 15, 24, 27, 43], we present **ViHallu**, a **Vi**sion-Centric **Hallu**cination mitigation framework that enhances visual-semantic alignment through **Visual Variation Image Generation** and **Visual Instruction Construction**. At its core, ViHallu introduces a novel generation approach that combines text guidance with segmentation mask control to produce high-quality visual variation images that conform to specified captions and preserve the global structure of original images. The generated samples differ from their original counterparts only in regions corresponding to the modified textual caption. As shown in Figure 1, when replacing *"brown horse"* with *"chestnut mare"* in the image caption, only the target region transforms while maintaining structural consistency with the original image. **In the visual variation image generation process, we place objects in uncommon contextual settings where they rarely appear, creating counterfactual interventions [26]. Fine-tuning on these samples helps models rely on visual features rather than contextual expectations, reducing misleading correlations between frequently co-occurring objects. This improves the model's ability to make judgments based on visual evidence instead of statistical associations.** To leverage these samples effectively, we construct tailored visual instructions for paired original and variation images, enabling LVLMs to learn fine-grained discriminative features and enhance visual-semantic alignment through high-quality image-instruction pairs.

Our main contributions are summarized as follows:

- The **ViHallu** propose a novel visual variation image generation approach that integrates text guidance and segmentation mask control, producing samples that exhibit controlled visual alterations. while maintaining the overall image structure.
- To the best of our knowledge, **ViHallu** is the first to construct tailored instruction data for visual variation images, establishing a novel visual instruction data construction paradigm.
- We release **ViHallu-Instruction**, a visual instruction dataset specifically designed for fine-grained visual-semantic alignment. This dataset includes carefully curated visual variation images and high-quality instruction data to facilitate research in hallucination mitigation for LVLMs.
- Extensive empirical evaluations show that **ViHallu** significantly mitigates hallucination in LVLMs and enhances model robustness.

2 Related Work

2.1 Large Vision-Language Model

Large Language Models (LLMs) [2, 25, 33] have demonstrated remarkable performance across various Natural Language Processing (NLP) tasks. Similarly, LVLMs, which leverage the capabilities of LLMs, have undergone rapid development and shown significant advantages in multimodal tasks. The training of LVLMs typically comprises two critical phases: pre-training and visual instruction tuning. In the initial pre-training phase, the primary objective is to achieve vision-language alignment through large-scale image-text pair datasets. In the visual instruction tuning phase, the core objective is to enhance the model's ability to comprehend user instructions and complete specific tasks. Numerous visual instruction datasets are generated through the collaboration between LVLMs and LLMs. For instance, LLaVA [21] converts images into text descriptions with bounding box information and utilizes text-only GPT-4 [25] to generate new text data. Recently, with the emergence of the more powerful multimodal model GPT-4V [1], many works have adopted GPT-4V to generate data of higher quality, as exemplified by LVIS-Instruct4V [35] and ALLaVA [3].

2.2 LVLMs Hallucination

The hallucination problem in LVLMs has attracted much attention. Prior studies [12, 18, 36, 36, 41] have predominantly adopted text-centric approaches to mitigate hallucination in LVLMs. For instance, HACL [12] mitigates hallucination in LVLMs by introducing hallucinatory captions as hard negative text samples. HA-DPO [44] reframes hallucination mitigation as a preference optimization task by introducing hallucination-aware textual response pairs, training models to discriminate between hallucinated text and faithful text for the same visual input. However, these text-based approaches primarily focus on strengthening text generation capabilities while lacking enhancement of the model's visual understanding capabilities. This is particularly problematic when LVLMs need to distinguish subtle visual features. Several studies [11, 29, 31, 34] have demonstrated that incorporating model-generated negative images can enhance LVLMs multimodal compositional reasoning capabilities, suggesting the importance of the visual modality in improving model performance. Inspired by these advances and recognizing the limitations of purely text-centric approaches, we propose ViHallu,

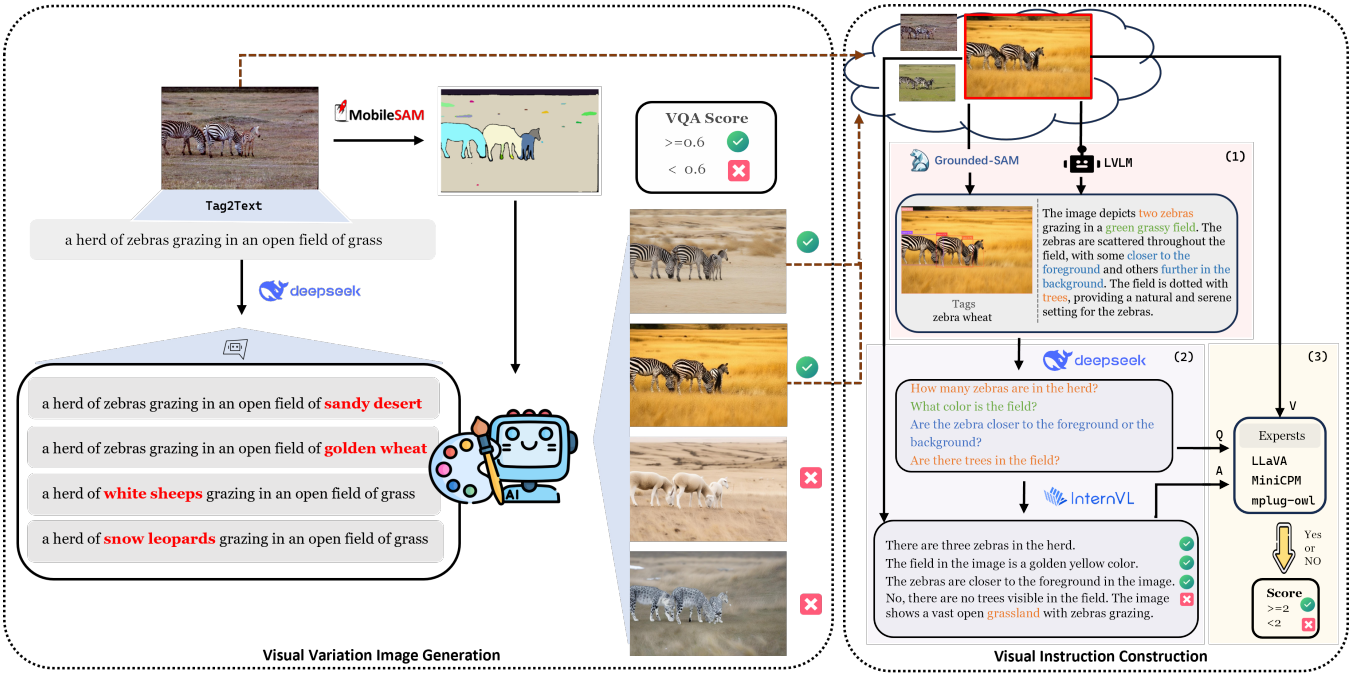

Figure 2: Overview of our framework ViHallu. The left shows the visual variation image generation process: (1) original image caption and segmentation mask generation, (2) caption editing through concept substitution, and (3) visual variation image generation with quality assessment. The right shows the instruction construction process: (1) detailed caption and object tag generation, (2) QA pair generation, and (3) QA pair quality assessment. Different types of hallucinations are indicated by distinct colors: Object Hallucination, Attribute Hallucination, and Relation Hallucination

.

a flexible framework that strengthens visual-semantic alignment through the generation of visual variation image and visual instruction data. Unlike previous work that primarily focuses on textual domain, ViHallu emphasizes the visual domain by generating high-quality visual variation images.

3 Methodology

This section delineates two phases of ViHallu. Visual variation image generation leverages advanced LLM and controllable T2I model to produce high-quality visual variation images. Visual instruction generation constructs instruction data based on the images.

3.1 Visual Variation Image Generation

This section introduces the visual variation image generation phase, which encompasses three primary modules: 1) original image caption generation and segmentation mask extraction, 2) caption editing through concept substitution, and 3) visual variation image generation and quality evaluation.

Image Caption and Segmentation Mask. To obtain image captions and segmentation masks from the original images, we employ Tag2Text [10] and MobileSAM [42] models. Tag2Text model extracts image tags and employs them to guide VLM in generating comprehensive image captions. MobileSAM demonstrates excellent zero-shot transfer capabilities, enabling precise object segmentation within images. By inputting the original image into MobileSAM, we obtain the segmentation mask image as shown in Figure 2, providing the foundation for visual variation image generation.

Caption Editing. To facilitate precise modifications in image captions for visual variation image generation, we propose a LLM-based caption editing mechanism. The editing mechanism generates variant captions that differ from the original captions exclusively in specified target object categories or attributes, while maintaining all other aspects of the original caption unchanged. The target object can encompass both individual object and background scene. To guide the model in generating new captions that meet these requirements, we design a specialized prompt template, as shown in Figure 3. In the implementation, considering comprehensive performance, we utilize the DeepSeek-chat V2 model [7] to execute the caption editing task. The primary purpose of this mechanism is to provide controlled inputs for generating visual variation images by precisely specifying which elements should be altered while preserving other contextual information. **Notably, this mechanism can integrate objects into tail scenes where they rarely appear, introducing counterfactual interventions**, as exemplified by the first caption on the left of Figure 2 *"a herd of zebras grazing in an open field of sandy desert"*. Through such a caption, our T2I model produces corresponding images. During fine-tuning, this approach of combining rarely co-occurring scenes and objects helps reduce spurious correlations between frequently co-occurring objects [41].

Replace one word in the given list with more detailed and visually relevant alternatives,and put the enriched discription to replace the object in the provided sentence.please notice that only replace one object.
Return only the updated sentences, each separated by a new line.

Example:
Words:
bird
Sentence:
A bird is sitting on a tree.
Answer:
A majestic eagle is sitting on a tree.
A colorful parrot is sitting on a tree.
A small sparrow is sitting on a tree.
A large hawk is sitting on a tree.

Input:
Words:
grass
Sentence:
A herd of zebras grazing in an open field of **grass**

Output:
A herd of zebras grazing in an open field of **green meadow**
A herd of zebras grazing in an open field of **golden wheat**
A herd of zebras grazing in an open field of **brown tundra**
A herd of zebras grazing in an open field of **sandy desert**

Figure 3: Illustration of caption editing prompt design. The prompt gives a example of caption editing and takes a target word and an original caption as input, guiding the model to generate enriched variations by replacing the target object with more detailed alternatives. The example demonstrates how a basic reference "grass" is transformed into diverse, attribute-rich descriptions (such as "green meadow" and "golden wheat").

By pairing objects with unexpected scenes to create semantic inconsistencies, the model is forced to rely on visual evidence rather than statistical priors during the inference process.

Visual variation Image Generation. Through a text-guided and segmentation-mask-controlled generation model, target images are produced that maintain structural consistency while featuring controlled local variations. Utilizing previously obtained segmentation masks and captions as dual inputs: the masks ensure structural layout and object positioning in target images, while the target caption (differing only in specific object categories and attributes) guides content generation. We employ ControlNet++ [27] as the image generation model. Through the dual input of segmentation masks and target captions, this model accurately generates the desired images while maintaining natural visual coherence, with modified regions seamlessly integrated without artifacts such as edge discontinuities or composition errors. To ensure the quality of generated target images, we implement quality assessment using the VQAScore metric [17] , which evaluates image-text alignment through the Visual Question Answering (VQA) model. A quality threshold is established by filtering out generated images with scores below 0.6, retaining only high-scoring images to maintain the generation quality standard. The filtered target images, together with the original images, constitute the image set.

3.2 Visual Instruction Generation

This section presents the high-quality visual instruction generation phase, which comprises three modules: 1) detailed description and object tag generation, 2) Question-Answer generation, and 3) quality assessment.

Detailed Description and Object Tag Generation. To construct fine-grained visual instruction data, we leverage LVLM's image-to-text capabilities to produce detailed image descriptions. Due to LVLM's inherent hallucination tendencies, the generated descriptions may contain hallucinations, as shown on the right of Figure 2, different types of hallucinations are indicated by distinct color. And the generated descriptions may not comprehensively cover all objects in the image, particularly in complex scenes where descriptions tend to focus on primary objects. To ensure that questions cover all objects, we employ Grounded-SAM [13, 23, 30] for object detection and segmentation, extracting object tags to guarantee comprehensive coverage of all objects in the image.

Question-Answer Generation. In the question generation phase, we leverage DeepSeek-chat V2 to process detailed descriptions and object tags to generate questions. Detailed descriptions provide contextual information and the presence of hallucinations in LVLM-generated descriptions is helpful for the detection and removal of hallucinations. **When the LLM processes a description containing hallucinations, it naturally exhibits a tendency to generate questions about these hallucinated elements**. As shown on the right of Figure 2, when a description incorrectly states there are trees in the image, this triggers questions such as *"Are there trees in the field? "*. Object tags ensure comprehensive coverage by identifying all objects in the images, including those absent from descriptions. The DeepSeek-chat V2 model generates questions based on these inputs, targeting areas where LVLMs are prone to hallucinate. To generate detailed and accurate answer for each question, we utilize InternVL2.5 [5] as the answer generation model, which responds to generated questions based on the actual image content rather than the potentially hallucinated description. Detailed prompt templates are provided in the Appendix.

Quality Assessment. To evaluate the quality of Question-Answer (QA) pairs, we employ a panel of LVLM experts (*e.g.*, LLaVA-1.5 [19], MiniCPM-V 2.6 [38], and mPLUG-OWL3 [39]). This multi-model consensus approach mitigates biases that might be present in individual models and enhances the robustness of our quality assessment process. Each QA pair and corresponding image are inputted to these expert models, which assess the alignment between the answers and the image content, producing binary outputs: *"Yes"* for aligned content and *"No"* for misaligned content. The experts models evaluate whether the answer accurately reflects what is visible in the image, checking for both factual correctness and comprehensiveness. We retain QA pairs endorsed by at least two expert models and discard those rejected by two or more experts. This majority voting mechanism ensures that only high-quality QA pairs are included in the final dataset. More details on the quality assessment methodology are provided in the Appendix.

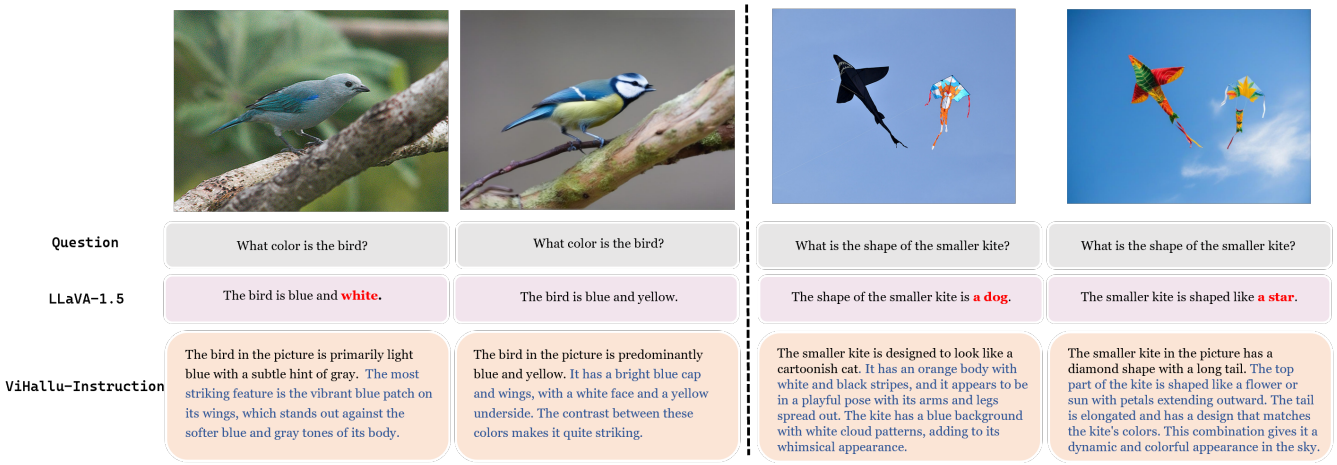

Figure 4: Samples from the ViHallu-Instruction dataset. The top row displays two pairs of images: original images (left) and their corresponding visual variation images (right). Below are exemplar QA pairs from ViHallu-Instruction alongside responses from the baseline LLaVA-1.5 model. For clarity, incorrect responses are highlighted in red, while detailed responses are emphasized in blue.

4 Experiments

In this section, we present the ViHallu-Instruction dataset construction, baseline models, evaluation metrics, and implementation details.

4.1 Implementation Setup

The following describes the experimental setup, including dataset construction and baseline models.

Dataset Construction. The ViHallu-Instruction dataset consists of 6,770 images paired with approximately 50k (±10k) tailored instructions for the baseline models. Representative samples from ViHallu-Instruction are shown in Figure 4. Through the generation of visual variation images, 7,209 images are created and evaluated using LLaVA-1.5-13B for VQA scoring. This quality assessment yields 5,051 high-quality visual variation images, which are combined with their 1,719 corresponding original images to form the final dataset of 6,770 images. The visual instruction generation process relies on detailed image descriptions generated by the baseline model, specifically targeting hallucinations in model-generated descriptions to facilitate hallucination correction during fine-tuning. For question generation, the DeepSeek-chat V2 model extracts seven questions per image from the descriptions. To enhance the model's ability to distinguish between original and visual variation samples, a subset of questions generated from original samples is applied to visual variation images. As shown in Figure 4, the visual differences between the original and visual variation samples naturally lead to distinct answers for the same questions. During fine-tuning, these different responses help the model better identify distinctions between visual variation and original samples, thus mitigating fine-grained hallucinations. Answer generation utilizes InternVL-2.5-38B to produce responses based on image content. Quality assessment by three expert models filters out approximately 20% of the answers. More construction details and samples of ViHallu-Instruction are provided in the Appendix.

Model Baselines . For comprehensive evaluation, we employ three representative open-source LVLMs: (1) LLaVA 1.5 (7B) [19]; (2) MiniGPT4 v2 [4]; (3) Qwen2-VL (7B) [37].

4.2 Implementation Details

For fine-tuning, we construct training data using ViHallu, which maintains the same image content while incorporating model-specific visual instruction data tailored for each LVLM.

LLaVA-1.5. For LLaVA-1.5-7B, the full parameter fine-tuning is carried out with a batch size per device of 16. The implementation utilizes a cosine annealing learning rate scheduler with an initial learning rate of 2e-5.

MiniGPT-4 v2. The MiniGPT-4 v2 is fine-tuned with LoRA, with the rank set to 64 and α to 16, utilizing linear projection layers as the sole trainable modules. The optimization process employs a linear warmup with cosine annealing learning rate schedule at 1e-5.

Qwen2-VL. The Qwen2-VL model is fine-tuned using full-parameter tuning while keeping the vision tower frozen. Only the multimodal projection layers are trained, with a batch size of 4 and gradient accumulation steps of 1. The optimization process employs a cosine annealing learning rate schedule at 1e-5 with a 0.1 warmup ratio.

4.3 Evaluation Metric

The following presents the evaluation metrics employed in experiments.

POPE[40], the Polling-based Object Probing Evaluation, evaluates object hallucination in LVLMs about object presence in images through balanced Yes/No questions (i.e., 50% vs. 50%). The benchmark implements three progressively challenging negative sampling strategies: random selection from arbitrary objects, popular selection from high-frequency objects, and adversarial selection from scene-relevant co-occurring objects. Based on MSCOCO,

Table 1: Results of LVLMs on *random, popular,* and *adversarial* settings of POPE dataset. The best results are shown in bold.

Setting	Model	Accuracy (%)	F1 Score (%)
Random	LLaVA-1.5	87.90	87.35
	w/ViHallu	**90.13**	**89.59**
	MiniGPT4 V2	82.30	84.07
	w/ViHallu	**85.23**	**85.66**
	Qwen2-VL	90.67	89.92
	w/ViHallu	**91.07**	**90.44**
Popular	LLaVA-1.5	87.23	86.37
	w/ViHallu	**88.40**	**87.98**
	MiniGPT4 V2	79.97	82.30
	w/ViHallu	**85.20**	**85.59**
	Qwen2-VL	89.33	88.64
	w/ViHallu	**89.40**	**88.86**
Adversarial	LLaVA-1.5	84.43	84.28
	w/ViHallu	**84.97**	**84.96**
	MiniGPT4 V2	73.07	77.61
	w/ViHallu	**75.77**	**78.43**
	Qwen2-VL	**87.27**	**86.74**
	w/ViHallu	86.93	86.61

A-OKVQA, and GQA datasets, POPE comprises 27,000 question-answer pairs, with performance evaluated using Accuracy, Precision, Recall, and F1 Score metrics.

LLaVA-Bench[21] is an evaluation suite designed to assess LVLMs' capabilities across diverse visual tasks. The benchmark includes 24 images (including indoor and outdoor scenes, memes, paintings, sketches, etc.) with 60 questions spanning three categories: conversation (simple QA), detailed description, and complex reasoning. This benchmark evaluates models' ability to interpret various visual content in challenging tasks and generalizability to novel domains.

MMHal-Bench[32] is a specialized benchmark dataset designed to evaluate hallucination phenomena in MLLMs. The dataset comprises 96 image-question pairs, ranging in 8 question categories × 12 object topics.The eight question categories cover various types of hallucination, including object attributes, counting, spatial relations, etc. By leveraging the GPT-4o model for response analysis and scoring, MMHal-Bench provides a rigorous and practical evaluation framework for assessing the accuracy and reliability of LVLMs in real-world applications.

5 Results

This section presents a comprehensive evaluation of ViHallu through extensive experiments. We compare models fine-tuned with ViHallu against both their original baseline and other approaches for hallucination mitigation, demonstrating the effectiveness of our method across multiple benchmark tasks. Furthermore, additional analyses validate the reliability of our approach.

Table 2: Results of hallucination mitigation methods on POPE dataset. We report the average accuracy and F1 score of POPE. The best results are shown in bold.

Method	Accuracy (%)	F1 Score (%)
LLaVA-1.5 [19]	86.52	86.00
LLaVA-1.5$_{w/HA-DPO}$[44]	86.63	86.87
LLaVA-1.5$_{w/HACL}$[12]	87.69	87.26
LLaVA-1.5$_{w/VH}$[9]	/	84.80
LLaVA-1.5$_{w/RAR}$[28]	87.16	86.43
LLaVA-1.5$_{w/ViHallu(Ours)}$	**87.83**	**87.51**
MiniGPT-4 v2 [4]	78.45	81.33
MiniGPT-4 v2$_{w/ViHallu(Ours)}$	**82.07**	**83.23**
Qwen2-VL [37]	89.09	88.43
Qwen2-VL$_{w/ViHallu(Ours)}$	**89.13**	**88.63**

5.1 Evaluation on Benchmarks

Results on POPE. Experimental results on POPE under the *random, popular,* and *adversarial* settings are summarized in Table 1. Fine-tuning with ViHallu demonstrates significant reduction in hallucination for MiniGPT-4 v2, achieving accuracy improvements of 2.93%, 5.23%, and 2.70% on *random, popular,* and *adversarial* sets, respectively, accompanied by substantial improvements in F1 scores across all settings. For LLaVA-1.5, which exhibits relatively fewer hallucinations in its baseline version due to visual instruction fine-tuning during initial training, fine-tuning with ViHallu yields further performance improvements. The model achieves accuracy improvements of 2.23%, 1.17%, and 0.54% on *random, popular,* and *adversarial* sets respectively, along with enhanced F1 scores. Qwen2-VL exhibits the strongest performance among all baseline models, with initial accuracy scores of 90.67%, 89.33%, and 87.27% on *random, popular,* and *adversarial* settings respectively. After fine-tuning with ViHallu, Qwen2-VL shows further improvements in the *random* and *popular* settings, with accuracy increasing by 0.40% and 0.07% respectively. Corresponding F1 scores also improve by 0.52% and 0.22%. In the *adversarial* setting, Qwen2-VL shows a slight decrease in performance, with accuracy decreasing by 0.34% and F1 score by 0.13%. This small decrease can be attributed to minor bias effects that emerge during fine-tuning. Despite this decrease, Qwen2-VL's performance after fine-tuning still significantly outperforms both baseline and fine-tuned versions of the other models in all settings, demonstrating its exceptional robustness against object hallucinations.

The performance improvements across models on POPE after fine-tuning with ViHallu demonstrate enhanced detection of object hallucinations. Particularly noteworthy is the improvement in the *adversarial* setting for MiniGPT-4 v2 and LLaVA-1.5, where negative samples are drawn based on co-occurrence frequencies with true objects, which is traditionally the most challenging scenario for hallucination detection. The improvements in this setting validate our hypothesis that **ViHallu effectively mitigates hallucination by introducing counterfactual samples to reduce spurious correlations between frequently co-occurring objects**. The

Table 3: Results of LVLMs on LLaVA-Bench (In-the-Wild) dataset. The best results are shown in bold.

Model	Conversation	Detail description	Complex reasoning	All
LLaVA-1.5	58.3	49.6	69.5	59.3
LLaVA-1.5$_{w/ViHallu}$	**62.3**	**57.0**	**77.0**	**67.7**
MiniGPT4 V2	43.5	42.8	64.6	52.8
MiniGPT4 V2$_{w/ViHallu}$	**53.9**	**48.2**	**73.8**	**61.5**
Qwen2-VL	68.5	69.0	70.6	69.6
Qwen2-VL$_{w/ViHallu}$	**87.5**	**76.0**	**86.3**	**84.0**

Table 4: Results of LVLMs on MMHal-Bench dataset. The best results are shown in bold.

Model	Overall Score↑	Hallucination Rate↓
LLaVA-1.5	1.93	0.70
LLaVA-1.5$_{w/ViHallu}$	**2.11**	**0.66**
MiniGPT4 V2	1.54	0.76
MiniGPT4 V2$_{w/ViHallu}$	**1.91**	**0.70**
Qwen2-VL	2.94	0.51
Qwen2-VL$_{w/ViHallu}$	**3.64**	**0.48**

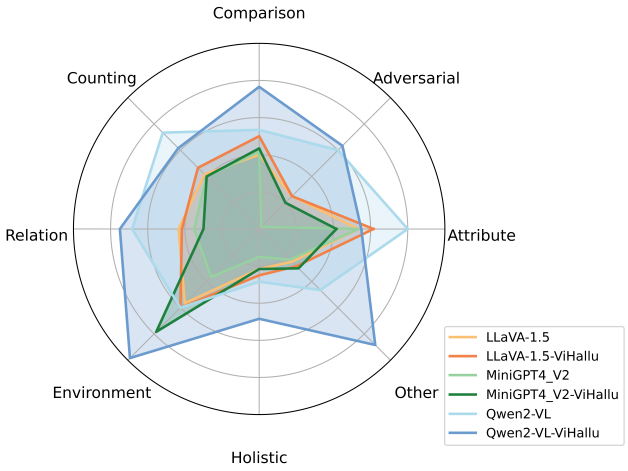

Figure 5: Performance comparison of different models before and after ViHallu fine-tuning across eight hallucination categories on MMHal-Bench.

slight decrease for Qwen2-VL in this setting suggests that it already has a strong capability to identify counterfactual examples, making the introduction of such samples largely ineffective.

Additionally, we compare ViHallu with other hallucination mitigation methods, with all approaches based on LLaVA-1.5-7B to ensure fair comparison, as shown in Table 2. The results show that ViHallu outperforms other methods, achieving the highest accuracy and F1 score on the POPE benchmark.

Results on LLaVA-Bench. To assess the impact of ViHallu on models' comprehensive capabilities, we conduct evaluations using LLaVA-Bench, a non-hallucination-focused benchmark. As shown in Table 3, the models demonstrate improved overall performance after fine-tuning with ViHallu. Specifically, MiniGPT-4 V2 exhibits performance improvements of 10.4, 5.4, and 9.2 points in *conversation*, *detailed description*, and *complex reasoning categories*, respectively. LLaVA-1.5 also shows enhancements of 4.0, 7.4 and 7.5 points. Most notably, Qwen2-VL shows the most substantial improvements with gains of 19.0, 7.0, and 15.7 points across the three evaluation categories, resulting in an impressive 14.4 point increase in overall performance. The results demonstrate that ViHallu effectively enhances models' conversational capabilities and fine-grained understanding through fine-tuning, achieved by incorporating visual variation samples and detailed instructions. These consistent improvements across all baseline models confirm that the ViHallu successfully strengthens visual-semantic alignment, enhancing models' ability to accurately interpret and reason about visual content.

Results on MMHal-Bench. Experimental results on MMHal-Bench are summarized in Table 4. For LLaVA-1.5, the application of ViHallu improves the overall score from 1.93 to 2.11, representing a 9.3% increase in performance. Simultaneously, its hallucination rate decreases from 0.70 to 0.66, achieving a 5.7% reduction in erroneous outputs. MiniGPT4 V2 demonstrates even more substantial improvements with ViHallu enhancement, with its overall score rising from 1.54 to 1.91, along with a significant reduction in hallucination rate from 0.76 to 0.70. Qwen2-VL shows the most impressive performance both before and after enhancement. Its baseline score of 2.94 already outperforms other models, and after applying ViHallu, this score increases to 3.64. Similarly, its already low hallucination rate further decreases from 0.51 to 0.48, representing a 5.9% reduction. The Figure 5 provides additional insights into model performance across eight hallucination categories. All models with ViHallu enhancement display expanded polygons in the radar chart, indicating improvements across most evaluation dimensions, while all enhanced models show notable improvements in previously challenging areas like *Adversarial* and *Environment* categories. These visualization results further confirm that ViHallu effectively enhances multimodal understanding and reduces various types of hallucinations across different model architectures.

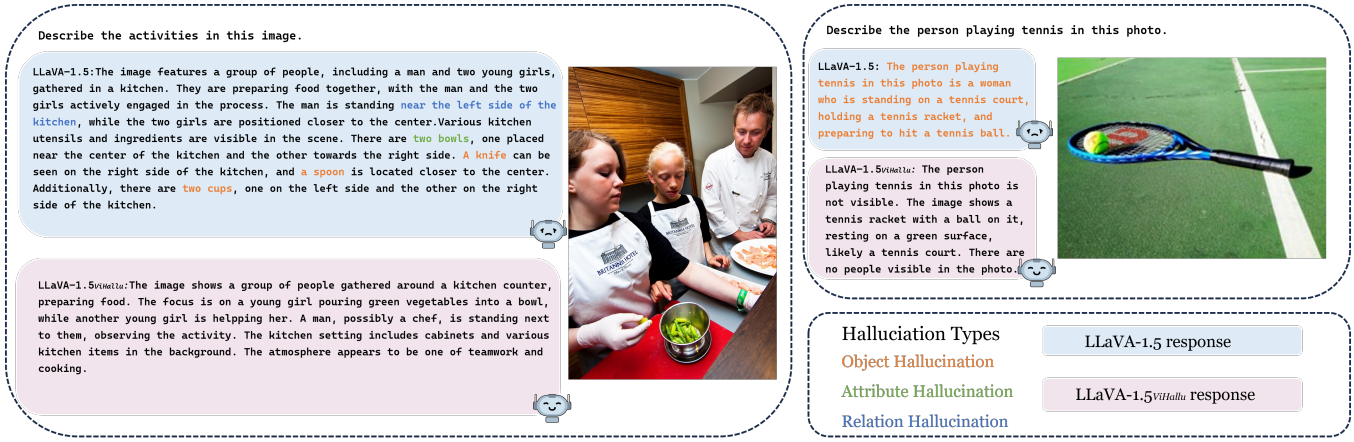

Figure 6: Response comparison between baseline LLaVA-1.5 and LLaVA-1.5$_{ViHallu}$ (LLaVA-1.5 fine-tuned with ViHallu) on two samples from MMHal-Bench dataset.

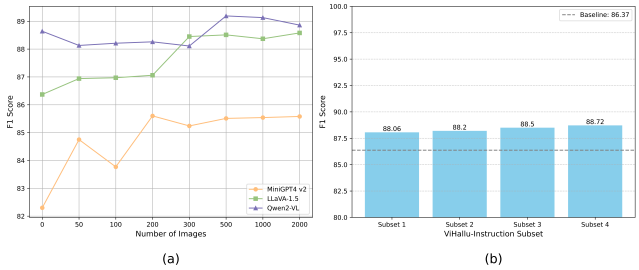

Figure 7: (a) F1 score on POPE benchmark of LVLMs fine-tuned with varying sizes of ViHallu-Instruction subsets. (b) F1 score on POPE benchmark of LLaVA-1.5 fine-tuned on four independent subsets of ViHallu-Instruction.

5.2 Further Analysis

Case Study. Figure 6 presents a comparison of responses between the baseline LLaVA-1.5 and LLaVA-1.5$_{ViHallu}$ fine-tuned with ViHallu. In the left example, the LLaVA-1.5 hallucinates multiple non-existent objects (such as cups, spoons, and knives) and makes spatial relationship errors (describing the man on the right as being on the left). In contrast, LLaVA-1.5$_{ViHallu}$ produces concise, hallucination-free response while effectively preventing interference from frequently co-occurring objects. In the right example, when asked about a non-existent *"person"* in the image, the LLaVA-1.5 generates completely inaccurate responses, incorrectly describing a tennis-playing scene. However, LLaVA-1.5$_{ViHallu}$ successfully identifies the absence of people and accurately notes the specific detail of the ball on the racket. **These results demonstrate that ViHallu effectively reduces the model's reliance on language priors and significantly mitigates hallucination phenomena**. Further comparative examples illustrating model responses before and after fine-tuning with ViHallu can be found in the Appendix.

Impact of the dataset size. To evaluate the effect of fine-tuning dataset size on model performance, varying subsets are sampled from ViHallu-Instruction, ranging from 50 to 2,000 images with their corresponding instructions (averaging 7-8 instructions per image). The LVLMs are fine-tuned using these sampled datasets and evaluated on the POPE benchmark. As shown in Figure 7 (a), all LVLMs exhibit a general upward trend in performance that gradually plateaus as the amount of fine-tuning data increases. **This trend demonstrates that ViHallu effectively mitigates hallucination in LVLMs with increasing dataset size**.

Consistency analysis with different subsets. To assess the consistency of ViHallu-Instruction, four independent subsets are extracted from the dataset, each comprising 4,201 instructions paired with approximately 500 distinct images. Using these subsets for fine-tuning LLaVA-1.5, evaluations on the POPE benchmark reveal consistent performance improvements. As shown in Figure 7 (b), all fine-tuned models demonstrate substantial improvements over the baseline, with variations in the F1 score within 0.7 between subsets. This minimal variance suggests **the robustness of ViHallu-Instruction across different dataset samples**.

6 Conclusion

In this paper, we present ViHallu, a vision-centric hallucination mitigation framework that enhances visual-semantic alignment through visual variation image generation and visual instruction construction. ViHallu addresses a critical limitation in existing hallucination mitigation methods by introducing visual variation samples focuses on fine-grained visual differences. Extensive experiments across multiple benchmarks and LVLMs demonstrate that ViHallu effectively enhances models' fine-grained visual understanding while significantly reducing hallucination tendencies. Furthermore, ViHallu-Instruction dataset provides a valuable resource for future research in visual-semantic alignment and hallucination mitigation.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684* (2024).

[4] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478* (2023).

[5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271* (2024).

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500 [cs.CV] https://arxiv.org/abs/2305.06500

[7] DeepSeek-AI. 2024. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. arXiv:2405.04434 [cs.CL]

[8] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778* (2023).

[9] Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024. Visual hallucinations of multi-modal large language models. *arXiv preprint arXiv:2402.14683* (2024).

[10] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. 2023. Tag2Text: Guiding Vision-Language Model via Image Tagging. *arXiv preprint arXiv:2303.05657* (2023).

[11] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. 2022. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2785–2795.

[12] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27036–27046.

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *arXiv:2304.02643* (2023).

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[15] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. 2025. ControlNet ++: Improving Conditional Controls with Efficient Consistency Feedback. In *European Conference on Computer Vision*. Springer, 129–147.

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

[17] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2025. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*. Springer, 366–384.

[18] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved Baselines with Visual Instruction Tuning.

[20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26296–26306.

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

[22] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253* (2024).

[23] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).

[24] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4296–4304.

[25] OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt

[26] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. 2020. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems* 33 (2020), 857–869.

[27] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. 2023. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147* (2023).

[28] Xiaoye Qu, Qiyuan Chen, Wei Wei, Jishuo Sun, and Jianfeng Dong. 2024. Alleviating hallucination in large vision-language models with active retrieval augmentation. *arXiv preprint arXiv:2408.00555* (2024).

[29] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. 2023. Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6967–6977.

[30] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:2401.14159 [cs.CV]

[31] Ugur Sahin, Hang Li, Qadeer Khan, Daniel Cremers, and Volker Tresp. 2024. Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5563–5573.

[32] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning Large Multimodal Models with Factually Augmented RLHF. (2023).

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

[34] Haonan Wang, Minbin Huang, Runhui Huang, Lanqing Hong, Hang Xu, Tianyang Hu, Xiaodan Liang, Zhenguo Li, Hong Cheng, and Kenji Kawaguchi. 2023. Boosting visual-language models by exploiting hard samples. *arXiv preprint arXiv:2305.05208* (2023).

[35] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574* (2023).

[36] Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*. Springer, 32–45.

[37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).

[38] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800* (2024).

[39] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models. arXiv:2408.04840 [cs.CV] https://arxiv.org/abs/2408.04840

[40] Kun Zhou Jinpeng Wang Wayne Xin Zhao Yifan Li, Yifan Du and Ji-Rong Wen. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. https://openreview.net/forum?id=xozJw0kZXF

[41] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2024. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12944–12953.

[42] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. 2023. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289* (2023).

[43] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

[44] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839* (2023).

[45] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754* (2023).