Evaluating and Mitigating Object Hallucination in Large Vision-Language Models: Can They Still See Removed Objects?

Yixiao He, Haifeng Sun, Pengfei Ren, Jingyu Wang, Huazheng Wang, Qi Qi, Zirui Zhuang†, Jing Wang†
State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications
{heyixiao,hfsun,rpf,wangjingyu,wanghz,qiqi8266}@bupt.edu.cn
{zhuangzirui,wangjing}@bupt.edu.cn

Abstract

Large Vision-Language Models (LVLMs) have a significant issue with object hallucinations, where researchers have noted that LVLMs often mistakenly determine objects as present in images where they do not actually exist. Some recent studies evaluate the occurrence of object hallucinations by asking LVLMs whether they see objects that do not exist in input images. However, we observe that these evaluation methods have some limitations, such as the objects being questioned potentially having little relevance to the image. In this paper, we introduce a more challenging benchmark for evaluating object hallucinations by removing objects from images and then asking the model whether it can still see the removed objects. Our evaluation result reveals that LVLMs suffer from severe hallucinations, as they often still claim to see the removed objects. Through our analysis, we find that biases in training result in LVLMs lacking guidance on learning about the absence of objects, which in turn leads to a lack of ability to determine that objects do not exist in images. To address this issue, we further propose oDPO, a direct preference optimization objective based on visual objects. By guiding LVLMs to learn to determine the existence of objects, oDPO effectively alleviates object hallucinations. It achieves more competitive results than other hallucination mitigation approaches across multiple object hallucination benchmarks and enhances the performance of LVLMs in various vision-language tasks.

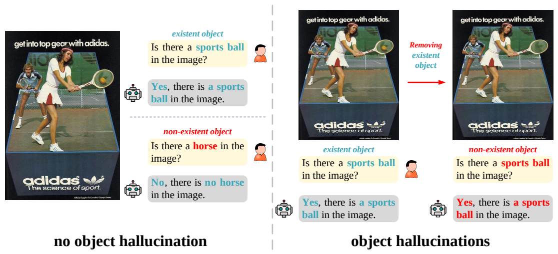

Figure 1: Comparison between existing evaluation benchmarks and ROHE. This example is taken from Fu et al. (2023). The non-existent objects sampled by existing benchmarks may lack challenge for LVLMs and fail to evaluate object hallucination. In contrast, ROHE reveals potential object hallucination by removing existent object in the image.

1 Introduction

With the advancement of Large Language Models (LLMs) (OpenAI, 2023; Anil et al., 2023; Touvron et al., 2023) and the emergence of powerful pre-trained vision-language models (Radford et al., 2021; Caron et al., 2021; Oquab et al., 2024; Woo et al., 2023), several Large Vision-Language Models (LVLMs) have achieved remarkable performance in vision-language tasks such as visual question answering and image captioning (Li et al., 2023a; Dai et al., 2023; Chen et al., 2024a; Liu et al., 2024c; Bai et al., 2023). Despite their impressive performance across various tasks, LVLMs still suffer from severe *object hallucination* issue, which impedes their ability to describe image information at the object level, greatly reducing the reliability of their responses (Rohrbach et al., 2018; Li et al., 2023c; Zhou et al., 2024b).

Recent studies convert the evaluation of object hallucination into a binary discrimination task (Li et al., 2023c; Hu et al., 2023; Fu et al., 2023; Wang et al., 2023a). These studies typically design visual questions about objects (e.g., *"Is there a cat in the image?"*) and prompt LVLMs to provide correct answers (*"yes"* or *"no"*). However, we observed that the non-existent objects they choose for questioning may not be significantly relevant to the image, thereby failing to reveal object hallucinations in LVLMs. As illustrated on the left side of Figure 1, the non-existent *horse* being questioned is not relevant to the tennis-playing scene, making it easy for the model to correctly determine

†Corresponding Author.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6841–6858
April 29 - May 4, 2025

the absence of a *horse*. Additionally, a *"yes"* response from LVLMs does not necessarily indicate the absence of hallucinations. As illustrated on the right side of Figure 1, even when the *sports ball* is removed from the image, the LVLM still responds with *"yes."* This suggests that the LVLM exhibits hallucinations concerning the *sports ball*, which existing methods have overlooked.

To uncover object hallucinations that existing methods have neglected and to provide guidance for mitigating them, we introduce **ROHE** (**R**emoved **O**bject **H**allucination **E**valuation benchmark). As shown on the right side of Figure 1, ROHE utilizes LaMa (Suvorov et al., 2022) to remove existent objects from images. These modified images typically retain other objects and visual backgrounds closely associated with the removed objects, providing a highly challenging test of LVLMs' ability to determine object existence. Furthermore, ROHE considers the model free of hallucinations only if it correctly determines that the object exists in the original image and does not exist in the modified image. This approach uncovers the hallucinations that have been overlooked due to LVLMs' tendency to answer *"yes"* (Li et al., 2023c; Zhou et al., 2024b; Leng et al., 2024).

To ensure the quality of the evaluation, we manually selected the constructed data. Specifically, ROHE comprises 5,504 high-quality evaluation examples (examples in Appendix A), effectively assessing LVLMs' hallucinations across different object categories. We evaluated several representative LVLMs, and the results in Table 1 indicate that LVLMs experience significant hallucinations when confronted with removed objects. Although these LVLMs effectively determine object exists in the image, they struggle to determine the absence of the same object after it has been removed.

In addition to addressing this, we further propose the **o**bject-based **D**irect **P**reference **O**ptimization objective (**oDPO**), a multimodal direct preference optimization (DPO) objective (Rafailov et al., 2023). Unlike existing DPO approaches that construct text-only preference responses (Yu et al., 2024; Li et al., 2023b; Zhou et al., 2024a; Pi et al., 2024; Sarkar et al., 2024), oDPO samples the most important object in the conversation and removes it from the image. oDPO encourages LVLMs to prefer the original image, thereby enhancing their ability to determine the absence of the removed object and reducing associated hallucinations. Extensive experiments (as shown in Figure 2) show

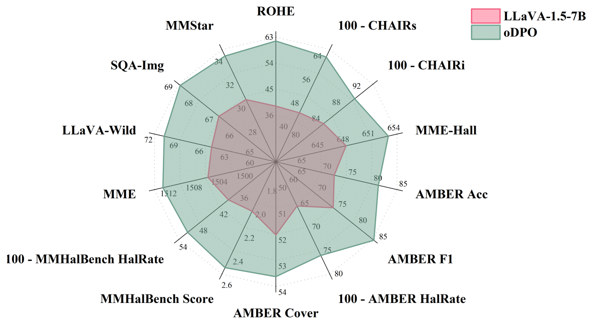

Figure 2: Performance comparison between the model optimized by our proposed approach oDPO and the base model LLaVA-1.5-7B (Liu et al., 2024c) on multiple various vision-language benchmarks. Our approach is effective in improving the performance on various tasks.

that oDPO not only effectively reduces object hallucinations but also improves the performance of LVLMs across various vision-language tasks.

Our contributions can be summarized in three key aspects. (1) We introduce a challenging object hallucination evaluation benchmark called **ROHE** and construct the evaluation data through manual selection. (2) We evaluate several representative LVLMs, revealing the severity of object hallucinations. (3) We propose **oDPO** to mitigate object hallucinations, and experimental results demonstrate the effectiveness of our approach.

2 The Proposed ROHE Benchmark

In this section, we introduce ROHE (§2.1) and the process of constructing the evaluation data (§2.2). We then evaluate representative LVLMs using ROHE (§2.3) and discuss the results (§2.4).

2.1 Overview of ROHE

Description. We devise ROHE to provide a more challenging evaluation of object hallucinations. ROHE utilizes LaMa (Suvorov et al., 2022) to remove existent objects from images. We refer to the original image as the positive image and the image with the object removed as the negative image. To maintain consistency with existing methods (Li et al., 2023c; Fu et al., 2023; Wang et al., 2023a), ROHE adopts a binary question-answering approach to prompt LVLMs to answer *"yes"* or *"no"*, such as *"Is there a cat in the image?"*. For each pair of images, ROHE uses the same question to ask the LVLM whether it sees the object in the positive image or negative image. ROHE requires the LVLM to determine not only objects in the positive image (answering *"yes"*) but also their

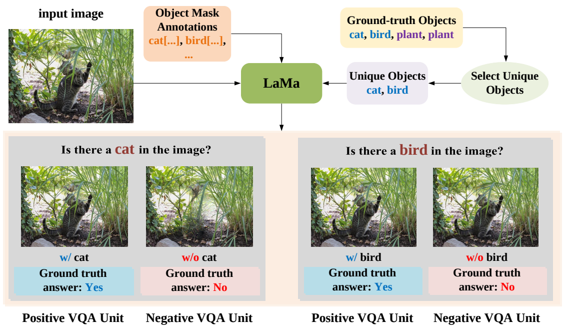

Figure 3: The pipeline of ROHE. Given an input image with ground-truth objects and their corresponding masks, ROHE sequentially removes unique objects using LaMa (Suvorov et al., 2022). Then, ROHE constructs the positive and negative VQA units using images with (**w/**) and without (**w/o**) the object, respectively.

absence in the negative image (answering *"no"*).

Definition. Given an input image v, an existent object o, and the corresponding object mask m, the image obtained by using LaMa (Suvorov et al., 2022) to remove the object o is denoted as v_{ro}. The images v and v_{ro} differ only in the content within the mask m, while the content outside the mask m remains unchanged. The evaluation unit constructed by ROHE can be described as follows:

$$(\langle v, q(o), a\rangle, \langle v_{ro}, q(o), a_{ro}\rangle) \tag{1}$$

where $q(o)$ is the question about the object o based on the prompt template while a and a_{ro} represent answers to the questions when given v and v_{ro} respectively. Here, a is always *"yes"* and a_{ro} is always *"no"*. We refer to $\langle v, q(o), a\rangle$ as the positive VQA unit and $\langle v_{ro}, q(o), a_{ro}\rangle$ as the negative one.

Pipeline. Figure 3 illustrates the ROHE pipeline. First, ROHE selects objects that are uniquely present in the image and then uses LaMa (Suvorov et al., 2022) to remove these objects. Subsequently, ROHE constructs positive VQA units using images containing the objects and negative VQA units using images without the objects. Each VQA unit comprises both a positive and a negative unit concerning the same object. LVLMs are expected to respond *"yes"* to the positive unit and *"no"* to the negative one.

Metrics. ROHE reports two scores: *acc* and *acc+*. The *acc* score represents the proportion of correctly answered question in the positive VQA unit; while the *acc+* score reflects the proportion

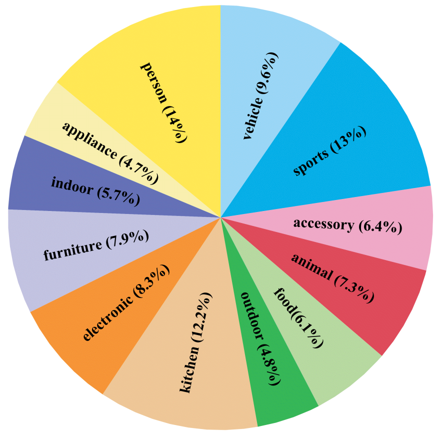

Figure 4: Statistics of our evaluation data.

of correctly answered question in both the positive VQA unit and negative VQA unit.

2.2 Evaluation Data Construction

Dataset. Since most LVLMs are trained using the MSCOCO dataset (Lin et al., 2014), they are expected to exhibit strong recognition capabilities for objects within MSCOCO. However, differences in data splits between the 2014 and 2017 versions of MSCOCO may lead to potential data leakage issues when using the validation set from MSCOCO 2014 for evaluation. To ensure that the evaluation is not out-of-distribution (OOD) and to prevent data leakage, we choose the MSCOCO 2017 validation set to construct our evaluation data.

Manual Selection. We discover that the ROHE evaluation data constructed using LaMa (Suvorov et al., 2022) might still contain incomplete removal of visual information about the objects. To ensure a high-quality evaluation, we conducted manual filtering based on two guidelines: first, to confirm the complete removal of the object, ensuring no visual traces remain in the areas filled by LaMa; second, to verify that humans can determine the object is absent from the negative image. We selected 5,504 high-quality evaluation data units, and Figure 4 provides statistics of our evaluation data.

2.3 Evaluation Settings

We investigate object existence hallucinations in the following representative LVLMs: LLaVA-1.5 (Liu et al., 2024c), LLaVA-1.6 (Liu et al., 2024c), InstructBlip (Dai et al., 2023), Qwen-VL-Chat (Bai et al., 2023), LLaVA-MOF (Tong et al., 2024), VW-LMM (Peng et al., 2024), Monkey-Chat (Li et al., 2024), and SPHINX (Lin et al., 2023). To maintain

supercategory	metrics	LLaVA-1.5-7B	LLaVA-1.5-13B	LLaVA-1.6-7B	LLaVA-1.6-13B	LLaVA-1.6-34B	InstructBlip-7B	InstructBlip-13B	Qwen-VL-Chat	LLaVA-MOF	VW-LMM-Vicuna	Monkey-Chat	SPHINX	SPHINX-1k
vehicle	acc+	46.84	34.39	64.13	64.87	**74.91**	22.30	11.90	61.34	35.69	44.24	74.16	57.06	69.33
	acc	99.81	**100.0**	99.44	98.33	95.54	95.54	98.70	94.61	99.63	**100.0**	87.36	97.03	97.40
sports	acc+	25.51	8.73	42.97	33.29	47.75	4.37	1.36	35.74	9.96	19.65	**56.34**	11.05	36.56
	acc	99.86	99.86	99.45	99.18	98.23	99.32	99.86	97.68	99.86	**100.0**	93.45	99.59	99.18
accessory	acc+	16.62	9.14	45.15	45.98	**62.88**	31.30	9.42	34.63	11.91	9.14	54.02	21.88	43.21
	acc	**100.0**	**100.0**	96.12	96.40	89.20	86.43	95.57	93.63	**100.0**	**100.0**	91.97	99.45	95.84
animal	acc+	74.57	58.19	88.26	82.15	88.26	40.34	27.63	75.55	49.88	72.13	**91.20**	63.57	84.84
	acc	**100.0**	99.51	99.51	98.53	99.76	98.04	99.27	98.04	99.02	99.76	97.31	99.02	99.51
food	acc+	53.39	36.20	69.23	70.14	**75.57**	36.65	16.29	68.78	37.10	45.70	75.11	48.87	57.47
	acc	99.55	**100.0**	98.19	95.48	95.48	95.02	99.55	94.12	99.10	99.55	90.95	97.29	97.74
outdoor	acc+	22.79	15.81	55.88	50.37	70.59	15.44	1.47	45.22	19.49	12.13	**73.53**	31.62	56.62
	acc	**100.0**	**100.0**	97.43	**100.0**	95.22	95.22	**100.0**	95.96	**100.0**	**100.0**	89.71	98.16	99.63
kitchen	acc+	30.26	18.57	49.12	48.68	**64.77**	27.49	4.97	41.81	18.13	25.15	54.68	35.53	50.00
	acc	99.42	**100.0**	96.93	97.66	91.67	90.94	99.56	83.48	99.12	99.56	70.61	97.95	96.78
electronic	acc+	26.34	13.28	49.46	43.47	62.10	22.91	5.78	39.40	13.70	18.20	**66.81**	30.84	46.25
	acc	**100.0**	**100.0**	98.50	99.14	97.00	96.57	98.72	94.65	99.79	**100.0**	83.94	98.72	98.29
furniture	acc+	36.91	28.41	57.05	54.14	**65.55**	19.02	5.15	54.14	30.65	29.75	62.86	43.85	55.93
	acc	99.55	**99.78**	98.43	97.99	93.96	96.20	99.55	91.72	**99.78**	**99.78**	76.51	97.54	98.43
indoor	acc+	28.66	14.95	51.40	57.63	61.99	25.23	7.79	45.48	15.58	19.31	**65.42**	28.97	46.73
	acc	**100.0**	**100.0**	97.51	97.82	95.64	94.08	99.38	93.46	**100.0**	**100.0**	85.05	99.07	97.51
appliance	acc+	12.50	8.33	37.88	33.71	43.56	11.36	3.03	36.36	9.47	8.71	**55.30**	21.97	29.92
	acc	**100.0**	**100.0**	98.11	98.86	97.35	97.35	99.62	93.56	**100.0**	**100.0**	76.14	98.48	98.86
person	acc+	70.39	61.75	83.99	80.94	83.23	47.65	16.65	60.23	58.58	70.78	**86.66**	52.60	62.52
	acc	**99.87**	99.75	99.62	99.75	99.36	98.09	99.62	98.73	**99.87**	**99.87**	90.60	99.49	99.62
total	acc+	39.21	27.53	58.81	55.89	67.13	25.78	9.25	49.58	27.40	34.08	**68.15**	37.59	53.67
	acc	99.82	**99.89**	98.46	98.46	95.93	95.53	99.18	94.11	99.67	99.87	86.01	98.58	98.29

Table 1: Results of the ROHE evaluation. The results in **bold** and underlined represent the best and the second-best results, respectively.

consistency with previous work (Fu et al., 2023; Li et al., 2023c), we also use *"Is there a/an {obj} in the image?"* as the prompt template. We leave more details in Appendix B.

2.4 Evaluation Results

Table 1 presents the evaluation results, indicating that while LVLMs effectively determine the presence of objects in positive VQA units, most of them fail to determine the absence of the same objects in negative VQA units. Instead, they still to claim that they see those objects.

The overall results show that LVLMs with enhanced visual resolution performed better, suggesting that inputting more detailed visual tokens provides these models with sufficient fine-grained visual information, aiding them in better learning and perceiving visual objects. In contrast, InstructBLIP (Dai et al., 2023) achieved the lowest $acc+$ score, possibly due to the limited visual information extracted by Q-Former (Li et al., 2023a), which restricts the language model's access to sufficient object-level visual details. Additionally, LVLMs exhibit more hallucinations for categories such as outdoor and appliance, while showing fewer hallucinations for categories like person and animal.

Overall, the evaluated LVLMs still exhibit significant object hallucinations when confronted with removed objects, suggesting that they significantly lack the ability to determine the absence of objects. We observe that during training, LVLMs are instructed to learn what objects are present in an image, but there is considerably less focus on learning what objects are absent. This imbalance likely leads to their inability to determine the absence of objects, thereby resulting in severe hallucinations.

3 The Proposed oDPO Approach

To address the issue revealed in §2.4, we propose oDPO, an object-based DPO objective designed to enhance LVLM' ability to determine the existence of objects, thereby mitigating hallucinations.

3.1 Preliminaries

Preference optimization aims to align the model's behavior with human behavior through fine-tuning. Typically, given a text input x, an image input v, and an output text response y, a model π_θ parameterized by θ can produce a conditional distribution $\pi_\theta(y \mid x, v)$. The model is encouraged to maximize

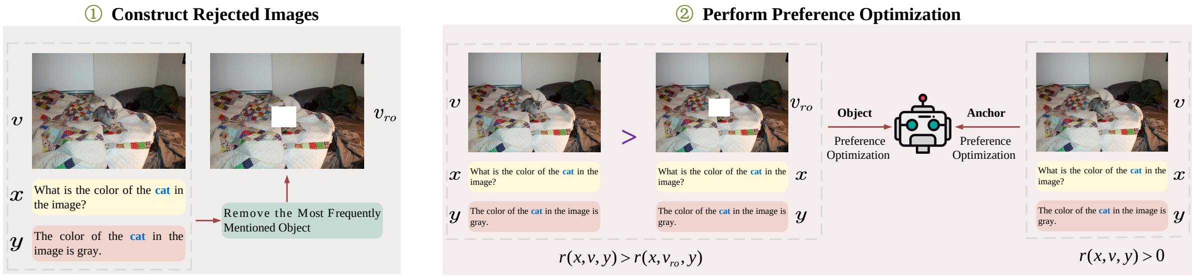

Figure 5: Overview of oDPO. The oDPO process is divided into two steps: constructing rejected images and performing preference optimization.

the average reward of output $r(x, v, y)$. To avoid over-optimization (Gao et al., 2023), it is necessary to control the divergence between π_θ and its reference model π_{ref} (π_θ and π_{ref} are initialized from the same checkpoint). Thus, the overall objective loss is typically formulated as follows:

$$\mathcal{L}_{\text{PO}} = -\log \sigma(r(x, v, y) - \beta \log \frac{\pi_\theta(y|x,v)}{\pi_{\text{ref}}(y|x,v)}) \quad (2)$$

where β is a hyperparameter that controls the divergence between π_θ and π_{ref}, and $\sigma(\cdot)$ is the sigmoid function. Recently, DPO (Rafailov et al., 2023) simplifies the above process by maximizing the difference between the chosen reward $r(x, v, y_w)$ and the rejected reward $r(x, v, y_l)$. Following the Bradley-Terry model (Bradley and Terry, 1952), the optimization objective becomes:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma(\beta \log \frac{\pi_\theta(y_w|x,v)}{\pi_{\text{ref}}(y_w|x,v)} - \beta \log \frac{\pi_\theta(y_l|x,v)}{\pi_{\text{ref}}(y_l|x,v)}) \quad (3)$$

3.2 Object-based Optimization Objective

To mitigate the severe object hallucinations, we propose **oDPO** (**o**bject-based **D**irect **P**reference **O**ptimization objective). Unlike previous work (Zhao et al., 2023; Li et al., 2023b; Zhou et al., 2024a), oDPO is a multimodal optimization objective based on visual objects. As illustrated in Figure 5, given a text input x, an image input v, and an output text response y, oDPO removes the most frequently mentioned object o using its mask and obtain the rejected image input v_{ro}. Here, $r(x, v, y)$ represents the chosen reward, and $r(x, v_{\text{ro}}, y)$ represents the rejected reward. Then, the preference optimization objective is formulated as:

$$\mathcal{L}_{\text{roDPO}} = -\log \sigma(\beta \log \frac{\pi_\theta(y|x,v)}{\pi_{\text{ref}}(y|x,v)} - \beta \log \frac{\pi_\theta(y|x,v_{\text{ro}})}{\pi_{\text{ref}}(y|x,v_{\text{ro}})}) \quad (4)$$

Inspired by Wang et al. (2024a), we employ anchor preference optimization to ensure that the chosen reward consistently remains at a high value. The anchored objective is formulated as follows:

$$\mathcal{L}_{\text{AncPO}} = -\log \sigma(\beta \log \frac{\pi_\theta(y \mid x, v)}{\pi_{\text{ref}}(y \mid x, v)}) \quad (5)$$

Then the total preference optimization objective is

$$\mathcal{L}_{\text{oDPO}} = \mathcal{L}_{\text{roDPO}} + \gamma \mathcal{L}_{\text{AncPO}} \quad (6)$$

where γ controls the influence of the anchored objective.

4 Experiment

4.1 Experimental Setups

Training Data. The Silkie dataset (Li et al., 2023b) contains 80K preference data, from which we selected 19K examples constructed by LLaVA-Instruct-150K (Liu et al., 2024c) for training. It is important to note that oDPO utilizes only the chosen responses constructed by Silkie and does not require the use of rejected responses.

Base Models. Following related work (Zhou et al., 2024a), we applied oDPO on LLAVA-1.5 (7B and 13B) (Liu et al., 2024c). To compare oDPO with standard DPO (Rafailov et al., 2023), we also implemented standard DPO using the same training data. Apart from the differences in optimization objectives, all other settings are identical.

Implementation Details. We set the learning rate to 1e-7, used a cosine learning rate scheduler with a warmup ratio of 0.03, and set the default value of γ to 1. All models were trained for only one epoch, and all experiments were conducted on one A100 80GB GPU. More details can be found in Appendix C.

4.2 Main Results

Performance on ROHE. We compare oDPO with other approaches (Leng et al., 2024; Yue et al.,

	vehicle	sports	accessory	animal	food	outdoor	kitchen	electronic	furniture	indoor	appliance	person	**total**
LLaVA-1.5-7B (Liu et al., 2024c)	46.84	25.51	16.62	74.57	53.39	22.79	30.26	26.34	36.91	28.66	12.50	70.39	39.21
+ VCD (Leng et al., 2024)	39.96	19.65	17.45	75.55	51.13	23.90	27.05	23.34	33.78	27.73	12.50	74.46	37.46
+ EOS (Yue et al., 2024)	48.33	25.65	13.02	71.39	50.23	19.49	26.17	25.70	35.12	23.36	12.12	75.10	38.24
+ DPO (Rafailov et al., 2023)	49.07	27.29	15.51	75.06	55.66	20.59	28.80	23.77	36.47	27.10	12.12	69.50	38.94
+ oDPO (ours)	**69.89**	**50.61**	**48.48**	**86.80**	**72.40**	**55.88**	**52.78**	**53.96**	**59.28**	**54.21**	**38.64**	**82.59**	**61.65**
LLaVA-1.5-13B (Liu et al., 2024c)	34.39	8.73	9.14	58.19	36.20	15.81	18.57	13.28	28.41	14.95	8.33	61.75	27.53
+ VCD (Leng et al., 2024)	26.77	7.64	11.36	58.68	33.48	18.38	20.03	11.99	27.74	16.82	9.47	65.18	27.51
+ DPO (Rafailov et al., 2023)	35.50	9.96	8.86	60.88	36.20	14.34	19.30	14.13	28.86	16.20	9.85	62.39	28.34
+ oDPO (ours)	**58.74**	**24.97**	**24.65**	**77.02**	**52.94**	**34.19**	**37.28**	**31.05**	**42.73**	**33.02**	**20.45**	**75.86**	**44.71**

Table 2: Results on ROHE. We report acc+ scores and provide the complete results in Appendix D. The best results are shown in **bold**.

	Object HalBench		MME-Hall	AMBER						MMHalBench	
	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	Score ↑	CHAIR ↓	Cover ↑	HalRate ↓	Cog ↓	Acc ↑	F1 ↑	Score ↑	HalRate ↓
LLaVA-1.5-7B (Liu et al., 2024c)	53.3	15.6	648.3	7.6	51.8	35.6	4.3	71.5	74.1	2.02	0.61
+ VCD (Leng et al., 2024)	53.3	15.7	604.7	6.9	50.6	32.2	3.7	72.0	74.8	2.12	0.54
+ EOS (Yue et al., 2024)	41.7	12.7	606.7	5.3	49.1	**23.5**	**2.0**	71.4	73.1	2.03	0.59
+ HA-DPO (Zhao et al., 2023)	43.7	12.0	618.3	6.5	49.8	30.1	3.2	74.2	78.0	1.97	0.60
+ POVID (Zhou et al., 2024a)	40.7	10.2	591.7	5.2	50.2	27.9	3.0	78.5	81.9	2.23	0.54
+ HALVA[†] (Sarkar et al., 2024)	41.4	11.7	**665.0**	6.6	53.0	32.2	3.4	-	83.4	2.25	0.54
+ RLHF-V[‡] (Yu et al., 2024)	-	-	-	5.7	49.7	27.3	2.6	-	80.9	2.08	0.60
+ V-DPO[†] (Xie et al., 2024)	-	-	-	5.6	49.7	27.3	2.7	-	81.6	2.16	0.56
+ mDPO[†] (Wang et al., 2024a)	35.7	9.8	-	**4.4**	52.4	24.5	2.4	-	-	2.39	0.54
+ DPO (Rafailov et al., 2023)	50.7	14.9	641.7	7.3	**54.1**	38.5	4.1	70.7	73.1	2.23	0.58
+ oDPO (ours)	**34.3**	**9.5**	653.3	4.6	53.4	25.1	2.4	**80.2**	**84.1**	**2.50**	**0.49**
LLaVA-1.5-13B (Liu et al., 2024c)	49.3	14.6	643.3	6.8	52.0	31.7	3.5	71.3	73.1	2.38	0.53
+ VCD (Leng et al., 2024)	47.7	13.2	601.7	6.7	51.3	31.0	3.5	71.5	73.5	2.40	0.51
+ DPO (Rafailov et al., 2023)	51.7	13.3	646.7	7.1	**54.1**	36.0	3.9	71.7	73.7	2.48	0.52
+ oDPO (ours)	**34.7**	**9.8**	**660.0**	**4.3**	52.1	**23.1**	**2.2**	**79.3**	**82.2**	**2.74**	**0.45**

Table 3: Results on object hallucination. We report sentence-level and object-level scores (CHAIR_s and CHAIR_i) on Object HalBench (Rohrbach et al., 2018), overall score on MME-Hall (Fu et al., 2023). For AMBER (Wang et al., 2023a), we report CHAIR scores, object coverage (Cover), hallucination rate (HalRate) and cognition (Cog) in generation task, along with Acc and F1 scores of discrimination task. We also report the overall score and hallucination rate (HalRate) on MMHalBench (Sun et al., 2024). The best and second-best results are shown in **bold** and underlined, respectively. †: We directly report the results from their papers. ‡: results are from Xie et al. (2024).

2024; Rafailov et al., 2023). Table 2 and Table 11 (provided in Appendix D) present the evaluation results. Although other approaches aim to reduce object hallucinations in LVLMs, they struggle to improve LVLMs' ability to determine the absence of removed objects. In contrast, oDPO effectively enhances their ability to determine the existence of visual objects through preference optimization based on visual objects, significantly mitigating hallucinations on ROHE.

Performance on Object Hallucination. To ensure that oDPO effectively mitigates object hallucinations in LVLMs, we conduct evaluations on four widely used object hallucination benchmarks: Object HalBench (Rohrbach et al., 2018), MME-Hall (Fu et al., 2023), AMBER (Wang et al., 2023a) and MMHalBench (Sun et al., 2024). Please refer to Appendix E for details. Table 3 demonstrates the effectiveness of oDPO in reducing hallucinations. Compared to other approaches, oDPO consistently exhibits stable and superior performance in various object hallucination tasks. It is worth noting that some approaches reduce the object coverage in image descriptions generated by LVLMs, which while potentially alleviating object hallucinations, also diminish the richness of descriptions. In contrast, oDPO reduces hallucinated objects while ensuring that LVLMs can richly describe the image content.

4.3 Analysis and Discussion

How does oDPO perform on general vision-language tasks? We further evaluate oDPO on four popular general vision-language benchmarks: MME (Fu et al., 2023), LLaVA-Wild (Chen et al., 2024a), SQA-Img (Lu et al., 2022), and MMStar (Chen et al., 2024b). The results in Figure 6 show that oDPO outperforms the base model across these general benchmarks, suggesting that oDPO mitigates hallucinations without deteriorating the performance of LVLMs on other tasks.

How does oDPO perform when using different training data? As shown in Table 4, we explore

	ROHE		MME-Hall	Object HalBench		AMBER						MMHalBench	
	acc+ ↑	acc ↑	Score ↑	CHAIR_s ↓	CHAIR_i ↓	CHAIR ↓	Cover ↑	HalRate ↓	Cog ↓	Acc ↑	F1 ↑	Score ↑	HalRate ↓
LLaVA-1.5-7B (Liu et al., 2024c)	39.21	**99.82**	648.3	53.3	15.6	7.6	51.8	35.6	4.3	71.5	74.1	2.02	0.61
+ oDPO (Silkie-19K)	61.65	98.49	653.3	**34.3**	**9.5**	**4.6**	**53.4**	25.1	**2.4**	80.2	84.1	**2.50**	**0.49**
+ oDPO (LLaVA-17K)	**63.70**	98.42	**661.7**	43.0	12.1	5.0	51.0	**24.8**	2.7	**80.5**	**84.5**	2.48	0.51
LLaVA-1.5-13B (Liu et al., 2024c)	27.53	**99.89**	643.3	49.3	14.6	6.8	52.0	31.7	3.5	71.3	73.1	2.38	0.53
+ oDPO (Silkie-19K)	44.71	99.58	**660.0**	**34.7**	**9.8**	**4.3**	**52.1**	**23.1**	**2.2**	79.3	82.2	**2.74**	**0.45**
+ oDPO (LLaVA-17K)	**47.15**	99.53	**660.0**	42.7	11.6	5.0	51.4	24.4	2.6	**80.0**	**83.0**	2.70	0.46

Table 4: The results of oDPO using different training data. LLaVA-17K: 17K examples randomly sampled from LLaVA-Instruct-150K (Liu et al., 2024c); Silkie-19K: 19K examples sampled from Silkie (Li et al., 2023b). The best results are denoted in **bold**.

	ROHE		MME-Hall	Object HalBench		AMBER						MMHalBench	
	acc+ ↑	acc ↑	Score ↑	CHAIR_s ↓	CHAIR_i ↓	CHAIR ↓	Cover ↑	HalRate ↓	Cog ↓	Acc ↑	F1 ↑	Score ↑	HalRate ↓
LLaVA-1.6-13B (Liu et al., 2024d)	55.89	**98.46**	**660.0**	30.0	10.3	8.6	62.0	50.5	4.2	81.2	84.9	3.09	0.46
+ DPO (Rafailov et al., 2023)	43.95	98.13	648.3	40.3	8.8	7.6	**63.1**	49.2	4.5	69.9	71.1	3.40	0.42
+ oDPO (ours)	**63.06**	90.75	650.0	**27.7**	**7.2**	**5.7**	59.0	**33.9**	**2.9**	**82.7**	**86.8**	**3.42**	**0.33**

Table 5: Results on LLaVA-1.6-13B. The best results are shown in **bold**.

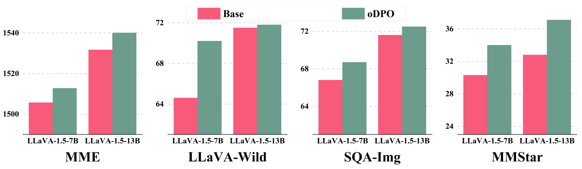

Figure 6: Results on general vision-language tasks. We report the scores of base models (**Base**) and oDPO-enhanced models (**oDPO**) on four benchmarks: MME (Fu et al., 2023), LLaVA-Wild (Chen et al., 2024a), SQA-Img (Lu et al., 2022), and MMStar (Chen et al., 2024b).

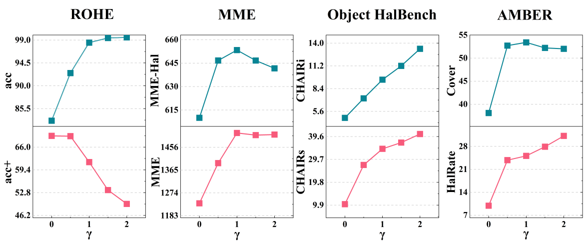

Figure 7: Impact of different γ values. We report the results of two primary metrics specific to each of the four benchmarks: ROHE, MME (Fu et al., 2023), Object HalBench (Rohrbach et al., 2018), and AMBER (Wang et al., 2023a). The base model is LLaVA-1.5-7B.

the performance of oDPO with different training data. oDPO is effective in different training data.

How does oDPO perform on LVLM that support high resolution? Table 5 shows the performance of oDPO on LLaVA-1.6-13B (Liu et al., 2024d), which compared to LLaVA-1.5 (Liu et al., 2024c), supports dynamic high resolution. Although oDPO slightly reduces object coverage and MME-Hal scores, it effectively mitigates object hallucination across different benchmarks. In contrast, the standard DPO not only fails to reduce object hallucination but also exacerbates it in some aspects.

How does γ affect the performance of oDPO? As shown in Figure 7, we investigate the impact of different γ values on the performance of oDPO. It is observed that a small γ value significantly reduces object hallucinations. However, this reduction is accompanied by a decline in performance on other tasks and a suppression of response diversity. To balance these effects, we set the γ value to 1, aiming to mitigate object hallucinations without compromising performance in other areas.

Why oDPO performs better than standard DPO and other baselines? During the pre-training process, there is little direct guidance for model to learn how to determine the existence of an object. Although RLHF methods based on textual preferences construct fine-grained preference pairs, they do not directly guide the model to learn why the chosen response aligns with the input image and why the preference for the rejected response should be reduced. In contrast, oDPO directly guides the model to increase its preference when the input image contains the key object, and to decrease it when the object is absent.

4.4 Fine-grained Results

We further provide fine-grained results on MMHal-Bench (Sun et al., 2024) in Table 6. Although oDPO slightly decreases scores in the relation and other categories, it surpasses standard DPO in all other categories. Notably, in the adversarial category, oDPO boosts the base model's score by 163%. These findings highlight the benefits of

	overall	attribute	adversarial	comparison	counting	relation	environment	holistic	other
LLaVA-1.5-7B (Liu et al., 2024c)	2.02	3.00	1.17	1.83	2.25	**2.00**	3.08	1.75	1.08
+ DPO (Rafailov et al., 2023)	2.23	3.25	1.67	2.17	2.00	1.83	3.17	**2.17**	**1.58**
+ oDPO (ours)	**2.50**	**3.33**	**3.08**	**2.42**	**2.33**	1.75	**4.00**	**2.17**	0.92

Table 6: Fine-grained results on MMHalBench (Sun et al., 2024). The best results are denoted in **bold**.

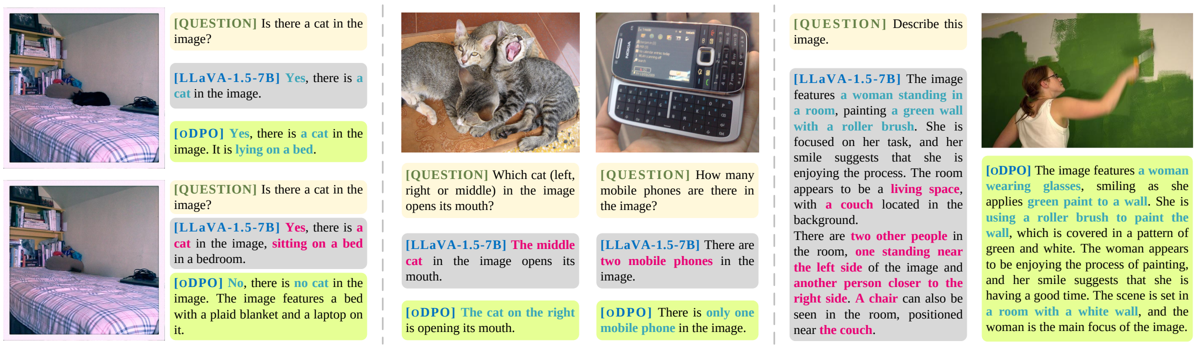

Figure 8: Qualitative results of oDPO. The left, middle and right figures are from ROHE, MMHalBench (Sun et al., 2024) and AMBER (Wang et al., 2023a), respectively. It can be observed that oDPO significantly reduces hallucination and enhances the model's ability to describe detailed information in images across different tasks.

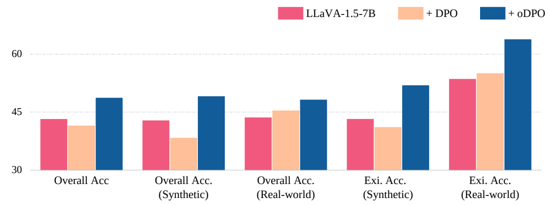

Figure 9: Results of oDPO on AutoHallusion (Wu et al., 2024b).

	Seen			Unseen		
	Wild ↑	Hom. ↑	Het. ↑	Wild ↑	Hom. ↑	Het. ↑
Default Multi-Object						
LLaVA-1.5-7B (Liu et al., 2024c)	24.94	58.05	7.76	16.35	37.71	4.72
+ DPO (Rafailov et al., 2023)	23.69	54.88	8.00	14.40	33.40	4.70
+ oDPO (ours)	**25.89**	**61.25**	**9.49**	**19.45**	**46.92**	**5.45**
Single-Object						
LLaVA-1.5-7B (Liu et al., 2024c)	30.28	61.15	13.08	25.32	52.49	10.24
+ DPO (Rafailov et al., 2023)	30.66	62.00	13.01	25.84	52.78	10.41
+ oDPO (ours)	**31.25**	**63.75**	**13.21**	**26.04**	**54.57**	**10.65**
Student-Forcing						
LLaVA-1.5-7B (Liu et al., 2024c)	2.50	5.98	0.95	2.22	3.77	1.25
+ DPO (Rafailov et al., 2023)	3.30	7.74	1.43	2.84	4.63	**1.56**
+ oDPO (ours)	**3.84**	**8.84**	**1.80**	**3.49**	**6.72**	1.30
Teacher-Forcing						
LLaVA-1.5-7B (Liu et al., 2024c)	3.32	7.89	1.68	3.32	6.63	1.56
+ DPO (Rafailov et al., 2023)	4.24	10.31	1.97	3.51	7.38	1.61
+ oDPO (ours)	**4.65**	**10.37**	**2.30**	**4.28**	**9.06**	**1.81**

Table 7: Results of oDPO on ROPE (Chen et al., 2024c). The best results are denoted in **bold**.

oDPO across various fine-grained scenarios.

4.5 Qualitative Study

Figure 8 presents qualitative examples. oDPO effectively mitigates hallucinations across different tasks. Furthermore, compared to the base model, the oDPO-enhanced model generally provides more detailed descriptions of the images, suggesting that oDPO effectively enhances LVLMs' visual understanding and reasoning capabilities.

4.6 Results on More Complex Scenarios

To further evaluate oDPO's performance on more complex scenarios, we conduct additional experiments on AutoHallusion (Wu et al., 2024b) and ROPE (Chen et al., 2024c). The results in Figure 9 and Table 7 demonstrate that oDPO effectively alleviates hallucination issues of the base model across different scenarios.

5 Related Work

Evaluations of Object Hallucinations in LVLMs. Currently, the evaluation methods for object hallucinations are primarily divided into two categories (Liu et al., 2024b): evaluation in generation task and in discrimination task. Evaluation in generation task, typically uses hand-designed pipelines (Rohrbach et al., 2018; Zhai et al., 2023; Lee et al., 2024) or LLM-based methods (Liu et al., 2024a; Sun et al., 2024; Gunjal et al., 2024; Wang et al., 2023b) to locate the hallucinatory parts in the LVLM's responses and calculate the proportion and score of hallucinations. Evaluation in discrimination task aims to evaluate the performance of LVLMs in judging objects. They usually design visual questions about objects (e.g., *"Is there a cat in the image?"*) and prompt LVLMs, expecting them to provide correct answers (*"yes"* or *"no"*). They usually employ three different approaches to choose the objects for questioning: manual design (Fu et al., 2023), handcrafted pipelines (Li et al., 2023c; Wang et al., 2023a), or GPT generation (Hu et al., 2023; Lovenia et al., 2023). ROHE also evaluates object hallucinations in discrimination task. It not only evaluates the ability of LVLMs to determine an object exists in the image but also assesses their ability to determine the absence of the same object after it has been removed.

Mitigation of Object Hallucinations in LVLMs. To mitigate object hallucinations in LVLMs, some studies have focused on constructing more robust datasets or designing specific training strategies during the pre-training stage (Sun et al., 2024; Jiang et al., 2024; Liu et al., 2024a; Yue et al., 2024). Other approaches have utilized specific decoding strategies (Leng et al., 2024; Zhu et al., 2024; Wang et al., 2024b) or directly corrected the responses of LVLMs (Zhou et al., 2024b; Wu et al., 2024a). Recently, researchers have performed preference alignment in LVLMs by collecting human preference data (Sun et al., 2024; Yu et al., 2024) or collecting the preferences from advanced LLMs (Zhao et al., 2023; Li et al., 2023b; Zhou et al., 2024a; Sarkar et al., 2024). Wang et al. (2024a) introduces a multimodal direct preference optimization objective that constructs the rejected image by cropping the original image. oDPO also uses the multimodal DPO objective, but it constructs the rejected image by removing objects from the original image. Additionally, oDPO does not use the rejected responses from the preference dataset; instead, it focuses on preference optimization based on visual objects and aims for LVLMs to learn to prefer the original image. This enhances their ability to determine the absence of removed objects and reduces hallucinations related to them.

6 Conclusion

In this paper, we introduce ROHE, which designed to evaluate object hallucinations by removing objects from images. Our evaluation results reveal that LVLMs still suffer from severe hallucinations, as they often struggle to determine the absence of removed objects. To address this, we propose oDPO, an object-based DPO objective designed to guide LVLMs to learn to determine the existence of objects. We conducted extensive experiments and the results demonstrate that oDPO not only enhances LVLMs' ability to determine the existence of objects but also improves their performance on various vision-language tasks, particularly in reducing object hallucinations.

Limitations

Although we have conducted extensive exploration and experiments, this work still has many limitations. First, we only evaluated object hallucinations through binary question-answering, it does not allow us to assess the overall hallucination performance of LVLMs. Second, due to budget and resource constraints, we developed the benchmark only on the MSCOCO 2017 validation dataset (Lin et al., 2014). Third, we have evaluated only some open-source LVLMs and have not yet assessed closed-source LVLMs or the latest LVLMs. In addition, due to the limitations of LaMa (Suvorov et al., 2022), the synthesized images may contain unrealistic artifacts. Finally, owing to computational resource constraints, although we have conducted experiments on several baseline LVLMs and training datasets, it is challenging for us to explore the performance of oDPO on larger-scale LVLMs, e.g., LLaVA-1.6-34B (Liu et al., 2024d).

Ethics Statement

In this work, we use LaMa (Suvorov et al., 2022) to generate images based on the MSCOCO dataset (Lin et al., 2014). It is important to acknowledge that the generated images may contain counterfactual or fake information. Researchers can employ ROHE to evaluate object hallucination in LVLMs, but should be cautious about applying the fake images in the benchmark to other purposes to avoid causing social interference.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62201072, Grant U23B2001, Grant 62171057, Grant 62101064, Grant 62001054, and Grant 62071067; in part by the Ministry of Education and China Mobile Joint Fund under Grant MCM20200202 and Grant MCM20180101; in part by the BUPT-China Mobile Research Institute Joint Innovation Center.

References

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE.

Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. 2024a. Visual instruction tuning with polite flamingo. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17745–17753. AAAI Press.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024b. Are we on the right way for evaluating large vision-language models? *CoRR*, abs/2403.20330.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David F. Fouhey, and Joyce Chai. 2024c. Multi-object hallucination in vision-language models. *CoRR*, abs/2407.06192.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18135–18143. AAAI Press.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023. CIEM: contrastive instruction evaluation method for better instruction tuning. *CoRR*, abs/2309.02301.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 27026–27036. IEEE.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2024. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*

(Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 391–404. Association for Computational Linguistics.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13872–13882. IEEE.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023b. Silkie: Preference distillation for large visual language models. *CoRR*, abs/2312.10665.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26753–26763. IEEE.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. 2023. SPHINX: the joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *CoRR*, abs/2311.07575.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024b. A survey on hallucination in large vision-language models. *CoRR*, abs/2402.00253.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024d. Llava-next: Improved reasoning, ocr, and world knowledge.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (NOPE) to measure object hallucination in vision-language models. *CoRR*, abs/2310.05338.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Tianshuo Peng, Zuchao Li, Lefei Zhang, Hai Zhao, Ping Wang, and Bo Du. 2024. Multi-modal autoregressive modeling via visual words. *CoRR*, abs/2403.07720.

Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. 2024. Strengthening multimodal large language model with bootstrapped preference optimization. *CoRR*, abs/2403.08730.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hanoona Abdul Rasheed, Muhammad Maaz, Sahal Shaji Mullappilly, Abdelrahman M. Shaker, Salman H. Khan, Hisham Cholakkal, Rao Muhammad Anwer, Eric P. Xing, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2024. Glamm: Pixel grounding large multimodal model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13009–13018. IEEE.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4035–4045. Association for Computational Linguistics.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö. Arik, and Tomas Pfister. 2024. Mitigating object hallucination via data augmented contrastive tuning. *CoRR*, abs/2405.18654.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 13088–13110. Association for Computational Linguistics.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 3172–3182. IEEE.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9568–9578. IEEE.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mdpo: Conditional preference optimization for multimodal large language models. *CoRR*, abs/2406.11839.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023a. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR*, abs/2311.07397.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. 2023b. Evaluation and analysis of hallucination in large vision-language models. *CoRR*, abs/2308.15126.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15840–15853. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. 2023. Convnext V2: co-designing and scaling convnets with masked autoencoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 16133–16142. IEEE.

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. 2024a. Logical closed loop: Uncovering object hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6944–6962. Association for Computational Linguistics.

Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan L. Boyd-Graber, Tianyi Zhou, and Dinesh Manocha. 2024b. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 8395–8419. Association for Computational Linguistics.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. V-DPO: mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 13258–13273. Association for Computational Linguistics.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *CoRR*, abs/2403.04652.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2024. RLHF-V: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13807–13816. IEEE.

Zihao Yue, Liang Zhang, and Qin Jin. 2024. Less is more: Mitigating multimodal hallucination from an EOS decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11766–11781. Association for Computational Linguistics.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2023. Halleswitch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *CoRR*, abs/2310.01779.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *CoRR*, abs/2311.16839.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024a. Aligning modalities in vision large language models via preference fine-tuning. *CoRR*, abs/2402.11411.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024b. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024. IBD: alleviating hallucinations in large vision-language models via image-biased decoding. *CoRR*, abs/2402.18476.

A Examples of ROHE

Figure 10 and 11 show some examples of ROHE.

B Details of the Evaluation on ROHE

All experiments in this work were conducted using the PyTorch framework (Paszke et al., 2019) and incorporated capabilities from HuggingFace's Transformers library (Wolf et al., 2019). The experiments were conducted using an NVIDIA A100 GPU and an Intel Xeon Silver 4210R CPU. We used the settings in Table 8. Here are the details of the LVLMs we evaluated:

Hyperparameters	
do_sample	False
num_beams	1
top_p	1
top_k	None
temperature	0

Table 8: Hyperparameter settings of ROHE.

- **LLaVA** (Chen et al., 2024a; Liu et al., 2024c,d): We evaluated LLaVA-1.5-7B, LLaVA-1.5-13B, LLaVA-1.6-7B, LLaVA-1.6-13B, and LLaVA-1.6-34B. Notably, LLaVA-1.6 supports dynamic high-resolution capabilities for higher image resolutions. LLaVA-1.6-34B is based on Hermes-Yi-34B (Young et al., 2024), and the other models are based on Vicuna (Chiang et al., 2023).

- **InstructBlip** (Dai et al., 2023): We evaluated InstructBlip-7B and InstructBlip-13B which are based on Vicuna (Chiang et al., 2023).

- **Qwen-VL-Chat** (Bai et al., 2023): We evaluated Qwen-VL-Chat which is based on Qwen-7B (Bai et al., 2023).

- **LLaVA-MOF** (Tong et al., 2024): This model is an improved version of LLaVA-1.5-13B (Liu et al., 2024c), enhancing the visual perception by mixing CLIP-VIT (Radford et al., 2021) and DINOv2-VIT (Oquab et al., 2024).

- **VW-LMM** (Peng et al., 2024): VW-LMM uses the same training dataset as LLaVA-1.5 (Liu et al., 2024c), but constructs visual words to introduce visual supervisory information. To compare results with LLaVA-1.5, we evaluated VW-LMM-Vicuna-7B.

- **Monkey-Chat** (Li et al., 2024): Monkey-Chat uses Qwen-7B (Bai et al., 2023) as its foundational model and is capable of processing images with resolutions up to 1344×896 pixels through a super-resolution method.

- **SPHINX** (Lin et al., 2023): SPHINX employs four visual encoders, CLIP-VIT (Radford et al., 2021), CLIP-ConvNext (Woo et al., 2023), DINOv2-VIT (Oquab et al., 2024), and Q-former (Li et al., 2023a), to extract visual features, thereby enhancing visual perception by combining visual features. We evaluated two versions, SPHINX and SPHINX-1k, where SPHINX takes a low-resolution image of 224×224 as input while SPHINX-1k handles an image resolution of 448×448 by averaging four sub-images into 1,445 visual tokens by cropping the images.

C More Implementation Details of oDPO

All experiments were conducted using the PyTorch framework (Paszke et al., 2019) and incorporated capabilities from HuggingFace's Transformers library (Wolf et al., 2019). The experiments were conducted using an NVIDIA A100 GPU and an Intel Xeon Silver 4210R CPU. We adapted LoRA fine-tuning (Hu et al., 2022). The details of training hyperparameters used in oDPO is presented in Table 9.

Hyperparameters	
lora rank	128
lora alpha	256
mm projector lr	1e-5
batch size	1
learning rate	1e-7
warmup decay	0.
warmup ratio	0.03
learning rate scheduler	Cosine
max length	1024

Table 9: Training hyperparameters used in oDPO.

D Complete Results on ROHE

Table 11 provides the complete results on ROHE.

E Details of Evaluation Benchmarks

- **Object HalBench** (Rohrbach et al., 2018): Object HalBench is a widely used method for evaluating object hallucination in image descriptions. It typically reports two object hallucination scores: sentence-level and object-level CHAIR scores, referred to as CHAIR_s and CHAIR_i, respectively. They can be formulated as:

$$\text{CHAIR}_\text{S} = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}, \quad (7)$$

$$\text{CHAIR}_\text{i} = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}. \quad (8)$$

  We use 300 images randomly sampled by Yu et al. (2024) from MSCOCO (Lin et al., 2014) along with their corresponding prompts as the evaluation examples. The detection of objects in the LVLMs' responses is conducted using an exact match approach.

- **MME-Hall** (Fu et al., 2023): MME-Hall is the hallucination subset of the MME benchmark (Fu et al., 2023), including four object-related subtasks: existence, count, position, and color. It effectively evaluates the hallucination in LVLMs on discrimination tasks. Each subtask has a total score of 200, making the overall score for MME-Hall is 800.

- **AMBER** (Wang et al., 2023a): AMBER is an LLM-free object hallucination benchmark that effectively evaluates the hallucination performance of LVLMs on both generative and discrimination tasks. For generative tasks, we report CHAIR scores, object coverage (Cover), hallucination rate (HalRate), and cognition (Cog). For discrimination tasks, we report accuracy (Acc) and F1 scores.

- **MMHalBench** (Sun et al., 2024): MMHalBench is an object hallucination evaluation benchmark that utilizes GPT-4 (OpenAI, 2023) to assist in scoring. It effectively evaluates the quality and degree of hallucination in LVLMs' responses. MMHalBench reports the overall score (with a maximum of 6) and the hallucination rate (HalRate). It is important to note that the default evaluation GPT model, gpt-4-0314, is currently inaccessible, so we use gpt-4-0613 for the evaluation.

F Evaluation of Closed-Source Models and Grounding LVLMs

Apart from the grounding LVLMs like SPHINX, SPHINX-1k, and Qwen-VL-Chat already reported in Table 1, we have conducted further evaluations of GLaMM (Rasheed et al., 2024) (an open-source Grounding LVLM) and Qwen-VL-Plus (Bai et al., 2023) (a leading closed-source model known for its strong performance on grounding tasks). The results are shown in Table 10. It is important to note that for a fair comparison, we do not provide GLaMM with additional Region Input, which may be the reason for its suboptimal performance on ROHE. Although GLaMM performs excellently on grounding tasks, its performance on ROHE is not satisfactory. In contrast, Qwen-VL-Plus achieves remarkable performance.

supercategory	GLaMM		Qwen-VL-Plus	
	acc	acc+	acc	acc+
vehicle	100.0	5.58	92.94	71.38
sports	100.0	0.00	92.77	55.39
accessory	100.0	0.00	95.01	58.17
animal	100.0	3.91	99.27	86.80
food	100.0	2.26	95.48	61.54
outdoor	100.0	2.21	95.96	66.18
kitchen	100.0	0.00	82.75	65.64
electronic	100.0	2.57	88.65	67.67
furniture	99.11	12.30	84.12	61.74
indoor	100.0	2.49	85.36	61.37
appliance	98.86	11.74	89.02	52.27
person	99.75	21.47	97.20	82.59
total	99.84	6.03	91.41	67.17

Table 10: Evaluation results of leading closed-source models and grounding LVLMs on ROHE.

		LLaVA-1.5-7B					LLaVA-1.5-13B			
		base	VCD	EOS	DPO	oDPO	base	VCD	DPO	oDPO
vehicle	acc+	46.84	39.96	48.33	49.07	**69.89**	34.39	26.77	35.50	**58.74**
	acc	99.81	93.87	**100.0**	**100.0**	99.44	**100.0**	94.42	**100.0**	99.81
sports	acc+	25.51	19.65	25.65	27.29	**50.61**	8.73	7.64	9.96	**24.97**
	acc	**99.86**	97.14	**99.86**	**99.86**	98.77	**99.86**	98.77	99.73	99.73
accessory	acc+	16.62	17.45	13.02	15.51	**48.48**	9.14	11.36	8.86	**24.65**
	acc	**100.0**	98.34	**100.0**	**100.0**	99.45	**100.0**	99.17	100.0	99.72
animal	acc+	74.57	75.55	71.39	75.06	**86.80**	58.19	58.68	60.88	**77.02**
	acc	**100.0**	97.80	99.76	**100.0**	99.76	**99.51**	98.29	**99.51**	**99.51**
food	acc+	53.39	51.13	50.23	55.66	**72.40**	36.20	33.48	36.20	**52.94**
	acc	**99.55**	98.19	**99.55**	**99.55**	97.29	**100.0**	98.19	**100.0**	99.10
outdoor	acc+	22.79	23.90	19.49	20.59	**55.88**	15.81	18.38	14.34	**34.19**
	acc	**100.0**	98.90	**100.0**	**100.0**	98.90	**100.0**	99.63	**100.0**	**100.0**
kitchen	acc+	30.26	27.05	26.17	28.80	**52.78**	18.57	20.03	19.30	**37.28**
	acc	99.42	93.13	**99.56**	**99.56**	96.20	**100.0**	94.30	**100.0**	99.12
electronic	acc+	26.34	23.34	25.70	23.77	**53.96**	13.28	11.99	14.13	**31.05**
	acc	**100.0**	97.00	99.79	**100.0**	98.50	**100.0**	97.64	**100.0**	**100.0**
furniture	acc+	36.91	33.78	35.12	36.47	**59.28**	28.41	27.74	28.86	**42.73**
	acc	99.55	94.18	**99.78**	**99.78**	98.21	99.78	95.08	**100.0**	98.66
indoor	acc+	28.66	27.73	23.36	27.10	**54.21**	14.95	16.82	16.20	**33.02**
	acc	**100.0**	97.20	**100.0**	**100.0**	99.07	**100.0**	98.75	**100.0**	**100.0**
appliance	acc+	12.50	12.50	12.12	12.12	**38.64**	8.33	9.47	9.85	**20.45**
	acc	**100.0**	97.73	**100.0**	**100.0**	97.73	**100.0**	99.24	**100.0**	99.62
person	acc+	70.39	74.46	75.10	69.50	**82.59**	61.75	65.18	62.39	**75.86**
	acc	99.87	95.93	99.62	**100.0**	98.86	99.75	96.19	**99.87**	99.75
total	acc+	39.21	37.46	38.24	38.94	**61.65**	27.53	27.51	28.34	**44.71**
	acc	99.82	96.18	99.80	**99.89**	98.49	99.89	97.06	**99.91**	99.58

Table 11: Complete results on ROHE. The best results are shown in **bold**.

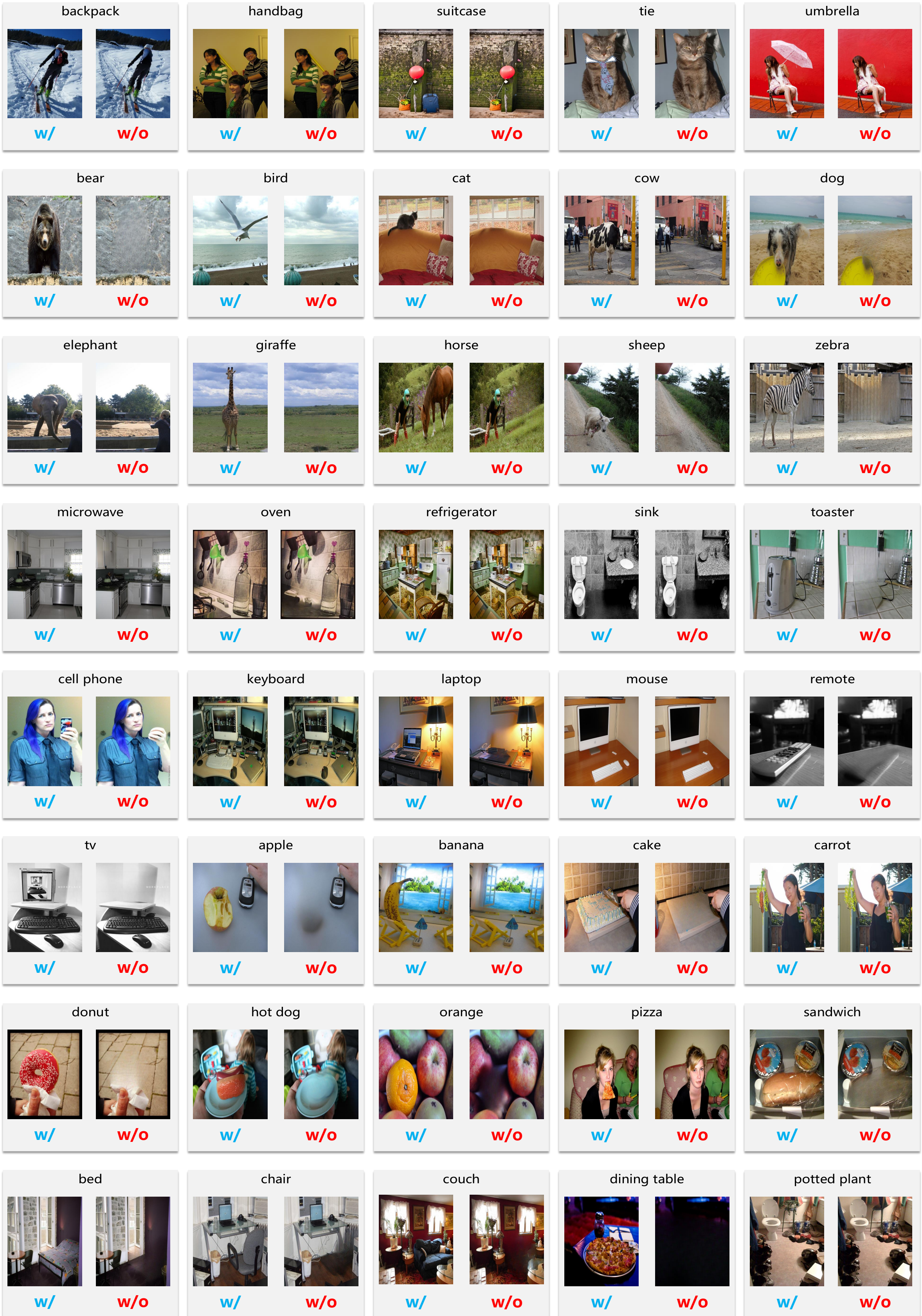

Figure 10: Examples of ROHE (Part I). The positive images (with objects) are labeled as w and the negative images (without objects) are labeled as w/o.

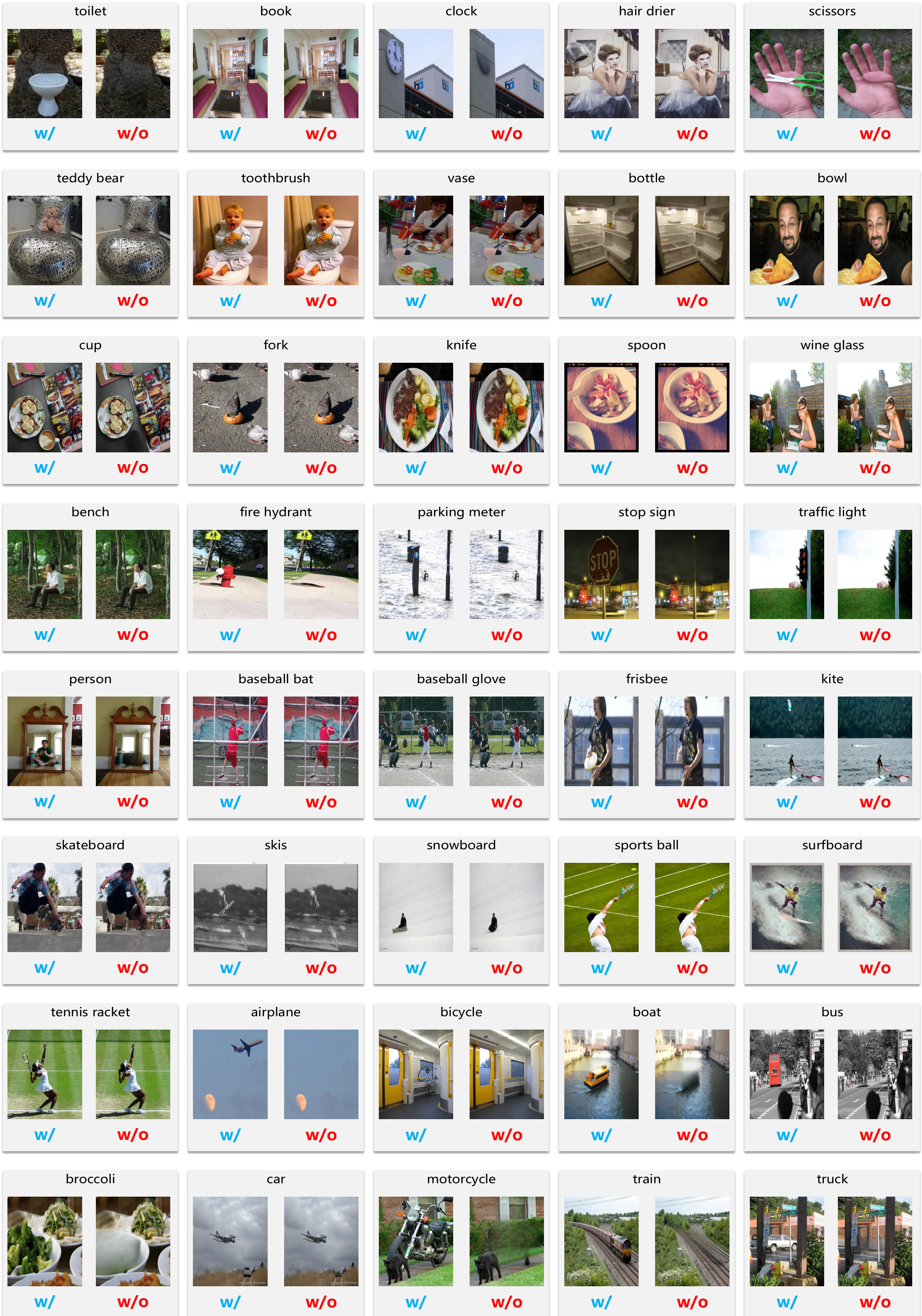

Figure 11: Examples of ROHE (Part II). The positive images (with objects) are labeled as **w** and the negative images (without objects) are labeled as **w/o**.