Detecting and Mitigating Hallucination in Large Vision Language Models via Fine-Grained AI Feedback

Wenyi Xiao[1*], Ziwei Huang[1*], Leilei Gan[1†], Wanggui He[2]
Haoyuan Li[2], Zhelun Yu[2], Fangxun Shu[2], Hao Jiang[2], Linchao Zhu[3]
[1]School of Software Technology, Zhejiang University
[2]Alibaba Group
[3]College of Computer Science and Technology, Zhejiang University
{wenyixiao, leileigan}@zju.edu.cn,

Abstract

The rapidly developing Large Vision Language Models (LVLMs) still face the *hallucination phenomena* where the generated responses do not align with the given contexts, significantly restricting the usages of LVLMs. Most previous work detects and mitigates hallucination at the coarse-grained level or requires expensive annotation (e.g., labeling by human experts or proprietary models). To address these issues, we propose detecting and mitigating hallucinations in LVLMs via fine-grained AI feedback. The basic idea is that we generate a small-size sentence-level hallucination annotation dataset by proprietary models, whereby we train a detection model which can perform sentence-level hallucination detection. Then, we propose a detect-then-rewrite pipeline to automatically construct preference dataset for hallucination mitigation training. Furthermore, we propose differentiating the severity of hallucinations, and introducing a Hallucination Severity-Aware Direct Preference Optimization (HSA-DPO) which prioritizes the mitigation of critical hallucination in LVLMs by incorporating the severity of hallucinations into preference learning. Extensive experiments on hallucination detection and mitigation benchmarks demonstrate that our method sets a new state-of-the-art in hallucination detection on MHaluBench, surpassing GPT-4V and Gemini, and reduces the hallucination rate by 36.1% on AMBER and 76.3% on Object HalBench compared to the base model.

Code — https://github.com/Mr-Loevan/HSA-DPO

1 Introduction

Large Language Models (LLMs) (OpenAI 2023a; Touvron et al. 2023; OpenAI 2023b; Jiang et al. 2023) have marked a significant milestone in the development of natural language processing and have been further extended to encompass multi-modality data, such as language and vision, leading to the emergence of Large Vision Language Models (LVLMs) (OpenAI 2023c; Liu et al. 2024b; Team et al. 2023; Bai et al. 2023). Despite the remarkable performance of LVLMs across a broad spectrum of vision-language tasks (Chen et al. 2023a; Liu et al. 2023; Zhu et al. 2024; Chen et al. 2023b; Dai et al. 2024; Guan et al. 2024a,b), LVLMs still grapple with the phenomena of *hallucination*, wherein the completions do not align with the given contexts. In other words, the generated responses contain incorrect objects, attributes and relations concerning the vision and language inputs, thereby significantly restricting the utility of LVLMs (Liu et al. 2024c; Yin et al. 2023a).

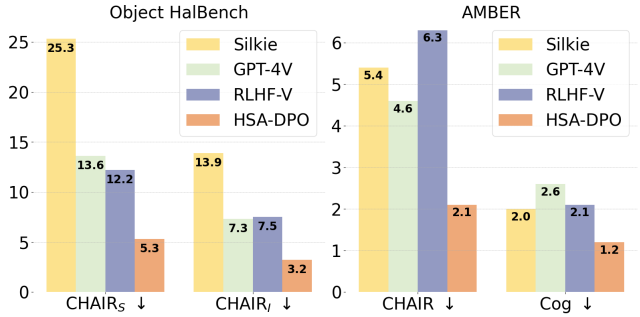

Figure 1: Comparison of HSA-DPO (red) with state-of-the-art models in mitigating hallucinations (Silkie, GPT-4V, RLHF-V) on Object HalBench and AMBER benchmarks. Notably, HSA-DPO outperforms state-of-the-art models in all metrics.

Research on addressing hallucinations in LVLMs can be primarily categorized into hallucination detection and mitigation. Hallucination detection aims to identify the presence of hallucinations in the LVLM outputs for preventing potential malicious usages (Li et al. 2023b; Wang et al. 2023; Rohrbach et al. 2018; Sun et al. 2023; Liu et al. 2024a; Chen et al. 2024). Hallucination mitigation aims to enable LVLMs to generate more faithful responses and can be mainly divided into training-free and training-based approaches (Yin et al. 2023b; Huang et al. 2023; Sun et al. 2023; Zhao et al. 2023; Yu et al. 2024; Gunjal, Yin, and Bas 2024; Zhou et al. 2024a). Training-free approaches address potential hallucinations by post-processing the outputs of LVLMs (Yin et al. 2023b; Huang et al. 2023; Zhou et al. 2024b; Han et al. 2024). While not requiring additional training costs, training-free approaches tend to reduce the inference speed. On the other hand, training-based approaches seek to reduce

*These authors contributed equally.
†Corresponding author.

hallucinations in LVLMs through further instruction fine-tuning (Liu et al. 2024a; Yue, Zhang, and Jin 2024) or preference learning (Sun et al. 2023; Yu et al. 2024; Li et al. 2023a; Zhao et al. 2023; Gunjal, Yin, and Bas 2024) on specifically constructed datasets. Some recent studies (Li et al. 2023a; Zhao et al. 2023; Gunjal, Yin, and Bas 2024) have exploited feedback from powerful closed-source LVLMs for improving the fidelity of LVLMs responses.

However, despite the aforementioned efforts, several challenges persist in detecting and mitigating hallucinations in LVLMs. First, the preference data is generally at response-level (Sun et al. 2023; Li et al. 2023a; Zhao et al. 2023; Gunjal, Yin, and Bas 2024), which is sub-optimal for thoroughly detecting and mitigating hallucinations. Second, constructing preference data for training-based mitigation approaches requires expensive annotations either by human experts (Sun et al. 2023; Yu et al. 2024) or proprietary commercial models (Li et al. 2023a; Zhao et al. 2023; Gunjal, Yin, and Bas 2024; Chen et al. 2024; Yu et al. 2023), especially if fine-grained annotation is involved. Lastly, existing studies often treat all hallucinations equally, leading to scenarios where less significant hallucinations are addressed, while more critical ones are neglected. For example, in certain scenarios, compared to incorrect color descriptions of objects, addressing the hallucinatory description of non-existent objects should be prioritized.

To address these issues, in this work, we propose detecting and mitigating hallucinations in LVLMs via fine-grained AI feedback. As shown in Figure 2, our framework consists of three key components: (1) **Fine-Grained AI Feedback.** The initial step involves generating a small-scale, sentence-level hallucination annotation dataset by proprietary models. Beyond merely detecting hallucinations, we meticulously craft prompts to collect detailed feedback on the type, severity and rationale of each hallucination. Compared to coarse-grained feedback, this sentence-level granularity ensures more precise and thorough hallucination detection. (2) **Fine-Grained AI Feedback for Hallucination Detection.** The next step proposes training a hallucination detection model using this fine-grained AI feedback, enabling it to perform sentence-level hallucination detection across primary types (e.g., object, attribute, and relationship). This step also introduces an automatic pipeline for constructing preference dataset where given a hallucinatory response, the detection model first identifies hallucinations within each sentence of the response. Based on the detected hallucinations, a rewriting model then revises the hallucinatory response into non-hallucinatory one, forming the <chosen_answer, rejected_answer> pair. This pipeline enables us to more cost-effectively annotate a large-scale preference dataset for training mitigation models. The underlying insight behind this approach aligns with the concept of scalable oversight which aims to train machines to assist humans in accurately evaluating model output (Bai et al. 2022; Lee et al. 2023; Ganguli et al. 2023; McAleese et al. 2024). (3) **Hallucination Severity-Aware DPO.** Lastly, we propose differentiating the severity of hallucinations, and introduce a Hallucination Severity-Aware Direct Preference Optimization (HSA-DPO). HSA-DPO incorporates hallucination severity into preference learning for prioritizing the mitigation of critical hallucinations.

We conduct extensive experiments on a range of hallucination detection and mitigation benchmarks and the experimental results have demonstrated the effect of the proposed method. For hallucination detection, our detection model achieves new state-of-the-art results on MHaluBench, surpassing GPT-4V and Gemini. As shown in Figure 1, for hallucination mitigation, HSA-DPO improves the base LVLM by reducing the Hallucination Rate on AMBER by 36.1% and CHAIR_S on Object HalBench by 76.3%. These results demonstrate the effectiveness of fine-grained AI feedback and the proposed HSA-DPO.

2 Related Work

In this section, we give a brief introduction of related studies on LVLMs hallucination detection and mitigation.

2.1 Detecting Hallucination in LVLMs

Current approaches for hallucination detection mainly focus on utilizing the abilities of off-the-shelf tools, such as closed-source LLMs, LVLMs or visual tools. GAVIE (Liu et al. 2024a) employs GPT-4 to facilitate the evaluation of object hallucinations. Zhao et al. (2023) introduce the sentence-level hallucination metric SHR, which harnesses GPT-4 to determine the presence of hallucinations in LVLM outputs. UNIHD leverages GPT-4V (OpenAI 2023c) or Gemini (Team et al. 2023) to extract verifiable claims from the generations of LVLMs, and then uses visual tools for hallucination detection. Compared to previous studies, our detection model is trained on fine-grained feedback from proprietary LVLMs, which covers main hallucination types (i.e., object, attribute, and relationship). It can also evaluate the severity of hallucinations and provide detailed reasons.

2.2 Mitigating Hallucination in LVLMs

Hallucination mitigation can be mainly divided into training-free and training-based approaches (Yin et al. 2023b; Huang et al. 2023; Sun et al. 2023; Zhao et al. 2023; Yu et al. 2024; Gunjal, Yin, and Bas 2024; Jing and Du 2024). Training-free approaches address potential hallucinations by post-processing the outputs of LVLMs (Yin et al. 2023b; Huang et al. 2023), thereby tending to reduce the inference speed of LVLMs. Instead, the latter reduce hallucinations in LVLMs via further training, such as instruction fine-tuning (Liu et al. 2024a) or preference learning (Sun et al. 2023; Yu et al. 2024; Li et al. 2023a; Zhao et al. 2023; Gunjal, Yin, and Bas 2024; Jing and Du 2024). Our work belongs to preference learning which biases LVLMs to favor the non-hallucinatory responses (Zhao et al. 2023; Gunjal, Yin, and Bas 2024; Sun et al. 2023; Yu et al. 2024). LLaVA-RLHF (Sun et al. 2023) is the first to train an LVLM to align with human preference. RLHF-V (Yu et al. 2024) manually collects segment-level human preference and conduct dense DPO over the human feedback to reduce hallucinations. POVID (Zhou et al. 2024a) constructs preference dataset by inserting textual hallucinations and distorting input images. Silkie (Li et al. 2023a) exploits feedback from various

LVLMs for constructing preference dataset, but the feedback are coarse-grained. In this study, we propose a pipeline for automatically constructing preference dataset and introduce the hallucination severity-aware preference learning for prioritizing the mitigation of critical hallucinations.

3 Methodology

In this section, we begin by introducing how to gather fine-grained AI feedback in §3.1. Following this, we detail how this fine-grained AI feedback is used for detecting and mitigating LVLM hallucinations in §3.2 and §3.3, respectively.

3.1 Fine-Grained AI Feedback Generation

Before introducing the method for gathering fine-grained AI feedback, we first detail the process of generating hallucinatory responses, upon which the collection of AI feedback is performed.

Hallucinatory Response Generation We investigate hallucination in the tasks of Detailed Description Generation (DDG) and Visual Complex Reasoning (VCR) following Liu et al. (2024a). These tasks require the LVLM to generate longer detailed description or complex reasoning response given the visual-language content, making the LVLM more susceptible to produce hallucinations. Note that our method is not limited to the two tasks but can be extended to other visual-language tasks.

We choose the Visual Genome (VG) (Krishna et al. 2017) and Silkie (Li et al. 2023a) dataset for constructing the DDG and VCR prompts, respectively. The images in VG are content-rich and associated with bounding boxes which specify various objects, attributes of each object, and spatial relationships within image content. These detailed annotations can help obtain more accurate AI feedback. Specifically, given a target LVLM M, for DDG, a randomly selected image from VG and an instruction from the instruction set in RLHF-V(Yu et al. 2024) are used as the prompt for M to generate a potentially hallucinatory response. Randomly choosing instruction from the instruction set injects randomness and harnesses the model to discern intricate image details. For VCR, we randomly select a <image, question> pair from the Silkie dataset as the prompt to generate a potentially hallucinatory response.

We denote the set of generated hallucinatory responses as $\mathcal{D}_{\text{hal}} = \{(x_i, \hat{y}_i)\}_{i=1}^{N}$ where $\hat{y}_i$ is the hallucinatory response, x_i is the corresponding prompt and N is the size of $\mathcal{D}_{\text{hal}}$.

Fine-Grained Hallucination Annotation via GPT-4/GPT-4V. Given the collected hallucinatory dataset $\mathcal{D}_{\text{Hal}}$, we can now gather fine-grained AI feedback upon it using GPT-4 and GPT-4V. The motivation behind this is that manually annotating large-scale datasets at fine-grained level is time-consuming, costly, and challenging.

Specifically, for each VCR hallucinatory sample $(x_i, \hat{y}_i)$ in $\mathcal{D}_{\text{Hal}}$, we input it into GPT-4V to generate fine-grained AI feedback at a rigorous sentence-level. For DDG, we provide $(x_i, \hat{y}_i)$ along with the associated verbal object bounding boxes in the VG dataset to GPT-4 as these additional annotations enable more accurate AI feedback compared to relying solely on the images. The used instruction prompt to generate the fine-grained AI feedback is shown in Figure 1 and 2 of the supplementary material. The obtained feedback is a six-tuple $(x_i, \hat{y}_i^j, h_{i,\text{type}}^j, h_{i,\text{R}}^j, \text{HS}_i^j, \text{HS}_{i,\text{R}}^j)$ where $\hat{y}_i^j$ is the j-th sentence of $\hat{y}_i$, $h_{i,\text{type}}^j$ is the hallucination type, $h_{i,\text{R}}^j$ is the reason that explains why $\hat{y}_i^j$ is considered a hallucination, HS_i^j is the hallucination severity score used to differentiate the effect of different hallucinations and $\text{HS}_{i,\text{R}}^j$ is the reason for hallucination severity score HS_i^j. The provided reasons $h_{i,\text{R}}^j$ and $\text{HS}_{i,\text{R}}^j$ improve the explainability of the hallucination detection process. Compared to coarse-grained feedback, this sentence-level granularity ensures a thorough hallucination detection.

For $h_{i,\text{type}}^j$, we consider the following types of hallucinations: (i) <object> for object hallucinations, such as perceiving physical entities that are not actually present; (ii) <relationship> for relationship hallucinations, such as giving inaccurate description of the relationship between objects; (iii) <attribute> for attribute hallucinations, such as inaccurate perceptions of the characteristics of objects. For the hallucination severity score HS_i^j, we define the following Likert-style ratings: (i) Minor (1 point): the hallucination concerns a minor detail and does not significantly affect the overall portrayal of the scene; (ii) Moderate (2 points): the hallucination involves a noticeable detail that is incorrect within the context of the scene, yet the overall comprehension of the scene is maintained; (iii) Major (3 points): the hallucination introduces a significant error or an entirely fabricated element that fundamentally alters the viewer's understanding of the scene. These scores facilitate the assessment of hallucination severity and are further incorporated into the preference learning(See §3.3) to prioritize the mitigation of critical hallucinations. We denote the set of fine-grained AI feedback as $\mathcal{D}_{\text{faif}}$.

3.2 Hallucination Detection via Fine-Grained AI Feedback

With the collected fine-grained AI feedback dataset $\mathcal{D}_{\text{faif}}$, we can then train a hallucination detection model using open-source LVLMs. Training an open-source model for hallucination detection offers the following merits. First, it enables perform fine-grained LVLMs hallucination detection with severity scores but not relies on proprietary models at lower cost. Second, the detected fine-grained hallucinations can be further used to construct a preference dataset, as discussed in §3.3.

Formally, given the sentence-level training dataset $\mathcal{D}_{\text{faif}} = \{(x^i, \hat{y}^i, h_{\text{type}}^i, h_{\text{R}}^i, \text{HS}^i, \text{HS}_{\text{R}}^i)\}_{i=1}^{M}$, we train fine-grained hallucination detection model $M_{\text{det}}(;\theta)$, parameterized by θ, by minimizing the negative log likelihood loss:

$$\mathcal{L}_{\text{DET}}(\theta) = -\sum_{i=1}^{M}\sum_{t=1}^{T} \log M_{det}(g_t^i | x^i, \hat{y}^i, g_{1:t}^i) \quad (1)$$

where g^i is concatenation of $h_{\text{type}}^i, h_{\text{R}}^i, \text{HS}^i, \text{HS}_{\text{R}}^i$. Note that for non-hallucinated sentences, $h_{\text{type}}^i, h_{\text{R}}^i, \text{HS}^i, \text{HS}_{\text{R}}^i$ are set

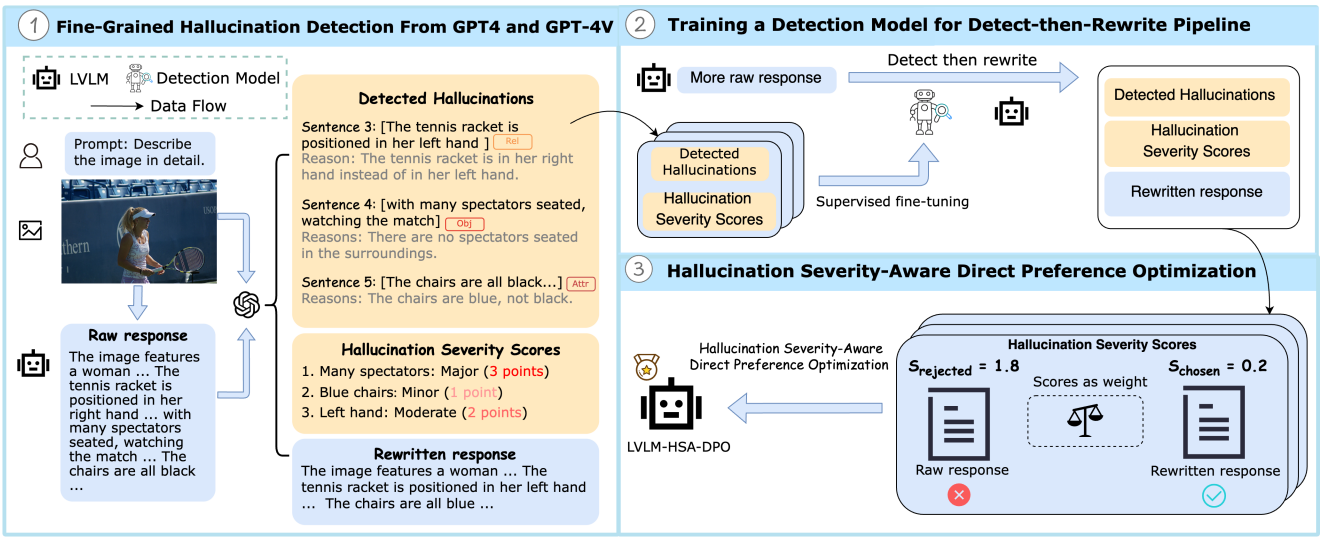

Figure 2: Our work consists of three components: § 3.1 fine-grained hallucination detection from GPT-4/GPT-4V; § 3.2 hallucination detection model for detect-then-rewrite preference dataset construction pipeline; and § 3.3 hallucination severity-aware direct preference optimization.

to <No hallucination, None, 0, None>, respectively. Our preliminary experiments indicate that the ratio of hallucinated to non-hallucinated data significantly impacts the performance of the detection model. We tested multiple ratios and found that a final ratio of 1:1.2 provides optimal results.

3.3 Hallucination Mitigation via Fined-Grained AI Feedback

With the built hallucination detection model, we next introduce an automated method for constructing the preference dataset.

Detect-then-Rewrite Pipeline for Automatic Preference Dataset Construction. To reduce the expensive annotation costs either caused by human experts or proprietary models (Sun et al. 2023; Li et al. 2023a; Gunjal, Yin, and Bas 2024), we propose a detect-then-write pipeline for automatic preference dataset construction, which allows for cost-effectively fine-grained feedback annotation at scale.

Specifically, the detect-then-rewrite pipeline consists of the hallucination detection model, M_{det}, and one rewriting model, M_{wri}. Given a prompt and its hallucinatory response $(x, \hat{y})$, the detection model first identifies a set of hallucination $\mathcal{H} = \{(h^j_{\text{type}}, h^j_{\text{R}}, \text{HS}^j, \text{HS}^j_{\text{R}})\}_{j=1}^{|\hat{y}|}$. Then, using $\hat{y}$ and $\mathcal{H}$ as the input prompt, M_{det} rewrites $\hat{y}$ into a non-hallucinatory response y. The specific rewriting prompt is provided in section A.1 of the supplementary material. In practice, we choose an open-source LVLM LLaVA as the rewriting model M_w, as it not only has demonstrated the impressive instruction-following and rewriting capability in our pilot experiments, but also offers a way to further reduce the annotation cost. We denote this preference dataset as $\mathcal{D}_{\text{pref}} = \{(x_i, \hat{y}_i, y_i, \mathcal{H}_i)\}_{i=1}^{N}$ where N is the dataset size.

Connection to Scalable Oversight. Compared with previous studies, this preference dataset construction pipeline enables us to more budget-friendly annotate a large-scale preference dataset for training mitigation models. The underlying insight behind this approach is closely correlated with the concept of *scalable oversight*, which aims to train machines to assist humans in supervising models by critiquing the model's output (Bai et al. 2022; Lee et al. 2023; Ganguli et al. 2023; McAleese et al. 2024) or decomposing complex problems into simpler sub-problems(Leike et al. 2018; Lightman et al. 2024). In this work, we break down the complicated fine-grained labeling process into two steps: first, detecting (critiquing) hallucinations in the generated completion, and then rewriting them into non-hallucinated ones. Moreover, we leverage the capabilities of current open-source LVLMs by employing them as hallucination detection and rewriting experts, thereby reducing the cost of providing a large-scale supervisions. Notably, our pipeline does not require any ground truth datasets, making it not only cost-effective but also practical for curtain scenarios where labeled datasets may be unavailable.

Hallucination Severity-Aware Direct Preference Optimization. With the automatically constructed preference dataset, we can now perform preference learning for hallucination mitigation. In particular, we choose the offline preference optimization method Direct Preference Optimization (DPO) (Rafailov et al. 2024; Xiao et al. 2024) as it is more stable and efficient compared to online RLHF methods(Ouyang et al. 2022; Schulman et al. 2017). The learning objective of DPO is directly formulated over the the policy model $\pi_\theta(y|x)$ and a reference model $\pi_{\text{ref}}(y|x)$:

$$\mathcal{L}_{\text{DPO}} = -\mathop{\mathbb{E}}_{(x,y_w,y_l)\in\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \tag{2}$$

where y_l and y_w are the rejected and chosen answer. σ is the logistic function. β is the hyper-parameter controlling the deviation from $\pi_{\text{ref}}(y|x)$. Action score $\log \pi(y|x)$ is the response generation likelihood.

As can be observed, the standard DPO loss $\mathcal{L}_{\text{DPO}}$ treats all pairwise preference responses equally, making more severe hallucinations (e.g., description of non-existent objects) not being greater considered compared with other hallucinations (e.g., incorrect color descriptions of objects). To tackle this limitation, we present Hallucination Severity-Aware Direct Preference Optimization (HSA-DPO), which incorporates hallucination severity into the preference optimization for mitigating critical hallucinations with higher priority. Specifically, we begin with aggregating the sentence-wise hallucination severity scores in $\mathcal{H}_i$ as the response-level severity score:

$$S^i_{\text{AVG}} = \frac{1}{T}\sum_{j=1}^{T} \text{HS}^j_i \tag{3}$$

where T is the number of sentences in the response which helps prevent potential reward hacking introduced by the length bias. This strategy bears similarity to our fine-grained hallucination detection, which can alleviate the difficulty of directly assessing the hallucination severity of the whole response. Subsequently, we adaptively assign this severity score to the implicit reward model of DPO to ensure responses with more severe hallucinations receive stronger penalties for correction:

$$\mathcal{L}_{\text{MIT}} = -\sum_{i=1}^{|\mathcal{D}_{\text{pref}}|} \log \sigma(\beta[\log \frac{\pi_\theta(y_i|x_i)}{\pi_{\text{ref}}(y_i|x_i)} - S^i_{\text{AVG}} \log \frac{\pi_\theta(\hat{y}_i|x_i)}{\pi_{\text{ref}}(\hat{y}_i|x_i)}]) \tag{4}$$

4 Experiments

In this section, we evaluate the efficacy of our method for detecting and mitigating hallucinations in LVLMs.

4.1 Datasets and Metrics

We introduce the datasets and metrics for evaluating hallucination detection and mitigation. For hallucination detection, we use the following benchmarks: (1) **MHaluBench**(Chen et al. 2024) is a newly established benchmark for detecting hallucination in both image-to-text and text-to-image settings as binary classification task. We adopt the image-to-text part to evaluate our detection model. (2) To evaluate various types of hallucination and their severity, we manually labeled a dataset named **MFHaluBench**, which includes object, attribute, and relationship hallucinations along with a human-annotated severity score for each segment. We evaluate MFHaluBench using binary classification to identify hallucinated segments and multi-class classification to distinguish between different types of hallucinations.

For hallucination mitigation, we use the following benchmarks: (1) **Object HalBench** (Rohrbach et al. 2018) is a widely adopted benchmark for evaluating object hallucination in DDG. Following Yu et al. (2024), we use CHAIR_S(i.e., the percentage of responses that contain hallucinations) and CHAIR_I (i.e., the percentage of hallucinated object mentions among all object mentions) as the evaluation metrics. (2) **AMBER** (Wang et al. 2023) consists of generative and discriminative parts, focusing on common objects and pitfall objects which easily cause hallucinations. We use the generative part of AMBER and report the following metrics: CHAIR, Cover, Hal and Cog. More details about these metrics can be found in section C.1 of the supplementary material. (3) **MMHal-Bench** (Sun et al. 2023) is a benchmark for evaluating object hallucination by using GPT-4 to compare the model output with the annotated response. We report overall score rated by GPT-4 and the hallucination rate. (4) **POPE** (Li et al. 2023b) is an object hallucination evaluation benchmark by testing LVLMs in a form of question answering. We choose the Adversarial part of POPE and report its F1 scores.

In addition to the above hallucination detection and mitigation benchmarks, we also adopt the widely used **LLava Bench in the wild**(Liu et al. 2024b) to evaluate the multimodal capabilities after mitigating training. We also introduce **Hallucination Severity Score**: a metric to evaluate the hallucination severity of model response. Severity scores ranges from 0 to 3. For complete definitions of these scores, refer to section A.2 of the supplementary material.

Methods	Levels	Average			
		Acc.	P	R	Mac.F1
Gemini with Self-Check	Claim	74.74	75.80	75.68	74.74
	Segment	75.11	73.89	77.44	73.85
Gemini with UNIHD	Claim	77.41	77.76	77.99	77.39
	Segment	78.68	75.97	78.64	76.74
GPT-4V with Self-Check	Claim	79.25	79.02	79.16	79.08
	Segment	80.80	77.80	78.30	78.04
GPT-4V with UNIHD	Claim	81.91	81.81	81.52	81.63
	Segment	84.60	82.77	80.89	81.71
Our Detection Model	Claim	**85.60**	**85.46**	**85.79**	**85.52**
	Segment	**86.94**	**87.73**	**82.88**	**85.23**

Table 1: Experimental results of MhaluBench on Image-to-Text hallucination detection. The results for Gemini and GPT-4V are sourced from the UNIHD(Chen et al. 2024).

Model	Binary				Multi
	P	R	ACC	F1	ACC
GPT-4V 2shot	59.7	98.7	63.3	74.4	40.6
LLaVA-1.6-34B 2shot	55.5	100	56.7	71.4	36.7
Our detection model	87.8	88.8	87.3	88.2	74.3

Table 2: Experimental results of MFHaluBench. Details of Multi refer to section B.3 of the supplementary material.

4.2 Baselines

For hallucination detection, we compare our method with GPT-4V(OpenAI 2023c), Gemini(Team et al. 2023) and UNIHD(Chen et al. 2024) following the experiment settings

Model	Object HalBench		AMBER				MMHal-Bench		LLaVA Bench	POPE Adv.
	$CHAIR_S\downarrow$	$CHAIR_I\downarrow$	CHAIR ↓	Cover. ↑	Hal.↓	Cog.↓	Overall ↑	Resp. ↓	Overall ↑	F1 ↑
LRV	32.3	22.3	-	-	-	-	-	-	-	-
POVID	48.1	24.4	7.3	49.5	31.1	3.7	2.08	0.56	-	81.6
InstructBLIP	25.9	14.3	8.8	52.2	38.2	4.4	2.14	0.58	-	78.4
Qwen-VL-Chat	36.0	21.3	6.6	53.2	31.0	2.9	2.89	0.43	79.8	82.8
LLaVA-1.5	46.3	22.6	7.8	51.0	36.4	4.2	2.42	-	72.5	84.5
LLaVA-RLHF	38.1	18.9	7.7	52.1	39.0	4.4	2.53	0.57	76.9	80.5
RLHF-V	12.2	7.5	6.3	46.1	25.1	2.1	2.81	0.49	59.7	-
GPT-4V	13.6	7.3	4.6	67.1	30.7	2.6	3.49	0.28	-	-
Silkie	25.3	13.9	5.4	55.8	29.0	2.0	3.01	0.41	84.9	82.1
DPO										
w/ Qwen-VL	14.3	8.0	3.8	53.2	19.7	1.8	2.98	0.38	82.0	82.6
w/ LLaVA-1.5	6.7	3.6	2.8	47.8	15.5	1.6	2.58	0.50	79.3	84.5
HSA-DPO										
w/ Qwen-VL	11.0	5.5	3.7	52.4	19.0	1.6	3.07	0.34	82.4	82.9
w/ LLaVA-1.5	**5.3**	**3.2**	**2.1**	47.3	**13.4**	**1.2**	2.61	0.48	80.5	84.9

Table 3: Main experimental results on hallucination mitigation. Note that LLaVA Bench denotes LLaVA Bench in the wild(Liu et al. 2024b). POPE Adv. denotes POPE Adversarial(Li et al. 2023b).

of (Chen et al. 2024).

For hallucination mitigation, we adopt a range of competitive hallucination mitigation methods as baselines: (1) InstructBLIP(Dai et al. 2024); (2) LLaVA 1.5(Liu et al. 2023); (3) Qwen-VL-Chat(Bai et al. 2023); (4) GPT-4V(OpenAI 2023c); (4) LRV (Liu et al. 2024a); (5) LLaVA-RLHF (Sun et al. 2023); (6) RLHF-V(Yu et al. 2024); (7) Silkie(Li et al. 2023a); (8) POVID (Zhou et al. 2024a).

4.3 Main Results

We report main results with respect to hallucination detection and mitigation, respectively.

Hallucination Detection. Table 1 and 2 report the main results on hallucination detection benchmarks. We can draw the following conclusions. First, our detection model achieves the state-of-the-art results on MHaluBench on average, outperforming GPT-4V and Gimini. Specifically, at the claim level, our detection model relatively surpasses UNIHD by 4.7% in Mac. F1 score and GPT-4V Self-Check 2-shot by 8.1% in Mac. F1 score. This improvement is consistently observed at the segment level as well. Second, on MFHaluBench, our detection model achieves an F1 score of 88.2% in binary classification, and an accuracy of 74.3% in Multi (fine-grained classification), outperforming GPT-4V 2-shot and LLaVA-1.6-34B 2-shot.

Hallucination Mitigation. Table 3 reports the main results on hallucination mitigation benchmarks. We can draw the following conclusions. First, HSA-DPO achieves state-of-the-art results on Object HalBench, outperforming the leading-edge closed-source LVLMs like GPT-4V. Second, HSA-DPO reduces the hallucination of LLaVA-1.5, our base model, by 76.3% for $CHAIR_S$ on Object HalBench and by 36.1% for Hal on AMBER. Third, compared to models trained on coarse-grained AI feedback (Silkie) or human-labelled fine-grained dataset (RLHF-V), our method gives superior results in hallucination mitigation. This demonstrates the effectiveness of using fine-grained AI feedback for detecting and mitigating hallucinations in LVLMs. Fourth, after preference learning, HSA-DPO is not affected by *alignment tax* (Askell et al. 2021; Ouyang et al. 2022) and maintains its multi-modality capabilities, as evidenced by the results on the overall metric of MMHal-Bench and LLaVA Bench in the wild. Lastly, to investigate the effect of our method on different base models, we train HSA-DPO on Qwen-VL-Chat(Bai et al. 2023), where the hallucinated responses are generated by Qwen-VL-Chat. Both DPO and HSA-DPO results in Table 3 indicate that our model gives a significant improvement over Qwen-VL-Chat, which demonstrate that our methodology can be applied to various LVLMs for mitigating hallucinations.

Methods	ObjHal		AMBER				HS↓
	C.s↓	C.i↓	C.↓	Cover.↑	Hal.↓	Cog.↓	
Ours	5.3	3.2	2.1	47.3	13.4	1.2	0.60
w/o Detection	42.1	20.3	7.6	52.1	32.4	4.0	0.79
w/o HSA	6.7	3.6	2.8	47.8	15.5	1.6	0.65
w/o FineGrained	17.0	8.9	5.0	53.6	27.3	1.7	0.67

Table 4: Ablation results on Object HalBench (ObjHal) and AMBER. HS denotes Hallucination Severity rated by GPT-4V. C.s, C.i, C. are short for $CHAIR_S$, $CHAIR_I$, CHAIR.

4.4 Ablations

We conduct ablation studies to validate the effectiveness of each design of our method. The results are reported in Table 4, where "w/o Detection" represents instead of detecting hallucinations by the detection model, we directly rewrite the hallucinated response into non-hallucinated one to construct preference dataset. "w/o HSA" represents that we do not incorporate hallucination severity into preference opti-

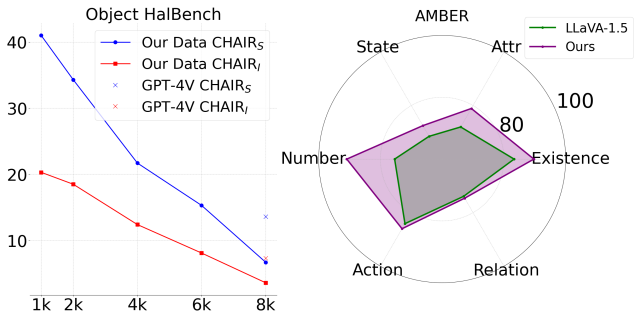

Figure 3: Effect of scaling preference dataset (Figure A) and different hallucination types (Figure B).

Feedback	Train Time	Efficiency	Cost
Human	0	20 s+	$4800
GPT-4V	0	8 s	$1600
Our pipeline	5.7 h	5 s	$600

Table 5: Annotation Efficiency and Cost. Efficiency is annotation time per sample and Cost is for collecting 16k preference datapoints.

mization and use the vanilla DPO. "w/o FineGrained" means we use coarse-grained GPT-4V feedback.

From the table, we have the following conclusion. First, without detection model identifying hallucinations, model performance seriously degrades on both benchmarks, highlighting the necessity of our detect-then-rewrite pipeline. Second, HSA-DPO demonstrate robust improvement over vanilla DPO in mitigating hallucinations across all hallucination metrics, relatively reducing by 20.8% for CHAIR$_S$ on Object HalBench, by 13.5% for Hal on AMBER and by 7.6% on Hallucination Severity. Third, fine-grained feedback outperforms coarse-grained one from GPT-4V, relatively reducing by 68.8% for CHAIR$_S$ on Object HalBench, by 50.9% for Hal on AMBER, which reveals the efficacy of fine-grained granularity.

4.5 Analyses

Annotation Efficiency and Cost. Our method constructs preference dataset using AI feedback through a detect-then-rewrite pipeline. Table 5 compares the annotation efficiency and cost with those of GPT-4V and human annotation. While our method needs 5.7 hours to train models, it only requires 25% and 62.5% of the time needed by human annotation and GPT-4V, respectively. Regarding annotation cost, our method incurs only $600, which is 12.5% and 37.5% of the cost needed by human annotation and GPT-4V, respectively. These results underscore our approach's superior efficiency and cost-effectiveness compared to other methods.

Effect of Scaling Preference Dataset. We conduct experiments to investigate the effect of scaling the preference dataset on Object HalBench. As depicted in Figure 3, we observe that with the increasing size of dataset, the hallucination rate of LLaVA with HSA-DPO shows a rapid and consistent decrease. When the size of the training dataset reaches 8k, our model surpasses GPT-4V on both CHAIR$_S$ and CHAIR$_I$, highlighting the value of our pipeline in enabling cost-effective annotation of preference datasets.

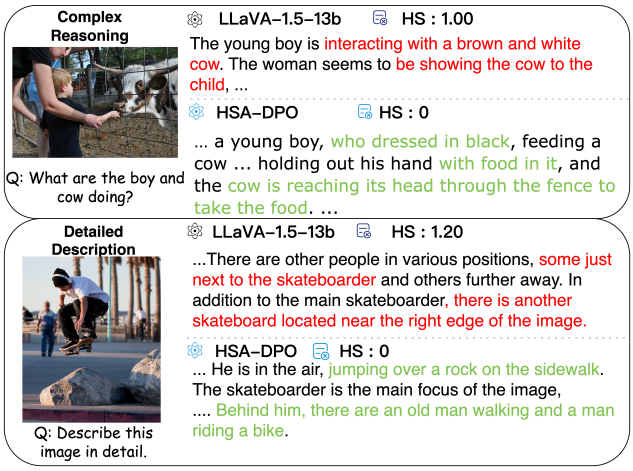

Figure 4: Qualitative results of different models on VCR and DDG. Correct answers, factual hallucinations are highlighted in red and green respectively.

Effect on Different Hallucination Types. To evaluate the effect of our method on different hallucination types, we conduct experiments on Amber benchmark and report the F1 scores on all types, i.e., object existence, attribute, state, number, action and relation. Figure 3 shows HSA-DPO with LLaVA-1.5 outperforms LLaVA-1.5 across all hallucination types. We also find that our method are more effective in improving number and existence types (object hallucinations) . However, we observe limited improvement in the relation and action types (relationship hallucinations). This is likely due to the scarcity of relationship hallucinations in preference dataset, as well as the inherent complexity of addressing relationship hallucinations compared to object ones.

4.6 Case Studies

In Figure 4, we qualitatively compare model performance on VCR and DDG. In VCR case, LLaVA struggles to recognize the key action of the boy feeding the cow. After HSA-DPO training, model accurately answers the question with more details. For DDG, LLaVA makes serious errors about nearby people and skateboard. However, model with HSA-DPO provides a precise description, accurately identifying relationships and objects in the image.

5 Conclusion

In this work, we propose detecting and mitigating hallucinations in LVLMs via fine-grained AI feedback. We begin with generating sentence-level hallucination annotation dataset via AI feedback. Then, a detect-then-rewrite pipeline is used to more cost-effectively construct preference dataset at scale. Lastly, HSA-DPO is introduced to incorporates hallucination severity into preference learning.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62441605), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C01142), and Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies.

References

Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023a. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*.

Chen, X.; Wang, C.; Xue, Y.; Zhang, N.; Yang, X.; Li, Q.; Shen, Y.; Gu, J.; and Chen, H. 2024. Unified Hallucination Detection for Multimodal Large Language Models. *arXiv preprint arXiv:2402.03190*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Muyan, Z.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2023b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Ganguli, D.; Askell, A.; Schiefer, N.; Liao, T. I.; Lukošiūtė, K.; Chen, A.; Goldie, A.; Mirhoseini, A.; Olsson, C.; Hernandez, D.; et al. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Guan, R.; Li, Z.; Tu, W.; Wang, J.; Liu, Y.; Li, X.; Tang, C.; and Feng, R. 2024a. Contrastive Multiview Subspace Clustering of Hyperspectral Images Based on Graph Convolutional Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–14.

Guan, R.; Tu, W.; Li, Z.; Yu, H.; Hu, D.; Chen, Y.; Tang, C.; Yuan, Q.; and Liu, X. 2024b. Spatial-Spectral Graph Contrastive Clustering with Hard Sample Mining for Hyperspectral Images. *IEEE Transactions on Geoscience and Remote Sensing*, 1–16.

Gunjal, A.; Yin, J.; and Bas, E. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18135–18143.

Han, Z.; Bai, Z.; Mei, H.; Xu, Q.; Zhang, C.; and Shou, M. Z. 2024. Skip \n: A Simple Method to Reduce Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2402.01345*.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2023. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jing, L.; and Du, X. 2024. FGAIF: Aligning Large Vision-Language Models with Fine-grained AI Feedback. *arXiv preprint arXiv:2404.05046*.

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

Lee, H.; Phatale, S.; Mansoor, H.; Lu, K.; Mesnard, T.; Bishop, C.; Carbune, V.; and Rastogi, A. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; and Legg, S. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

Li, L.; Xie, Z.; Li, M.; Chen, S.; Wang, P.; Chen, L.; Yang, Y.; Wang, B.; and Kong, L. 2023a. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2024. Let's Verify Step by Step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2024a. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

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

McAleese, N.; Pokorny, R. M.; Uribe, J. F. C.; Nitishinskaya, E.; Trebacz, M.; and Leike, J. 2024. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00215*.

OpenAI. 2023a. ChatGPT.

OpenAI. 2023b. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2023c. GPT-4V(ision) system card.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

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

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, J.; Wang, Y.; Xu, G.; Zhang, J.; Gu, Y.; Jia, H.; Yan, M.; Zhang, J.; and Sang, J. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Xiao, W.; Wang, Z.; Gan, L.; Zhao, S.; He, W.; Tuan, L. A.; Chen, L.; Jiang, H.; Zhao, Z.; and Wu, F. 2024. A Comprehensive Survey of Direct Preference Optimization: Datasets, Theories, Variants, and Applications. *arXiv preprint arXiv:2410.15595*.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2023a. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Yin, S.; Fu, C.; Zhao, S.; Xu, T.; Wang, H.; Sui, D.; Shen, Y.; Li, K.; Sun, X.; and Chen, E. 2023b. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Yu, T.; Hu, J.; Yao, Y.; Zhang, H.; Zhao, Y.; Wang, C.; Wang, S.; Pan, Y.; Xue, J.; Li, D.; et al. 2023. Reformulating vision-language foundation models and datasets towards universal multimodal assistants. *arXiv preprint arXiv:2310.00653*.

Yu, T.; Yao, Y.; Zhang, H.; He, T.; Han, Y.; Cui, G.; Hu, J.; Liu, Z.; Zheng, H.-T.; Sun, M.; et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13807–13816.

Yue, Z.; Zhang, L.; and Jin, Q. 2024. Less is More: Mitigating Multimodal Hallucination from an EOS Decision Perspective. arXiv:2402.14545.

Zhao, Z.; Wang, B.; Ouyang, L.; Dong, X.; Wang, J.; and He, C. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Zhou, Y.; Cui, C.; Rafailov, R.; Finn, C.; and Yao, H. 2024a. Aligning Modalities in Vision Large Language Models via Preference Fine-tuning. arXiv:2402.11411.

Zhou, Y.; Cui, C.; Yoon, J.; Zhang, L.; Deng, Z.; Finn, C.; Bansal, M.; and Yao, H. 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.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.