Adaptive Hallucination Alleviation in Multimodal Large Language Models: From Strategic Data Selection to Severity-Guided Training

Yuanyi Xu[1], Xiangru Zhu[1], Sihang Jiang[1*], Zhixu Li[2], Bei Yang[3], Xiaoxiao Xu[3], Yanghua Xiao[1*], Wei Wang[1]

[1]Shanghai Key Laboratory of Data Science, College of Computer Science and Artificial Intelligence, Fudan University
[2]School of Information, Renmin University of China
[3]Alibaba Group
yyxu24@m.fudan.edu.cn, {xrzhu19, shawyh, weiwang1}@fudan.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) have recently achieved strong performance across a variety of multimodal tasks. However, they still suffer from various forms of hallucination, which hinder their practical deployment. Prior approaches often struggle to efficiently construct high-quality hallucination-related samples and to process them in a fine-grained manner, resulting in limited effectiveness in hallucination alleviation. To address this issue, we propose a data sampling strategy that selects samples better suited for hallucination-oriented training, thereby enhancing training effectiveness. In addition, we introduce a quantitative method for measuring hallucination severity and assign individualized weights to training samples accordingly. Building on this, we present Hallucination-Differentiated Direct Preference Optimization (HD-DPO), a novel preference optimization framework. During fine-tuning, HD-DPO incorporates these weights into both the formulation of customized loss functions and the modulation of localized visual attention, enabling fine-grained optimization. Experimental results demonstrate that our method outperforms existing fine-tuning strategies across multiple benchmarks and generalizes well to diverse MLLM architectures, effectively reducing hallucination rates and enhancing overall model performance.

Introduction

Multimodal Large Language Models (MLLMs) have achieved astonishing results in cross-modal understanding and generation tasks (Achiam et al. 2023; Bai et al. 2023; Liu et al. 2023). Despite their impressive capabilities in visual understanding, MLLMs remain susceptible to "hallucinations", where generated responses contradict image content or objective facts (Bai et al. 2024; Liu et al. 2024b). Those hallucinations manifest in various forms—such as erroneous objects, attributes, locations, actions, quantities, or miscellaneous errors, limiting the further application of MLLMs (Chen et al. 2024a; Yu et al. 2024).

The generation of MLLM hallucinations is mainly attributed to the excessive reliance on pretrained textual corpora during inference, which can overshadow genuine visual cues and lead to incorrect outputs (Cui et al. 2023; Pi et al. 2024). This issue stems from a substantial modality gap between text and vision: insufficient alignment amplifies pretrained Large Language Model (LLM) biases and diminishes attention to visual inputs (Jiang et al. 2024).

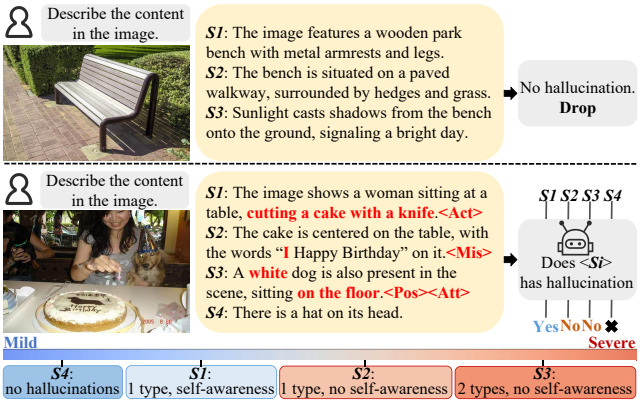

Figure 1: Filtering of image samples and distinction of the severity of hallucinations (sentence-level). Hallucinations and their types detected by GPT-4V are highlighted in red.

To reduce hallucinations, existing work has enhanced the cross-modal alignment capabilities of MLLMs through preference-based optimization techniques such as Direct Preference Optimization (DPO) (Rafailov et al. 2023). These methods differ in how they construct preference data. Some approaches rely on manual annotation to correct model outputs (Gunjal, Yin, and Bas 2024; Yu et al. 2024), while others leverage GPT to synthesize data by either injecting errors into responses to generate negative samples or correcting errors to obtain positive samples (Zhao et al. 2023; Fu et al. 2024).

However, these methods uniformly treat all samples during both pre-processing and training, leading to two key limitations. First, **not all collected samples constitute high-quality training data**. Many images, due to reasons such as trivial scene elements, only produce sparse and mild hallucinations when described by the model. Fine-tuning on such samples is inefficient and yields limited performance improvement. Second, **the varying severity of hallucinations is rarely considered during training**. When mild and

*Corresponding authors.

severe errors are treated equally, the model fails to focus extra effort on fixing the most critical hallucinations, allowing some errors to remain unaddressed after training.

To overcome these issues, we propose a two-stage strategy to better alleviate hallucinations in MLLMs, as illustrated in Figure 1. First, we design a quantitative criterion to select images with severe hallucinations, retaining unstable samples prone to hallucinations while eliminating stable ones unlikely to produce errors. Specifically, we inject noise into each image and compare the MLLM's responses before and after noise injection. Samples that severely deteriorate in both image-text matching and text perplexity are flagged as highly biased and chosen for training. These samples help construct strongly contrasting positive (non-hallucinatory)-negative (hallucinatory) response pairs, thereby enhancing the effectiveness of preference optimization. Second, we introduce a severity-aware hallucination quantification scheme. By analyzing hallucination types and applying a self-check mechanism, we compute a severity score for response sentence. This enables differentiated treatment of errors during training, allowing the model to allocate greater attention to more severe hallucinations and their corresponding visual regions. As a result, the model can better concentrate efforts to eliminate stubborn hallucinations and further reduce the residue of hallucinations.

Based on this strategy, we propose a severity-guided preference optimization approach. Compared with vanilla DPO, our approach refines both the loss computation and visual attention allocation. We refer to this method as Hallucination-Differentiated Direct Preference Optimization (HD-DPO).

Overall, our contributions are as follows:

- We design an effective sampling strategy that selects images prone to severe hallucinations from both textual and visual modalities. These samples are more suitable for hallucination alleviation, improving the training effect.
- We formalize hallucination severity at sentence-level and develop a practical metric for its quantification. This metric provides a comprehensive assessment by integrating both the hallucination categories and the internal cognition of model, without the reliance on human annotation.
- We introduce Hallucination-Differentiated Direct Preference Optimization (HD-DPO), a novel optimization method that assigns differentiated weights across visual and textual modalities during training, and demonstrate its efficiency through comprehensive experiments.
- We evaluate our approach on four representative MLLMs and conduct experiments across popular hallucination alleviation benchmarks. The results demonstrate that models fine-tuned with our data filtering and HD-DPO yield superior hallucination alleviation effect.

Related Work

Hallucination alleviation techniques have merged into two main paradigms: decoding-based and fine-tuning-based approaches. Decoding-based methods require no additional training and incur minimal overhead. OPERA (Huang et al. 2024) posits that low-information tokens (e.g., punctuation) can mislead subsequent predictions and trigger hallucinations, it therefore applies a retrospection–reallocation strategy to dynamically adjust attention weights during decoding. Visual Context Distillation (VCD) (Leng et al. 2024) contrasts output distributions from original and perturbed visual inputs to reduce reliance on statistical biases. HALC (Chen et al. 2024b) employs an adaptive focal-contrast mechanism to identify the optimal visual context for each potentially hallucinated token and a matching-based beam search to enforce global text–image consistency.

Fine-tuning-based strategies alleviate hallucinations by improving data collection and designing loss functions that bias models toward factual outputs. Gunjal, Yin, and Bas (2024) curate a dataset with segment-level human annotations for fine-grained model adjustment. Zhou et al. (2024a) inject noise into images and use GPT-4 to generate hallucinated responses, exposing the model to a broader error spectrum. Xie et al. (2024) extend text-contrastive fine-tuning to an image-contrastive framework, using visual cues to guide preference optimization. Ouali et al. (2024) leverage CLIP to score MLLM outputs, treating high and low scoring texts as positive and negative examples, respectively, and demonstrate substantial performance gains.

In contrast to prior work, our study emphasizes rigorous data filtering to exclude low-quality examples unsuited for hallucination alleviation fine-tuning. Furthermore, while Xiao et al. (2025) mentions use GPT to roughly classify the severity of hallucinations, we develop a comprehensive sentence-level severity metric and integrate it directly into our fine-tuning pipeline, achieving efficient and effective hallucination reduction.

Method

Task Formulation

Our objective is to fine-tune the MLLM using a preference dataset $\mathcal{D}$ in order to minimize hallucinations in the generated outputs. The overall structure of the proposed method is illustrated in Figure 2. We construct the dataset as $\mathcal{D} = \left\{\left(t^i, x^i, y_p^i, y_d^i, s^i\right)\right\}_{i=1}^{N}$, where t^i denotes the input text prompt, x^i is the image, y_p^i is the preferred (non-hallucinatory) response, y_d^i is the dispreferred (hallucinatory) response, and s^i is the associated sample weight score.

Image Selection

Existing preference datasets contain a large number of web-sourced images, many of which are visually simple and stable. Such images rarely induce significant hallucinations, making it difficult to construct high-quality preferred and dispreferred description pairs. Therefore, selecting unstable images that are prone to severe hallucinations can improve the efficiency of training. Inspired by settings in Leng et al. (2024), given the original images and prompts, we apply a Gaussian mask to each image to introduce visual uncertainty and accentuate the model's inherent text bias. By comparing the model's responses to the original and masked images, we identify and retain samples that exhibit significant deviations, thereby constructing a more focused and efficient

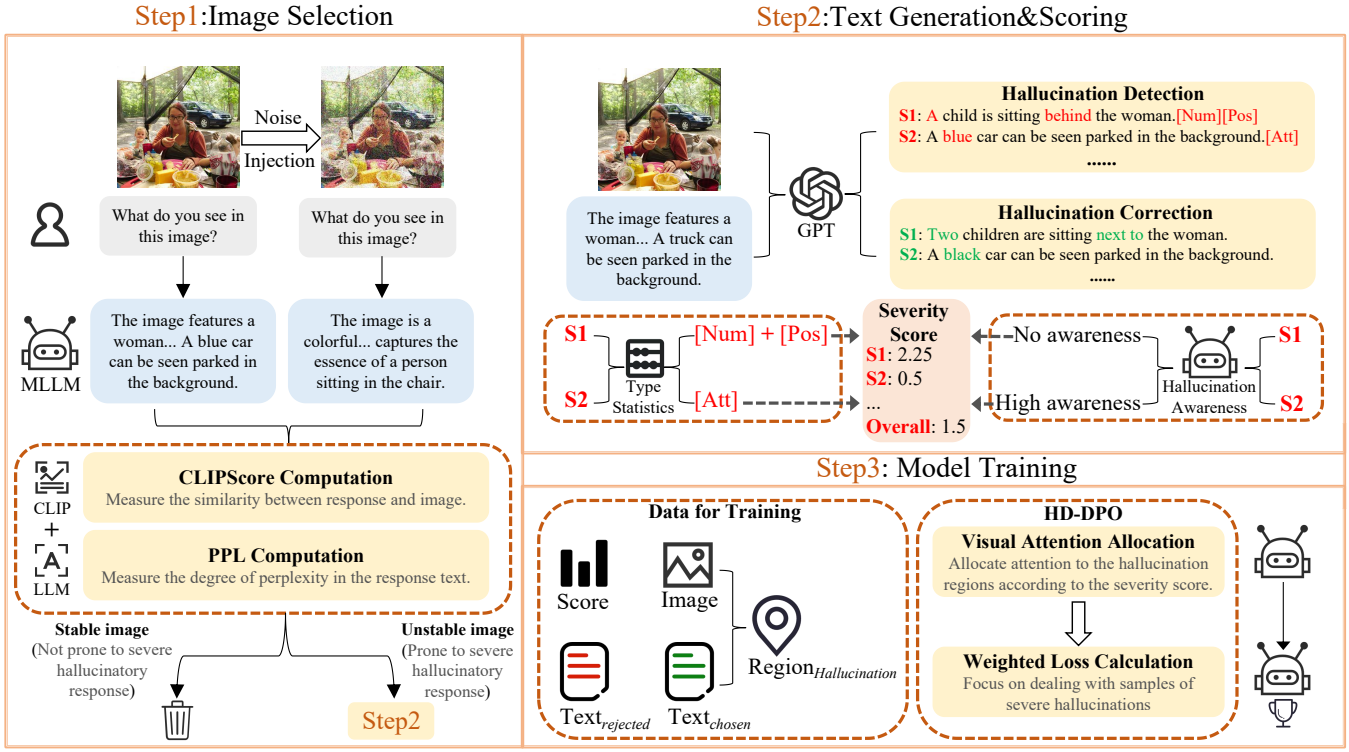

Figure 2: Overview of our framework. In the first step, we use CLIPScore and PPL metrics to screen suitable data samples. The second step is to annotate and correct the responses generated by the MLLM using GPT-4V. In the third step, we fine-tune MLLM using HD-DPO approach to alleviate the occurrence of hallucinations.

training dataset.

We aim to comprehensively evaluate images from two perspectives: the confidence level of the generated text, and the alignment between text and image. Denoting the original image and its noisy counterpart as x and $\dot{x}$, respectively, with corresponding model-generated descriptions y and $\dot{y}$, we compute the **Perplexity (PPL)** of y and $\dot{y}$ using the LLM component of the MLLM. PPL measures a language model's predictive ability over a token sequence, where lower PPL denotes greater confidence. Samples exhibiting a large increase in PPL after noise injection indicate that the model's description relies on unstable or ambiguous visual cues and is thus prone to hallucination. Such examples are particularly informative for fine-tuning, as they reveal the model's vulnerabilities and offer opportunities to teach it to produce more robust, grounded descriptions.

To evaluate image–text alignment, we adopt the **CLIPScore** (Hessel et al. 2021) metric (abbreviated to CLIP-S), computing CLIP-S(y, x) and CLIP-S$(\dot{y}, x)$. Higher CLIP-S reflects stronger semantic consistency. When comparing y to $\dot{y}$, a pronounced drop in CLIP-S implies that slight perturbations can distort the model's descriptions. This signals weak visual-semantic grounding, which is a key indicator of hallucination. Training with these samples encourages the model to maintain consistent, image-faithful outputs even under partially unreliable inputs.

Finally, we define a comprehensive score to assess each sample's suitability for hallucination alleviation fine-tuning:

$$Score_{sample} = \frac{PPL(\dot{y}) - PPL(y)}{PPL(y)} + \frac{CLIP\text{-}S(y,x) - CLIP\text{-}S(\dot{y},x)}{CLIP\text{-}S(y,x)} \quad (1)$$

PPL and CLIP-S are normalized to a dimensionless scale, making them comparable across different ranges of values. The resulting $\text{Score}_{\text{sample}}$ captures the stability and alignment consistency of the model when describing x. A large increase in PPL indicates low confidence in certain details or semantic boundaries, causing the model to vacillate among multiple plausible narratives and produce hallucinations. Conversely, a substantial drop in CLIP-S reveals over-reliance on subtle visual cues, which may not generalize and can trigger inaccurate descriptions derived from pretrained text biases. Thus, samples with higher $\text{Score}_{\text{sample}}$ are more prone to severe, content-violating hallucinations and expose critical weaknesses in the model's visual understanding. We therefore prioritize these cases for fine-tuning. After scoring, we selected the images that are most likely to induce severe hallucination descriptions to construct our dataset.

Text Generation&Scoring

Due to factors such as model size, MLLMs often generate descriptive text for x^i that contains hallucinations, denoted

as y_d^i. However, compared to this binary judgment of hallucination presence or absence, we argue that introducing a graded assessment of hallucination severity can provide richer annotations for each sample, thereby enabling more fine-grained processing during subsequent model training. To this end, we first use GPT-4V to correct y_d^i to obtain y_p^i, and then comprehensively assess the hallucination severity of y_d^i by jointly considering its observable hallucinatory information and the internal perception of MLLM.

Specifically, we feed (x^i, y_d^i) into GPT-4V with instructions to perform sentence-level hallucination detection, analysis, and correction based on the visual content of x^i. During detection, GPT-4V assigns one or more of six hallucination categories to each sentence:

- **Object**: Hallucination of the object itself, such as fabricating an object or mistaking one object for another.
- **Attribute**: Hallucination of object attributes, such as incorrect descriptions of object size or color.
- **Position**: Hallucination of the object spatial position, such as incorrectly describing the relative positional relationship between two objects.
- **Action**: Hallucination of object action, such as misunderstanding what is happening to an object.
- **Number**: Hallucination of object counting, such as miscounting the number of a certain object as less or more.
- **Miscellaneous type**: Hallucinations that do not belong to other types, such as optical character recognition errors.

For the remaining two tasks, hallucination analysis directs GPT-4V to justify why each identified error qualifies as a hallucination. This explanation both refines detection accuracy and supplies evidence for computing sentence-level hallucination weights. In hallucination correction, GPT-4V minimally edits the original response y_d^i to eliminate the detected errors, yielding the corrected response y_p^i. Manual evaluation shows that the detection and correction accuracy reaches 93.5%, confirming the feasibility of the approach.

Hallucination severity varies across samples, and samples exhibiting more severe hallucinations should be assigned higher severity scores. To quantify sentence-level severity, we introduce a dual-query self-check mechanism inspired by recent MLLM self-evaluation methods (Huo et al. 2024; Yan et al. 2024). First, we present MLLM with the image x^i and the jth hallucinatory sentence hs_j^i from y_d^i, and ask whether hs_j^i contains a hallucination. If recognized, we assign a self-check score of 0.5, indicating a mild hallucination. For sentences not flagged in this step, we augment the prompt with the analysis information provided by GPT-4V and query MLLM itself again. Sentences identified here receive a score of 1 (moderate hallucination), while those still undetected are scored 1.5 (severe hallucination). Tests show that overly close score intervals fail to adequately emphasize severe hallucinations, whereas excessively large intervals cause the model to neglect mild ones. The 0.5/1.0/1.5 configuration strikes an effective balance, and we therefore adopt it.

Another design choice motivated by empirical observations is that hallucination severity in a sentence is positively correlated with the number of hallucinations present. However, directly counting individual hallucinations can be ambiguous, as multiple errors related to the same object may be conflated. To address this, we compute the number of distinct types of hallucinations present in a sentence, rather than the total count. This approach offers a clearer and more practical measure of hallucination severity. Empirically, 6.1% of hallucinating sentences contain more than one type of hallucinations. We assign a base hallucination with count score of 1 for a single type and add 0.5 for each additional type. Furthermore, since "Object" errors represent the most fundamental category (Li et al. 2023), any sentence containing an "Object" hallucination receives a $\alpha\times$ multiplier on its count score.In our paper, we set α to 1.2 by comparing the occurrence probability of "Object" type with other types. After obtaining the self_check score and hallucination_number score, we multiply them to obtain the final weight score of a sentence and denote it as ws_j.

Next, we use the following formula to calculate the score s corresponding to a sample:

$$s = \frac{\sum_{j=1}^{K} t_j \cdot ws_j}{\sum_{j=1}^{K} t_j}. \tag{2}$$

where t_j is the number of tokens in the jth hallucinative sentence, and K is the total number of hallucinative sentences. In this way, we have obtained the final form of data $\mathcal{D} = \left\{\left(t^i, x^i, y_p^i, y_d^i, s^i\right)\right\}_{i=1}^{N}$.

Model Training

We aim to leverage the obtained severity scores to guide the model in placing greater emphasis on more severe hallucinations, while focusing on visual regions that are more susceptible to hallucination. This further reduces residual hallucinations. So during training, we explicitly incorporate the severity scores into the loss computation, and implicitly integrate them into the visual attention allocation process. This enables differentiated treatment of samples with varying levels of hallucination severity.

We adopt Direct Preference Optimization (DPO) (Rafailov et al. 2023) as our preference tuning method, which directly maximizes the reward gap between preferred and dispreferred responses while optimizing the output probabilities of the model. The formula used to calculate the maximum likelihood objective of DPO is as follows:

$$\begin{aligned}\mathcal{L}_{DPO} = -\mathbb{E}_{(t,x,y_p,y_d)\sim D}[&\log \sigma \\ &\left(\beta \log \frac{\pi_\theta\left(y_p \mid t, x\right)}{\pi_{\text{ref}}\left(y_p \mid t, x\right)} - \beta \log \frac{\pi_\theta\left(y_d \mid t, x\right)}{\pi_{\text{ref}}\left(y_d \mid t, x\right)}\right)]\end{aligned} \tag{3}$$

where π_{ref} is the reference model and π_θ is the policy model.

DPO treats all samples equally in its loss calculation, preventing severe hallucinations from receiving additional focus and allowing them to persist post-training. To overcome this, we propose Hallucination-Differentiated DPO (HD-DPO), which augments DPO with both explicit and implicit weighting, respectively for the sample itself and the local region in the sample image. After dataset construction,

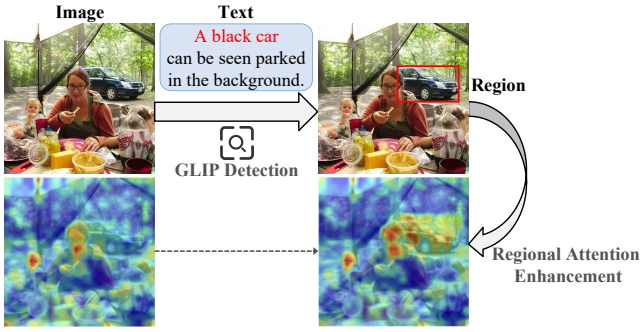

Figure 3: An example of implicit improvement. The local area corresponding to the detected target "car" receives more attention by multiplying the weighted score.

each sample is assigned a hallucination severity score s^i. We integrate s^i into the standard DPO objective, enabling the model to de-emphasize minor hallucinations and concentrate its capacity on alleviating the most severe errors. The improved formula is as follows:

$$\mathcal{L}_{HD-DPO(explicit)} = -\mathbb{E}_{(t,x,y_p,y_d,s)\sim D}[\log \sigma \left(\beta \log \frac{\pi_\theta\left(y_p \mid t, x\right)}{\pi_{\text{ref}}\left(y_p \mid t, x\right)} - s \cdot \beta \log \frac{\pi_\theta\left(y_d \mid t, x\right)}{\pi_{\text{ref}}\left(y_d \mid t, x\right)}\right)] \quad (4)$$

we refer to the above method as explicit improvement, as it directly incorporates severity scores into the DPO objective.

In contrast, the implicit improvement is not reflected in the loss function, but in the visual attention allocation during the image encoding stage. Specifically, before training on image x^i, we extract corrected sentences from the positive response y_p^i and use them as textual guidance. With the powerful vision-language pre-trained model GLIP (Li et al. 2022), we perform object detection on x^i to localize regions corresponding to these sentences. These detected regions are prone to hallucinations when describing the visual content of x^i. When encoding x^i with the image encoder, we amplify attention weights for each region in proportion to the sample's severity score s^i. Let M denote the number of image patches produced by the image encoder, $E = [E_0; E_1; \ldots; E_M]$ represent the matrix of patch embeddings, where E_0 is the embedding of the [*CLS*] token, and $\mathcal{R}$ denotes the set of patch indices corresponding to the detected regions. The embeddings are then modulated as follows:

$$\widetilde{E}_k = \begin{cases} s^i E_k, & k \in \mathcal{R}, \\ E_k, & k \notin \mathcal{R}, \end{cases} \quad k = 0, 1, \ldots, M \quad (5)$$

This strategy guides the model to focus more on these critical regions in subsequent self-attention computations. By biasing the model toward these critical regions, we encourage more accurate visual grounding and further reduce hallucination generation. An example of implicit improvement is shown in Figure 3.

Experiments

Benchmarks. We collect the following 4 different hallucination benchmarks for evaluation:

1. **AMBER** (Wang et al. 2023) is a fully automated benchmark requiring no LLM-based annotation, covering both generative and discriminative tasks and providing fine-grained hallucination metrics across existence, attribute, and relation dimensions.
2. **MMHal-Bench** (Sun et al. 2024) is a specialized evaluation suite of 96 challenging image–question pairs spanning eight categories, which penalizes hallucinations via human-augmented RLHF alignment and includes ground-truth answers and image content.
3. **CHAIR** (Rohrbach et al. 2018) is a free-form image-captioning dataset for object hallucination detection, measuring the rate of non-existent object mentions to assess the factual grounding of multimodal generation.
4. **HallusionBench** (Guan et al. 2024) is a discriminative benchmark comprised of 346 images and 1,129 controlled yes/no question pairs, designed to disentangle language hallucination and visual illusion failure modes and quantitatively analyze model biases and error types.

To more comprehensively measure the performance of MLLMs, we additionally introduce a comprehensive MLLM benchmark, MMBench (Liu et al. 2024c), to holistically assess MLLM' perception and reasoning abilities by converting free-form outputs into predefined choices and applying strict quality-control schemes for robust evaluation.

Baselines. We use LLaVA-1.5-7B as our main backbone model. In addition, we also evaluate other widely used baseline models including Qwen-VL (Bai et al. 2023), Instructblip (Dai et al. 2023), and InternLM-XComposer2 (Dong et al. 2024), and compare our HD-DPO with other preference-tuning methods based on LLaVA, including RLHF (Sun et al. 2024), CSR (Zhou et al. 2024b), POVID (Zhou et al. 2024a) and V-DPO (Xie et al. 2024).

Implementation Details. Considering the widespread adoption of the LLaVA-Instruct-150K dataset (Liu et al. 2024a; Park et al. 2025), we utilize its "detail_23k" subset as our source of image–prompt pairs. In the Image Selection stage, we follow the scale of dataset in Yu et al. (2024), filtering our training set from 23k to 5k samples. Training proceeded for 2 epochs with a batch size of 64, a learning rate of 2e-6, zero weight decay, a LoRA rank of 64, a dropout rate of 0.05, and a beta value of 0.1.

Main Results

The results in Table 1 demonstrate that HD-DPO substantially enhances LLaVA-1.5-7B across both generative and discriminative hallucination benchmarks. Notably, we train exclusively on descriptive data—without any QA examples—yet observe marked gains in discrimination tasks. This confirms that HD-DPO fundamentally strengthens the model's visual–semantic alignment and reduces hallucinations by improving its overall visual understanding.

In generative evaluations, HD-DPO achieves a 20.9% relative improvement on MMhal-Bench, indicating a markedly better ability to avoid hallucinations in open-ended visual question answering. On CHAIR (both CHAIR_S and

Model/Method	AMBER				MMHal-Bench		CHAIR		HallusionBench	MMBench
	Acc.	F1	CHAIR ↓	Score	Hal. ↓	Score	$CHAIR_S$ ↓	$CHAIR_I$ ↓	aAcc.	Score
LLaVA-1.5	72.3	75.6	7.6	84.0	0.58	2.15	47.3	24.3	41.82	63.3
+RLHF	73.4	77.8	7.2	85.3	0.56	2.18	39.2	20.2	38.15	63.2
+CSR	74.1	77.4	3.8	86.8	0.55	2.21	19.6	7.4	41.97	**64.1**
+POVID	73.2	76.4	5.6	85.4	0.52	2.28	33.2	10.3	42.05	63.7
+V-DPO	76.9	81.6	5.6	88.0	0.52	2.26	16.7	6.2	**46.20**	63.8
+HD-DPO(ours)	**79.7**	**83.2**	**2.0**	**90.6**	**0.41**	**2.60**	**4.1**	**2.8**	45.46	64.0

Table 1: Performance comparison between HD-DPO and other baselines on different benchmarks.

Model/Method	AMBER				MMHal-Bench		CHAIR		HallusionBench	MMBench
	Acc.	F1	CHAIR ↓	Score	Hal. ↓	Score	$CHAIR_S$ ↓	$CHAIR_I$ ↓	aAcc.	Score
Qwen-VL	81.0	86.3	6.7	89.8	0.44	2.82	39.7	22.0	35.27	60.8
+HD-DPO	**81.6**	**86.7**	**3.3**	**91.7**	**0.33**	**3.09**	**10.3**	**5.8**	**36.07**	**61.9**
InstructBLIP	78.1	83.2	9.4	86.9	0.59	2.06	21.0	12.7	31.23	38.4
+HD-DPO	**78.9**	**84.3**	**3.7**	**90.3**	**0.47**	**2.31**	**4.2**	**4.0**	**31.99**	**40.5**
InternLM-XComposer2	87.1	90.0	4.0	93.0	0.33	3.29	9.3	6.1	35.98	78.1
+HD-DPO	**88.7**	**91.5**	**3.4**	**94.0**	**0.30**	**3.37**	**5.0**	**3.3**	**37.13**	**78.3**

Table 2: Performance comparison of models before and after HD-DPO training.

$CHAIR_I$), HD-DPO reduces the rates of hallucinated sentences and objects to 4.1 and 2.8, respectively, reflecting enhanced accuracy in free-form image descriptions.

For discrimination tasks, HD-DPO attains an accuracy of 79.7% and an F1 score of 83.2% on AMBER, setting a new state-of-the-art. Although HD-DPO does not outperform specialized QA-optimized approach V-DPO on the more challenging HallusionBench, it still surpasses most fine-tuning baselines, achieving a 3.64-point improvement. On the comprehensive MMBench, HD-DPO's performance is only marginally below the best existing method, demonstrating significant gains in both visual cue perception and complex semantic reasoning. These results further validate the effectiveness of our method in alleviating hallucinations and bolstering overall model competence.

To validate the generality of our approach, we apply HD-DPO to three additional open-source MLLMs, namely Qwen-VL, InstructBLIP, and InternLM-XComposer2, and report the results in Table 2. Although the magnitude of improvements differ by architecture, HD-DPO consistently reduces hallucinations across all models. Qwen-VL, initially strong in discriminative alignment, sees its largest gains on generative benchmarks (e.g., CHAIR, MMHal-Bench) once HD-DPO emphasizes highly hallucinated examples. InstructBLIP is the model with the weakest initial performance, it benefits from severity-weighted optimization, achieving a balanced enhancement in both fluency and factuality that surpasses what pure generation or discrimination tuning can deliver alone. InternLM-XComposer2, which combines robust cross-modal fusion with reasoning capabilities but originally lagged in fine-grained discrimination, gains most on discrimination benchmarks (e.g., AMBER, HallusionBench) by penalizing severe hallucinations more heavily. These results confirm that HD-DPO is broadly applicable and effective across diverse MLLM architectures.

Analysis

Ablation Study To verify the effectiveness of each step, we design a series of ablation experiments. Starting from applying the original DPO with random 5k data, we successively add the step of image selection, apply the explicit HD-DPO, and apply the complete HD-DPO. The results are shown in Table 3, from which we can draw the following conclusions: Firstly, adding a data filtering step when building the dataset has improved the metrics on all benchmarks. On AMBER, accuracy rises by 0.8 and F1 by 1.1. On HallucsionBench, all accuracy increases by 0.57. This confirms that data screening based on CLIP-S and PPL outperforms random sampling or off-the-shelf dataset usage, enhancing discrimination without altering the optimization method. Secondly, after modifying DPO to explicit HD-DPO, the model's metrics on the generation task have significantly improved. CHAIR on AMBER drops by 0.4, and on MMhal-Bench the hallucination rate and overall score improve by 0.05 and 0.14, respectively. It is indicated that introducing the severity of hallucinations into preference optimization can alleviate various types of hallucinations represented by object hallucinations. Thirdly, the visual attention allocation mechanism enables the model to focus more on key objects during training and ignore irrelevant information, thereby making it easier to distinguish and avoid hallucinations. This is reflected in an increase of 0.06 in AMBER score and a decrease of 0.03 in the hallucination rate on MMhal-Bench. The overall score on MMhal-Bench has also increased by 0.09. A slight dip on MMBench suggests that attention reallocation can sometimes overlook relevant cues, but this trade-off is minor relative to the overall gains.

Effects of Hallucination Categories To assess the impact of individual hallucination categories, we assembled 1k samples for each type and fine-tuned LLaVA-1.5 separately. Table 4 reports results on three benchmarks. Dif-

Model/Method	AMBER				MMHal-Bench		CHAIR		HallusionBench	MMBench
	Acc.	F1	CHAIR ↓	Score	Hal. ↓	Score	$CHAIR_S$ ↓	$CHAIR_I$ ↓	aAcc.	Score
LLaVA-1.5	72.3	75.6	7.6	84.0	0.58	2.15	47.3	24.3	41.82	63.3
+DPO(w/o *IS*)	77.6	80.8	2.6	89.1	0.51	2.32	6.1	4.2	43.72	63.7
+DPO	78.4	81.9	2.5	89.7	0.49	2.37	5.6	4.1	44.29	63.9
+HD-DPO(explicit)	78.8	82.1	2.1	90.0	0.44	2.51	4.5	3.0	44.64	**64.3**
+HD-DPO	**79.7**	**83.2**	**2.0**	**90.6**	**0.41**	**2.60**	**4.1**	**2.8**	**45.46**	64.0

Table 3: Ablation study on the influence of Image Selection and HD-DPO. *IS* refers to Image Selection.

Model/Method	AMBER				MMHal-Bench		CHAIR	
	Acc.	F1	CHAIR ↓	Score	Hal. ↓	Score	$CHAIR_S$ ↓	$CHAIR_I$ ↓
LLaVA-1.5	72.3	75.6	7.6	84.0	0.58	2.15	47.3	24.3
+HD-DPO(Miscellaneous type)	73.0	76.6	5.7	85.5	0.55	2.21	38.6	21.3
+HD-DPO(Number)	73.6	77	5.9	85.6	0.54	2.22	40.7	22.4
+HD-DPO(Action)	73.4	77.1	6.4	85.4	0.55	2.23	43.0	24.5
+HD-DPO(Position)	73.7	77.3	5.8	85.8	**0.53**	2.28	42.0	23.0
+HD-DPO(Attribute)	**74.0**	**77.7**	5.5	**86.1**	0.58	2.16	37.7	20.0
+HD-DPO(Object)	73.5	77.3	**5.1**	**86.1**	**0.53**	**2.30**	**36.7**	**19.5**

Table 4: A comparative analysis of hallucination categories in preference optimization.

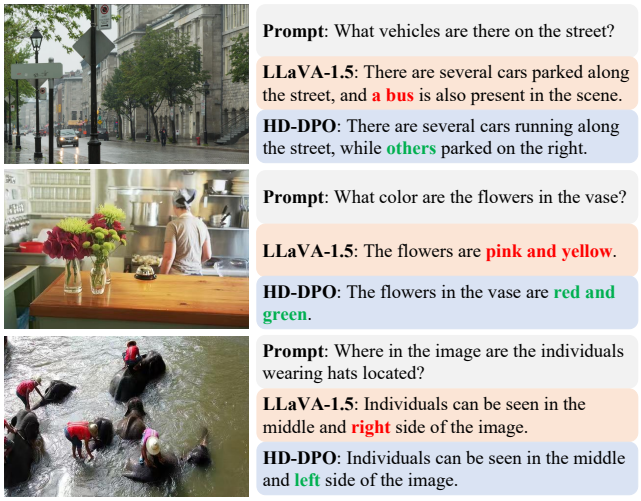

Figure 4: Case study on different types of hallucinations.

ferent hallucination types yield distinct effects: "Miscellaneous", "Action", and "Number" errors exert relatively minor influence. From a visual perspective, Gestalt principles (Todorovic 2008) suggest the human visual system prioritizes overall structure and spatial relationships before actions or quantities. Cognitively, spatial and attribute information are encoded earlier and more stably than actions or counts, which require object recognition and high-level semantic inference (Damasio and Tranel 1993). Accordingly, we term these three categories "Secondary Types", whereas "Object", "Attribute", and "Position" are "Primary Types".

Among primary types, "Position" errors yield significant improvement on MMhal-Bench, indicating that fine-tuning helps the model decouple "where" from "what", reducing spatial mismatches and enhancing spatial reasoning. "Attribute" hallucinations produce the largest gain on AMBER, suggesting that training on subtle attribute discrepancies sharpens the model's alignment of visual features with attribute vocabulary. "Object" hallucinations have the greatest overall impact, only inferior to "Attribute" type in the discrimination sub-task of AMBER, which underscoring object recognition as the cornerstone of visual-language understanding. Enhancing basic object classification thus propagates benefits across downstream tasks, validating "Object" as the most critical hallucination category to address.

Case Study To further validate HD-DPO, we select some QA pairs targeting the three primary hallucination types, namely "Object", "Attribute", and "Position", and compare LLaVA-1.5's outputs before and after HD-DPO fine-tuning. As shown in Figure 4, in the top example, HD-DPO removes the hallucinated object "bus" and generates a more precise description of the car's action. In the middle example, it corrects the car colors from pink and yellow to red and green. In the bottom example, HD-DPO enables the model to answer a descriptive question correctly by recognizing the spatial relationship between the individuals. These cases demonstrate that HD-DPO substantially reduces diverse hallucination types and improves performance across various tasks.

Conclusion

In this work, we first introduce a data sampling strategy to select suitable data for the training of hallucination alleviation through the calculation of CLIP-S and PPL. Then, we propose HD-DPO, calculate the severity of hallucination in the responses of MLLMs from multiple dimensions, and introduce it into loss calculation and attention allocation during training, enabling the model to pay more attention to severe samples. Extensive experiments conducted on multiple benchmarks have demonstrated that our method successfully reduces multiple types of hallucinations in MLLMs.

Acknowledgements

This work is supported by Alibaba Group through the Alibaba Innovation Research Program.

References

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

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, Z.; Wang, P.; Xiao, T.; He, T.; Han, Z.; Zhang, Z.; and Shou, M. Z. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Chen, X.; Wang, C.; Xue, Y.; Zhang, N.; Yang, X.; Li, Q.; Shen, Y.; Liang, L.; Gu, J.; and Chen, H. 2024a. Unified Hallucination Detection for Multimodal Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3235–3252.

Chen, Z.; Zhao, Z.; Luo, H.; Yao, H.; Li, B.; and Zhou, J. 2024b. HALC: object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning*, 7824–7846.

Cui, C.; Zhou, Y.; Yang, X.; Wu, S.; Zhang, L.; Zou, J.; and Yao, H. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 49250–49267.

Damasio, A. R.; and Tranel, D. 1993. Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences*, 90(11): 4957–4960.

Dong, X.; Zhang, P.; Zang, Y.; Cao, Y.; Wang, B.; Ouyang, L.; Wei, X.; Zhang, S.; Duan, H.; Cao, M.; et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Fu, Y.; Xie, R.; Sun, X.; Kang, Z.; and Li, X. 2024. Mitigating Hallucination in Multimodal Large Language Model via Hallucination-targeted Direct Preference Optimization. *arXiv preprint arXiv:2411.10436*.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14375–14385.

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.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13418–13427.

Huo, F.; Xu, W.; Zhang, Z.; Wang, H.; Chen, Z.; and Zhao, P. 2024. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*.

Jiang, C.; Xu, H.; Dong, M.; Chen, J.; Ye, W.; Yan, M.; Ye, Q.; Zhang, J.; Huang, F.; and Zhang, S. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27036–27046.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13872–13882.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10965–10975.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 292–305.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

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

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, 216–233. Springer.

Ouali, Y.; Bulat, A.; Martinez, B.; and Tzimiropoulos, G. 2024. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *European Conference on Computer Vision*, 395–413. Springer.

Park, Y.; Lee, D.; Choe, J.; and Chang, B. 2025. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6434–6442.

Pi, R.; Han, T.; Xiong, W.; Zhang, J.; Liu, R.; Pan, R.; and Zhang, T. 2024. Strengthening multimodal large language model with bootstrapped preference optimization. In *European Conference on Computer Vision*, 382–398. Springer.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4035–4045.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.; Wang, Y.-X.; Yang, Y.; et al. 2024. Aligning Large Multimodal Models with Factually Augmented RLHF. In *Findings of the Association for Computational Linguistics ACL 2024*, 13088–13110.

Todorovic, D. 2008. Gestalt principles. *Scholarpedia*, 3(12): 5345.

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

Xiao, W.; Huang, Z.; Gan, L.; He, W.; Li, H.; Yu, Z.; Shu, F.; Jiang, H.; and Zhu, L. 2025. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25543–25551.

Xie, Y.; Li, G.; Xu, X.; and Kan, M.-Y. 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*, 13258–13273.

Yan, H.; Zhu, Q.; Wang, X.; Gui, L.; and He, Y. 2024. Mirror: A multiple-perspective self-reflection method for knowledge-rich reasoning. *arXiv preprint arXiv:2402.14963*.

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.

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 preprint arXiv:2402.11411*.

Zhou, Y.; Fan, Z.; Cheng, D.; Yang, S.; Chen, Z.; Cui, C.; Wang, X.; Li, Y.; Zhang, L.; and Yao, H. 2024b. Calibrated self-rewarding vision language models. *Advances in Neural Information Processing Systems*, 37: 51503–51531.