DA-CPO: Difficulty-Aware Curriculum Preference Optimization for Mitigating Hallucinations in Large Vision-Language Models

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) have made significant progress in multimodal understanding but still suffer from visual hallucinations caused by over-reliance on language priors. Existing DPO-based mitigation methods either rely on coarse response-level preferences or uniformly optimize fine-grained preference samples, overlooking localized hallucination signals and sample difficulty. We propose **DA-CPO** (**D**ifficulty-**A**ware **C**urriculum **P**reference **O**ptimization), a progressive preference optimization framework that combines sentence-level hallucination supervision with difficulty-aware curriculum learning. DA-CPO diagnoses hallucinated sentences, revises responses to construct fine-grained preference samples, estimates sample difficulty from multiple aspects, and progressively optimizes LVLMs. Experiments across multiple benchmarks show consistent hallucination mitigation while preserving general multimodal capabilities. Moreover, DA-CPO achieves consistent reductions across hallucination metrics with only a relatively small amount of preference data.

1 Introduction

Recent advances in LVLMs have substantially improved visual instruction following, visual question answering, and multimodal reasoning (Alayrac et al., 2022; Dai et al., 2023; Li et al., 2023a; Roumeliotis and Tselikas, 2023). However, LVLMs still frequently suffer from visual hallucinations, generating plausible but factually incorrect content that contradicts visual inputs (Leng et al., 2024; Bai et al., 2024). This issue is often caused by the dominance of language priors over visual evidence during generation (Rohrbach et al., 2018; Min et al., 2025). Without explicit supervision favoring image-grounded responses over fluent yet unsupported alternatives, LVLMs tend to assign high likelihood to seemingly plausible hallucinated outputs (Xie et al., 2024).

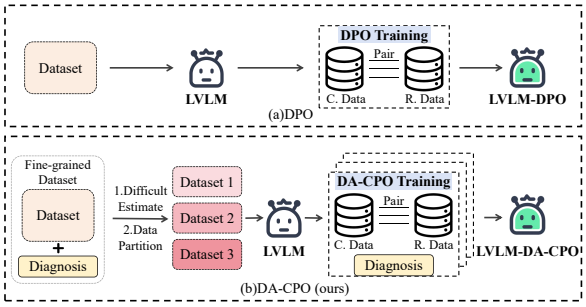


Figure 1: Comparison of standard DPO training strategy and our proposed DA-CPO. C. Data denotes the chosen data, while R. Data denotes the rejected data.

To mitigate visual hallucinations, recent methods based on Direct Preference Optimization (DPO) (Rafailov et al., 2023) (e.g., HA-DPO (Zhao et al., 2023), POVID (Zhou et al., 2024), and RLAIF-V (Yu et al., 2025)) align LVLMs by optimizing preferences between visually faithful and hallucinated responses. However, such response-level preferences provide coarse supervision: they compare responses as a whole and assign undifferentiated importance to all tokens, making it hard to target hallucination-related content.

To overcome the coarse supervision of response-level DPO, recent fine-grained DPO-based methods introduce more localized preference supervision for hallucination mitigation. HSA-DPO (Xiao et al., 2025) incorporates sentence-level hallucination feedback and severity-aware preference learning, while HDPO (Fu et al., 2025) constructs hallucination-targeted preference pairs for different hallucination sources. Nevertheless, as illustrated in Figure 1(a), these methods typically treat all preference samples uniformly during optimization, without considering sample-level difficulty differences. This uniform training strategy introduces challenging samples at early optimization stages,

impairing training stability and limiting the effectiveness of hallucination mitigation.

To address both limitations, we propose DA-CPO (**D**ifficulty-**A**ware **C**urriculum **P**reference **O**ptimization), a framework that constructs fine-grained hallucination-aware preference supervision and accounts for sample-level difficulty differences. As shown in Figure 1(b), unlike existing methods that treat all preference samples uniformly during optimization, DA-CPO estimates sample difficulty and progressively introduces preference samples. The framework proceeds in three steps. First, we perform hallucination diagnosis and assessment-guided response revision to construct informative fine-grained preference samples. Second, we estimate sample difficulty and organize the data into stage-specific curriculum sets. Finally, we progressively optimize the LVLM from easy to hard stages.

We evaluate DA-CPO across multiple standard hallucination benchmarks. Results show that DA-CPO consistently mitigates visual hallucinations while preserving general multimodal capabilities. Further analyses validate the proposed objectives and difficulty-aware curriculum, demonstrating the stability and effectiveness of progressive preference optimization. Our main contributions are summarized as follows:

- We propose **DA-CPO**, a difficulty-aware curriculum preference optimization framework that constructs fine-grained hallucination-aware preference supervision and progressively optimizes LVLMs for hallucination mitigation.
- Extensive experiments across multiple hallucination benchmarks show that DA-CPO consistently mitigates visual hallucinations while preserving general multimodal capabilities, without requiring large-scale training data.
- DA-CPO is broadly applicable to LVLMs with different model scales and can be integrated into preference optimization pipelines, highlighting its potential for robust and general hallucination mitigation.

2 Related Work

DPO-based Hallucination Mitigation. Visual hallucination has been widely studied in LVLMs, including object hallucination, hallucination-oriented evaluation, and language-prior-driven generation errors (Rohrbach et al., 2018; Li et al., 2023b; Wang et al., 2023; Guan et al., 2024; Leng et al., 2024). To mitigate this issue, preference-based alignment methods have been increasingly explored, with DPO serving as a simple and effective alternative to reward-model-based RLHF (Rafailov et al., 2023; Sun et al., 2024; Yu et al., 2024). Existing DPO-based methods construct preference data from hallucination-aware responses, modality-aligned feedback, AI-generated feedback, vision-guided contrastive pairs, or CLIP-based preference ranking to encourage visually consistent outputs over hallucinated ones (Zhao et al., 2023; Zhou et al., 2024; Yu et al., 2025; Xie et al., 2024; Ouali et al., 2024). Other works further improve multimodal preference optimization by introducing conditional preference modeling, image-contrast preferences, or on-policy data construction (Wang et al., 2024; Yang et al., 2025). Although effective, these methods mainly focus on preference data construction or multimodal preference modeling, while often lacking localized supervision that explicitly identifies hallucination-related content within responses.

Fine-grained DPO-based Hallucination Mitigation. To provide more informative preference supervision, recent studies further explore fine-grained DPO-based methods for hallucination mitigation. Fine-grained feedback and correctional supervision have been shown effective for identifying localized hallucinations and improving multimodal response faithfulness (Yu et al., 2024). HSA-DPO (Xiao et al., 2025) incorporates hallucination severity into preference learning based on sentence-level AI feedback, HDPO (Fu et al., 2025) constructs hallucination-targeted preference pairs according to different hallucination sources, and EMPO (Wu et al., 2025) introduces entity-centric multimodal preference optimization to improve fine-grained visual grounding. These methods enhance preference supervision beyond response-level comparisons. However, they typically treat preference samples uniformly during optimization, overlooking sample difficulty differences and limiting the stability of preference learning.

3 Methodology

In this work, we propose **DA-CPO**, a curriculum preference optimization framework for hallucination mitigation. As illustrated in Figure 2, DA-CPO constructs fine-grained preference samples, partitions these samples by difficulty, and progressively

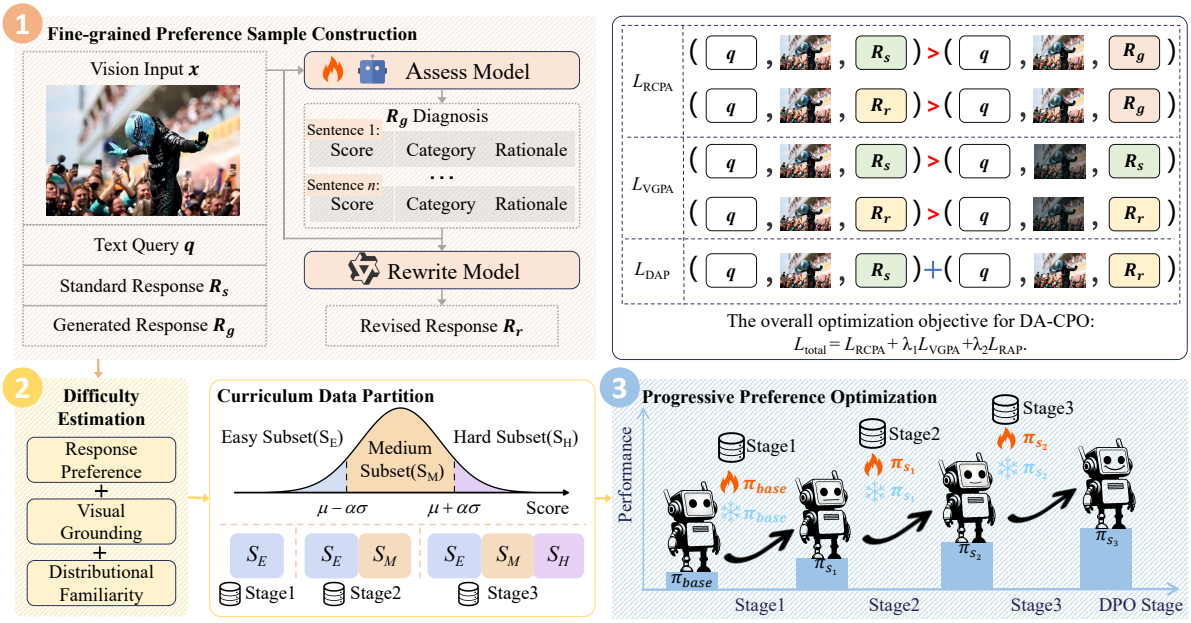


Figure 2: Overview of **DA-CPO**. The framework first constructs fine-grained preference samples through sentence-level hallucination diagnosis and response revision, then partitions them by difficulty to form curriculum subsets, and finally performs progressive preference optimization from easy to hard samples.

optimizes the LVLM to improve visual faithfulness and mitigate hallucinations.

3.1 Fine-grained Preference Sample Construction

To support curriculum preference optimization, we first construct fine-grained preference samples. We extract image-instruction pairs and human-corrected standard responses from the RLHF-V dataset. Given an image x and an instruction q, the base LVLM M_{base} generates a candidate response R_{g} conditioned on (x, q). We further use the corresponding standard response R_{s} as a high-quality reference, which helps the assessor identify unsupported content in R_{g} and provides reliable supervision for constructing preference pairs.

Hallucination Assessor Training and Diagnosis. To obtain reliable preference supervision, we train a hallucination assessor M_{assess} by fine-tuning LLaVA-v1.5-7B. After training, M_{assess} assesses R_{g} conditioned on (x, q, R_{s}). Specifically, the assessor produces sentence-level diagnostic annotations:

$$M_{\text{assess}}(x, q, R_{\text{s}}, R_{\text{g}}) = \{(s_j, c_j, e_j)\}_{j=1}^{n}, \quad (1)$$

where n denotes the total number of sentences in R_{g}, and j is the sentence index. Each (s_j, c_j, e_j) corresponds to the j-th sentence, where s_j denotes a hallucination score on a 1–4 scale, c_j denotes the error category among {correct, image_recognition_error, language_comprehension_error}, and e_j denotes the explanatory rationale. The diagnostic annotation set is denoted as:

$$\mathcal{A} = M_{\text{assess}}(x, q, R_{\text{s}}, R_{\text{g}}). \quad (2)$$

Assessment-guided Response Revision. Based on $\mathcal{A}$, we use Qwen2.5-VL-7B as the rewriting model M_{rewrite} to minimally correct hallucinated sentences and generate the revised response R_{r} while preserving the original response distribution. We adopt Qwen2.5-VL-7B as M_{rewrite} due to its strong instruction-following capability, which facilitates controlled revisions guided by the diagnostic rationales from M_{assess}.

Finally, each instance in the DPO training set is organized as a fine-grained preference sample $\mathcal{P}$:

$$\mathcal{P} = (x, q, R_{\text{s}}, R_{\text{r}}, R_{\text{g}}, \mathcal{A}). \quad (3)$$

3.2 Difficulty-based Curriculum Partitioning

After constructing $\mathcal{P}$, DA-CPO estimates a difficulty score for each sample before preference optimization and organizes the training data into a progressive curriculum.

Difficulty Estimation. For each $\mathcal{P}$, we compute three complementary difficulty scores. D_{r} captures response preference difficulty by measuring whether high-quality responses can be clearly

distinguished from the generated hallucinated response. D_{v} captures visual grounding difficulty by comparing response likelihoods under the original and perturbed images. D_{f} captures distributional familiarity by measuring how well high-quality responses fit the reference policy distribution. First, let π_{ref} denote the frozen M_{base} used as the reference policy. For each response $R \in \{R_{\mathrm{s}}, R_{\mathrm{r}}, R_{\mathrm{g}}\}$, we compute its length-normalized log-likelihood:

$$\bar{\ell}_{\mathrm{ref}}(R \mid x, q) = \frac{1}{|R|} \sum_{t=1}^{|R|} \log \pi_{\mathrm{ref}}(r_t \mid x, q, r_{<t}), \tag{4}$$

where $|R|$ denotes the number of tokens in response R, and r_t denotes the t-th token in R.

We define the reference margin between two responses as:

$$\begin{aligned} \Delta_{\mathrm{r}}(x, q, R_1, R_2) = \bar{\ell}_{\mathrm{ref}}(R_1 \mid x, q) \\ - \bar{\ell}_{\mathrm{ref}}(R_2 \mid x, q). \end{aligned} \tag{5}$$

Accordingly, D_{r} is computed from the reference margins between high-quality responses and the generated response. We use the softplus function to obtain a smooth and non-negative difficulty measure:

$$\mathrm{softplus}(z) = \log(1 + \exp(z)). \tag{6}$$

Then D_{r} is defined as:

$$\begin{aligned} D_{\mathrm{r}} = \frac{1}{2}\Big[\mathrm{softplus}\big(-\Delta_{\mathrm{r}}(x, q, R_{\mathrm{s}}, R_{\mathrm{g}})\big) \\ + \mathrm{softplus}\big(-\Delta_{\mathrm{r}}(x, q, R_{\mathrm{r}}, R_{\mathrm{g}})\big)\Big]. \end{aligned} \tag{7}$$

Smaller margins, including negative ones, lead to larger values, indicating that the corresponding preference relations are harder to learn.

Second, we estimate D_{v} by comparing response likelihoods under the original image x and the patch-shuffled image $\tilde{x}$. Compared with Gaussian noise, patch shuffling preserves low-level image statistics while disrupting spatial structures and object-level evidence, making the likelihood gap better reflect visual grounding. For $R \in \{R_{\mathrm{s}}, R_{\mathrm{r}}\}$, its visual margin is defined as:

$$\Delta_{\mathrm{v}}(x, q, R) = \bar{\ell}_{\mathrm{ref}}(R \mid x, q) - \bar{\ell}_{\mathrm{ref}}(R \mid \tilde{x}, q). \tag{8}$$

Accordingly, D_{v} is computed from the visual margins of high-quality responses under the original and perturbed images:

$$\begin{aligned} D_{\mathrm{v}} = \frac{1}{2}\Big[\mathrm{softplus}\big(-\Delta_{\mathrm{v}}(x, q, R_{\mathrm{s}})\big) \\ + \mathrm{softplus}\big(-\Delta_{\mathrm{v}}(x, q, R_{\mathrm{r}})\big)\Big]. \end{aligned} \tag{9}$$

Smaller visual margins, including negative ones, lead to larger D_{v} values, indicating that the response likelihood is less sensitive to visual perturbation and thus the sample has higher visual grounding difficulty.

Third, we compute D_{f}, which measures how well high-quality responses fit the reference policy distribution:

$$D_{\mathrm{f}} = -\frac{1}{2}\Big[\bar{\ell}_{\mathrm{ref}}(R_{\mathrm{s}} \mid x, q) + \bar{\ell}_{\mathrm{ref}}(R_{\mathrm{r}} \mid x, q)\Big]. \tag{10}$$

Larger D_{f} values indicate lower likelihoods of high-quality responses under the reference policy, suggesting higher distributional unfamiliarity and a greater need for preservation during optimization.

Finally, we normalize the three difficulty scores over the training set using z-score normalization:

$$\widetilde{D}_k = \frac{D_k - \mu_k}{\sigma_k}, \quad k \in \{\mathrm{r}, \mathrm{v}, \mathrm{f}\}, \tag{11}$$

where μ_k and σ_k denote the mean and standard deviation of D_k over the training set, respectively. We then aggregate the normalized scores into the overall difficulty score:

$$d = \widetilde{D}_{\mathrm{r}} + \widetilde{D}_{\mathrm{v}} + \widetilde{D}_{\mathrm{f}}. \tag{12}$$

Z-score normalization places the three scores on a comparable scale, preventing any single component from dominating the overall difficulty due to differences in numerical range. A larger d indicates that the preference sample is harder for DPO training.

Curriculum Data Partitioning. Using d, we partition the training set into three curriculum subsets, denoted as $\mathcal{S}_E$, $\mathcal{S}_M$, and $\mathcal{S}_H$, corresponding to easy, medium, and hard samples, respectively. We then construct stage-specific training sets in a cumulative manner:

$$\begin{aligned} \mathcal{T}_1 &= \mathcal{S}_E, \\ \mathcal{T}_2 &= \mathcal{S}_E \cup \mathcal{S}_M, \\ \mathcal{T}_3 &= \mathcal{S}_E \cup \mathcal{S}_M \cup \mathcal{S}_H. \end{aligned} \tag{13}$$

Preference optimization is progressively performed on $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$. This cumulative design retains previously learned easier samples in later stages, helping prevent forgetting while progressively introducing more challenging preference samples.

3.3 Progressive Preference Optimization

After constructing the curriculum subsets, DA-CPO progressively optimizes M_{base} over stage-specific training sets from easy to hard. At each stage, the model is trained with three complementary objectives that jointly correct hallucinated responses, strengthen visual grounding, and preserve high-quality responses.

Response Correction Preference Alignment (RCPA). RCPA aims to suppress hallucinated content by leveraging both the standard response and the revised response as positive supervision. Given $\mathcal{P}$, RCPA constructs two response preference pairs conditioned on (x, q): $(R_{\text{s}}, R_{\text{g}})$ and $(R_{\text{r}}, R_{\text{g}})$, where the first response in each pair is preferred. For a preference pair (R_a, R_b) conditioned on (x, q), we define its DPO loss as:

$$\mathcal{L}_{\text{DPO}}(x, q, R_a, R_b) = -\log \sigma \left(\beta \log \frac{\pi_\theta(R_a \mid x, q)}{\pi_{\text{ref}}(R_a \mid x, q)} - \beta \log \frac{\pi_\theta(R_b \mid x, q)}{\pi_{\text{ref}}(R_b \mid x, q)} \right), \tag{14}$$

where π_θ denotes the trainable policy, and β controls the strength of preference optimization.

To assign stronger supervision to hallucinated sentences, we apply hallucination-aware token weighting to the pair $(R_{\text{r}}, R_{\text{g}})$, where token weights are derived from hallucination scores in $\mathcal{A}$. The hallucination-weighted DPO loss is defined as:

$$\mathcal{L}^{\text{hw}}_{\text{DPO}}(x, q, R_{\text{r}}, R_{\text{g}}) = -\log \sigma \left(\beta \log \frac{\pi^{\text{hw}}_\theta(R_{\text{r}} \mid x, q)}{\pi^{\text{hw}}_{\text{ref}}(R_{\text{r}} \mid x, q)} - \beta \log \frac{\pi^{\text{hw}}_\theta(R_{\text{g}} \mid x, q)}{\pi^{\text{hw}}_{\text{ref}}(R_{\text{g}} \mid x, q)} \right). \tag{15}$$

The RCPA objective is then formulated as:

$$\mathcal{L}_{\text{RCPA}} = \mathbb{E}_{(x,q,R_{\text{s}},R_{\text{r}},R_{\text{g}})\sim\mathcal{D}} \left[\mathcal{L}_{\text{DPO}}(x, q, R_{\text{s}}, R_{\text{g}}) + \mathcal{L}^{\text{hw}}_{\text{DPO}}(x, q, R_{\text{r}}, R_{\text{g}}) \right]. \tag{16}$$

Visual Grounding Preference Alignment (VGPA). VGPA aims to strengthen the model's reliance on genuine visual evidence and reduce visually unsupported generation. Given $\mathcal{P}$, VGPA constructs visual preference pairs by contrasting the same high-quality response under the original image x and the perturbed image $\tilde{x}$, where the original-image condition is preferred. For R_{s}, we define the visual DPO loss as:

$$\mathcal{L}_{\text{vDPO}}(x, \tilde{x}, q, R_{\text{s}}) = -\log \sigma \left(\beta \log \frac{\pi_\theta(R_{\text{s}} \mid x, q)}{\pi_{\text{ref}}(R_{\text{s}} \mid x, q)} - \beta \log \frac{\pi_\theta(R_{\text{s}} \mid \tilde{x}, q)}{\pi_{\text{ref}}(R_{\text{s}} \mid \tilde{x}, q)} \right). \tag{17}$$

For R_{r}, we further apply visual-aware token weighting according to the error categories in $\mathcal{A}$. Specifically, tokens associated with image-recognition errors receive stronger weights. Based on this weighting, the visual-weighted DPO loss is defined as:

$$\mathcal{L}^{\text{vw}}_{\text{vDPO}}(x, \tilde{x}, q, R_{\text{r}}) = -\log \sigma \left(\beta \log \frac{\pi^{\text{vw}}_\theta(R_{\text{r}} \mid x, q)}{\pi^{\text{vw}}_{\text{ref}}(R_{\text{r}} \mid x, q)} - \beta \log \frac{\pi^{\text{vw}}_\theta(R_{\text{r}} \mid \tilde{x}, q)}{\pi^{\text{vw}}_{\text{ref}}(R_{\text{r}} \mid \tilde{x}, q)} \right). \tag{18}$$

The VGPA objective is then formulated as:

$$\mathcal{L}_{\text{VGPA}} = \mathbb{E}_{(x,\tilde{x},q,R_{\text{s}},R_{\text{r}})\sim\mathcal{D}} \left[\mathcal{L}_{\text{vDPO}}(x, \tilde{x}, q, R_{\text{s}}) + \mathcal{L}^{\text{vw}}_{\text{vDPO}}(x, \tilde{x}, q, R_{\text{r}}) \right]. \tag{19}$$

Reference-anchored Preservation (RAP). Relative preference optimization constrains the margin between chosen and rejected responses, but does not explicitly preserve the likelihood of high-quality responses. RAP addresses this by anchoring R_{s} and R_{r} to the reference policy. Specifically, we use a sample-level preservation weight derived from D_{f} on the policy log-ratio, encouraging sufficient support for high-quality responses during optimization. The RAP objective is then formulated as:

$$\mathcal{L}_{\text{RAP}} = -\mathbb{E}_{(x,q,R_{\text{s}},R_{\text{r}})\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi^{\text{f}}_\theta(R_{\text{s}} \mid x, q)}{\pi^{\text{f}}_{\text{ref}}(R_{\text{s}} \mid x, q)} \right) + \log \sigma \left(\beta \log \frac{\pi^{\text{f}}_\theta(R_{\text{r}} \mid x, q)}{\pi^{\text{f}}_{\text{ref}}(R_{\text{r}} \mid x, q)} \right) \right]. \tag{20}$$

The overall optimization objective of DA-CPO is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RCPA}} + \lambda_1 \mathcal{L}_{\text{VGPA}} + \lambda_2 \mathcal{L}_{\text{RAP}}, \tag{21}$$

where λ_1 and λ_2 control the contributions of visual preference alignment and reference-anchored preservation, respectively.

4 Experiments

By comparing our results against state-of-the-art methods, we demonstrate the effectiveness and superiority of the DA-CPO framework in hallucination mitigation across multiple benchmarks.

Method	AMBER				ObjectHal		MMHal-Bench		General QA	
	CHAIR↓	Cover↑	HalRate↓	Cog↓	CHAIRs↓	CHAIRi↓	Score↑	HalRate↓	HallusionBench↑	ScienceQA↑
LLaVA-v1.5-7B(Liu et al., 2024)§	8.40	52.10	35.70	4.00	52.33	17.21	2.15	0.65	16.70	53.20
+ LLaVA-RLHF(Sun et al., 2024)§	9.70	53.20	46.60	5.30	57.48	15.34	1.96	0.69	16.20	55.10
+ HALVA(Sarkar et al., 2025)	6.90	52.80	33.20	3.50	41.40	11.70	2.12	0.59	-	-
+ HA-DPO(Zhao et al., 2023)§	7.80	52.10	35.60	4.20	53.43	14.15	1.89	0.65	16.60	55.80
+ mDPO(Wang et al., 2024)	4.40	52.40	24.50	2.40	35.70	9.80	2.39	0.54	-	-
+ SCPO(Li et al., 2025)	3.30	**53.60**	19.50	1.40	20.30	11.60	2.47	0.44	-	-
+ RLHF-V(Yu et al., 2024)§	8.30	52.20	41.80	4.50	56.00	15.20	1.93	0.67	-	-
+ RLAIF-V(Yu et al., 2025)§	3.00	50.40	16.20	1.00	16.00	**3.70**	3.00	**0.38**	17.58	57.71
+ POVID(Zhou et al., 2024)§	7.40	51.30	34.30	3.80	49.97	15.08	2.08	0.60	16.90	56.20
+ OPA-DPO(Yang et al., 2025)§	2.70	43.90	11.50	0.80	15.33	5.26	2.78	0.46	-	-
+ **DA-CPO (ours)**	**2.60**	43.10	**10.50**	**0.50**	**8.67**	4.20	**3.28**	0.41	**20.88**	**66.98**
LLaVA-v1.5-13B(Liu et al., 2024)§	6.70	51.20	31.20	3.40	51.01	23.63	2.53	0.56	14.28	63.06
+ LLaVA-RLHF(Sun et al., 2024)§	7.70	52.30	38.60	4.00	43.92	11.51	2.34	0.61	**18.80**	63.80
+ RLHF-V(Yu et al., 2024)§	6.30	46.10	25.10	2.10	43.92	11.51	2.81	0.49	-	-
+ HALVA(Sarkar et al., 2025)	6.40	**52.60**	30.40	3.20	45.40	12.80	2.58	0.45	-	-
+ OPA-DPO(Yang et al., 2025)§	2.40	48.30	12.80	0.90	15.98	5.26	3.14	**0.37**	-	-
+ HSA-DPO(Yang et al., 2025)§	**2.10**	47.30	13.40	1.20	-	-	2.61	0.48	-	-
+ **DA-CPO (ours)**	2.40	43.20	**9.90**	**0.80**	**13.67**	**5.25**	**3.46**	0.40	15.82	**70.80**

Table 1: Main results on hallucination mitigation and multimodal QA. Results marked with § are reproduced by us under the same pipeline; the others are taken from the original papers. Best results are in bold; second-best results are underlined.

4.1 Experiment Setup

Models and Datasets. We apply DA-CPO to two LVLMs with distinct parameter sizes: LLaVA-v1.5-7B and LLaVA-v1.5-13B, both using CLIP ViT-L-336px as the vision encoder. To obtain sentence-level supervision, we process the fine-grained annotations from the OPA-DPO dataset and use them to fine-tune M_{assess} based on LLaVA-v1.5-7B. We further reserve 490 annotated samples as a held-out validation set to compare the assessment performance of M_{assess} with generalist MLLMs. We then use Qwen2.5-VL-7B as M_{rewrite} to generate revised responses according to the sentence-level diagnostic results. For preference optimization, We construct training samples based on the RLAIF-V (Yu et al., 2025) dataset by extracting image–instruction pairs and their corresponding standard responses. For each pair, M_{base} generates a candidate response R_{g}, while M_{assess} provides sentence-level hallucination scores, error categories, and diagnostic rationales to construct revised responses and difficulty-based curriculum partitions. Finally, we obtain a training dataset containing 4,503 fine-grained preference samples.

Benchmarks and evaluation. We evaluate hallucination mitigation and general multimodal reasoning on Object HalBench(Rohrbach et al., 2018), MMHal-Bench(Sun et al., 2024), AMBER(Wang et al., 2023), HallusionBench(Guan et al., 2024), and ScienceQA(Lu et al., 2022). All benchmarks are evaluated with their official prompts and scripts. For Object HalBench, MMHal-Bench, and HallusionBench, we use GPT-4.1 for the required LLM-based judgment. AMBER and ScienceQA do not require LLM-based judgment. For reproduced baselines, we apply the same inference and evaluation pipeline as DA-CPO; for methods without public checkpoints, we report the numbers from the original papers. For each evaluation setting, we follow the official evaluation protocol, run the evaluation three times, and report the averaged results.

Implementation Details. We adopt the LLaVA-v1.5 architecture as the backbone. The policy model is optimized using the AdamW optimizer with a learning rate of 1e-6 and a global batch size of 32. For the patch shuffling strategy, the shuffling ratio γ is set to 0.3 to effectively disrupt spatial semantics. Regarding the optimization objective, the scaling coefficients are fixed at $\lambda_1 = 0.2$ and $\lambda_2 = 1$. In our equations, we set $\beta = 0.1$ for all preference optimization objectives. The model is trained progressively across three stages, with each stage inheriting optimized weights from its predecessor. All experiments and evaluations are conducted on NVIDIA H100 SXM5(80G) GPUs.

4.2 Main Results

Table 1 summarizes the main results on hallucination mitigation and multimodal QA.

Hallucination benchmarks. DA-CPO delivers the clearest gains on hallucination-oriented benchmarks. On Object HalBench, CHAIRs and CHAIRi drop from 52.33 and 17.21 to 8.67 and 4.20 for the 7B policy, and from 51.01 and 23.63 to 13.67 and 5.70 for the 13B policy. DA-CPO therefore achieves the best listed Object HalBench

results on both backbones. On MMHal-Bench, the score and HalRate improve from 2.15 and 0.65 to 3.28 and 0.41 for 7B, and reach 3.46 and 0.40 for 13B. On AMBER, DA-CPO also significantly reduces hallucination while maintaining generative quality: CHAIR drops from 8.40 to 2.6 for 7B and from 6.70 to 2.4 for 13B.

General multimodal QA. DA-CPO also improves general multimodal QA performance. On HallusionBench, accuracy rises from 16.70 to 20.88 for the 7B policy and from 14.28 to 15.82 for the 13B policy. On ScienceQA, image accuracy improves significantly from 53.20 to 66.98 for 7B, and from 63.06 to 70.80 for 13B.

Reliability Analysis of the Hallucination Assessor. The reliability of M_{assess} is crucial, as its sentence-level hallucination scores and error categories are used to construct the weighted DPO objectives, while its diagnostic rationales guide the revision of hallucinated responses. We evaluate M_{assess} on a held-out validation set of 490 samples and compare it with several zero-shot generalist MLLMs. As shown in Table 2, our fine-tuned assessor achieves 81.92% Exact Accuracy, indicating its ability to exactly match the reference hallucination labels. It also obtains an F1 score of 0.88, showing strong balance between precision and recall in identifying hallucination-related errors. In addition, the low MAE of 0.17 demonstrates that its predicted hallucination severity scores are close to the reference scores. These results show that the fine-tuned assessor substantially outperforms generalist models and provides reliable fine-grained supervision for DA-CPO.

Method	Exact Acc↑	Hal Acc↑	MAE↓	F1↑
LLaVA-v1.5-7B	46.09	63.03	1.06	0.39
LLaVA-v1.5-13B	46.91	65.96	0.96	0.47
Qwen-VL-Chat	43.13	64.27	1.00	0.62
Our Training Model	**81.92**	**90.55**	**0.17**	**0.88**

Table 2: Reliability evaluation of the hallucination assessor on the held-out validation set.

Progressive Improvement across Curriculum Stages. To further analyze the effect of curriculum learning, we visualize the performance of DA-CPO across different training stages. As shown in Figure 3 and Figure 4, DA-CPO exhibits a clear progressive improvement from Stage 1 to Stage 3 on hallucination-oriented metrics. This trend demonstrates that the proposed curriculum learning strategy stabilizes preference optimization and enhances hallucination mitigation by gradually expanding training from easy samples to more challenging ones.

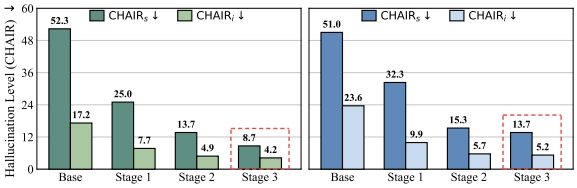


Figure 3: Object HalBench results across DPO training stages using LLaVA-v1.5-7B as the base model.

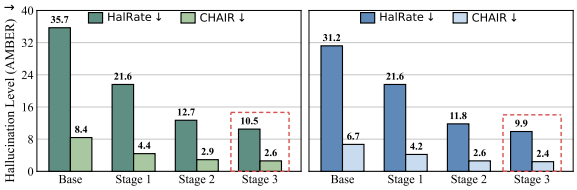


Figure 4: AMBER results across DPO training stages using LLaVA-v1.5-13B as the base model.

4.3 Ablation Study

Effectiveness of Optimization Components. To assess the contribution of each optimization component, we conduct ablation experiments on LLaVA-v1.5-7B. As shown in Table 3, RCPA directly suppresses hallucinated responses by preferring standard and revised responses, leading to clear reductions in hallucination metrics. VGPA further enhances visual grounding through original–perturbed image preference pairs, bringing additional improvements. RAP anchors high-quality responses to the reference policy; removing it causes a clear degradation on POPE, indicating its role in preserving discriminative capability. Overall, these results show that RCPA, VGPA, and RAP are complementary for hallucination mitigation.

Components			Hallucination Metrics			
RCPA	VGPA	RAP	AMBER ↓	Obj Hal ↓	POPE Acc ↑	POPE F1 ↑
			8.40	52.33	**84.97**	**83.89**
✓		✓	3.1	16.00	83.20	80.75
✓	✓		2.9	14.67	81.56	78.44
✓	✓	✓	**2.6**	**8.67**	84.07	82.14

Table 3: Ablation study on the core optimization components of DA-CPO.

Ablation on Curriculum Training Strategies. We compare the proposed cumulative curriculum with three alternative training strategies. *No Curriculum (Mixed)* ignores sample difficulty and

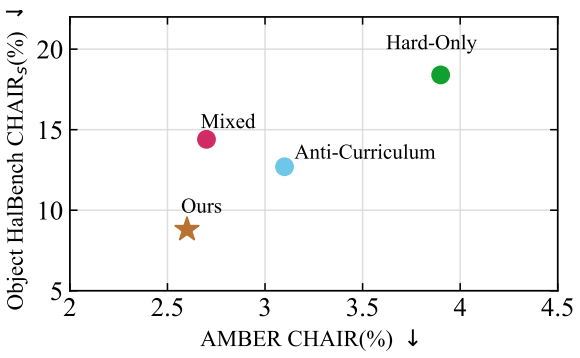


Figure 5: Performance comparison of different curriculum training strategies on AMBER and Object HalBench.

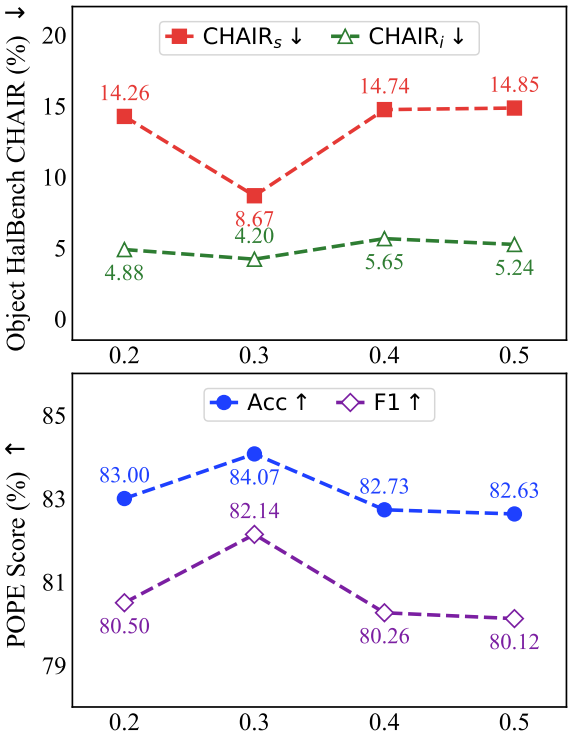


Figure 6: Impact of the patch-shuffling ratio on Object HalBench and POPE.

trains the model on all preference samples simultaneously. *Anti-Curriculum* follows the reverse order of curriculum learning, starting from hard samples and then gradually introducing medium and easy samples. *Hard-Only* uses only the hard subset throughout training, removing easier samples from the optimization process. As shown in Figure 5, all alternative strategies yield inferior results. This suggests that ignoring sample difficulty, reversing the learning order, or relying only on difficult samples can impair stable preference optimization. In contrast, the cumulative curriculum first builds reliable preference alignment with easier samples and then progressively incorporates more challenging samples, leading to better hallucination mitigation.

Analysis of Patch Shuffling Ratio. We analyze the effect of the patch-shuffling ratio $\gamma \in \{0.2, 0.3, 0.4, 0.5\}$ used to construct the perturbed image $\tilde{x}$. As shown in Figure 6, $\gamma = 0.3$ achieves the best overall trade-off, yielding the lowest CHAIRs and CHAIRi on Object HalBench and the highest Accuracy and F1 on POPE. Smaller ratios preserve excessive spatial information and weaken visual contrast, while larger ratios introduce stronger perturbations that can distort visual evidence. Therefore, we adopt $\gamma = 0.3$ for patch shuffling in our experiments.

5 Conclusion

In this work, we present DA-CPO, a difficulty-aware curriculum preference optimization framework for mitigating visual hallucinations in LVLMs. DA-CPO addresses the limitations of prior DPO-based methods by integrating diagnostic preference construction with curriculum learning. It constructs hallucination-aware preference samples and progressively organizes them according to sample difficulty. By introducing samples from easy to hard, DA-CPO stabilizes preference learning and strengthens visual-faithfulness alignment. Experiments on multiple hallucination benchmarks demonstrate consistent improvements. Notably, DA-CPO achieves effective hallucination mitigation with relatively small-scale training data, offering a data-efficient solution.

Limitations

DA-CPO effectively mitigates hallucinations but remains challenged in scenarios requiring fine-grained visual perception and complex relational reasoning. When images contain dense visual information, cluttered scenes, small objects, or subtle spatial relations, LVLMs can still produce visually unsupported responses. These failures suggest that hallucination mitigation remains constrained by the visual perception and relational reasoning capabilities of the base LVLM. In addition, DA-CPO relies on a hallucination assessor and a rewriting model to construct fine-grained preference samples, which introduces additional data construction cost before training. Nevertheless, DA-CPO consistently improves hallucination mitigation across multiple benchmarks, and its difficulty-aware curriculum provides a stable and effective optimization

strategy for LVLM alignment.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Yuhan Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong Li. 2025. Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16563–16577.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 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*, pages 14375–14385.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.

Yuanshuai Li, Yuping Yan, Junfeng Tang, Yunxuan Li, Zeqi Zheng, and Yaochu Jin. 2025. Mitigating visual hallucinations via semantic curriculum preference optimization in mllms. *arXiv preprint arXiv:2509.24491*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

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. *Advances in neural information processing systems*, 35:2507–2521.

Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee, and Kyomin Jung. 2025. Mitigating hallucinations in large vision-language models via summary-guided decoding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4183–4198.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. 2024. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *European Conference on Computer Vision*, pages 395–413. Springer.

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. *Advances in neural information processing systems*, 36:53728–53741.

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*, pages 4035–4045.

Konstantinos I Roumeliotis and Nikolaos D Tselikas. 2023. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Arik, and Tomas Pfister. 2025. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. In *International Conference on Learning Representations*, volume 2025, pages 73796–73846.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8078–8088.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and 1 others. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Jiulong Wu, Zhengliang Shi, Shuaiqiang Wang, Jizhou Huang, Dawei Yin, Lingyong Yan, Min Cao, and Min Zhang. 2025. Mitigating hallucinations in large vision-language models via entity-centric multimodal preference optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19456–19472.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. 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, pages 25543–25551.

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*, pages 13258–13273.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. 2025. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10610–10620.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 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*, pages 13807–13816.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, and 1 others. 2025. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19985–19995.

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

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

A Appendix

A.1 Weight Assignment for DPO Losses

This section details the weight assignment used in the three DPO losses. For RCPA, the hallucination-aware token weights are determined by s_j, as shown in Table 4. For VGPA, the visual-aware token weights are determined by c_j, as shown in Table 5. Tokens within the same sentence share the same update weight.

Hallucination Score s_j	**Weight** w_{hw}	**Severity**
4	2.5	Not at all
3	2.0	Minor
2	1.5	Major
1	1.0	Totally

Table 4: Score-to-weight mapping for the hallucination-weighted DPO loss.

Weight w_{vw}	**Error Category** c_j
1.0	correct
1.0	language_comprehension_error
2.0	image_recognition_error

Table 5: Category-to-weight mapping for the visual-weighted DPO loss.

For RAP, we assign a sample-level preservation weight according to the normalized D_{f}:

$$w^{\mathrm{f}} = 1 + \sigma(\widetilde{D}_{\mathrm{f}}), \tag{22}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\widetilde{D}_{\mathrm{f}}$ is the normalized distributional familiarity difficulty score. A larger $\widetilde{D}_{\mathrm{f}}$ indicates that R_{s} and R_{r} are less supported by the reference policy, and therefore receive stronger preservation during optimization.

A.2 Qualitative Comparison

Figures 7 and 8 present qualitative comparisons on POPE, AMBER, and HallusionBench. Green and red text indicate correct predictions and hallucinations, respectively. Compared with LLaVA-v1.5, RLAIF-V, and OPA-DPO, DA-CPO generates more visually faithful responses and reduces visually unsupported content.

A.3 Error Analysis

Figures 9 and 10 present error analysis on challenging samples across different methods. Although DA-CPO generally produces more visually faithful responses than LLaVA-v1.5, RLAIF-V, and OPA-DPO, hallucinations can still occur in cases involving complex or dense visual information, such as cluttered scenes and maps. Errors also remain when the question requires fine-grained spatial reasoning, contact relation judgment, small-object recognition, or geometric comparison. These examples indicate that further improving fine-grained visual perception and relational reasoning remains important for hallucination mitigation.

A.4 Prompt for Training the Hallucination Assessor

To train M_{assess}, we use the fine-grained sentence-level annotations processed from the OPA-DPO dataset. The complete prompt template is shown in Table 6.

A.5 Prompt for Response Revision

We use the following prompt with Qwen2.5-VL-7B for assessment-guided response revision. Given the image, instruction, generated response, and sentence-level diagnostic annotations produced by M_{assess}, the rewriting model minimally revises hallucinated sentences to generate R_{r}. The prompt template is shown in Table 7.

A.6 Dataset and Model Licensing

Datasets. For preference data construction, we use the RLAIF-V dataset (Yu et al., 2025), from which we extract image–instruction pairs and standard responses. RLAIF-V is released under the CC BY-NC 4.0 license. Based on the extracted image–instruction pairs, we generate candidate responses with M_{base}, obtain sentence-level diagnostic annotations using M_{assess}, and construct revised responses with M_{rewrite}.

To train the hallucination assessor M_{assess}, we use fine-grained sentence-level annotations processed from the OPA-DPO dataset. We follow the release terms of the original OPA-DPO resources when using these annotations for assessor training.

For evaluation, we use AMBER (Wang et al., 2023), Object HalBench (Rohrbach et al., 2018), POPE (Li et al., 2023b), MMHal-Bench (Sun et al., 2024), HallusionBench (Guan et al., 2024), and ScienceQA (Lu et al., 2022). We follow the official evaluation protocols and release conditions of these benchmarks.

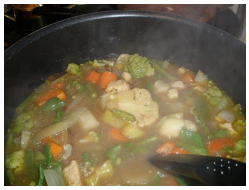
Models. We conduct experiments with publicly released LVLMs and follow their corresponding licenses. LLaVA-v1.5-7B and LLaVA-v1.5-13B (Liu et al., 2024) are released under the Apache-2.0 license and are used as the

base LVLMs. We further fine-tune LLaVA-v1.5-7B to obtain the hallucination assessor M_{assess}. For assessment-guided response revision, we use Qwen2.5-VL-7B as M_{rewrite}, which is released under the Apache-2.0 license. All external checkpoints are used in accordance with their original release terms.

POPE

Question: Is there a bowl in the image?
Ground truth answer: No.
Base Model: Yes.
RLAIF-V: Yes.
OPA-DPO: Yes.
DA-CPO: No.

Visual Input



AMBER

Question: Is there a dog in this image?
Ground truth answer: No.
Base Model: Yes, a dog is on the beach in this figure.
RLAIF-V: Yes. There is a dog on the beach.
OPA-DPO: Yes. The dog lies on the beach.
DA-CPO: No, there is no dog in this figure.

Visual Input

HallusionBench

Question: Are there any ducks in this figure?
Ground truth answer: Yes, there is one duck in this figure.
Base Model: No, there is no duck in this figure.
RLAIF-V: No. There is no duck in this figure.
OPA-DPO: No. No duck can be found in the image.
DA-CPO: Yes, there is one duck can be observed in this figure.

Visual Input

Figure 7: Qualitative comparison among LLaVA-v1.5-7B, RLAIF-V, OPA-DPO, and DA-CPO.

POPE

Question: Is there a person in the image?
Ground truth answer: No.
Base Model: Yes.
RLAIF-V: Yes.
OPA-DPO: Yes.
DA-CPO: No.

Visual Input

AMBER

Question: Is there a ship in this image?
Ground truth answer: No.
Base Model: Yes, there is a ship in this image.
RLAIF-V: Yes. There is a ship in the image.
OPA-DPO: Yes. There is a ship can be observed in this image.
DA-CPO: No, there is no ship can be observed in this figure.

Visual Input

HallusionBench

Question: Is there an orange cat in this figure?
Ground truth answer: Yes, there is an orange cat in this figure.
Base Model: No, there is no orange cat in this figure.
RLAIF-V: No. There is no cat in this figure.
OPA-DPO: No. Only fried chicken can be seen in the picture.
DA-CPO: Yes, an orange cat is visible in the lower-right part of this figure.

Visual Input

Figure 8: Qualitative comparison among LLaVA-v1.5-13B, RLAIF-V, OPA-DPO, and DA-CPO.

POPE

Question: Is there a vase in the image?
Ground truth answer: No.
Base Model: Yes.
RLAIF-V: Yes.
OPA-DPO: Yes.
DA-CPO: Yes.

Visual Input

AMBER

Question: Is there direct contact between the woman and beach?
Ground truth answer: No.
Base Model: Yes, there is.
RLAIF-V: Yes. She touches the beach directly.
OPA-DPO: Yes. Direct contact exists between her and the beach.
DA-CPO: Yes, there is direct contact between the woman and the beach.

Visual Input

HallusionBench

Question: According to the image, does Chile cover part of the Amazon Basin?
Ground truth answer: No. According to the image, Chile does not cover part of the Amazon Basin.
Base Model: Yes, Chile covers part of the Amazon Basin.
RLAIF-V: Yes, Chile covers a portion of the Amazon Basin.
OPA-DPO: Yes, part of the Amazon Basin lies within Chile.
DA-CPO: Yes, Chile covers part of the Amazon Basin.

Visual Input

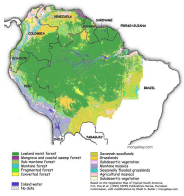


Figure 9: Error analysis on challenging samples, with DA-CPO trained from LLaVA-v1.5-7B.

POPE

Question: Is there a bicycle in the image?
Ground truth answer: No.
Base Model: Yes.
RLAIF-V: Yes.
OPA-DPO: Yes.
DA-CPO: Yes.

Visual Input

AMBER

Question: Is there direct contact between the frisbee and grass?
Ground truth answer: No.
Base Model: Yes, there is direct contact between the frisbee and the grass.
RLAIF-V: Yes, the frisbee is in direct contact with the grass.
OPA-DPO: Yes, the frisbee touches the grass directly.
DA-CPO: Yes, the frisbee is on the grass and in direct contact with it.

Visual Input

HallusionBench

Question: Is the red line on the left different length as the red line on the right? Yes or No
Ground truth answer: No, the red line on the left is not the same length as the red line on the right.
Base Model: Yes.
RLAIF-V: Yes.
OPA-DPO: Yes.
DA-CPO: Yes.

Visual Input

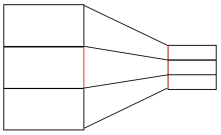


Figure 10: Error analysis on challenging samples, with DA-CPO trained from LLaVA-v1.5-13B.

Prompt for Training the Hallucination Assessor

You are an expert evaluator for multimodal responses. Given an image, an instruction, a human-corrected standard response, and a generated response, your task is to diagnose hallucinations in the generated response at the sentence level.

Input:
Image: {image}
Instruction: {instruction}
Standard response: {standard response}
Generated response: {generated response}

Evaluation criteria:
Evaluate each sentence in the generated response according to its accuracy and relevance to the image and instruction. Use the image as the primary evidence and the standard response as a human-corrected reference.

Assign a score from 1 to 4 to each sentence:
4: The sentence is accurate and relevant, requiring no adjustment.
3: The sentence is mostly correct but requires minor adjustment.
2: The sentence contains significant changes or errors, such as an incorrect object recognition or an incorrect relationship between objects.
1: The sentence is completely irrelevant or incorrect, with no relation to the image or the instruction.

Assign one error category to each sentence:
correct: The sentence is accurate or requires only minor adjustment.
image_recognition_error: The sentence contains incorrect visual interpretation, such as misidentified objects, attributes, counts, actions, or relationships.
language_comprehension_error: The image is correctly understood, but the sentence contains language-level misunderstanding or semantic inconsistency with the instruction.

For each sentence, provide a concise rationale explaining the assigned score, error category, and necessary correction.

Output format:
Return your answer in JSON format. Each item should correspond to one sentence in the generated response:

```
[
  {
    "sentence_id": 1,
    "sentence": "<the original sentence>",
    "score": <1/2/3/4>,
    "error_category": "<correct/image_recognition_error/language_comprehension_error>",
    "rationale": "<brief explanation>"
  },
  ...
]
```

Do not rewrite the generated response. Only provide sentence-level diagnostic annotations.

Table 6: Prompt template used for training M_{assess}.

Prompt for Response Revision

You are an expert assistant for revising multimodal responses. Given an image, an instruction, a generated response, and sentence-level diagnostic annotations, your task is to revise the generated response by correcting hallucinated or inaccurate content.

Input:
Image: {image}
Instruction: {instruction}
Generated response: {generated response}
Diagnostic annotations: {diagnostic annotations}

Revision guidelines:
Use the image as the primary evidence. The diagnostic annotations provide sentence-level hallucination scores, error categories, and rationales.
Revise only the sentences identified as hallucinated or inaccurate. Sentences marked as `correct` should remain unchanged unless a minimal adjustment is necessary for fluency or consistency.
For sentences with `image_recognition_error`, correct the visual content according to the image, including objects, attributes, counts, actions, spatial relations, and other visual details.
For sentences with `language_comprehension_error`, correct the semantic inconsistency or misunderstanding with respect to the instruction.
The revised response should remove unsupported visual claims while preserving the style, structure, and wording of the generated response whenever possible. Avoid adding new information unless it is necessary to answer the instruction correctly.

Output format:
Return your answer in JSON format:

```
{
"revised_response": "<the revised response>"
}
```

Table 7: Prompt template used for M_{rewrite}.