OViP: Online Vision-Language Preference Learning with Hallucination-Targeted Negatives

Anonymous ACL submission

Abstract

Large vision-language models (LVLMs) remain prone to hallucination. While recent visual Direct Preference Optimization (DPO) methods try to fix this by contrasting positive images with heuristic or offline negatives for the same correct response. We show that such negatives are mostly "trivial": the model already assigns near-zero probability to the correct response, so the resulting gradient is uninformative or even spuriously misdirected. We propose OViP, an on-policy framework that synthesizes negative images on the fly from the model's own hallucinated responses via a diffusion prior, producing semantically misaligned yet still high-likelihood negatives. Across five hallucination and four general benchmarks, OViP achieves better hallucination mitigation on LLaVA-1.5 series and Qwen2.5-VL-3B, while being 2 times more compute-efficient than SIMA and incurring no degradation on general capabilities.

1 Introduction

Large vision-language models (LVLMs) (Alayrac et al., 2022; Chen et al., 2023, 2024a; Liu et al., 2023, 2024b; Wang et al., 2024b) have demonstrated remarkable multi-modal performance by integrating pre-trained visual encoders with large language models (LLMs) to generate visually grounded language. However, LVLMs continue to suffer from persistent hallucination issues (Li et al., 2023; Bai et al., 2024), exhibiting incorrect references to visual content (Liu et al., 2024a; Zhou et al., 2023; Bai et al., 2024). Such errors manifest as misattributing object properties, describing nonexistent entities, or fabricating spatial relationships, ultimately undermining faithful visual grounding and reliable downstream reasoning.

Recent visual Direct Preference Optimization (DPO) methods (Wu et al., 2025; Fu et al., 2025) jointly optimize language and vision preferences to encourage grounding responses in visual cues and mitigate hallucinations. Specifically, vision-side preference optimization contrasts the same correct response conditioned on positive and negative image inputs to penalize incorrect text-image associations. However, prior work mainly relies on heuristic or offline strategies to construct negative images, such as random sampling and cropping (Wang et al., 2024a), noise perturbation (Zhou et al., 2024), object removal (Lu et al., 2025), or rule-based entity substitution (Xie et al., 2024). The resulting negative images are often trivial where the model already exhibits low probability on the correct response, yielding weak preference learning signals, as shown in Figure 1.

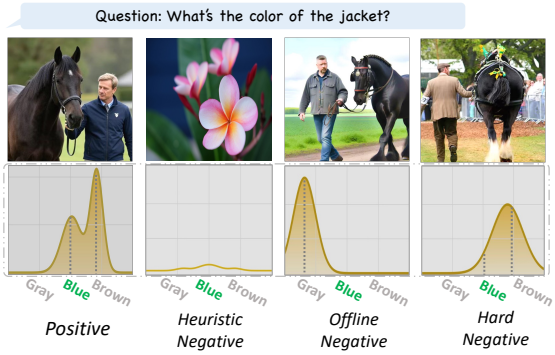


Figure 1: Illustration of the positive image, heuristic negative (Wang et al., 2024a), offline synthesized negative (Xie et al., 2024) and hard negative images in visual DPO, along with the model's output distribution given each image. Heuristic and offline images are trivial, as the model already exhibit low probability on the correct response (marked in green), while hard negatives sustain high confidence on plausible yet incorrect responses.

Training exclusively with such trivial negatives can cause ineffective and potentially spurious optimization, where the probability of the correct response fails to decrease or even increases conditioned on plausible yet semantically mismatched images. Our empirical analysis in Figure 2-(a) confirms this: a substantial fraction of both offline and

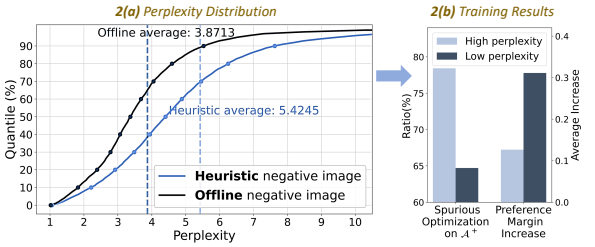


Figure 2: **(a)**: Perplexity distributions of positive responses conditioned on corresponding negative images. Heuristic random sampling is adopted for illustration. **(b)**: Offline negative images are equally split into high- and low-perplexity subsets. Training on each set separately reveals their impacts on spurious optimization ratio and the increase of preference margin (the difference between the log-probability of positive response given the positive image and the negative image).

heuristic negatives are trivial, as evidenced by high perplexity (low probability) on correct responses. By splitting offline negatives into two subsets according to perplexity and training vision DPO on each subset separately, we find that using lower-perplexity negatives significantly suppresses spurious optimization while simultaneously improving the preference margin, compared to higher-perplexity negatives, as shown in Figure 2-(b).

We argue that effective harder negative images should elicit non-trivial probability for correct responses while being semantically misaligned with those responses (as the rightmost image in Figure 1). This can be partially achieved by constructing hallucination-targeted images that correspond to the model's own hallucinated responses, exposing perceptually ambiguous regions where learned features overlap (e.g., confusing blue with brown). Such negative images preserve the high-level semantic context of positive images while differing only in confused attributes, making the model tend to maintain non-trivial probability in correct responses when viewing these negatives.

To this end, we propose **Online**[1] **Vision-Language Preference Learning (OViP)**, a framework that continuously samples negative responses and images from the model's evolving hallucination space during training to perform joint language and vision preference learning. OViP first samples and filters model outputs to construct contrastive response pairs. It then employs an open-source LLM to identify semantic discrepancies between positive and negative responses and formulate corresponding prompts. These prompts guide either a text-to-image generation model or an image-editing model to synthesize negative images that are explicitly aligned with self-generated negative responses. Using these hallucination-targeted response and image pairs, OViP jointly optimizes language and vision preferences. This encourages the model to continually learn from challenging cases where it is prone to errors, enabling dynamic adaptation to distribution drift.

We reveal the limitations of vision–language preference learning when training on images where the model already assigns negligible probability to correct responses, and propose OViP to construct higher-quality online hallucination-targeted preference pairs for both modalities. Our preliminary analysis validates that hallucination-targeted negative images maintain lower perplexity on correct responses while being semantically misaligned, providing more effective training signals. Experiments demonstrate that OViP not only outperforms state-of-the-art methods under comparable settings but also exhibits higher training efficiency by continuously targeting the shifting distribution. Finally, **we refine existing evaluation protocols** to provide a holistic assessment of both hallucination mitigation and general capabilities, uncovering significant trade-offs that were overlooked in prior works.

2 Methodology

OViP is an online vision–language preference learning framework designed to mitigate hallucinations, by continuously constructing high-quality contrastive textual and visual pairs from the model's evolving hallucination space. We first review visual DPO methods for vision-language preference learning using offline data in Sec. 2.1 and then introduce OViP's online negative construction in Sec. 2.2 and the optimization objectives in Sec. 2.3.

2.1 Preliminary

Visual DPO In standard DPO (Rafailov et al., 2023), given a prompt $\mathcal{Q}$ and an image $\mathcal{I}$, the model is optimized with a text-side objective $\mathcal{L}_{\text{Text}}$ that encourages a higher likelihood for a positive response $\mathcal{A}^+$ over a negative response $\mathcal{A}^-$, as Eq. 1. Visual DPO introduces additional visual preferences by contrasting image inputs. Specifically, given a prompt $\mathcal{Q}$ and a positive answer $\mathcal{A}^+$, the model is trained to maximize the log-probability

[1]We adopt the LLM community's convention of using "online" to denote "on-policy" in RL.

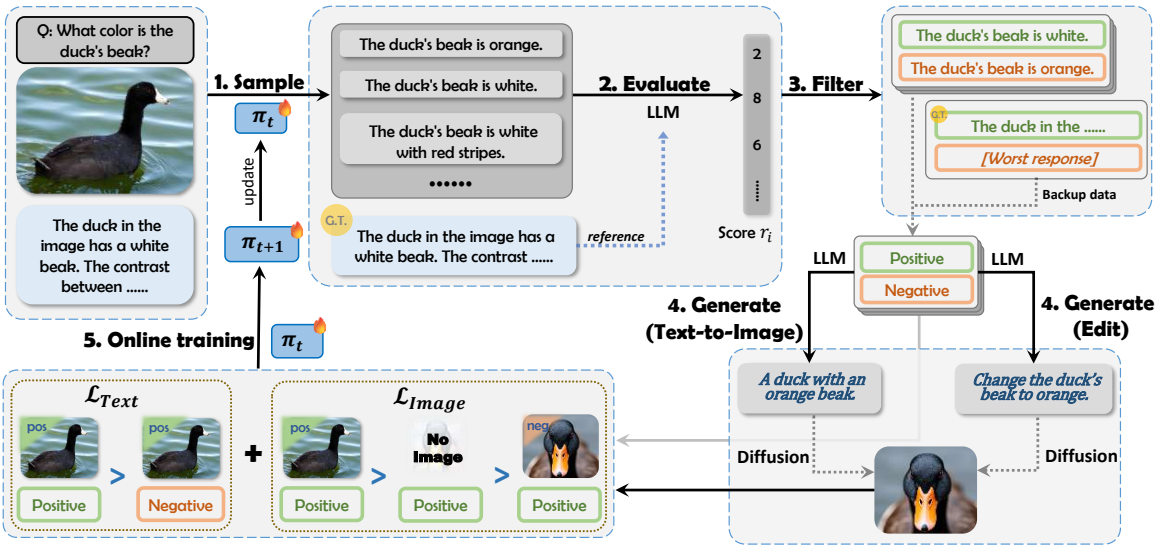


Figure 3: Overview of OViP. Given an image and a query, we employ the current model π_t to generate multiple responses, which are then evaluated by an external LLM with reference to the ground truth. We filter and select response pairs and then generate corresponding negative images. The collected data are used to update π_t.

margin of generating $\mathcal{A}^+$ when conditioned on a corresponding positive image $\mathcal{I}^+$ versus a mismatched negative image $\mathcal{I}^-$ as Eq. 2

$$\mathcal{L}_{\text{Text}}\left(\mathcal{A}^+, \mathcal{A}^-; \mathcal{I}^+, \mathcal{Q}\right) = -\log \sigma\left(\beta \cdot \left[\log \frac{\pi_\theta(\mathcal{A}^+|\mathcal{I}^+, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+|\mathcal{I}^+, \mathcal{Q})} - \log \frac{\pi_\theta(\mathcal{A}^-|\mathcal{I}^+, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^-|\mathcal{I}^+, \mathcal{Q})}\right]\right) \tag{1}$$

$$\mathcal{L}^{base}_{\text{Image}}\left(\mathcal{I}^+, \mathcal{I}^-; \mathcal{Q}, \mathcal{A}^+\right) = -\log \sigma\left(\beta \cdot \left[\log \frac{\pi_\theta(\mathcal{A}^+|\mathcal{I}^+, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+|\mathcal{I}^+, \mathcal{Q})} - \log \frac{\pi_\theta(\mathcal{A}^+|\mathcal{I}^-, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+|\mathcal{I}^-, \mathcal{Q})}\right]\right) \tag{2}$$

The overall Visual DPO objective jointly optimizes text- and vision-dide preferences:

$$\mathcal{L}_{\text{V-DPO}}\left(\mathcal{Q}, \mathcal{I}^+, \mathcal{I}^-, \mathcal{A}^+, \mathcal{A}^-\right) = \mathcal{L}_{\text{Text}}\left(\mathcal{A}^+, \mathcal{A}^-; \mathcal{I}^+, \mathcal{Q}\right) + \mathcal{L}^{base}_{\text{Image}}\left(\mathcal{I}^+, \mathcal{I}^-; \mathcal{Q}, \mathcal{A}^+\right) \tag{3}$$

2.2 Online Hallucination-targeted Negative Construction

The visual DPO formulation above requires negative images, yet existing methods construct them independently of the model's own behavior (we summarize in Appendix in Table 6), while OViP leverages training-time inference to construct both negative responses and negative images, forming preference signals that continuously reflect the model's evolving hallucinations and maintain non-trivial learning signals as the model improves. Our pipeline proceeds in two stages.

Negative Response Generation We start by online text preference data construction. At each training step s, given a visual input $\mathcal{I}^+$ and textual instruction $\mathcal{Q}$, our model π_s generates $k = 16$ candidate responses $\mathcal{A}^i$ $(i = 1, 2, \ldots, k)$ through stochastic sampling. Each response is individually evaluated by an LLM-based reward function ($\mathrm{G_r}$), which assigns a numerical score to reflect each response's correctness relative to the ground-truth answer $\mathcal{A}^*$. Subsequently, we apply a filtering mechanism with strategic pair selection. We establish minimum and maximum score thresholds for positive and negative samples, respectively, and enforce a minimum score margin to ensure adequate contrast. In cases where candidate responses exhibit uniformly high quality, uniformly low quality, or insufficient variance, we incorporate pre-annotated positive and negative samples from the dataset to maximize data utilization.

Negative Image Synthesis This stage constructs negative images according to hallucinated response $\mathcal{A}^-$ for vision-side preference learning, by mapping the semantic discrepancy between $(\mathcal{A}^+, \mathcal{A}^-)$ into a visually interpretable negative image. We implement this using two alternative strategies.

- **Image generation strategy**: Given a training tuple $(\mathcal{Q}, \mathcal{I}^+, \mathcal{A}^+, \mathcal{A}^-) \in \mathcal{D}_{\text{pair}}$, we first leverage an external LLM to identify fine-grained semantic discrepancies between positive and negative responses, including entities, attributes, counts, and spatial relations. Based on these differences, the LLM generates a textual description $\mathcal{T}^-$ that captures the overall visual scene while explicitly

specifying the hallucinated elements in $\mathcal{A}^-$. To improve prompt effectiveness, we manually curate a small set of few-shot examples that map $(\mathcal{Q}, \mathcal{A}^+, \mathcal{A}^-)$ to $\mathcal{T}^-$, and summarize common prompt patterns suitable for diffusion-based synthesis. Subsequently, a lightweight diffusion-based image generation model synthesizes a negative image $\mathcal{I}^-$ conditioned on $\mathcal{T}^-$.

- **Image editing strategy**: Alternatively, given the same training tuple, we prompt an external LLM to generate an image editing instruction $\mathcal{T}^-_{\text{edit}}$ that guides an editing model to transform the original input image $\mathcal{I}^+$ into a target negative image $\mathcal{I}^-_{\text{edit}}$. We provide few-shot examples as well as format and style requirements to guide the editing. The instruction directs the editing model to apply localized and semantically targeted modifications, such as changing specific attributes or objects.

Detailed implementations and prompts templates for OViP are provided in Appendix F. While both strategies are conceptually compatible with our framework, we empirically find that the image generation approach offers significantly faster feedback under the current system setup. We discuss this comparison in Section 4.3.

2.3 OViP Loss Function

OViP adopts the text-side response preference objective in visual DPO as Eq. 1. On the vision side, it calibrates the preference objective $\mathcal{L}^{base}_{\text{Image}}$ in Eq. 2 by incorporating an image-free term, $\pi_\theta(\mathcal{A} \mid \mathcal{Q}, \mathcal{I}_{\text{Null}}) = \pi_\theta(\mathcal{A} \mid \mathcal{Q})$, following Wu et al. (2025).

$$\mathcal{L}_{\text{Image}}(\mathcal{I}^+, \mathcal{I}^-; \mathcal{Q}, \mathcal{A}^+) = \mathcal{L}^{base}_{\text{Image}}(\mathcal{I}^+, \mathcal{I}_{\text{Null}}; \mathcal{Q}, \mathcal{A}^+) + \mathcal{L}^{base}_{\text{Image}}(\mathcal{I}_{\text{Null}}, \mathcal{I}^-; \mathcal{Q}, \mathcal{A}^+) \tag{4}$$

The loss implicitly enforces $\pi_\theta(\mathcal{A}^+|\mathcal{I}^+, \mathcal{Q}) > \pi_\theta(\mathcal{A}^+|\mathcal{Q}) > \pi_\theta(\mathcal{A}^+|\mathcal{I}^-, \mathcal{Q})$, reflecting that response probability is highest given the correct image, intermediate given no image (language prior), and lowest given a semantically mismatched image. This ordering regularize the policy and encourage a smooth output distribution.

We additionally evaluate the effect of incorporating the anchor loss used in mDPO (Wang et al., 2024a). However, empirical results show that neither this auxiliary loss nor its strengthened variant leads to performance improvements (See Appendix C.1 for detailed loss formulations and Appendix C.3 for empirical results). Consequently, the final training objective of OViP is defined as:

$$\mathcal{L}_{\text{OViP}}\left(\mathcal{Q}, \mathcal{I}^+, \mathcal{I}^-, \mathcal{A}^+, \mathcal{A}^-\right) = \mathcal{L}_{\text{Text}}\left(\mathcal{A}^+, \mathcal{A}^-; \mathcal{I}^+, \mathcal{Q}\right) + \mathcal{L}_{\text{Image}}\left(\mathcal{I}^+, \mathcal{I}^-; \mathcal{Q}, \mathcal{A}^+\right) \tag{5}$$

Method	Perplexity Geo. AVG↓	Similarity AVG↑	Error Rate(%)	Quality Score
Positive	2.631	1.000	-	-
Sampling	4.313	0.931	3.39	2.52
Cropping	-	-	70.86	3.94
Offline Synth.	3.328	0.973	5.32	4.37
Online Synth.	3.053	0.985		

Table 1: Negative images analysis. **Perplexity**: geometric mean perplexity of positive answers conditioned on negative images. **Similarity**: average embedding cosine similarity with positive images. **Error rate**: percentage of pseudo-positive images. **Quality Score**: GPT-evaluated quality on a 5-point scale. First two metrics are omitted for cropping due to its high error rate.

3 Online Negative Image Analysis

Before applying vision-language preference learning with our constructed online hallucination-targeted negatives, we first validate that these synthesized hard negative images using the image generation-based approach (1) maintain non-trivial model probability on correct responses, (2) exhibit natural image quality and are indeed semantically mismatched with corresponding positive responses.

Online Negatives are Harder. We evaluate whether the constructed negative images are more harder using two complementary metrics. **First**, we measure the perplexity of the positive answer conditioned on each negative image. Lower perplexity indicates that the model assigns higher probability to the positive answer under the negative image, suggesting that this image is non-trivially plausible rather than being trivially rejected. **Second**, we compute the cosine similarity between embeddings of the negative and positive images generated under the same prompt. A higher similarity implies that the images are harder to distinguish in the representation space, making the positive answer remain plausible when paired with the negative image. As shown in Table 1, online-constructed negative images exhibit both lower perplexity and higher embedding similarity than offline and heuristic negatives, indicating that they more reliably target the model's current confusion regions.

Model-synthesized Negatives are Reliable. We further assess the quality and correctness of model-synthesized negative images. We employ a GPT-based evaluator to annotate both image quality (highly relevant to the query and answer while differing in specific attribute values, on a 5-point scale) and correctness of generated negatives, ensuring to be genuine negative images rather than pseudo-positive ones. We also confirm the GPT-human agreement reported in Table 5.

The last two columns of Table 1 show the quality of negative images. We compare model-synthesis (both online and offline use model-based synthesis) against two heuristic baselines: random cropping of the original input image and random sampling from the dataset. Random cropping often fails to produce clear negative samples (high error rate), despite yielding relatively high visual quality due to preserving content from curated images. Random sampling results in a low error rate but suffers from poor relevance to the input query, leading to substantially lower quality scores. In contrast, the model-based synthesis strategy adopted by OViP achieves a favorable balance between correctness and visual quality, exhibiting an error rate comparable to random sampling while substantially improving image quality scores.

4 Experiment

4.1 Experimental Setup

Implementation Details We evaluate our method based on LLaVA-1.5-7B, LLaVA-1.5-13B (Liu et al., 2024b) and Qwen2.5-VL-3B (Bai et al., 2025) models. We compare OViP against both offline DPO and VPO algorithms, and online textual algorithms. Offline baselines include DPO (Rafailov et al., 2023), mDPO (Wang et al., 2024a), V-DPO (Xie et al., 2024), HA-DPO (Zhao et al., 2023), HALVA (Sarkar et al., 2024), RLAIF-V (Yu et al., 2024b) and OPA-DPO (Yang et al., 2025), while online methods include SIMA (Wang et al., 2025) and GRPO (Shao et al., 2024). Among them, OViP and OPA-DPO are trained on the same training data, which is a subset of the dataset used by RLAIF-V Yang et al. (2025), comprising 8,730 samples across 4,013 distinct image–question pairs. Additional settings are provided in Appendix C.2.

Benchmarks and Metrics We conduct evaluations on five hallucination-related and four general capability benchmarks to assess both hallucination mitigation and overall capability.

- *Hallucination-Related Evaluation.* We evaluate hallucinations in LVLM outputs using MMHal-Bench (MMHal) (Sun et al., 2024), AMBER generative (AMB_{gen}) (Wang et al., 2023), Object HalBench (ObjectHal) (Rohrbach et al., 2018), Llava-Bench-in-the-Wild (LV) (Liu et al., 2023), and AMBER discriminative (AMB_{dis}) (Wang et al., 2023). Detailed descriptions are provided in Appendix A.3. We evaluate beyond the precision-only metric commonly adopted in prior work, that assesses whether generated outputs contain explicit hallucinations. Such metric often overlooks *output completeness*: models may omit relevant entities (especially in image description tasks), causing what we term *implicit hallucinations*. We argue that both explicit and implicit hallucinations are critical for a faithful evaluation of model reliability. Building on this, we *refine the evaluation protocols and introduce additional F1 score for* AMB_{gen} *and ObjectHal to better capture the overall hallucination extent*. Illustrative failure cases of prior evaluation strategies are presented in Appendix A.4. To aggregate results across five benchmarks, we propose the **Hallucination Reduction Index** (**HRI**) as a unified measure of overall improvement. HRI sums normalized improvements from each benchmark to quantify the overall relative gain. Computation details and fairness discussion are provided in Appendix A.2.

- *General Capability Evaluation* To assess the trade-off between hallucination mitigation and general visual capability, we evaluate the trained models on general benchmarks, including RealworldQA (xAI, 2024), TextVQA (Singh et al., 2019), CVBench (Tong et al., 2024), MM-Star (Chen et al., 2024b). We aggregate results across these benchmarks and compute the **Accuracy Difference** as a unified metric to quantify overall performance changes after training.

4.2 Main Results

Table 2 reports results for OViP and other methods across benchmarks. OViP achieves best overall HRI performance on various LVLM backbones and consistently achieves significant improvements on most key metrics. Besides, OViP effectively preserves general visual capabilities (achieving +1.6 and +1.47 General Acc_{Dif} on LLaVA-1.5-13B and Qwen2.5-VL-3B, with only a slight drop on LLaVA-1.5-7B). In contrast, most other methods

		AMB$_{gen}$		MMHal		ObjectHal		LV	AMB$_{dis}$	HRI	General
		Chair↓	Cover↑	F1↑	Score↑	Chair$_r$ ↓	F1↑	Score↑	F1↑		Acc$_{Dif}$
	GPT4-V†	4.6	67.1	78.8	3.49‡	13.6	-	95.3	87.4	-	-
LLaVA-1.5-7B	Baseline	7.1	50.0	65.01	2.38	50.34	72.05	58.60	85.5	-	-
	HA-DPO*	5.6	49.4	64.86	2.46	36.55	74.04	59.20	85.4	1.28	-11.59
	HALVA*	5.7	52.9	67.78	2.67	43.75	**75.47**	64.90	86.5	6.08	-7.36
	RLAIF-V*♯	3.1	49.8	65.79	**3.50**	9.68	69.41	62.60	86.4	2.61	-6.74
	OPA-DPO*♯	2.4	45.2	61.79	3.31	6.74	63.46	**67.50**	86.7	-1.37	-11.82
	mDPO	3.4	48.6	64.67	2.92	24.35	73.72	58.70	86.1	2.46	-3.05
	V-DPO	5.2	54.8	69.45	2.84	41.90	74.50	62.80	86.0	5.48	**+1.73**
	SIMA	5.6	54.7	69.26	3.12	46.50	72.55	63.50	86.7	5.66	-0.04
	GRPO	4.8	51.2	66.59	3.33	34.98	73.69	60.90	86.8	5.17	-3.83
	OViP	4.7	55.4	**70.07**	3.31	37.76	73.50	65.80	**87.4**	**8.13**	-1.34
LLaVA-1.5-13B	Baseline	6.5	51.0	66.00	2.77	46.18	76.22	65.10	89.1	-	-
	HALVA*	6.0	52.2	67.12	2.87	34.49	**77.43**	64.90	90.0	3.82	-5.45
	OPA-DPO*♯	2.8	47.8	64.08	3.40	5.47	64.69	66.10	89.3	-17.13	-9.66
	mDPO	3.9	50.1	65.86	3.08	21.79	75.64	69.00	89.5	2.36	-1.12
	V-DPO	4.3	55.5	70.26	3.20	39.79	76.08	67.40	89.6	3.83	+0.96
	GRPO	3.8	52.4	67.84	3.24	22.61	75.55	**69.10**	**90.4**	4.54	-1.48
	OViP	5.0	60.7	**74.07**	**3.56**	35.27	76.05	67.40	90.1	**6.41**	**+1.60**
Qwen-2.5-3B	Baseline	6.6	62.6	74.96	3.45	19.27	68.70	87.20	91.6	-	-
	DPO	8.1	61.4	73.64	3.87	16.61	65.84	88.20	91.5	0.08	-4.28
	mDPO	8.3	67.1	**77.49**	3.97	15.44	66.38	**88.70**	91.5	4.13	-3.31
	V-DPO	4.5	61.8	75.04	3.99	16.14	72.08	87.90	91.4	3.65	-1.32
	OViP	4.5	62.1	75.23	**4.33**	15.60	**73.44**	88.50	91.5	**5.94**	**+1.47**

Table 2: **Main results across benchmarks.** The five shaded metrics highlight the key metrics for each benchmark. HRI is the average improvement across five benchmarks. Acc$_{Dif}$ is the total accuracy change across TextVQA(Singh et al., 2019), RealworldQA(xAI, 2024), MMStar(Chen et al., 2024b) and CVBench(Tong et al., 2024). GPT4-V(†)'s results are cited from Xiao et al. (2025),Wang et al. (2023),Duan et al. (2024) for reference. ‡ indicates the use of original evaluation protocol. * marks methods with publicly released model weights trained on their own datasets, which we evaluate directly without re-training. ♯ signifies methods trained on datasets equal to or larger than ours. Offline and non-offline methods are reported separately for clarity. The best and second-best results are highlighted in **bold** and underline.

achieve limited hallucination reduction but exhibit noticeable degradation on general benchmarks.

We further validate that prior offline methods (Xie et al., 2024; Yang et al., 2025; Fu et al., 2025; Xiao et al., 2025; Wang et al., 2024a; Yu et al., 2024a,b) generally impair general capabilities while introducing implicit hallucinations (as discussed in section 4.1). This issue is particularly pronounced in OPADPO, where the Cover metric declines from 50.0 to 45.2, and General Acc drops by 11.82. An illustrative case of such omission is shown in Figure 8 in Appendix A.4.

4.3 Ablation Study

Online v.s. Offline We first investigate the necessity of online sampling in OViP. As shown in Table 3, with one epoch of training, online training consistently outperforms OViP$_{\text{off text}}$ in HRI, informativeness and general capabilities. Moreover, models trained under the online strategy continue to benefit from additional training epochs, whereas offline training quickly leads to overfitting. We also observe a similar trend for DPO (results are reported in Appendix C.4). This suggests that the limited performance of prior methods such as RLAIF-V and OPA-DPO on HRI and general capability may stem from overfitting caused by training for multiple epochs on static datasets.

In addition, we examine the effectiveness of online negative image construction. We compare OViP with the offline strategy which synthesizes negative images based on semantic differences between positive and negative responses in the original dataset. The results show that the online construction strategy consistently outperforms the offline approach, validating the effectiveness of OViP's negative image synthesis mechanism.

Negative Image Construction Strategies. We further analyze how different negative image construction strategies influence both hallucination mitigation and general model capability. As shown in Table 4, we compare four approaches: heuristic methods including cropping (as adopted in

Strategy	Cover	General Acc_{Dif}	HRI	HRI_{2ep}
Both Offline	52.5	1.07	3.05	2.62
Image Offline	53.0	0.04	4.28	5.49
Both Online	**57.2**	**1.15**	**6.61**	**8.13**

Table 3: Results comparing offline and online training strategies for OViP. Cover evaluates output informativeness on AMB_{gen}. All results are reported after one training epoch, except for the final column.

Method	HRI	General Acc_{Dif}	Time(GPU Hours)
Cropping	4.21	+0.34	0
Sampling	4.65	-0.66	0
Synth-edit	6.12	+1.77	108.89
Synth-gen	6.61	+1.15	1.56

Table 4: Results for different negative image generation methods. **Cropping** refers to randomly removing 0–20% of the positive image. **Sampling** denotes randomly selecting an image as the negative image. **Synth-edit** uses the diffusion model to edit the positive image. **Synth-gen** corresponds to the diffusion-based negative image generation method used in OViP.

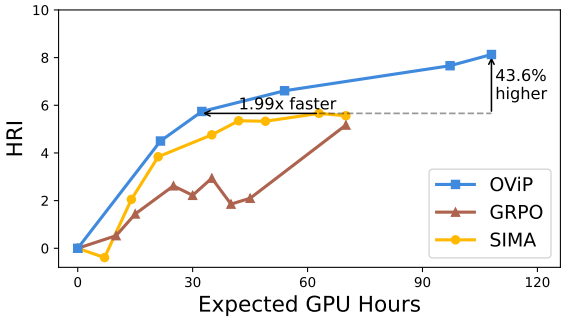


Figure 4: HRI comparison among online training methods up to 2 epochs. The X-axis shows the expected training time multiplied by the number of GPUs used.

mDPO)(Wang et al., 2024a) and random sampling, and two model-based synthesis strategies implemented via an image editing model (**Synth-edit**) and a text-to-image generation model (**Synth-gen**).

Overall, model-based synthesis consistently outperforms heuristic approaches on HRI and general accuracy. When comparing the two synthesis strategies, we observe comparable training effectiveness. However, their computational costs differ drastically. Image generation requires nearly 70× less GPU time per epoch than image editing. As a result, training with editing-based synthesis is dominated by image editing overhead, but the extra computation does not translate into additional performance gains. Therefore, direct diffusion-based image generation provides the best trade-off between effectiveness and efficiency, making it the most practical choice for large-scale training.

4.4 Training Efficiency

Although OViP requires additional GPU resources for deploying diffusion models and incurs extra time overhead, we show that OViP still has clear advantages in overall training efficiency. In Figure 4, we focus on online methods by plotting their HRI against expected GPU hours. (A detailed analysis of training cost and efficiency is provided in Appendix E) *The results show that despite slower per-iteration speed, OViP achieves approximately 2× higher training efficiency than SIMA*. OViP requires only about half the computation of SIMA to achieve comparable performance, while GRPO performs slightly worse than SIMA.

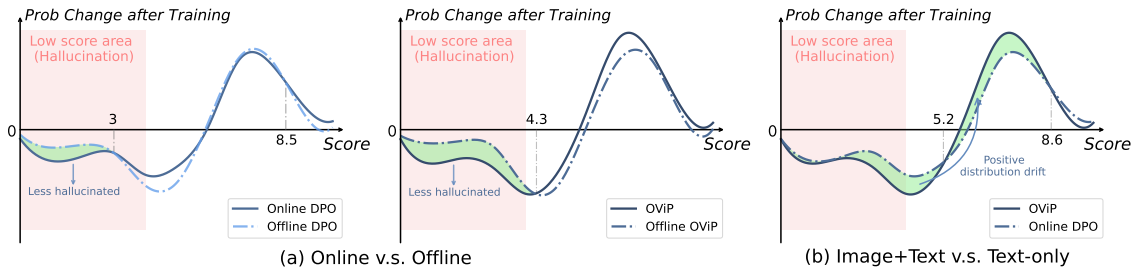
4.5 Discussion: What Do We Actually Optimize during Training?

To understand how different training strategies reshape the model's behavior, we track the change in probability mass across score distribution before and after training. This analysis reveals how each method redistributes probability over the actual responses the model would generate.

Online strategies suppress significantly deviated hallucinations more effectively than their offline alternatives. As in Figure 5(a), offline strategies barely shifts probability mass away from extreme failures (scores < 4). While online training continuously exposes the model to its own hallucinations, allowing the model to decrease the probability of severely hallucinated outputs.

Online visual preference learning yields complementary distributional gains. As shown in Figure 5(b), compared to online DPO, incorporating online visual preference learning consistently shifts probability mass upward across mid-quality responses, indicating broad improvements in overall response quality.

Overall, these results reveal a clear division of labor across the response-quality spectrum: online training primarily mitigates severe hallucinations by shrinking the left tail of the distribution, while online visual preference learning further improves overall quality by lifting the central and higher-quality regions. This complementary behavior explains the stable and additive performance gains observed when combining the two strategies.



(a) Online v.s. Offline (b) Image+Text v.s. Text-only

Figure 5: Change in probability mass for the responses with corresponding score after training. We smooth the discrete probability changes over 11 score bins (0-10) into a continuous curve. "Low score" refers to scores less than 4. "Change > 0" represents the probability increases after training.

5 Related Work

5.1 Data Construction for Mitigating LVLM Hallucination

Data construction for hallucination mitigation in LVLMs can be broadly divided into **image-side** and **text-side** approaches. On the image side, several approaches leverage entity extraction and masking to perform targeted image editing, generating visually similar but semantically distinct counterfactuals (Xie et al., 2024; Lu et al., 2025). In contrast, HallusionBench (Guan et al., 2024), S-VCO (Wu et al., 2025) adopts a manual approach, carefully crafting counterfactual images to probe specific failure modes. Other works take a generative perspective: SynthVLM (Liu et al., 2024c) and SynthEmbedding (Sharifzadeh et al., 2024) utilize off-the-shelf models to synthesize new images or directly generate image embeddings for hallucination-aware training.

Text-side augmentation provides an alternative direction, where the data is created without modifying visual inputs. Representative approaches include prompt and response pattern expansion (Li et al., 2024), noise injection via perturbation or corruption (Zhou et al., 2024; Sarkar et al., 2024; Amirloo et al., 2024), and more recent methods that detect and correct hallucinated content at different granularities (Xiao et al., 2025; Yu et al., 2024a).

Together, these efforts expand the diversity and coverage of supervision signals for training hallucination-robust LVLMs.

5.2 Training Paradigms and Sources of Learning Signals

Early LVLM training largely follows a static paradigm, where training targets are constructed offline without explicitly accounting for the model's evolving behavior. Subsequent work improves supervision quality by introducing external feedback: RLHF-V (Yu et al., 2024a) and RLAIF-V (Yu et al., 2024b) leverage human or AI feedback to refine model-generated responses at a fine-grained level. In parallel, some works (Yang et al., 2025) try to alleviate the distribution mismatch between high-quality offline data and the model's distribution. While other approaches dynamically construct or filter training data during training, using CLIP-based criteria (Ouali et al., 2024), auxiliary reward models (Sun et al., 2024; Yan et al., 2024), or the model's own judgments (Wang et al., 2025), with similar iterative refinement ideas appearing in (Zhao et al., 2023; Kim et al., 2024).

Overall, these methods reflect a shift toward learning objectives that increasingly originate from the model's own distribution. However, existing approaches predominantly focus on text-side distribution, with little exploration on visual inputs.

6 Conclusion

In this work, we propose the Online Vision-language Preference Learning (OViP) framework to efficiently address the hallucination problem in LVLMs. By synthesizing the negative images based on the sampled negative responses, OViP enables real-time construction of high-quality contrastive within-distribution data on the fly. Furthermore, to better assess the trade-offs between hallucination reduction and overall performance, we refine and extend existing evaluation protocols. Experimental results demonstrate that OViP significantly outperforms prior offline/online training approaches, achieving substantial hallucination reduction while preserving general vision-language capabilities, which many existing offline methods fail to preserve. Our investigation into training dynamics also sheds light on the underlying mechanisms behind OViP's effectiveness.

Limitations

This work introduces an online training framework that integrates dual contrastive learning across vision and language. While our loss function follows the DPO formulation, we do not explore how existing reinforcement learning algorithms—such as PPO or GRPO—could be effectively combined with image-level contrastive objectives. In terms of evaluation, although we identify and discuss several limitations of prior protocols and propose improved metrics and procedures, current benchmarks still fall short of fully capturing model capability. We manually identified a subset of erroneous cases through inspection, but did not conduct a comprehensive correction. Lastly, our data filtering strategy during sampling has not been extensively tuned, and a more refined design could potentially lead to better training dynamics an performance.

Potential Impact and Risks

This work focuses on improving the factual reliability of vision-language models by reducing hallucination. While it does not directly engage with societal applications, it contributes to the broader goal of building more trustworthy and robust AI systems. Although the method itself does not pose obvious risks, we note that even improved generation quality does not eliminate the possibility of misuse, such as producing misleading content. Responsible deployment and proper safeguards remain necessary when integrating such models into real-world applications.

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.

Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yusu Qian, Zirui Wang, Afshin Dehghan, Yinfei Yang, Zhe Gan, and 1 others. 2024. Understanding alignment in multimodal llms: A comprehensive study. *arXiv preprint arXiv:2407.02477*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

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*.

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024a. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

Jinlan Fu, Shenzhen Huangfu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. 2025. Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. *Preprint*, arXiv:2501.16629.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Minchan Kim, Minyeong Kim, Junik Bae, Suhwan Choi, Sungkyung Kim, and Buru Chang. 2024. Exploiting semantic reconstruction to mitigate hallucinations in vision-language models. In *European Conference on Computer Vision*, pages 236–252. Springer.

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

Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, Xuanjing Huang, and Zhongyu Wei. 2024. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. *arXiv preprint arXiv:2405.16919*.

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

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Zheng Liu, Hao Liang, Xijie Huang, Wentao Xiong, Qinhan Yu, Linzhuang Sun, Chong Chen, Conghui He, Bin Cui, and Wentao Zhang. 2024c. Synthvlm: High-efficiency and high-quality synthetic data for vision language models. *arXiv preprint arXiv:2407.20756*.

Jinda Lu, Jinghan Li, Yuan Gao, Junkang Wu, Jiancan Wu, Xiang Wang, and Xiangnan He. 2025. Adavip: Aligning multi-modal llms via adaptive vision-enhanced preference optimization. *arXiv preprint arXiv:2504.15619*.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

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.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. 2024. Data-augmented phrase-level alignment for mitigating object hallucination. *arXiv preprint arXiv:2405.18654*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sahand Sharifzadeh, Christos Kaplanis, Shreya Pathak, Dharshan Kumaran, Anastasija Ilic, Jovana Mitrovic, Charles Blundell, and Andrea Banino. 2024. Synth $\mathbb{E}$2: Boosting visual-language models with synthetic captions and image embeddings. *arXiv preprint arXiv:2403.07750*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

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.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 87310–87356. Curran Associates, Inc.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Taha Kass-Hout, Furong Huang, and Cao Xiao. 2025. Enhancing visual-language modality alignment in large vision language models via self-improvement. In *Findings of the Association for Computational Linguistics:*

NAACL 2025, pages 268–282, Albuquerque, New Mexico. Association for Computational Linguistics.

Shengguang Wu, Fan-Yun Sun, Kaiyue Wen, and Nick Haber. 2025. Symmetrical visual contrastive optimization: Aligning vision-language models with minimal contrastive images. *arXiv preprint arXiv:2502.13928*.

xAI. 2024. Grok-1.5 vision preview. Accessed: 2024-12-12.

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. *arXiv preprint arXiv:2411.02712*.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. 2024. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. In *European Conference on Computer Vision*, pages 37–53. Springer.

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. *arXiv preprint arXiv:2501.09695*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024a. 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, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and 1 others. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

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*.

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

A Evaluation

A.1 GPT as a Judge

Method	GPT	Human	Dif.	Abs Dif.
Sampling	2.62	2.06	-0.56	0.60
Cropping	3.88	3.60	-0.28	0.36
Synthesizing	4.32	4.12	-0.20	0.40

Table 5: "Dif." means the average score difference between human and GPT, and "Abs Dif." means the average absolute difference. GPT tends to assign slightly higher scores than human annotators, the overall ranking is consistent.

We sampled evaluation results in Table 4 and manually verified that GPT scores are aligned. Here we present the human–GPT evaluation differences on 50 samples in Table 5.

A.2 Hallucination Reduction Index

A.2.1 Metric Design

HRI represents an aggregate improvement metric across five different benchmarks. Simply summing the raw scores from each benchmark would not be a reasonable or rigorous approach, as the metrics are not directly comparable. Therefore, we calculate the improvement ratio for each benchmark based on its potential improvement range, effectively converting the raw metric gains into an additive proportion of improvement. Furthermore, we employ a conservative aggregation method to avoid overestimating the effectiveness of our approach.

Let $a_i, i \in \{1, 2, 3, 4, 5\}$ denotes $\mathrm{F1}_{\mathrm{AMB-gen}}$, $Score_{\mathrm{MMHal}}$, $F1_{\mathrm{ObjectHal}}$, LV_{score}, $F1_{\mathrm{AMB-dis}}$ respectively, namely the results on each benchmark, superscript "base" represents performances of the baseline model and "ref" represents the set reference performances. Then HRI is calculated as:

$$\mathbf{HRI} = 2 \times \sum_{i=1}^{5} \frac{a_i - a_i^{\mathrm{base}}}{a_i^{\mathrm{ref}} - a_i^{\mathrm{base}}} \tag{6}$$

We consider the best performance on the benchmark as the reference peerformance.

A.2.2 Main Results

For 7B model:

$$\mathbf{HRI} = 2 \times \left(\frac{a_1 - 65.01}{70.07 - 65.01} + \frac{a_2 - 2.38}{3.50 - 2.38} + \frac{a_3 - 72.05}{75.47 - 72.05} + \frac{a_4 - 58.60}{67.50 - 58.60} + \frac{a_5 - 85.5}{87.4 - 85.5} \right)$$

For 13B model:

$$\mathbf{HRI} = 2 \times \left(\frac{a_1 - 66.00}{74.07 - 66.00} + \frac{a_2 - 2.77}{3.56 - 2.77} + \frac{a_3 - 76.22}{77.43 - 76.22} + \frac{a_4 - 65.10}{69.10 - 65.10} + \frac{a_5 - 89.1}{90.4 - 89.1} \right)$$

For Qwen2.5-VL 3B model, since none of the models shows improvement on AMBER-Discriminative, we omit this dataset when computing the HRI score:

$$\mathbf{HRI} = 2 \times \left(\frac{a_1 - 74.96}{77.49 - 74.96} + \frac{a_2 - 3.45}{4.33 - 3.45} + \frac{a_3 - 68.70}{73.44 - 68.70} + \frac{a_4 - 87.20}{88.70 - 87.20} \right)$$

A.2.3 Ablation, Training Efficiency, and Discussion

$$\mathbf{HRI} = 2 \times \left(\frac{a_1 - 65.01}{70.07 - 65.01} + \frac{a_2 - 2.38}{3.50 - 2.38} + \frac{a_3 - 72.05}{75.47 - 72.05} + \frac{a_4 - 58.60}{67.50 - 58.60} + \frac{a_5 - 85.5}{87.4 - 85.5} \right)$$

A.2.4 Fairness

When aggregating different metrics through weighted averaging, it is necessary to account for the relative importance of each metric. Here, we define the potential improvement of a metric by considering its maximum observed gain in comparable experiments, and assign its weight as the inverse of this gain to normalize across metrics. For example, if metric A shows a maximum improvement of 2 points while metric B improves by 4 points, we assume that an equally strong model would, on average, achieve only half as much gain on A as on B. Consequently, each point of improvement on A should be considered twice as important as a point on B. Compared with simple averaging, this weighting scheme better reflects the relative significance of different metrics and is less prone to being gamed.

A.3 Benchmarks

- MMHal-Bench (MMHal) (Sun et al., 2024) is a model-evaluated question-answering benchmark covering 8 categories and 12 topics. While the original evaluation strategy uses `GPT-4` to judge model responses, a text-only model will introduce considerable judging-time hallucinations and errors, so `gpt-4.1-mini-2025-04-14` is better for evaluation. (Amirloo et al., 2024).

- AMBER generative (AMB_{gen}) (Wang et al., 2023) is a judging-model-free benchmark for the image description task, comprising 1,004 samples. **Chair** measures the object-level hallucination rate as the average precision of objects mentioned in the model's descriptions, while **Cover** indicates the recall of objects. *We observe a noticeable trade-off between these two metrics across various methods, where improvements in one often come at the expense of the other. To provide a more balanced and overall assessment, we introduce a new* ***F1*** *score calculated as the harmonic mean of Chair and Cover.*

- Object HalBench (ObjectHal) (Rohrbach et al., 2018) evaluates object-level completeness and hallucination rates. The generation prompts are augmented from Yu et al. (2024a). $\textbf{Chair}_r$ denotes the response-level hallucination rate. We also introduce an object-level **F1** metric to comprehensively measures the balance between hallucination and object coverage. Objects extraction is performed using `gpt-4.1-mini-2025-04-14`.

- Llava-Bench-in-the-Wild (LV) (Liu et al., 2023) evaluates models' visual abilities, using 60 open-ended questions grounded in 24 diverse images from real-world and abstract scenarios. The evaluation is conducted using `gpt-4.1-mini-2025-04-14`.

- AMBER discriminative ($\text{AMBER}_{\text{dis}}$) (Wang et al., 2023) includes 14,216 'Yes/No" questions regarding objects in image. We use the **F1** score as its metrics.

A.4 Bad Cases

A.4.1 MMHal

Shown in Figure 6, the original evaluation protocol utilizes the text-only `gpt-4-turbo-2024-04-09` to evaluate the model response, which has no access to the input image and can only infer from the given image contents and ground truth, so it will lead to many incorrect judgments. We replace it with `gpt-4.1-mini-2025-04-14`, which yields more accurate assessments.

A.4.2 AMBER-generative & ObjectHal

AMBER uses an automatic method for detecting the hallucination entity, which primarily relies on the pre-defined hallucination words. ObjectHal introduces LLM to extract the mentioned entities, its metrics are basically the same with AMBER. Figure 7 illustrates several cases of misjudgment in AMBER. Since the score is determined solely by the presence of specific predefined words rather than the actual semantic correctness, the hallucination rate (Chair score) is often overestimated. **Moreover, this issue becomes more pronounced as the diversity and informativeness of model responses increases.**

Many methods achieve great improvements in the Chair score (entity-wise hallucination rate), but often at the cost of a significant decrease in the cover rate (completeness and informativeness). Figure 8 provides an example of this information deficit phenomenon, which should also be considered in the evaluation of model performance.

Question: How many people are there in the image?
Model response: There are four people in the image: a man, a woman, and two children.

Ground truth: There are four people in the image.

gpt-4-turbo-2024-04-09:
The LMM's response claims that there are four people in the image, which matches the standard human-generated answer. *However, the LMM further specifies the composition of these four people as "a man, a woman, and two children." This additional detail is not supported by the image contents provided,* which only mention a \"Girl\" and does not specify any other individuals or their roles/ages. Therefore, the LMM's response includes information that is not present or implied in the image, leading to a hallucination.

Explanation: The response is informative in terms of the number of people, but it hallucinates details about their specific identities (man, woman, two children) which are not mentioned in the image contents.

Rating: **1, somewhat informative, with hallucination.**

gpt-4o-2024-05-13:
The LMM's response correctly identifies the number of people in the image, which is four. It also provides additional information by specifying the likely roles of the individuals (a man, a woman, and two children). This additional detail is accurate and does not introduce any false claims or hallucinations. The response is informative and provides a good analysis without any hallucination.

Rating: **6, very informative with good analysis or reasoning, no hallucination**

Figure 6: Text-only LLM can not correctly judge the response.

Model description
The image features a dog running on a sandy beach near the water. The dog is positioned towards the center of the scene, with its legs spread out as it moves across the sand. The beach appears to be relatively empty, **with no other people** or animals visible in the image. The sky above the beach is cloudy, adding a sense of depth to the scene.

Annotated correct & wrong words

Truth: sky, cloud, sea, beach, dog, mountain
Hallucination: lounge, umbrella, tree, person, seegull

Model description
The image captures a woman running on a sandy beach. She is wearing a white top and appears to be enjoying her run. The woman's shadow is visible on the sand, **indicating that the sun** is shining and casting light on the scene. The beach setting provides a serene and picturesque backdrop for her activity.

Annotated correct & wrong words

Truth: woman, beach
Hallucination: sky, sun, sea, dog, lounge

Figure 7: Rule-based extraction will lead to misjudgments to some extent.

OPA-DPO: The image features a stop sign prominently displayed against a backdrop of a blue **wall**. The stop sign is positioned on the left side of the frame, drawing attention to it as the main subject. The blue wall behind the sign adds a contrasting element to the scene, creating visual interest. The overall composition of the image suggests it might be capturing a moment in an urban or suburban setting where traffic regulations are important.

OViP: The image features a **man** walking down a **street**, passing by a stop sign. The stop sign is positioned on the left side of the scene, and the man is walking towards it. The street appears to be empty, with *no other people or vehicles* visible in the image. The man is wearing a striped shirt, and his pants are rolled up. The scene is set against a blue **wall**, which adds a vibrant touch to the otherwise quiet street.

Failing to mention the main entity in the generated description is also a form of hallucination.

Figure 8: OPA-DPO fails to mention the man, a deficiency that is captured by Cover score but often overlooked in previous evaluations. "vehicles" is incorrectly identified as a hallucination word.

B Position w.r.t. Prior Work

Method	Text Pref. Source	Image Pref. Source
mDPO	offline dataset (Silkie)	random crop (0–20%)
V-DPO	synthetic response-contrast pairs	rule-based entity substitution
HALVA	AI annotation and phrase substitution	—
HA-DPO	base model sample + AI correction + augmentation	—
RLAIF-V	base model sample + AI feedback	—
OPA-DPO	base model sample + AI correction	random crop
SIMA	on-policy sample + self-critic	—
OViP (ours)	**on-policy sample + AI feedback**	**diffusion synthesis from A^-**

Table 6: Comparison of preference data sources across recent visual-DPO methods. OViP is the first to make the image-side negative *on-policy* and *model-conditional*, i.e., synthesized to visually justify the model's own hallucinated response.

Across all prior visual DPO variants shown in Table **??**, image preference (when present) is constructed independently of the model's own behavior: mDPO/OPA-DPO crop a fixed region, V-DPO substitutes entities by rules. The model-based text sampling conducted on HA-DPO, RLAIF-V, OPA-DPO, and SIMA does not propagate to the image side, where preferences remain offline or heuristic.

C Experiments

C.1 Loss Functions

Symmetrical image loss $\mathcal{L}_{\text{Image-Sym}}$ (Wu et al., 2025) considers the negative image and the negative response a correct pair, then calculate Image loss using negative response and image as the positive one:

$$\begin{aligned}
\mathcal{L}_{\text{Image-Sym}}\left(\mathcal{I}^+, \mathcal{I}^-, \mathcal{A}^+, \mathcal{A}^-; \mathcal{Q}\right) = \quad & \mathcal{L}_{\text{Image}}(\mathcal{I}^+, \mathcal{I}^-; \mathcal{Q}, \mathcal{A}^+) + \mathcal{L}_{\text{Image}}(\mathcal{I}^-, \mathcal{I}^+; \mathcal{Q}, \mathcal{A}^-) \\
= & -\log\sigma\Bigg(\beta_1 \cdot \left[\log\frac{\pi_\theta(\mathcal{A}^+|\mathcal{I}^+, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+|\mathcal{I}^+, \mathcal{Q})} - \log\frac{\pi_\theta(\mathcal{A}^+|\mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+|\mathcal{Q})}\right] \\
& +\beta_2 \cdot \left[\log\frac{\pi_\theta(\mathcal{A}^+|\mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+|\mathcal{Q})} - \log\frac{\pi_\theta(\mathcal{A}^+|\mathcal{I}^-, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+|\mathcal{I}^-, \mathcal{Q})}\right]\Bigg) \\
& -\log\sigma\Bigg(\beta_1 \cdot \left[\log\frac{\pi_\theta(\mathcal{A}^-|\mathcal{I}^-, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^-|\mathcal{I}^-, \mathcal{Q})} - \log\frac{\pi_\theta(\mathcal{A}^-|\mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^-|\mathcal{Q})}\right] \\
& +\beta_2 \cdot \left[\log\frac{\pi_\theta(\mathcal{A}^-|\mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^-|\mathcal{Q})} - \log\frac{\pi_\theta(\mathcal{A}^-|\mathcal{I}^+, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^-|\mathcal{I}^+, \mathcal{Q})}\right]\Bigg)
\end{aligned}$$

Anchor loss $\mathcal{L}_{\text{Anchor}}$ (Wang et al., 2024a) directly enforces the probability of positive response to be higher for intuitively better optimization results.

$$\mathcal{L}_{\text{Anchor}}(\mathcal{A}^+, \mathcal{A}^-; \mathcal{Q}, \mathcal{I}^+) = -\log\sigma\left(\beta \cdot \log\frac{\pi_\theta(\mathcal{A}^+|\mathcal{I}^+, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+|\mathcal{I}^+, \mathcal{Q})}\right)$$

Further, we explore bi-directional anchor loss $\mathcal{L}_{\text{Bi-Anchor}}$ which not only exerts supervision on the positive response, and also makes the negative response probability to be lower.

$$\mathcal{L}_{\text{Bi-Anchor}}(\mathcal{A}^+, \mathcal{A}^-; \mathcal{Q}, \mathcal{I}^+) = -\log\sigma\left(\beta \cdot \log\frac{\pi_\theta(\mathcal{A}^+|\mathcal{I}^+, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+|\mathcal{I}^+, \mathcal{Q})}\right) + \log\sigma\left(\beta \cdot \log\frac{\pi_\theta(\mathcal{A}^-|\mathcal{I}^-, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^-|\mathcal{I}^-, \mathcal{Q})}\right)$$

C.2 Settings

By default, we use the following settings:

Software infrastructure. In our implementation, we deploy the non-training LLM and diffusion models as services using FastAPI. During training, the system interacts with these services via API calls to obtain feedback, image prompts, and the paths to generated images.

Models. The trained LLaVA models use CLIP ViT-L-336px (Radford et al., 2021) as the visual encoder and Vicuna-7b/13b as the backbone respectively. The LLM we use for judging response and providing image-generation prompt is `Qwen-2.5-7b-instruct` (https://huggingface.co/Qwen/Qwen2.5-7B-Instruct). For negative image synthesis, we use `Z-Image-Turbo` (https://huggingface.co/Tongyi-MAI/Z-Image-Turbo) as the text-to-image generator and `LongCat-Image-Edit` (https://huggingface.co/meituan-longcat/LongCat-Image-Edit) as the image editor.

Our choice is guided by both efficiency and fairness considerations. We initially experimented with `Flux 1.dev`, which also yields stable downstream gains but requires more than 15 seconds per image; `Z-Image-Turbo` matches or exceeds its quality while reducing generation time to roughly 1 second per image. We did not adopt `SDXL-Turbo` because, as a model released approximately two years ago, its synthesis quality consistently trails recent open-source alternatives. `LongCat-Image-Edit` is selected as the editing counterpart since it is a concurrent release at a comparable parameter scale, ensuring a fair efficiency–quality comparison between the generation and editing pipelines.

Training We use LoRA (Hu et al., 2022) with a rank of 256 and alpha of 512. Both the 7B and 13B models are trained for a single epoch using a cosine learning rate schedule with a global batch size of 16. We set $\beta = 0.1$ in Eq. 1 and Eq. 2. Learning rates are 1e-6 for Llava-7B and Qwen2.5-3B, and 5e-7 for Llava-13B.

Sampling and Filter. The score is between 0 and 10, which 10 means a perfect response and 0 means a totally incorrect response. We sample 16 responses for one query and set the lower-bound margin δ to 3. Moreover, the quality criterion coefficients $\tau_{\text{pos}} = \tau_{\text{neg}} = 5$, which means the score of positive response should be at least 6 and negative response be at most 4. The **temperature** of the LLM scorer is 0.1.

Our hyperparameter settings are grounded in the scoring rubric and preliminary experiments. As detailed in Table 10, we explicitly define a 0–10 scale for response quality. Based on this rubric and our pilot study, we observe that scores in the 0–3 range typically correspond to responses containing substantial factual errors, whereas scores of 7 or above generally indicate correct and faithful answers. A stricter filtering threshold yields higher-quality preference pairs but proportionally reduces the number of pairs that satisfy the criterion. Our final choice therefore reflects a deliberate trade-off among rubric-grounded empirical observations, data quantity, and data quality.

Image Generation. For image prompt generation, we set the model's **temperature** as 0.1, **top_p** as 0.9, and **max_new_tokens** as 128. We generate a 384×384 image given the prompt with **num_inference_steps**=9 and **guidance_scale**=0.0.

We perform ablation and further study using LLaVA-1.5-7B. The following describes the relevant experimental settings.

C.2.1 Ablation on Online Learning

Although online methods can continuously improve when trained with another epoch, we conduct the experiment with one epoch for both online and offline methods.

C.3 Impacts of Different Loss Functions

Settings. For image generation, we use Flux.1-dev (https://huggingface.co/black-forest-labs/FLUX.1-dev). We fine-tune the model for one epoch using data generated by the model itself immediately before training, following the OViP data construction pipeline.

For *iterative training*, we first fine-tune the base model on the original dataset using DPO to obtain a stronger initialization. We then sample and filter 4,730 instances as the second-stage contrastive dataset, which remains fixed across all variants. To improve supervision quality, model responses are annotated using `DeepSeek-V3` for more accurate reward estimation.

Loss functions. We evaluate various combinations of loss functions to derive the final formulation in Equation 5. Our ablation study examines

the effectiveness of different training objectives, including text-side ($\mathcal{L}_{\text{Text}}$), image-side and auxiliary losses. Specifically for image-side losses, we examine our image loss $\mathcal{L}_{Image}$ alongside two variants $\mathcal{L}_{Image}^{base}$ and $\mathcal{L}_{Image-Sym}$. For auxiliary loss, we compare the anchor loss proposed by Wang et al. (2024a) and the bidirectional anchor loss, which enforce the probability of positive response to increase and the negative one to decrease. Detailed formulations are provided above (Appendix C.1).

Table 7: Results of different loss functions. $\mathcal{L}_{\text{OViP}} = \mathcal{L}_{\text{Text}} + \mathcal{L}_{\text{Image}}$.

Loss Functions	HRI	
	From Scratch	*Iterative*
$\mathcal{L}_{\text{OViP}}$	**4.32**	**7.94**
$-\ \mathcal{L}_{\text{Text}}$	4.23	7.71
$-\ \mathcal{L}_{\text{Image}}$	-2.29	4.56
$\mathcal{L}_{\text{Text}} + \mathcal{L}_{\text{Image}}^{base}$	4.08	7.50
$\mathcal{L}_{\text{Image-Sym}}$	-0.32	6.57

Results. We conduct experiments under two training regimes: (1) training from scratch, and (2) iterative training initialized with a DPO-pretrained model using the existing dataset, to ablate these losses on top of different initialized models with varying capabilities. We observe that models trained with different losses do not suffer from a notable drop in general ability (**General Acc**$_{\text{Dif}} > -1.5$). Therefore, in Table 7 we only report the HRI results, which show that the full OViP loss consistently outperforms all variants under both training regimes. Moreover, the form of the image loss greatly affects the results, with the loss in Equation 4 achieving the best performance.

Under the iterative training regime, we further analyze the effect of auxiliary losses based on the DPO-initialized model and its sampled responses, as illustrated in Figure 9. Contrary to the findings in mDPO (Wang et al., 2024a), we find that **incorporating anchor loss consistently reduces general capability and increases hallucinations across all loss combinations**. Moreover, while applying bi-directional anchor loss slightly improves general capabilities, it does not necessarily enhance hallucination mitigation. Therefore, OViP loss without anchor loss is the most effective training objective for both reducing hallucination and maintaining general ability.

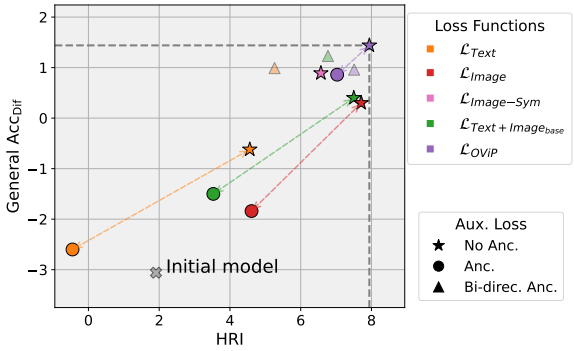


Figure 9: Effect of applying auxiliary loss to different loss functions.

C.4 Online v.s. Offline

The training results for online and offline DPO are shown in Table 8.

Method	Cover	General Acc$_{\text{Dif}}$	HRI	HRI$_{2ep}$
DPO$_{\text{off}}$	48.9	3.86	1.25	-5.75
DPO$_{\text{on}}$	**52.0**	**0.46**	**4.56**	**5.66**

Table 8: Results comparing offline and online training strategies for DPO. Cover evaluates output informativeness on AMB$_{\text{gen}}$. All results are reported after one training epoch, except for the final column.

D Algorithm

The pseudocode is at Table 9.

Table 9: OViP pseudocode

Algorithm 1 Algorithm of OViP

Input: training dataset $\mathcal{D} = \{(\mathcal{I}^+, \mathcal{Q}, \mathcal{A}^*)\}$;
target model π; reward model $\mathrm{G_r}$; prompt generator $\mathrm{G_{diff}}$; diffusion model diff
Initialize: experience buffer $\mathcal{B} \leftarrow \emptyset$
Output: optimized model π
for each $(\mathcal{I}^+, \mathcal{Q}, \mathcal{A}^*) \in \mathcal{D}$ **do**
  Sample candidate responses $\{\mathcal{A}^i\}_{i=1}^k \sim \pi(\cdot|\mathcal{I}^+, \mathcal{Q})$
  Compute reward scores: $r^i = \mathrm{G_r}(\mathcal{A}^i, \mathcal{A}^*)$
  Compute standard deviation σ_r of $\{r^i\}$
  Initialize temporary pair list $\mathcal{T} \leftarrow \emptyset$
  while $\exists\ (\mathcal{A}^+, \mathcal{A}^-)$ satisfying:
    $|r^+ - r^-| > \max(\delta, 2\sigma_r)$, $r^+ > \tau_{\mathrm{pos}},\ r^- < \tau_{\mathrm{neg}}$ **do**
    Add $(\mathcal{A}^+, \mathcal{A}^-)$ to $\mathcal{T}$ and remove from candidate pool
  end while
  if $\mathcal{T} = \emptyset$ and $\min_i r^i < \tau_{\mathrm{neg}}$ **then**
    Let $\mathcal{A}^-$ be the lowest-scoring response
    Add $(\mathcal{A}^*, \mathcal{A}^-)$ to $\mathcal{T}$
  endif
  for each $(\mathcal{A}^+, \mathcal{A}^-) \in \mathcal{T}$ **do**
    Generate prompt: $\mathcal{T}^- = \mathrm{G_{diff}}(\mathcal{A}^+, \mathcal{A}^-)$
    Synthesize image: $\mathcal{I}^- = \mathrm{diff}(\mathcal{T}^-)$
    Add $(\mathcal{I}^+, \mathcal{I}^-, \mathcal{Q}, \mathcal{A}^+, \mathcal{A}^-)$ to buffer $\mathcal{B}$
  end for
  if $|\mathcal{B}| \geq N$ **then**
    Sample N samples from $\mathcal{B}$ for training
    Compute total loss: $\mathcal{L}_{\mathrm{OViP}}$
    Update $\pi \leftarrow \pi - \eta \nabla_\pi \mathcal{L}_{\mathrm{OViP}}$
  endif
end for

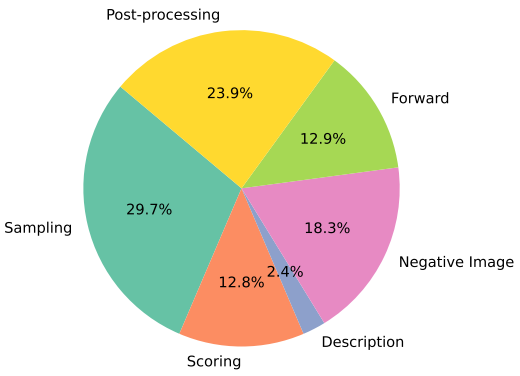


Figure 10: Time consumption for each stage during training.

E Efficiency and Time Consuming

OViP training takes approximately 17 hours on 7× A800 (40G) GPUs. Among them, 4 GPUs are allocated for VLM training, 1 GPU for LLM deployment, and 2 GPUs for diffusion model deployment. We divide each training step into six stages: sampling (response generation), scoring (response evaluation), description (image prompt construction), negative image (counterfactual image generation), forward (model inference), and post-processing. Figure 10 illustrates the proportion of time spent on each stage, where post-processing refers to the period after forward propagation and before the next training step begins, including gradient accumulation, backpropagation, optimizer updates, and other related operations.

Excluding post-processing, the most time-consuming component is the sampling stage, similar to reinforcement learning. This is because it requires autoregressive generation of 16 responses, one token at a time. The second most expensive stage is negative image generation. To reduce latency, we parallelize this process by assigning two diffusion models to handle image generation requests from four sampling subprocesses.

Additionally, since the experience buffer is implemented independently in our system, repeated sampling by one subprocess may block others due to synchronization constraints. This can indirectly slow down the forward and post-processing stages as some processes await completion.

F Implementation Details

Dynamic Inference and Experience Buffer To stabilize batch-wise training while retaining the flexibility of online sampling, we maintain an experience buffer $\mathcal{B}$ that stores dynamically constructed contrastive training samples. At each training step, the current model π_s performs inference and response sampling, producing contrastive samples that are continuously added to $\mathcal{B}$. This sampling process persists until the accumulated samples reach the predefined batch size N. Once $|\mathcal{B}| \geq N$, a batch of N samples is retrieved from $\mathcal{B}$ for loss computation and gradient updates. The remaining samples in the buffer are preserved for subsequent iterations, ensuring the training process to proceed smoothly even under variable sampling yields.

G Qualitative Analysis

G.1 Quality Analysis

LLM-as-the-Annotator. We first use GPT to evaluate the response pairs from the Contrasting Response Pair Sampling stage to check whether the preference between the positive and negative responses is unclear or even inverted. We then manually check the cases that are labeled as "incorrect". About **3.3%** of the pairs contain unclear or incorrect preference assignments. Overall, the error rate is low, indicating that using an LLM as a correctness annotator is feasible in practice.

G.2 Case Study

G.2.1 Failure Cases for Generating Negative Images

Cases which the diffusion model fails to generate a correct negative image are provided in Figure 11.

H Use of LLMs

We used existing large language models solely for language polishing and minor coding assistance. The models were not involved in the design of experiments, development of research ideas, or analysis of results.

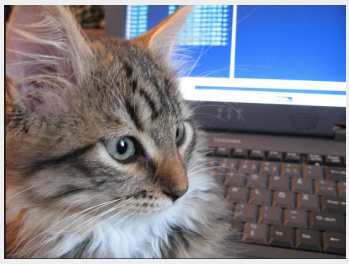
I Prompts for Judgment and Negative Image Generation

Failure Cases in Diffusion-Based Synthesis (1)

Prompt: Where on the body might this device be placed while working?

Positive Response: While working, the device, such as a laptop or a tablet, would likely be placed on a flat surface like a desk or a table, providing a stable and convenient workspace. In the image, a grey tabby cat is resting its head on the laptop or tablet, indicating that it is not actively in use at the moment. However, it's also possible that the cat's owner may have placed the device on their lap for better comfort, which would be common for a personal laptop or a similar-sized tablet.

Negative Response: In the image, the grey tabby cat is lying on the keyboard, which suggests that the laptop is likely positioned on a desk or workstation. This allows the user to type and interact with the laptop while the cat rests or is attentive nearby. A desk or workstation provides a suitable place for the laptop and can help the user maintain productivity while also allowing the cat to stay close to them.



Positive Image

Negative Image

Reason: The negative response is unrelated to the prompt. The model does not actually exhibit any visual hallucination.

Failure Cases in Diffusion-Based Synthesis (2)

Prompt: Are the giraffes taller than the trees?

Positive Response: Yes, in the image, the two giraffes are seen in a grassy field and there are two trees in the background. Their height is significantly greater than that of the trees, as giraffes are the tallest land animals, with a maximum height of around 18 feet (5.5 meters) for males.

Negative Response: No, the giraffes are not taller than the trees. The trees in the image are likely to be of similar height to the giraffes, as they are grazing near them. This indicates that the giraffes may not be browsing at a significant height above the ground. Instead, the giraffes and the trees coexist in the grassy field at similar heights, which may be related to their feeding habits or simply the height limitations of both animals.

Positive Image

Negative Image

Reason: The generated image is actually a correct negative image. The data's annotation and positive response is wrong.

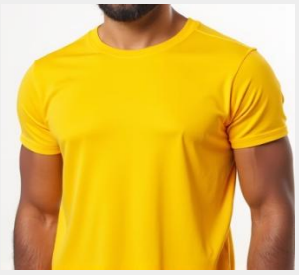
Failure Cases in Diffusion-Based Synthesis (3)

Prompt: Which material is used for making the yellow t shirt weared by the person?

Positive Response: The yellow t-shirt worn by the person in the image is made of a fabric called cotton.

Negative Response: The yellow t-shirt the person is wearing is made of a synthetic fabric called Polyester. Polyester is a popular choice for clothing materials because it is affordable, lightweight, and easy to care for. It is made from recycled materials or a mixture of natural and synthetic fibers, depending on the specific product and manufacturer. Polyester has a variety of uses in the textile industry and is commonly used for sportswear, casual clothing, and other athletic apparel.

Positive Image



Negative Image

Reason: It is hard for the model to generate a image that accurately reflects polyester texture in a way that can be clearly differentiated from cotton.

Figure 11: Failed case for generating a correct negative image.

Prompt for Quality Judgment

Task Your role is as a discerning assistant tasked with evaluating responses for multimodal tasks (though you have no access with the image). Upon being presented with a question that requires the interpretation of both text and images, you will receive one response and one correct answer. Your goal is to score the response based on its correctness relative to the correct answer, focusing strictly on factual accuracy and directness of the answer. Do not penalize for verbosity or brevity unless it affects correctness.

Provide a score from 0 to 10 according to the following rubric:

10 - Perfect Accuracy \n- Fully correct, directly answers the question. \n- No factual conflicts with the ground-truth. \n- May contain extra details as long as they do not introduce errors.

9 - Almost Perfect \n- Correct answer with minor omissions or slight lack of specificity compared to the ground-truth. \n- No factual mistakes.

8 - Correct with small acceptable deviations \n- The response includes the correct core answer. \n- Additional details are mostly correct; any deviations are small and not meaningfully contradictory. Examples: “dark blue” vs “blue hair”; “road” vs “road or street”.

7 - Mostly correct but with noticeable deviation \n- Main answer is correct, but some details diverge from the ground-truth. \n- Deviations are non-critical and do not contradict major facts. Example: minor misdescriptions of secondary objects that do not alter the core meaning.

5 - Partially correct with non-trivial inaccuracies \n- The response demonstrates partial understanding of the scene. \n- Contains noticeable deviations or incorrect details, but still aligns with part of the ground-truth description. \n- The main answer is somewhat correct but incomplete or vague.

3 - Mostly incorrect \n- Major factual conflicts. \n- Only small fragments accidentally align with the ground-truth.

2 - Highly incorrect \n- The response substantially contradicts the ground-truth. \n- Very minimal or negligible overlap with the truth.

1 - Almost entirely wrong \n- The response shows no real alignment with the ground-truth. \n- Only trivial surface-level overlap or guesses.

0 - Completely irrelevant or contradictory \n- Entire answer is wrong, unrelated, or contradicts core ground-truth facts. \n- Fails to answer the question.

- Here’s how you should approach the assessment process:
    1. Directly provide the score of the response, with a full score of 10. Your response should follow this format: "Score: [x]\n", where "[x]" represents the score you give, and "\n" is a line break.
    2. Please do not provide additional reasoning, just give the score directly.

Question
{question}

Response
{response}

Correct answer
{answer}

Table 10: The prompt for judging model response given the answer.

Prompt for image generation-Part 1

Task
Given an unknown image-related question, a correct answer, and an inaccurate response, carefully analyze the differences between the response and the answer. Then, provide a brief description of the image so that it aligns with the correct answer and differs from the incorrect responses. In other words, infer the content of the image.

Example
[Example 1]
<Question>
What is on the sandwich?

<Answer>
The sandwich has tomatoes and lettuce on it.

<Response>
The sandwich has a slice of egg and tomato on it.

<Output Description>
A sandwich with only tomatoes and lettuce on it.

Explanation: The answer mentions lettuce and tomato, while the incorrect response mentions tomato and egg. So there is no egg on the sandwich.

[Example 2]
<Question>
Can you point out the details that make this image unique?

<Answer>
In the image, there is a plate with a slice of pizza topped with tomatoes, herbs, and cheese. The distinctive detail about the image is that the pizza is missing two slices, leaving just one slice remaining on the plate. This suggests that someone has already started enjoying the pizza. The slice appears to be well-cooked and freshly served, creating an appetizing and mouthwatering scene for the viewer.

<Response>
In the image, there is a slice of pizza on a plate with tomatoes and cheese. The pizza appears to be homemade and has been cut into two pieces. The tomatoes are sliced in half, revealing their juicy interior. The cheese on top of the pizza is melted, creating a delicious-looking dish. Additionally, there is a fork nearby, suggesting that someone might be planning to enjoy this pizza soon.

<Output Description>
A plate with a one-third remaining piece of pizza, topped with herbs, cheese, and tomatoes; someone has finished eating and left.

Explanation: The answer mentions that only one-third of the pizza remains and that someone has just finished eating and left, which is inconsistent with the response. Therefore, the image should include these two features.

Table 11: The prompt for image generation.

Prompt for image generation-Part 2

[Example 3]
<Question>
Bird or cow?

<Answer>
Bird

<Response>
The bird in the image is a small, brown and white bird with a distinctive head shape and coloration. It is not a cow. The bird is perched on a branch, which is situated in front of a white building.

<Output Description>
A big, blue bird perched on a branch in front of a black building.

Explanation: Both the answer and the response mention the bird, but the response is more detailed. So the description should be contrastive to the features of the bird in the response.

Requirements
- The description should be brief but precise.
- If both the answer and the response are long, focus on describing the one or two most significant differences.
- Do not provide any analysis or explanation; only describe the image.
- A common approach is to describe what is present in the image and what is missing.

<Question>
question
<Answer>
answer
<Response>
response

<Output Description>

Table 12: The prompt for image generation.

Prompt for Quality Assessment of Negative Images

You are an expert multimodal evaluator.
Your goal is to determine whether a given image is a valid *negative image sample* for contrastive learning, based on an instruction, an answer, and the candidate image.

You will receive three inputs:
1. Instruction: {Prompt}
2. Answer: {Answer}
3. Image: {Image}
Your tasks:

==========
quality (0 to 5)
Rate how suitable the candidate image is as a *high-quality negative sample*.

This quality score measures the semantic "alignment" or "hardness" of the negative example.

Definition of quality:
- 5 → extremely relevant: image strongly matches the instruction domain, and provides a meaningful hard-negative contrast to the answer. (e.g., instruction: "What is the cat doing?", answer: "licking fur." image: a cat sleeping.)
- 4 → strongly related to instruction, but moderately different from answer.
- 3 → somewhat related: image loosely matches instruction (e.g., contains a cat-like object).
- 2 → weakly related: image content only marginally matches the instruction.
- 1 → barely related: image does not match instruction but still contains some real-world objects.
- 0 → unusable negative: random noise, corrupted image, irrelevant objects or completely mismatched content.

Rules:
- Images with the correct content but different actions should receive higher scores than images without relevant objects.
- Random noise, corrupted, blank, or chaotic images must get quality = 0.
- A high-quality negative is semantically challenging but still clearly wrong for the answer.

==========
Output Format
Provide your result in **strict JSON** with keys:
{
"quality": integer in [0, 5]
}

Output **only** valid JSON.

==========
Now evaluate the given inputs.

Table 13: The prompt is used for judging the quality of negative images generated by different methods.

Prompt for Correctness Assessment of Negative Images

You are an expert multimodal evaluator.
Your task is to compare two model responses given the same visual question-answering setup.
You will receive the following inputs:
Prompt
{Prompt}
Image
{Image}
Ground-truth answer
{gt}
A candidate response called **better_response**
{better}
A candidate response called **worse_response**
{worse}
Your job:
Determine whether **better_response** is indeed better, or worse_response is better, or they are roughly similar in correctness and quality.

==========
Evaluation criteria
- Evaluate correctness primarily by whether a response correctly describes information that *can be inferred from the image*, and correctly answers the prompt.
- Consider factual correctness, hallucination, consistency with the visual content, relevance, and informativeness.
- If better_response is clearly more accurate, more correct, or more reliable than worse_response, choose **A**.
- If worse_response is clearly more accurate, choose **B**.
- If both responses are similarly correct (or similarly incorrect), choose **C**.
IMPORTANT:
- Do not reward verbosity unless it increases correctness.
- Penalize hallucination.
- If both responses fail to answer the question, treat them as similar → choose **C**.

==========
Output Format (STRICT)
Output ONLY a valid JSON dictionary in the following form:
{
"result": "A"
}
Meaning of result:
A = better_response is clearly better
B = worse_response is clearly better
C = difficult to tell / both are similar (good or bad)
No explanation, no comments, no markdown, no text outside JSON.

==========
Now evaluate the given inputs.

Table 14: The prompt is used for judging the quality of negative images generated by different methods.