Mitigating Visual Hallucinations via Semantic Curriculum Preference Optimization in MLLMs

Abstract

Multimodal Large Language Models (MLLMs) have significantly improved the performance of various tasks, but continue to suffer from visual hallucinations, a critical issue where generated responses contradict visual evidence. While *Direct Preference Optimization* (DPO) is widely used for alignment, its application to MLLMs often fails to capture fine-grained semantic differences and encourages shortcut learning. To address these challenges, we propose *Semantic Curriculum Preference Optimization* (SCPO), a novel framework for MLLM alignment. SCPO employs a progressive, easy-to-hard curriculum built upon our *Semantic Curriculum Preference Pairs* dataset, which provides fine-grained semantic contrasts sorted by difficulty. This curriculum is trained with a dynamic reference model and a novel symmetric, bidirectional objective to facilitate simultaneous learning from both textual and visual preferences. To our knowledge, SCPO is the first framework to unify semantics, symmetry, and curriculum for MLLMs alignment, effectively mitigating visual hallucinations. Extensive experiments on LLaVA models across various scales and versions validate that SCPO demonstrates superior performance compared to baseline models on multiple hallucination benchmarks, reducing the hallucination rate by up to 62.9%. Moreover, evaluations on generalized benchmarks show that SCPO improves factuality while preserving general capabilities, with its performance remaining stable across general vision-language benchmarks.

1 Introduction

Multimodal large language models (MLLMs) (Liu et al., 2024a; OpenAI, 2023; Liu et al., 2024b) have demonstrated remarkable capabilities in tasks requiring fine-grained perception, such as visual question answering and image captioning (Yu et al., 2025). Despite these advances, their reliability is compromised by the persistent challenge of visual hallucination. During inference, MLLMs often rely too heavily on the semantic priors embedded in the language model, while failing to faithfully ground these priors in fine-grained visual evidence from the input image (Li et al., 2022; Kalai et al., 2025; Xu et al., 2024). This misalignment between language and vision produces outputs that contradict the actual visual content, thereby seriously limiting their applicability in high-stakes scenarios such as autonomous driving and medical image diagnosis (Yu et al., 2024a).

To address this challenge, a common approach is to align models with human preferences using frameworks like Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017). RLHF leverages human preferences to guide model optimization, typically by first collecting preference pairs and then employing reinforcement learning techniques such as Proximal Policy Optimization (PPO) to improve output quality and reliability (Schulman et al., 2017). Building on this, *Direct Preference Optimization (DPO)* (Rafailov et al., 2023) simplifies the pipeline by skipping the reward modeling and reinforcement learning stages, directly training on *"chosen–rejected"* preference pairs (Figure 1(a)). Its effectiveness in enhancing instruction-following and reducing undesirable outputs in language models (Touvron et al., 2023) has made it a leading paradigm for preference-based alignment.

Figure 1: Comparison of different DPO-based alignment strategies. *(a) Text-only Preference*: preference is defined over textual responses. *(b) Visually-Conditioned Textual Preference*: preference is defined over textual responses for a fixed image. *(c) Text-Conditioned Visual Preference*: preference is defined over images for a fixed response. *(d) Cross-Modal Symmetric Preference*: preference is defined symmetrically over complete image-text pairs.

Motivated by DPO's success in textual scenarios, recent work has extended it to MLLMs for finer-grained vision-language alignment. Initial efforts largely mirrored the unimodal paradigm, constructing preference pairs from textual responses, while the visual input remained fixed (Zhao et al., 2023; Yu et al., 2024a;b; Yang et al., 2025), as illustrated in Figure 1(b). In this setup, models primarily learn to distinguish factual descriptions from hallucinated ones, but visual evidence is not explicitly incorporated into preference judgment. To overcome this limitation, subsequent work introduced pairs of visual preferences by keeping the textual instruction and *"chosen"* response constant while generating a *"rejected"* image through perturbations of the original (Zhou et al., 2024). This design directly contrasts the original and perturbed images, allowing visual cues to be incorporated more explicitly into preference optimization (Figure 1(c)). Although incorporating visual preferences is a step forward, current methods are still constrained by several key limitations as discussed below.

However, these methods remain constrained by structural limitations. First, perturbation-based visual negatives frequently preserve high-level semantics, allowing models to identify superficial artifacts of image corruption rather than resolving semantic conflicts across modalities. Second, objectives that implicitly prioritize one modality allow preference constraints to be satisfied by amplifying modality-specific correlations, instead of establishing coherent cross-modal grounding. Finally, alignment is commonly performed in a static single-pass manner, where samples of vastly different difficulty are mixed together and the reference policy remains fixed as the model evolves. Such training dynamics hinder the development of stable grounding behaviors, especially when visual cues contradict strong language priors.

To address these challenges, we propose *Semantic Curriculum Preference Optimization (SCPO)*, a novel alignment framework designed to improve vision-language grounding systematically. SCPO integrates semantically challenging preference pairs, a symmetric and bidirectional optimization objective, and a dynamic curriculum-based training strategy. This unified approach enables robust hallucination mitigation while preserving general model capabilities. Our main contributions are as follows.

1. We construct the *Semantic Curriculum Preference Pairs (SCPP)*, a large-scale dataset for fine-grained vision-language alignment. SCPP is built via a novel pipeline that quantifies sample difficulty, enabling an easy-to-hard curriculum to build from foundational grounding to complex semantic reasoning.

2. We propose a novel *symmetric and bidirectional objective* for preference optimization. Designed to target visual neglect, this objective establishes a rigorous semantic alignment mechanism. Our objective first learns from complementary textual (Figure 1(b)) and visual (Figure 1(c)) preferences, and then enforces robust grounding through a symmetric framework (Figure 1(d)). This unified approach prevents the model from relying on shortcut learning and encourages a deeper semantic understanding.

3. We introduce the *SCPO framework*, which systematically integrates a difficulty-based curriculum with dynamic reference model updates for MLLM alignment. Extensive experiments demonstrate that SCPO significantly reduces hallucinations and establishes new state-of-the-art performance on multiple hallucination benchmarks.

2 RELATED WORK

Preference optimization has become a central paradigm for aligning multimodal models. Early multimodal DPO variants such as RLHF-V (Yu et al., 2024a) construct preference pairs by comparing textual responses under fixed visual inputs, focusing primarily on factuality rather than cross-modal grounding. To expose visual signals more explicitly, later works incorporate image-level negatives: mDPO (Wang et al., 2024), POVID (Zhou et al., 2024), and CHiP (Fu et al., 2025) generate rejected samples by perturbing the original image (e.g., noise injection or cropping). Such perturbation-based strategies mainly alter pixel-level statistics while preserving high-level semantics, offering limited supervision for resolving semantic discrepancies between vision and language.

Beyond data construction, recent work explores objective design and training dynamics. HA-DPO (Zhao et al., 2023) and HSA-DPO (Xiao et al., 2025) weight preference signals by hallucination severity, while HALVA (Sarkar et al., 2024) leverages detection-derived priors to emphasize visually grounded regions. OPA-DPO (Yang et al., 2025) addresses off-policy preference learning by aligning expert-revised responses toward the reference policy. Other approaches like S-VCO (Wu et al., 2025) and SymMPO (Liu et al., 2025) introduce symmetric constraints or synthetic counterfactuals. These approaches primarily focus on preference construction or loss formulation, whereas our work views alignment as a progressive process that combines fine-grained semantic preference signals, bidirectional grounding, and staged reference updates.

3 PRELIMINARIES

DPO (Rafailov et al., 2023) is derived from the KL-regularized reward maximization objective commonly used in RLHF:

$$\max_{\pi} \ \mathbb{E}_{x\sim\mathcal{D},y\sim\pi(y|x)}[r(x,y)] - \beta D_{\mathrm{KL}}\big(\pi(y|x) \,\|\, \pi_{\mathrm{ref}}(y|x)\big), \qquad (1)$$

where π denotes the policy to be optimized, π_{ref} is a fixed reference policy, $r(x,y)$ is a reward function, $\beta > 0$ is a regularization coefficient that controls the KL penalty, and $\mathcal{D}$ is a dataset of training prompts x. This objective allows for aligning a model with human preferences without requiring explicit reward modeling or reinforcement learning.

The optimal policy π^* induces a closed-form mapping from the policy to the reward:

$$r(x,y) = \beta \log \frac{\pi^*(y|x)}{\pi_{\mathrm{ref}}(y|x)} + \beta \log Z(x), \qquad (2)$$

where $Z(x)$ is a partition function that depends only on x.

Human preferences are typically observed through pairwise comparisons. Let y_w and y_l denote the chosen and rejected responses, respectively. These comparisons are modeled using the Bradley-Terry model:

$$p^*(y_w \succ y_l \mid x) = \sigma\big(r(x,y_w) - r(x,y_l)\big), \qquad (3)$$

where $\sigma(t) = \frac{1}{1+e^{-t}}$ is the sigmoid function. By substituting equation 2, the partition function $Z(x)$ cancels out, enabling DPO to directly compute the preference likelihoods from model probabilities without the need for explicit reward modeling.

$$r(x,y_w) - r(x,y_l) = \beta \log \frac{\pi^*(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)} - \beta \log \frac{\pi^*(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)}. \qquad (4)$$

Finally, by maximizing the likelihood of observed preferences in the dataset $\mathcal{D}$, we derive the DPO loss:

$$\mathcal{L}_{\mathrm{DPO}}(\pi;\pi_{\mathrm{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log \sigma\Big(\beta \log \frac{\pi(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)} - \beta \log \frac{\pi(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)}\Big)\right]. \qquad (5)$$

This formulation directly aligns the policy π with human preferences in a principled and efficient way, avoiding the instabilities and complexity associated with reinforcement learning.

4 METHODOLOGY

Our proposed SCPO framework, illustrated in Figure 2, consists of three core components. We first detail the construction of our *Semantic Curriculum Preference Pairs* (SCPP) dataset and its organization into a curriculum in Section 4.1. We then introduce the symmetric SCPO objective in Section 4.2. Finally, we describe the iterative alignment strategy with a dynamic reference model is described in Section 4.3.

Figure 2: Overview of the SCPO framework. **(a)** We first construct a preference dataset and partition it into an easy-to-hard curriculum based on semantic difficulty. **(b)** The model is then trained using our symmetric, bidirectional SCPO objective. **(c)** This training proceeds iteratively through the curriculum stages.

4.1 SEMANTICALLY SENSITIVE PREFERENCE DATA CONSTRUCTION

The construction of the SCPP dataset consists of two stages: (i) integration and refinement of public datasets, and (ii) curriculum-based partitioning into easy, medium, and hard sets.

4.1.1 STAGE 1: DATA INTEGRATION AND REFINEMENT

To enable fine-grained alignment, we construct the **SCPP** dataset, a large-scale collection where each sample consists of a prompt paired with two image-text pairs that differ by minimal but critical semantic attributes. The construction process integrates and refines several complementary public datasets. We leverage *RLHF-V* (Yu et al., 2024a) for its fine-grained textual feedback and *MVC* (Wu et al., 2025) for its minimally contrastive image pairs as the foundation of our dataset. To improve data quality and semantic diversity, we further incorporate *HumanEdit* (Bai et al., 2024) for its high-fidelity, human-guided visual edits, and *Winoground* (Thrush et al., 2022) for its challenging compositional reasoning examples. To ensure that all sources adhere to the required format, we developed a processing pipeline to generate missing textual annotations where necessary. The quality of this automated process was further validated by manual spot-checks. A detailed description of each data source and our processing pipeline is provided in Appendix B.

4.1.2 STAGE 2: CURRICULUM PARTITIONING

To enable a progressive learning trajectory, we partition the SCPP dataset into *Easy*, *Medium*, and *Hard* subsets. This requires a principled method to quantify the difficulty of each preference pair. As illustrated in Figure 2(a) and detailed in **Appendix C**, we propose a novel difficulty score derived from three complementary dimensions:

1. Semantic Proximity (s_{CLIP}). We measure high-level conceptual similarity using the cosine similarity between CLIP global embeddings (Radford et al., 2021): $s_{\text{CLIP}} = \langle \phi(I_w), \phi(I_l) \rangle$. High proximity indicates strong semantic distractors.

2. Structural Discrepancy (d_{OT}). To capture fine-grained visual differences often missed by CLIP, we utilize DINOv2 (Oquab et al., 2024), a self-supervised vision-only model. We quantify the structural distance between patch features $\psi(\cdot)$ using Optimal Transport (OT): $d_{\text{OT}} = \min_T \sum_{i,j} T_{ij} C_{ij},$

where C_{ij} is the cost matrix. This ensures that "hard" samples contain objectively verifiable visual differences (Tong et al., 2024).

3. MLLM Uncertainty ($\overline{\mathcal{H}}$). We quantify the model's subjective confusion via the average binary semantic entropy of its predictions.

Difficulty Aggregation. Since these metrics operate on different scales, we apply **Z-score normalization** to each component before summation. This places them on a comparable scale, avoiding hyperparameter tuning. The final score prioritizes pairs that are simultaneously ambiguous (high $\overline{\mathcal{H}}$), semantically confusing (high s_{CLIP}), and structurally distinct (high d_{OT}). The combination of high s_{CLIP} and d_{OT} is particularly challenging, as it forces the model to rely on genuine visual evidence (grounding) to resolve the conflict between high-level semantics and low-level details. The final score is computed as:

$$\text{Difficulty} = z_{\overline{\mathcal{H}}} + z_{s_{\mathrm{CLIP}}} + z_{d_{\mathrm{OT}}}. \tag{6}$$

Dataset Partitioning. Based on the computed difficulty scores, we partition the SCPP dataset into three curriculum stages. The *Easy* stage is composed of *5k* pairs primarily derived from the RLHF-V dataset. These samples, whose negative examples often involve simple visual perturbations, consistently yield lower difficulty scores. We retain them as the foundational stage to build the model's basic factual grounding, leveraging their high-quality, fine-grained textual feedback. The *Medium* stage is set to *9k* pairs, a size determined empirically to yield the best balance of performance and data diversity, as detailed in our analysis in Appendix C.3. The remaining *7k* pairs, which typically feature the most challenging semantic and structural conflicts, constitute the *Hard* stage, designed to refine the model's capabilities on nuanced distinctions. Some qualitative examples of these challenging "Hard" samples can be found in Section D, where we present specific instances to better illustrate the visual and semantic complexities involved.

4.2 SEMANTIC CURRICULUM PREFERENCE OPTIMIZATION OBJECTIVE

Our **SCPO Objective** is a novel loss function specifically tailored to mitigate visual hallucination by enforcing robust vision-language alignment. It is composed of two complementary components, which we detail below: *Cross-modal Complementary Optimization (CCO)* and *Cross-modal Symmetry Optimization (CSO)*.

Cross-modal Complementary Optimization (CCO). The CCO component learns preferences in both text and image modalities through two distinct objectives.

First, the **textual preference objective** trains the model to distinguish between correct and incorrect responses. Given a fixed image I_w, the model learns to prefer the chosen response y_w over the rejected response y_l:

$$\mathcal{L}_{\text{text}} = -\log \sigma \left(\beta \left[\log \frac{\pi_\theta(y_w \mid I_w, Q)}{\pi_{\text{ref}}(y_w \mid I_w, Q)} - \log \frac{\pi_\theta(y_l \mid I_w, Q)}{\pi_{\text{ref}}(y_l \mid I_w, Q)} \right] \right). \tag{7}$$

Second, the **visual preference objective** trains the model to identify the correct visual context. For a given correct response y_w, the model learns to prefer the matching image I_w over the mismatched image I_l:

$$\mathcal{L}_{\text{image}} = -\log \sigma \left(\beta \left[\log \frac{\pi_\theta(y_w \mid I_w, Q)}{\pi_{\text{ref}}(y_w \mid I_w, Q)} - \log \frac{\pi_\theta(y_w \mid I_l, Q)}{\pi_{\text{ref}}(y_w \mid I_l, Q)} \right] \right). \tag{8}$$

The total CCO loss is the sum of the objectives related to both textual and visual preferences:

$$\mathcal{L}_{\text{CCO}}\left(I_w, I_l, y_w, y_l, Q\right) = \mathcal{L}_{\text{text}}\left(y_w, y_l; I_w, Q\right) + \mathcal{L}_{\text{image}}\left(I_w, I_l; y_w, Q\right). \tag{9}$$

This design embodies the complementary nature of the two objectives. The model learns a *textual preference* for the correct response conditioned on the image, and simultaneously, a *visual preference* for the correct image conditioned on the response. This complementary preference optimization fosters a more robust cross-modal alignment.

Cross-modal Symmetry Optimization (CSO). While CCO learns direct preferences, it does not explicitly prevent the model from ignoring the visual input. To enforce a stronger visual grounding and mitigate shortcut learning, we introduce the Cross-modal Symmetry Optimization (CSO)

component, inspired by (Wu et al., 2025). CSO operates on complete image-text pairs, rewarding consistent pairs while penalizing contradictory ones. This is achieved through two core objectives.

First, the **matching objective** encourages the model to prefer generating a response y when conditioned on its matching image I, compared to generating it with no visual input (that is from language priors alone):

$$\mathcal{L}_{\text{match}}(I, y; Q) = -\log \sigma \left(\beta_1 \left[\log \frac{\pi_\theta(y \mid I, Q)}{\pi_{\text{ref}}(y \mid I, Q)} - \log \frac{\pi_\theta(y \mid Q)}{\pi_{\text{ref}}(y \mid Q)} \right] \right). \tag{10}$$

Second, the **contradiction objective** penalizes the model for associating a response y with a mismatched image I':

$$\mathcal{L}_{\text{contradict}}(I', y; Q) = -\log \sigma \left(\beta_2 \left[\log \frac{\pi_\theta(y \mid Q)}{\pi_{\text{ref}}(y \mid Q)} - \log \frac{\pi_\theta(y \mid I', Q)}{\pi_{\text{ref}}(y \mid I', Q)} \right] \right). \tag{11}$$

where β_1 and β_2 are scaling coefficients. The key insight of CSO is to apply these objectives symmetrically. The model is trained not only to match (I_w, y_w) and reject the mismatched pair (I_l, y_w), but also to match the pair (I_l, y_l) and reject (I_w, y_l). This symmetric application prevents the model from learning simple biases. The final CSO loss combines these four terms:

$$\mathcal{L}_{\text{CSO}}\left(I_w, I_l, y_w, y_l, Q\right) = \underbrace{\left(\mathcal{L}_{\text{match}}(I_w, y_w; Q) + \mathcal{L}_{\text{contradict}}(I_l, y_w; Q)\right)}_{\text{Alignment on } (I_w, y_w)}$$
$$+ \underbrace{\left(\mathcal{L}_{\text{match}}(I_l, y_l; Q) + \mathcal{L}_{\text{contradict}}(I_w, y_l; Q)\right)}_{\text{Alignment on } (I_l, y_l)} \tag{12}$$

Total SCPO Objective. The final SCPO objective is a weighted sum of the CCO and CSO components, creating a unified loss function for robust vision-language alignment:

$$\mathcal{L}_{\text{SCPO}} = \mathcal{L}_{\text{CCO}} + \lambda \, \mathcal{L}_{\text{CSO}}. \tag{13}$$

The hyperparameter λ balances the influence of the CSO component.

4.3 ITERATIVE ALIGNMENT WITH A DYNAMIC REFERENCE MODEL

A static reference model in DPO can struggle with curriculum learning, as difficult, off-policy samples from later stages may cause optimization instability due to large KL penalties or vanishing gradients (Appendix F). To address this, we introduce an iterative alignment strategy with a dynamic reference model, as visualized in Figure 3. The training process consists of three sequential stages on the $\mathcal{D}_{\text{easy}}$, $\mathcal{D}_{\text{medium}}$, and $\mathcal{D}_{\text{hard}}$ subsets. After completing stage t to produce an updated policy π_t, we dynamically reset the reference model for the subsequent stage $t + 1$ by setting $\pi_{\text{ref}}^{(t+1)} \leftarrow \pi_t$.

Figure 3: Visualization of our iterative alignment. *(left)* In a single step, the policy update is constrained by the KL divergence from the reference π_{ref}. *(right)* Our dynamic approach updates this reference at each stage of the easy-to-hard curriculum, enabling a stable and cumulative learning trajectory.

This staged dynamic approach serves two critical functions. First, it mitigates the distribution shift problem common in DPO-based methods. By resetting the reference model, the KL divergence penalty is effectively reset to zero at the start of each new stage:

$$\mathbb{E}_{x \sim \mathcal{D}_{t+1}} \left[D_{\text{KL}}\big(\pi_t(\cdot \mid x) \,\|\, \pi_{\text{ref}}^{(t+1)}(\cdot \mid x)\big) \right] = \mathbb{E}_{x \sim \mathcal{D}_{t+1}} \left[D_{\text{KL}}(\pi_t \| \pi_t) \right] = 0. \tag{14}$$

This prevents the policy from being overly constrained by a distant initial reference model, enabling more effective optimization. Second, this approach decomposes a single complex optimization problem into a series of more manageable steps. Theoretically, the final policy π_N after N stages approximates the optimal policy for the cumulative reward: $\pi_N \propto \pi_0 \exp(\frac{1}{\beta} \sum_{t=1}^N r_t)$ (derivation in

Appendix F.2). This ensures a smooth and effective learning trajectory, as the policy is always regularized against a closely related and progressively improving reference model. The complete staged alignment process is summarized as follows:

$$\pi_{\text{Base}} \xrightarrow[\pi_{\text{ref}}=\pi_{\text{Base}}]{\mathcal{D}_{\text{easy}},\mathcal{L}_{\text{SCPO}}} \pi_{\text{Easy}} \xrightarrow[\pi_{\text{ref}}=\pi_{\text{Easy}}]{\mathcal{D}_{\text{medium}},\mathcal{L}_{\text{SCPO}}} \pi_{\text{Medium}} \xrightarrow[\pi_{\text{ref}}=\pi_{\text{Medium}}]{\mathcal{D}_{\text{hard}},\mathcal{L}_{\text{SCPO}}} \pi_{\text{Hard}} \tag{15}$$

5 EXPERIMENT AND RESULTS

5.1 EXPERIMENTAL SETUP

Implementation Details. We implement and evaluate our SCPO framework on three LLaVA base models: *LLaVA-v1.5-7B*, *LLaVA-v1.6-7B*, and *LLaVA-v1.5-13B* (Liu et al., 2023; 2024a). All models utilize a CLIP ViT-L/14 vision encoder and a Vicuna LLM backbone. The alignment is performed on our *SCPP* dataset, which is partitioned into three curriculum stages: *Easy* (5k pairs), *Medium* (9k pairs), and *Hard* (7k pairs). We fine-tune all models for one epoch per stage with a batch size of *32* and a learning rate of *5e-7*. For our SCPO objective, we set the DPO coefficient $\beta = 0.1$ and the CSO coefficients $\beta_1 = \beta_2 = 0.1$. We set the CSO loss weight λ to *0.2* to balance the magnitudes of the CCO and CSO loss terms, and confirm its robustness through preliminary experiments (Appendix E.2). All LLaVA experiments were conducted on 4 NVIDIA H20 GPUs. A comprehensive breakdown of compute resources and hyperparameters settings is provided in Appendix E.

Evaluation Benchmarks. To provide a comprehensive evaluation, we assess the model performance on a suite of specialized hallucination benchmarks and general capability benchmarks..

HALLUCINATION BENCHMARKS. We employ four benchmarks to measure hallucination mitigation. **1) Object HalBench** (Rohrbach et al., 2018): A standard benchmark for assessing object hallucination. We evaluate across 300 instances, reporting hallucination rates at both the *sentence* (CHAIRs↓) and *object* levels (CHAIRi↓). **2) AMBER-Generative** (Wang et al., 2023), a judgment-model-free benchmark for open-ended image description with *1,004* samples, assesses factuality and coverage. We report five metrics: *object hallucination* (CHAIR↓), *object coverage* (Cover↑), *response-level hallucination* (HalRate↓), *cognitive hallucination* (Cog↓), and a composite *F1 score* (F1-Gen.↑).[1] For **3) AMBER-Discriminative** (Wang et al., 2023), which evaluates binary verification queries across *14,216* instances, we report overall *Accuracy* (Acc.↑) and *F1 score* (F1-Dis.↑). **4) MMHal Bench** (Sun et al., 2023): A question-answering benchmark with 96 images across 12 object categories. Following the official protocol, we use GPT-4 for evaluation and report the overall quality *Score* (Score↑) and *Hallucination Rate* (HalRate↓).

GENERAL CAPABILITY BENCHMARKS. To verify that our focus on factuality does not compromise general abilities, we evaluate our method on four standard benchmarks. **1) LLaVA-in-the-Wild** (Liu et al., 2024b): A benchmark with 60 open-ended questions to assess real-world conversational abilities. We report the relative quality score compared to GPT-4. **2) MMStar** (Chen et al., 2024): An elite vision-indispensable benchmark that filters out samples solvable by LLM priors alone, evaluating true visual grounding. **3) MMBench-CN** and **4) MMBench-ENG** (Liu et al., 2024c): Comprehensive multiple-choice benchmarks evaluating perception and reasoning in Chinese and English, respectively. We report the overall *accuracy* (Acc.↑).

Baselines. We compare SCPO against a wide range of state-of-the-art models to demonstrate its effectiveness. Our comparison includes powerful general-purpose MLLMs such as *GPT-4V* (OpenAI, 2023), *Qwen-VL-Chat* (Bai et al., 2023), and *MiniGemini* (Li et al., 2024b). We also benchmark against a suite of leading alignment methods based on the DPO framework, all specifically designed to mitigate hallucinations. This includes *Silkie* (Li et al., 2024a), standard *DPO* (Rafailov et al., 2023) and its multimodal variants like *mDPO* (Wang et al., 2024), as well as other advanced techniques such as *EOS* (Yue et al., 2024), *HA-DPO* (Zhao et al., 2023), *HSA-DPO* (Xiao et al., 2025), *RLHF-V (HD)* (Yu et al., 2024a), *HALVA* (Sarkar et al., 2025) and *symMPO* (Liu et al., 2025).

---

[1]The *F1-Gen.* score is not part of the original benchmark. We compute it as the harmonic mean of precision (1 - CHAIR) and recall (Cover) to provide a balanced metric for generative quality (Appendix H).

5.2 Main Results

Comparison with Leading MLLMs. We benchmark our SCPO-7B model against several powerful larger MLLMs in Table 1. The results reveal a clear trade-off between generative breadth and factual accuracy. Although leading models like GPT-4V excel in metrics measuring descriptive coverage, our significantly smaller model demonstrates superior performance in hallucination mitigation. Notably, SCPO-7B reduces the *CHAIRs* score by a relative *48.5%* compared to GPT-4V and achieves the highest accuracy on the *AMBER-Discriminative* benchmark. This pattern suggests that our alignment method effectively enhances factual consistency, enabling a smaller open-source model to achieve a state-of-the-art level of reliability.

Table 1: Performance comparison on hallucination benchmarks. Best results are in **bold**, second best are underlined. Parenthesized numbers indicate relative change vs. GPT-4V. Results are cited from: *(Jiang et al., 2024), ‡Xiao et al. (2025).

Model	Object Hal		AMBER-Generative					AMBER-Discriminative		MMHal Bench	
	CHAIRs↓	CHAIRi↓	CHAIR↓	Cover↑	HalRate↓	Cog↓	F1-Gen.↑	Acc.↑	F1-Dis.↑	Score↑	HalRate↓
GPT-4V *	13.6	7.3	4.6	67.1	30.7	2.6	**78.8**	83.4	87.4	**3.49**	**0.28**
Qwen-VL-Chat-10B ‡	36.0(↑22.4)	21.3(↑14.0)	6.6(↑2.0)	53.2(↓13.9)	31.0(↑0.3)	2.9(↑0.3)	-	81.9(↓1.5)	86.4(↓1.0)	2.89(↓0.60)	0.43(↑0.15)
Silkie-10B ‡	25.3(↑11.7)	13.9(↑6.6)	5.4(↑0.8)	55.8(↓11.3)	29.0(↓1.7)	2.0(↓0.6)	-	82.2(↓1.2)	87.6(↑0.2)	3.01(↓0.48)	0.41(↑0.13)
InstructBLIP-13B ‡	25.9(↑12.3)	14.3(↑7.0)	8.8(↑4.2)	52.2(↓14.9)	38.2(↑7.5)	4.4(↑1.8)	-	-	-	2.14(↓1.35)	0.58(↑0.30)
MiniGemini-34B *	14.5(↑0.9)	8.0(↑0.7)	-	-	-	-	-	82.6(↓0.8)	87.6(↑0.2)	3.08(↓0.41)	0.38(↑0.10)
LLaVA-v1.6-7B §	12.0(↓1.6)	6.8(↓0.5)	8.7(↑4.1)	61.1(↓6.0)	49.7(↑19.0)	4.2(↑1.6)	73.2(↓5.6)	82.8(↓0.6)	87.0(↓0.4)	2.80(↓0.69)	0.43(↑0.15)
+SCPO-7B (ours)	**7.0**(↓6.6)	**4.4**(↓2.9)	**4.5**(↓0.1)	60.2(↓6.9)	**27.1**(↓3.6)	**1.9**(↓0.7)	73.8(↓5.0)	**85.4**(↑2.0)	**89.2**(↑1.8)	3.16(↓0.33)	0.32(↑0.04)

Comparison with Alignment Baselines. We compare the SCPO framework with several alignment baselines across various LLaVA architectures, as shown in Table 2. The results highlight that SCPO consistently outperforms prior methods of all model sizes and nearly all hallucination-related metrics. To ensure a fair comparison, we include *DPO (SCPP)*, which applies the standard DPO objective to our data and training methods. While *DPO (SCPP)* yields competitive results, SCPO provides a substantial further improvement, reducing the *CHAIRs score* by 40.3% and the *AMBER HalRate* by 36.3% compared to *DPO (SCPP)*. On Object HalBench, SCPO reduces CHAIRs from 54.7 to 20.3 on LLaVA-v1.5-7B, yielding a 62.9% relative reduction. This confirms that the combination of our data, progressive curriculum, and objectives is essential for achieving state-of-the-art hallucination mitigation.

Table 2: Performance of SCPO against alignment baselines on various LLaVA models. Our method's performance is reported after each progressive curriculum stage. DPO (SCPP) denotes the standard DPO trained with our data and method. Baselines marked with § are our re-evaluations; others are cited from their respective papers: *(Wang et al., 2024), ‡(Sarkar et al., 2024), †(Fu et al., 2025), ♯(Yang et al., 2025). Best and second-best results are in **bold** and underlined. Parentheses show absolute change from the corresponding baseline.

Algorithm	Object Hal		AMBER-Generative					AMBER-Discriminative		MMHal Bench	
	CHAIRs↓	CHAIRi↓	CHAIR↓	Cover↑	HalRate↓	Cog↓	F1-Gen.↑	Acc.↑	F1-Dis.↑	Score↑	HalRate↓
LLaVA-v1.5-7B §	54.7	26.5	8.8	50.1	40.4	4.7	64.7	80.8	85.2	2.18	0.59
+DPO *	49.0	13.0	6.5	55.5	34.5	2.3	69.7	-	-	2.14	0.65
+mDPO *	35.7	**9.8**	4.4	52.4	24.5	2.4	67.7	-	-	2.39	0.54
+EOS ‡	40.2	12.3	5.1	49.1	22.7	2.0	64.7	-	75.6	2.03	0.59
+HA-DPO ‡	38.2	11.0	6.7	49.8	30.9	3.3	64.9	-	78.1	1.97	0.59
+HALVA §	52.7	26.1	8.3	51.0	37.1	4.2	65.5	81.1	85.5	2.19	0.58
+SymMPO §	20.4	10.3	5.2	49.5	25.4	3.0	65	81.4	86.9	2.63	0.51
+DPO (SCPP)	34.0	18.7	4.9	**58.2**	30.6	1.8	**72.2**	81.5	86.8	2.37	0.52
+Easy(ours)	27.7(↓27.0)	15.0(↓11.5)	4.0(↓4.8)	50.7(↑0.6)	21.7(↓18.7)	2.0(↓2.7)	66.4(↑1.7)	79.4(↓1.4)	85.6(↑0.4)	2.37(↑0.19)	0.54(↓0.05)
+Medium(ours)	25.3(↓29.4)	13.3(↓13.2)	**3.5**(↓5.3)	52.8(↑2.7)	21.3(↓19.1)	1.7(↓3.0)	68.3(↑3.6)	82.9(↑2.1)	87.3(↑2.1)	**2.73**(↑0.55)	0.48(↓0.11)
+Hard(ours)	**20.3**(↓34.4)	11.6(↓14.9)	**3.3**(↓5.5)	53.6(↑3.5)	**19.5**(↓20.9)	**1.4**(↓3.3)	69.0(↑4.3)	83.2(↑2.4)	87.3(↑2.1)	2.47(↑0.29)	**0.44**(↓0.15)
LLaVA-v1.6-7B §	12.0	6.8	8.7	61.1	49.7	4.2	73.2	82.8	87.0	2.80	0.43
+DPO †	11.0	6.6	5.9	61.0	38.9	3.0	74.0	-	87.4	2.70	0.44
+DPO (SCPP)	10.7	5.9	6.7	**62.5**	37.4	2.9	**74.9**	85.0	88.3	2.55	0.53
+Easy(ours)	8.3(↓3.7)	5.3(↓1.5)	4.4(↓4.3)	59.3(↓1.8)	28.8(↓20.9)	2.1(↓2.1)	73.2	82.3(↓0.5)	87.3(↑0.3)	2.64(↓0.16)	0.46(↑0.03)
+Medium(ours)	8.6(↓3.4)	5.3(↓1.5)	4.8(↓3.9)	60.3(↓0.8)	31.2(↓18.5)	2.2(↓2.0)	73.8(↑0.6)	85.1(↑2.3)	88.9(↑1.9)	2.94(↑0.14)	0.38(↓0.05)
+Hard(ours)	**7.0**(↓5.0)	**4.4**(↓2.4)	4.5(↓4.2)	60.2(↓0.9)	27.1(↓22.6)	1.9(↓2.3)	73.8(↑0.6)	85.4(↑2.6)	89.2(↑2.2)	3.16(↑0.36)	0.32(↓0.11)
LLaVA-v1.5-13B §	49.3	23.9	8.8	50.3	37.2	4.3	64.8	83.5	86.6	2.31	0.55
+HSA-DPO ♯	-	-	**2.1**	47.3	**13.4**	**1.2**	63.8	-	-	2.61	0.48
+RLHF-V (HD) ♯	-	-	6.3	46.1	25.1	2.1	61.8	-	-	2.81	0.49
+HALVA §	47.0	22.9	8.4	50.5	35.4	4.0	65.1	85.2	88.3	2.44	0.52
+SymMPO §	**18.3**	**10**	4.8	52.8	25.1	2.1	67.9	85	89.1	**2.85**	0.48
+DPO (SCPP)	34.0	17.5	4.9	**55.7**	27.9	1.8	**70.3**	86.2	82.8	2.42	0.55
+Easy(ours)	28.3(↓21.0)	14.4(↓9.5)	4.0(↓4.8)	51.3(↑1.0)	21.1(↓16.1)	1.9(↓2.4)	66.9(↑2.1)	85.6(↑2.1)	89.2(↑2.6)	2.52(↑0.21)	0.49(↓0.06)
+Medium(ours)	28.0(↓21.3)	13.6(↓10.3)	3.6(↓5.2)	53.4(↑3.1)	20.6(↓16.6)	1.2(↓3.1)	68.7(↑3.9)	87.3(↑3.8)	90.4(↑3.8)	2.84(↑0.53)	**0.39**(↓0.16)
+Hard(ours)	23.3(↓26.0)	11.5(↓12.4)	3.8(↓5.0)	53.1(↑2.8)	20.8(↓16.4)	1.3(↓3.0)	68.4(↑3.6)	87.2(↑3.7)	90.4(↑3.8)	2.63(↑0.32)	0.46(↓0.09)

Furthermore, our progressive curriculum design proves effective across all models, with performance steadily improving as training progresses through the Easy, Medium, and Hard stages. This

confirms that the easy-to-hard learning trajectory offers a clear advantage over static data mixtures in enhancing overall performance.

Evaluation on General Capabilities. A critical consideration for alignment is ensuring that gains in factuality do not come at the cost of general-purpose abilities. To verify this, we expanded our evaluation to four diverse benchmarks: *LLaVA-in-the-Wild, MMStar, MMBench-CN,* and *MMBench-EN*. As reported in Table 3, SCPO consistently maintains or improves performance in the final training stage. Notably, for LLaVA-v1.5-7B, the model achieves a *LLaVA-Wild* score of **61.5** after the final curriculum stage, marking a significant improvement over the baseline score of 55.7. Similarly, it im-

Table 3: Performance on general capability benchmarks. Best and second-best results are in **bold** and underlined. Parentheses show absolute change from the corresponding baseline.

Algorithm	LLaVA-wild	MMstar	MMBench-CN	MMbench-ENG
LLaVA-v1.5-7B [§]	55.7	0.327	58.2	65.1
+Easy(ours)	60.0(↑4.3)	0.327(-)	58.9(↑0.7)	65.5(↑0.4)
+Medium(ours)	**64.7**(↑9.0)	0.333(↑6e-3)	**60.0**(↑1.8)	**66.7**(↑1.6)
+Hard(ours)	61.5(↑5.8)	**0.334**(↑7e-3)	59.8(↑1.6)	**66.7**(↑1.6)
LLaVA-v1.6-7B [§]	71.4	0.352	**60.6**	68.0
+Easy(ours)	67.5(↓3.9)	0.348(↓4e-3)	59.5(↓1.2)	67.2(↓0.8)
+Medium(ours)	70.8(↓0.6)	0.353(↑1e-3)	59.5(↓1.1)	68.2(↑0.2)
+Hard(ours)	**72.6**(↑1.2)	**0.356**(↑4e-3)	59.4(↓1.2)	**68.3**(↑0.3)
LLaVA-v1.5-13B [§]	64.9	0.333	63.5	68.9
+Easy(ours)	66.1(↑1.2)	0.336(↑3e-3)	64.3(↑0.8)	69.6(↑0.7)
+Medium(ours)	**68.6**(↑3.7)	**0.349**(↑1.6e-2)	**65.1**(↑1.6)	**71.0**(↑2.1)
+Hard(ours)	67.4(↑2.5)	0.347(↑1.4e-2)	64.9(↑1.4)	70.4(↑1.5)

proves the *MMBench-CN* accuracy from 58.2 to **59.8**. On the 13B model, our method achieves consistent gains on the vision-indispensable *MMStar* benchmark, improving from 0.333 to **0.347**. This confirms that SCPO enhances genuine visual grounding rather than merely overfitting to hallucination patterns. Although minor fluctuations exist across different models, the overall trend demonstrates that SCPO effectively mitigates hallucinations while preserving, and often enhancing, the general multimodal capabilities of the model.

5.3 ABLATION STUDIES

Effectiveness of Objective Components. Our ablation of the SCPO objective in Table 4 demonstrates a clear synergistic effect between its components. While both the CCO and CSO objectives individually outperform the text-only baseline, their combination in the full SCPO objective is crucial for optimal performance. This unified approach achieves the best results across all metrics, reducing the *CHAIRs* score by *34.4%* and the *MMHal Bench HalRate* by *39.2%* relative to the text-only baseline. This confirms that the direct preference signals from CCO and the robust grounding enforced by CSO are complementary and essential for effective hallucination mitigation.

Table 4: Ablation on objective components. We evaluate the contribution of textual preference (L_{text}), image preference (L_{image}), and symmetric optimization (L_{CSO}). Best results are in **bold**.

Objective Components			Object Hallucination		MMHal Bench	
L_{text}	L_{image}	L_{CSO}	CHAIRs↓	CHAIRi↓	Score↑	HalRate↓
✓	✓	✓	**7.00**	**4.42**	**3.16**	**0.32**
✓	✓	✗	7.33	5.17	2.80	0.46
✓	✗	✗	10.67	5.91	2.55	0.53
✗	✗	✓	9.33	4.98	2.82	0.34

Table 5: Ablation on training strategy components. **Sort**: data sorted by difficulty. **Update**: dynamic reference model update. Best results are in **bold**.

Components		Object Hallucination		MMHal Bench	
Sort	Update	CHAIRs↓	CHAIRi↓	Score↑	HalRate↓
✓	✓	**7.00**	**4.42**	3.16	**0.32**
✓	✗	9.67	5.88	**3.22**	0.38
✗	✓	9.00	5.00	3.16	0.40
✗	✗	9.00	5.20	3.04	0.42

Effectiveness of the Curriculum Strategy. As shown in Table 5, our curriculum strategy's effectiveness stems from the synergy between data sorting (*Sort*) and dynamic reference model updates (*Update*). While each component provides moderate gains alone, their combination is critical. For example, adding sorting to dynamic updates further reduces the *CHAIRs score* by *22.2%*. Similarly, adding dynamic updates to sorting lowers the *MMHal Bench HalRate* by *13.9%*. This demonstrates that principled data scheduling and iterative model adaptation are mutually reinforcing, and both are essential to fully leverage our curriculum framework.

5.4 EFFECTIVENESS ON DIFFERENT ARCHITECTURES

To demonstrate the robustness and universality of SCPO, we extend our evaluation to **Muffin-13B** (Yu et al., 2023). Unlike LLaVA's CLIP-based architecture, Muffin employs a **BEiT-3** backbone, representing a distinct visual representation paradigm. As shown in Table 6, SCPO achieves significant improvements on Muffin-13B across all benchmarks. Notably, under the *Hard* setting, the *CHAIRs* score on Object HalBench decreases from 20.7 to 10.3, representing a reduction of

approximately *50%*. Similarly, the *HalRate* on AMBER drops from 33.6 to 11.1. These results confirm that our framework effectively generalizes to diverse MLLM architectures beyond LLaVA. A comprehensive breakdown of hyperparameters, compute resources, and configurations for additional architectures is provided in Appendix E.

Table 6: Performance of SCPO on Muffin-13B across multiple benchmarks. Best and second-best results are **bolded** and underlined. Parentheses indicate the absolute change from the baseline.

Algorithm	Object Hal		AMBER-Generative					AMBER-Discriminative		MMHal Bench	
	CHAIRs↓	CHAIRi↓	CHAIR↓	Cover↑	HalRate↓	Cog↓	F1-Gen.↑	Acc.↑	F1-Dis.↑	Score↑	HalRate↓
Muffin-13B	20.7	11.6	8.2	48.3	33.6	3.8	63.3	83.3	87.4	2.18	0.64
+Easy(ours)	18.0(↓2.7)	12.5(↑0.9)	6.5(↓1.7)	**48.4**(↑0.1)	26.2(↓7.4)	2.1(↓1.7)	**63.8**(↑0.5)	83.3(—)	87.6(↑0.2)	2.20(↑0.02)	0.64(—)
+Medium(ours)	16.0(↓4.7)	10.6(↓1.0)	5.1(↓3.1)	43.9(↓4.4)	17.9(↓15.7)	1.4(↓2.4)	60.0(↓3.3)	**83.4**(↑0.1)	**87.7**(↑0.3)	2.23(↑0.05)	0.60(↓0.04)
+Hard(ours)	**10.3**(↓10.4)	**7.6**(↓4)	**3.8**(↓4.4)	41.8(↓6.5)	**11.1**(↓22.5)	**0.8**(↓3.0)	58.3(↓5)	**83.4**(↑0.1)	87.5(↑0.1)	**2.28**(↑0.1)	**0.57**(↓0.07)

6 FURTHER ANALYSIS

Effectiveness of the Curriculum Order. To verify the importance of our progressive curriculum, we evaluate a reversed Hard → Medium → Easy training sequence. The performance trajectory, detailed in Table 7 and visualized in Figure 4, reveals an unstable and suboptimal learning path. Notably, performance on *Object HalBench* degrades as training progresses; the *CHAIRs score* increases from an initial *10.0* after the first stage to *12.3* in the final model. This suggests that exposing the model to complex examples early, without a proper foundation, can impair its ability to handle fine-grained tasks. We attribute this degradation to shortcut learning triggered by capacity mismatch (Gao et al., 2025). Without foundational grounding, exposure to semantically challenging samples encourages the model to minimize the loss by exploiting language priors rather than resolving fine-grained visual differences, resulting in unstable preference updates. This unstable behavior is in sharp contrast to the steady improvements observed with our proposed Easy-to-Hard curriculum (shown in Table 2 and Figure 4). This finding confirms that the order of data difficulty is a critical factor in our framework, and highlights the need to build from simpler to more complex examples for robust learning.

Table 7: Performance trajectory of the reversed curriculum. Best results are in **bold**.

Model Params	Training Data	Object HalBench		MMHal Bench	
		CHAIRs↓	CHAIRi↓	Score↑	HalRate↓
θ_0	—	12.0	6.8	2.80	0.43
θ_1	$\mathcal{D}_{hard}$	**10.0**	**6.57**	2.85	**0.46**
θ_2	$\mathcal{D}_{medium}$	11.3	6.61	**2.93**	**0.46**
θ_3	$\mathcal{D}_{easy}$	12.3	7.51	2.92	0.47

Figure 4: Performance Trajectories of the Proposed and Reversed Curricula. The solid blue line represents our Easy-to-Hard curriculum, while the dashed orange line represents the reversed order. The x-axis denotes the curriculum stages, and the y-axis shows the performance on four key metrics

7 CONCLUSION

In this work, we addressed the critical challenge of visual hallucinations in MLLMs by proposing SCPO, a novel alignment framework. SCPO's main contribution is a systematic approach that integrates three core innovations. First, we constructed the SCPP dataset, which provides semantically challenging preference pairs partitioned into a structured, easy-to-hard curriculum. Second, we designed a symmetric and bidirectional objective to enforce robust visual grounding. Finally, we introduced a curriculum-based training strategy where the model is trained progressively on data of increasing difficulty, coupled with a dynamic reference model that is updated at each stage. This unique combination of a data-driven curriculum and iterative alignment proves highly effective. Extensive experiments show that SCPO significantly reduces hallucinations across multiple benchmarks and model scales, establishing state-of-the-art performance while preserving general capabilities. Future work will explore adaptive online difficulty scheduling to better handle evolving model capabilities, alongside investigating more reliable online difficulty measures that robustly detect hallucinations while maintaining training stability.

ETHICS STATEMENT

We have read and adhered to the ICLR Code of Ethics. Our work does not involve human subjects or personally identifiable information. The datasets used are publicly available and properly cited. We have taken precautions to avoid harmful insights, and our methodology ensures fairness and avoids discrimination or bias. There are no conflicts of interest, and the research has been conducted in compliance with legal and ethical standards.

REPRODUCIBILITY STATEMENT

The **training code** for this work will be publicly available at the following **anonymous repository**: https://anonymous.4open.science/r/SCPO-Mitigating-Visual-Hallucinations-via-Semantic-Curriculum-Preference-Optimization-in-MLLMs-468D.

Due to the double-blind review process, the **datasets and models** will be made publicly available after the conclusion of the review process. We will provide all necessary documentation to support reproducibility, including the data processing pipeline and model descriptions.

REFERENCES

Jinbin Bai, Wei Chow, Ling Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Shuicheng Yan. Humanedit: A high-quality human-rewarded dataset for instruction-based image editing. *arXiv preprint arXiv:2412.04280*, 2024.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Jinlan Fu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, See-Kiong Ng, et al. Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chengqian Gao, Haonan Li, Liu Liu, Zeke Xie, Peilin Zhao, and zhiqiang xu. Principled data selection for alignment: The hidden risks of difficult examples. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=qut63YypaD.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 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*, pp. 14375–14385, 2024.

Songtao Jiang, Yan Zhang, Ruizhe Chen, Tianxiang Hu, Yeying Jin, Qinglin He, Yang Feng, Jian Wu, and Zuozhu Liu. Modality-fair preference optimization for trustworthy mllm alignment. *arXiv preprint arXiv:2410.15334*, 2024.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025. URL https://arxiv.org/abs/2509.04664.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6227–6246, 2024a.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Wenqi Liu, Xuemeng Song, Jiaxi Li, Yinwei Wei, Na Zheng, Jianhua Yin, and Liqiang Nie. Mitigating hallucination through theory-consistent symmetric multimodal preference optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=tIW29IpCwG`.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024c.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan O Arik, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. In *The Thirteenth International Conference on Learning Representations*, 2024.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan O Arik, and Tomas Pfister. Mitigating object hallucination in MLLMs via data-augmented phrase-level alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=yG1fW8igzP`.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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

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

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

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. 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, pp. 25543–25551, 2025.

Ziwei Xu, Sanjay Jain, and Mohan S Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *CoRR*, 2024.

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

Tao Yu, Yifan Zhang, Chaoyou Fu, Junkang Wu, Jinda Lu, Kun Wang, Xingyu Lu, Yunhang Shen, Guibin Zhang, Dingjie Song, et al. Aligning multimodal llm with human preference: A survey. *CoRR*, 2025.

Tianyu Yu, Jinyi Hu, Yuan Yao, Haoye Zhang, Yue Zhao, Chongyi Wang, Shan Wang, Yinxv Pan, Jiao Xue, Dahai Li, et al. Reformulating vision-language foundation models and datasets towards universal multimodal assistants. *arXiv preprint arXiv:2310.00653*, 2023.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 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*, pp. 13807–13816, 2024a.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv e-prints*, pp. arXiv–2405, 2024b.

Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 11766–11781, 2024.

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

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024.

APPENDIX OUTLINE

A ALGORITHM DETAILS

This section provides detailed pseudocode for our proposed SCPO framework. We first present the main training loop in Algorithm 1.

Algorithm 1 Semantic Curriculum Preference Optimization (SCPO)

Require: Base model π_{θ_0}, SCPP dataset $\mathcal{D}$, curriculum stages $S = \{\text{easy}, \text{medium}, \text{hard}\}$
Require: SCPO loss weights $\beta, \beta_1, \beta_2, \lambda$
1: Partition $\mathcal{D}$ into $\{\mathcal{D}_{\text{easy}}, \mathcal{D}_{\text{medium}}, \mathcal{D}_{\text{hard}}\}$ based on difficulty scores (Eq. 6)
2: Initialize policy model $\pi_\theta \leftarrow \pi_{\theta_0}$
3: Initialize reference model $\pi_{\text{ref}} \leftarrow \pi_{\theta_0}$
4: **for** stage in S **do**
5:    $\mathcal{D}_{\text{stage}} \leftarrow \text{get\_data\_for\_stage}(\mathcal{D}, \text{stage})$
6:    **for** each batch $(I_w, I_l, y_w, y_l, Q) \in \mathcal{D}_{\text{stage}}$ **do**
7:       **// Compute CCO Loss** (Eq. 9)
8:       $\mathcal{L}_{\text{text}} \leftarrow -\log \sigma(\beta[\log \frac{\pi_\theta(y_w|I_w,Q)}{\pi_{\text{ref}}(y_w|I_w,Q)} - \log \frac{\pi_\theta(y_l|I_w,Q)}{\pi_{\text{ref}}(y_l|I_w,Q)}])$
9:       $\mathcal{L}_{\text{image}} \leftarrow -\log \sigma(\beta[\log \frac{\pi_\theta(y_w|I_w,Q)}{\pi_{\text{ref}}(y_w|I_w,Q)} - \log \frac{\pi_\theta(y_w|I_l,Q)}{\pi_{\text{ref}}(y_w|I_l,Q)}])$
10:      $\mathcal{L}_{\text{CCO}} \leftarrow \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{image}}$
11:      **// Compute CSO Loss** (Eq. 12)
12:      $\mathcal{L}_{\text{match}}(I, y) \leftarrow -\log \sigma(\beta_1[\log \frac{\pi_\theta(y|I,Q)}{\pi_{\text{ref}}(y|I,Q)} - \log \frac{\pi_\theta(y|Q)}{\pi_{\text{ref}}(y|Q)}])$
13:      $\mathcal{L}_{\text{contradict}}(I', y) \leftarrow -\log \sigma(\beta_2[\log \frac{\pi_\theta(y|Q)}{\pi_{\text{ref}}(y|Q)} - \log \frac{\pi_\theta(y|I',Q)}{\pi_{\text{ref}}(y|I',Q)}])$
14:      $\mathcal{L}_{\text{CSO}} \leftarrow \mathcal{L}_{\text{match}}(I_w, y_w) + \mathcal{L}_{\text{contradict}}(I_l, y_w) + \mathcal{L}_{\text{match}}(I_l, y_l) + \mathcal{L}_{\text{contradict}}(I_w, y_l)$
15:      **// Total SCPO Loss** (Eq. 13)
16:      $\mathcal{L}_{\text{SCPO}} \leftarrow \mathcal{L}_{\text{CCO}} + \lambda \mathcal{L}_{\text{CSO}}$
17:      Update θ using $\nabla_\theta \mathcal{L}_{\text{SCPO}}$
18:   **end for**
19:   $\pi_{\text{ref}} \leftarrow \pi_\theta$ ▷ Dynamic reference model update
20: **end for**
21: **return** Final aligned model π_θ

B SCPP DATASET CONSTRUCTION DETAILS

This section provides a detailed description of the data sources and the processing pipeline used to construct our *Semantic Curriculum Preference Pairs (SCPP)* dataset, as mentioned in Section 4.1.

B.1 DATA SOURCES AND RATIONALE

Our dataset is built by aggregating and refining four complementary public datasets. The selection of these sources is motivated by the need to create a diverse and challenging set of preference pairs that cover both textual and visual semantic contrasts.

Foundational Datasets. We start with two foundational datasets to provide a broad base of preference signals.

- **RLHF-V** (Yu et al., 2024a): This dataset offers large-scale prompt-text-image triples with fine-grained human feedback on textual hallucinations. It serves as our primary source for learning textual preferences, where the model must distinguish between a factually correct and a hallucinated description for a given image.

- **MVC** (Wu et al., 2025): This dataset provides pairs of minimally contrastive images that differ in subtle but critical visual details (e.g., an object's color or count). It serves as our primary source for learning visual preferences, where the model must ground its response on the correct visual input.

Enhancement Datasets. Although the foundational datasets are useful, they have limitations. RLHF-V can contain noisy or low-quality images, and MVC's images are generated by diffusion

models, which may lack the fidelity of real-world photographs. To address these issues and increase the dataset's quality and diversity, we incorporate two additional sources:

- **HumanEdit** (Bai et al., 2024): This is a high-quality corpus of images edited by humans following specific instructions. It provides authentic and high-fidelity visual contrasts that are grounded in genuine human intent, serving as a source of realistic and challenging negative visual samples.

- **Winoground** (Thrush et al., 2022): This benchmark is specifically designed to test visio-linguistic compositional reasoning. Its image-text pairs often differ by a single relational or compositional word (e.g., "the cat is behind the box" vs. "the box is behind the cat"). We include it to enrich SCPP with examples that require a deeper level of semantic understanding.

B.2 LLM-based Text Generation for Preference Pairs

While datasets like *RLHF-V* and *MVC* provide a strong foundation, other high-quality visual sources such as *HumanEdit* and *Winoground* lack pre-existing, preference-formatted textual annotations (i.e., a prompt with corresponding chosen and rejected responses). To take advantage of these valuable visual assets, we developed an LLM-based pipeline to generate high-quality, contextually appropriate text for their image pairs.

We utilized GPT-5 to perform this task. The pipeline was designed with three primary goals:

- **Prompt Generation**: For each image pair, generate a concise and unambiguous question that targets the key semantic difference between them.

- **Response Generation**: Create a factually accurate "chosen" response that correctly describes the first image and a "rejected" response that correctly describes the second, ensuring both are brief and directly answer the generated question.

- **Quality Control**: Ensure the generated text is grammatically correct, stylistically consistent, and avoids ambiguous or subjective language. The process was carefully controlled to ensure all generated text is faithfully grounded in the visual content.

The core of this pipeline is a carefully designed prompt that guides the LLM to act as a data creator and quality assurance expert. The whole prompt set used in this process is provided in Figure 5 for transparency and reproducibility.

LLM Prompt for Quality Assurance and Annotation of Preference Pairs

You are an expert in data quality assurance for preference datasets.
You will be given a series of JSON entries, each containing three fields: "question", "chosen", and "rejected".
Your task is to evaluate each entry based on the following criteria:
1. **Clarity and Focus**: The question should be concise, clear, and unambiguous. It should highlight only one specific semantic difference.
  - **Invalid**: The question is vague, subjective, irrelevant, or covers multiple aspects.
2. **Accuracy of the "chosen" description**: The "chosen" answer should correctly describe the content of **image_0**.
3. **Accuracy of the "rejected" description**: The "rejected" answer should correctly describe the content of **image_1**.
4. **Concise and factual answers**: Both the "chosen" and "rejected" answers should directly respond to the question with brief and accurate factual statements.
5. **Grammar and Style**: Ensure that grammar and style are correct, with consistent tone across the entries.
6. **Avoid certain terms**: Do not mention terms such as "original", "edited", "before/after", or "image A/B".
For each entry, your output should be a JSON array. Each element should include:
- **"index"**: the index of the sample.
- **"is_valid"**: `true` if the entry meets all the criteria, or `false` otherwise.
- **"error_type"**: if invalid, specify one or more of the following categories: ["bad_question", "chosen_wrong", "rejected_wrong", "grammar_issue", "style_issue"].
- **"suggestion"**: if invalid, provide a corrected version of the entry.
Please do not provide any explanations outside of the JSON format.

Figure 5: Data refinement prompt pipeline.

C Difficulty Score Implementation Details

While the core mathematical definitions and motivations for the difficulty score are provided in Section 4.1.2 of the main text, this appendix details the specific algorithmic procedure and model configurations to ensure reproducibility.

C.1 Computation Algorithm

The step-by-step procedure for computing the difficulty score and partitioning the curriculum is outlined in **Algorithm 2**. Note that the standardization (Z-score normalization) is performed globally across the entire SCPP dataset.

Algorithm 2 Difficulty Score Computation for Preference Pairs

---

Require: Preference pairs dataset $\mathcal{D} = \{(I_w^{(i)}, I_l^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^N$
    Step 1: Feature Extraction
1: **for** each pair $(I_w^{(i)}, I_l^{(i)})$ in $\mathcal{D}$ **do**
2:     Compute MLLM mean entropy: $\overline{\mathcal{H}}^{(i)} \leftarrow \frac{1}{2}(H(I_w^{(i)}) + H(I_l^{(i)}))$
3:     Compute DINOv2 optimal transport distance: $d_{\mathrm{OT}}^{(i)}$
4:     Calculate CLIP cosine similarity: $s_{\mathrm{CLIP}}^{(i)} \leftarrow \langle \phi(I_w^{(i)}), \phi(I_l^{(i)}) \rangle$
5: **end for**

    Step 2: Standardization
6: **for** each feature $f \in \{\overline{\mathcal{H}}, d_{\mathrm{OT}}, s_{\mathrm{CLIP}}\}$ **do**
7:     Compute global statistics: μ_f, σ_f
8:     Normalize: $z_f^{(i)} \leftarrow (f^{(i)} - \mu_f)/\sigma_f$
9: **end for**

    Step 3: Aggregation
10: **for** each pair $i = 1, \ldots, N$ **do**
11:     $\text{Difficulty}^{(i)} \leftarrow z_{\overline{\mathcal{H}}}^{(i)} + z_{d_{\mathrm{OT}}}^{(i)} + z_{s_{\mathrm{CLIP}}}^{(i)}$
12: **end for**
13: **return** Difficulty scores $\{\text{Difficulty}^{(i)}\}_{i=1}^N$

C.2 Model Configurations

To compute the semantic and structural metrics, we utilize specific pre-trained vision encoders.

- **Semantic Proximity** (s_{CLIP}): We employ the `clip-vit-large-patch14-336` model (Radford et al., 2021) to extract global image embeddings.
- **Structural Discrepancy** (d_{OT}): We use the `dinov2-vit-large-patch14` model (Oquab et al., 2024) for extracting patch-level features. For the Optimal Transport calculation, we utilize the Sinkhorn algorithm with a regularization parameter of $\epsilon = 0.05$.

C.3 Determining the Size of the Medium Curriculum Stage

To determine the optimal size for the *Medium* curriculum stage, we conducted an empirical study by varying the number of medium-difficulty samples used for training, ranging from *1k* to *10k*. The results, summarized in Table 8, show a clear trend. As the data size increases from *1k* to *9k*, we observe a general improvement on most metrics, particularly in discriminative tasks. The model's performance on *AMBER-Discriminative* peaks at the *9k* data point, achieving the highest *Accuracy* and *F1 score*. While the lowest object hallucination (*CHAIR*) is observed at the *5k* mark, the *9k* model maintains a highly competitive hallucination rate while offering superior discriminative capabilities. Notably, increasing the data size further to *10k* leads to a slight degradation in several key metrics, including *Cover* and *Accuracy*, suggesting that *9k* represents an optimal trade-off between performance, data diversity, and training efficiency. Based on this analysis, we set the size of the *Medium* stage to *9k* pairs for all main experiments.

Table 8: Performance on AMBER with varying data sizes for the Medium curriculum stage. Best performance for each metric is highlighted in **bold**.

Data Size	AMBER-Generative				AMBER-Discriminative	
	CHAIR↓	Cover↑	HalRate↓	Cog↓	Acc.↑	F1-Dis.↑
1k	4.6	58.9	33.6	2.5	82.4	87.4
2k	4.8	59.4	36.4	2.8	82.5	87.4
3k	4.7	60.0	38.1	3.0	83.0	87.7
5k	**4.4**	60.1	33.6	2.5	83.8	88.2
7k	4.9	**60.3**	36.4	2.8	84.4	88.6
9k	4.8	**60.3**	**33.6**	**2.5**	**85.1**	**88.9**
10k	4.6	58.7	38.1	3.0	84.8	88.7

D QUALITATIVE EXAMPLES OF HARD PREFERENCE PAIRS

To provide a concrete understanding of the "Hard" samples in our SCPP dataset, we present examples in Figure 6. These pairs have high semantic similarity (s_{CLIP}), meaning they share similar high-level concepts (e.g., "mountain scenery" or "blooming flowers"), which can confuse models relying on language priors. They also exhibit high structural discrepancy (d_{OT}), with distinct visual differences (e.g., the presence of a balloon or butterfly) that are perceptually clear to humans. These examples serve as effective hard negatives for training the model's visual grounding.

Figure 6: Qualitative examples of "Hard" samples in our SCPP dataset. Each pair consists of two images that share high semantic similarity but contain distinct fine-grained visual differences, such as the presence of an object or attribute variations.

E IMPLEMENTATION DETAILS AND HYPERPARAMETERS

To ensure full reproducibility and transparency, we provide a comprehensive breakdown of the training configurations, compute resources, and hyperparameter sensitivity analysis used in our experiments.

E.1 TRAINING HYPERPARAMETERS AND COMPUTE RESOURCES

We utilize the AdamW optimizer with BF16 precision for all experiments. The detailed hyperparameter settings for LLaVA-family models and the Muffin-13B model are listed in Table 9.

Note on Muffin-13B: To accommodate the architectural differences of Muffin-13B (which uses a BEiT-3 encoder), we adjusted the global batch size to 128, while keeping the SCPO loss weights (including $\lambda = 0.2$) consistent with the LLaVA experiments. This demonstrates the robustness of our hyperparameter selection across different architectures.

Table 9: Detailed hyperparameters and training costs for all models.

Configuration	LLaVA Models (7B/13B)	Muffin-13B
Vision Encoder	CLIP ViT-L/14-336px	BEiT-3 Large
Optimizer	AdamW	AdamW
Precision	BF16	BF16
Learning Rate	5e-7	5e-7
LR Scheduler	Cosine	Cosine
Warmup Ratio	0.03	0.03
Global Batch Size	**32**	**128**
Gradient Accumulation	8 (per device batch 1)	16 (per device batch 1)
Epochs per Stage	1	1
SCPO Parameters		
DPO Beta (β)	0.1	0.1
CSO Beta (β_1, β_2)	0.1	0.1
CSO Weight (λ)	**0.2**	**0.2**
Compute & Time		
GPUs	$4 \times$ NVIDIA H20 (96GB)	$8 \times$ NVIDIA H20 (96GB)
Training Time (Easy)	$\sim$2 hours	$\sim$2 hours
Training Time (Medium)	$\sim$3 hours	$\sim$3 hours
Training Time (Hard)	$\sim$2.5 hours	$\sim$2.5 hours

E.2 SENSITIVITY ANALYSIS OF THE CSO WEIGHT λ

The hyperparameter λ in our SCPO objective balances the contributions of the complementary ($\mathcal{L}_{\text{CCO}}$) and symmetric ($\mathcal{L}_{\text{CSO}}$) loss components. To analyze the model's sensitivity to this parameter, we conducted a sweep for λ on LLaVA-v1.6-7B (Table 10).

The results demonstrate that SCPO is remarkably robust. Although $\lambda = 0.2$ yields a marginally better balance, performance remains highly stable within the range of [0.1, 0.3]. For instance, the *CHAIR* score only fluctuates between 4.5 and 4.6. This low sensitivity indicates that the benefits of our framework are not contingent on meticulous hyperparameter tuning. Based on this, we selected $\lambda = 0.2$ for all main experiments.

Table 10: Sensitivity analysis for the CSO weight λ on the AMBER benchmark (LLaVA-v1.6-7B). Performance remains highly stable. Best performance is highlighted in **bold**.

Setting (λ)	AMBER-Generative			AMBER-Discriminative	
	CHAIR↓	Cover↑	F1-Gen.↑	Acc.↑	F1-Dis.↑
0.1	4.6	59.3	73.14	85.2	89.1
0.2	**4.5**	**60.2**	**73.85**	**85.4**	**89.2**
0.3	4.6	60.0	73.67	**85.4**	**89.2**

F   Theoretical Analysis of Iterative Alignment

This appendix provides a concise motivation for our iterative alignment strategy with a *dynamic reference* (Section 4.3). We first analyze why a *fixed* reference can fail in the late stages of curriculum, then derive the *cumulative-reward* property of our approach.

F.1   Limitations of a Fixed Reference Model in DPO

A standard DPO setup with a fixed reference π_{ref} can struggle when preference pairs in later curriculum stages are deliberately *off-policy* for π_{ref} (the preferred response has tiny probability under the reference).

Infinite KL (in practice, arbitrarily large).   The objective regularized by the KL divergence (cf. Eq. 1) requires a finite divergence between π_θ and π_{ref}. If $\pi_{\text{ref}}(y_w \mid x, I) \approx 0$ yet the optimized policy must assign nonzero mass to y_w, the divergence

$$D_{\text{KL}}\big(\pi_\theta(\cdot \mid x, I) \,\|\, \pi_{\text{ref}}(\cdot \mid x, I)\big) \tag{16}$$

can become arbitrarily large, rendering the optimization effectively intractable in practice.

Vanishing Gradients (Logistic Saturation).   Let the per-pair DPO loss be $\ell(M) = -\log \sigma(\beta M)$ with margin

$$M_\pi(x, I) = \log \frac{\pi(y_w \mid x, I)}{\pi(y_l \mid x, I)} - \log \frac{\pi_{\text{ref}}(y_w \mid x, I)}{\pi_{\text{ref}}(y_l \mid x, I)}. \tag{17}$$

If the winner is severely off-policy under the fixed reference, i.e., $\pi_{\text{ref}}(y_w \mid x, I) \approx 0$ while $\pi_{\text{ref}}(y_l \mid x, I)$ is not, then

$$-\log \frac{\pi_{\text{ref}}(y_w \mid x, I)}{\pi_{\text{ref}}(y_l \mid x, I)} \quad \text{is a large positive term.} \tag{18}$$

However, if the new stage is initialized from the previous policy so that $\pi \approx \pi_{\text{ref}}$, the two log-ratios approximately cancel and

$$M_\pi(x, I) \approx 0 \quad \text{at stage start, even when } \pi_{\text{ref}}(y_w \mid x, I) \text{ is tiny.} \tag{19}$$

As training pushes the current log-ratio above the (tiny) reference log-ratio, $M_\pi(x, I)$ can grow large and positive, causing true saturation:

$$\frac{\partial \ell}{\partial M} = -\beta \, \sigma(-\beta M) \longrightarrow 0^- \quad \text{as } M \to +\infty. \tag{20}$$

For completeness, if instead $\pi_{\text{ref}}(y_l \mid x, I) \approx 0$ while $\pi_{\text{ref}}(y_w \mid x, I)$ is not, then $M_\pi(x, I) \to -\infty$ and

$$\frac{\partial \ell}{\partial M} \longrightarrow -\beta \quad \text{as } M \to -\infty, \tag{21}$$

i.e., the gradient does *not* vanish. In late curriculum stages, the problematic regime is the former (winner off-policy under a fixed reference), where M becomes very large and positive as π departs from π_{ref}. Our dynamic reference update mitigates this by keeping M in a moderate range at all stages.

F.2   Cumulative Reward with a Dynamic Reference

We now show that resetting the reference at each stage, $\pi_{\text{ref}}^{(t+1)} \leftarrow \pi_t$, yields *cumulative-reward* learning. Within stage t, we view the effective reward $r_t(x, I, y)$ as absorbing the weights from our CCO/CSO components. The DPO optimality condition implies the exponential family form:

$$\pi_t(y \mid x, I) \;\propto\; \pi_{\text{ref}}^{(t)}(y \mid x, I) \, \exp\!\Big(\tfrac{1}{\beta}\, r_t(x, I, y)\Big). \tag{22}$$

Let $\pi_0 = \pi_{\text{Base}}$. Then:

Stage 1 (Easy).

$$\pi_1 \;\propto\; \pi_0 \exp\!\Big(\tfrac{1}{\beta}\, r_1\Big). \tag{23}$$

Stage 2 (Medium).

$$\pi_2 \;\propto\; \pi_1 \, \exp\!\left(\tfrac{1}{\beta}\, r_2\right) \;\propto\; \pi_0 \, \exp\!\left(\tfrac{1}{\beta}(r_1 + r_2)\right). \tag{24}$$

Stage 3 (Hard).

$$\pi_3 \;\propto\; \pi_2 \, \exp\!\left(\tfrac{1}{\beta}\, r_3\right) \;\propto\; \pi_0 \, \exp\!\left(\tfrac{1}{\beta}(r_1 + r_2 + r_3)\right). \tag{25}$$

General N stages.   By induction,

$$\pi_N \;\propto\; \pi_0 \, \exp\!\left(\tfrac{1}{\beta} \sum_{t=1}^{N} r_t\right). \tag{26}$$

Thus, a single difficult optimization is decomposed into a sequence of tractable, on-policy stages whose rewards accumulate multiplicatively in the policy space. This also "resets" the KL at each stage boundary:

$$\mathbb{E}_{x,I} \, D_{\mathrm{KL}}\!\big(\pi_t(\cdot \mid x, I) \,\|\, \pi_{\mathrm{ref}}^{(t+1)}(\cdot \mid x, I)\big) = \mathbb{E}_{x,I} \, D_{\mathrm{KL}}(\pi_t \| \pi_t) = 0, \tag{27}$$

preventing cross-stage KL accumulation and helping avoid the saturation regime in Eq. equation 20.

G QUALITATIVE EXAMPLES OF MODEL RESPONSES

To provide a more intuitive understanding of our SCPO framework's effectiveness, we present several qualitative examples drawn from our evaluation benchmarks. These examples illustrate how the model's responses evolve through the curriculum stages and how the final model (π_{Hard}) successfully mitigates hallucinations present in earlier stages or baseline models. In the following figures, text highlighted in red indicates a hallucination or factual error, while text in green indicates a correct, factually grounded statement.

Figure 7: Example of mitigating descriptive hallucination. The initial model (π_{Easy}) hallucinates fine-grained details about the players' actions and focus. As the curriculum progresses, the model learns to be more faithful to the visual evidence, with the final model (π_{Hard}) correctly identifying that the fence obscures a clear view and avoids making unsubstantiated claims.

Figure 8: Examples of correcting factual errors. **(Top)** The model initially hallucinates the number of bicycles. Through curriculum learning, it corrects this counting error. **(Bottom)** The model initially makes an incorrect inference about the state of the toilet paper. The final model provides the correct, visually grounded answer.

H   METRIC COMPUTATION DETAILS

F1-Gen. Score for AMBER-Generative.   Evaluating hallucination mitigation often involves a trade-off: a model might reduce hallucinations simply by generating shorter, less descriptive responses, which would unfairly penalize its *Cover* score. The original AMBER benchmark provides separate metrics for hallucination (*CHAIR*) and coverage (*Cover*), but lacks a single score to evaluate this balance. To address this, we introduce the *F1-Gen.* score, which provides a holistic measure of generative quality by harmonizing these two competing factors.

We frame the task of object generation in terms of precision and recall.

- **Precision** reflects factuality. It is the fraction of generated objects that are correct, which we compute as $1 - CHAIR$. High precision indicates low hallucination.
- **Recall** reflects comprehensiveness. It is the fraction of ground-truth objects mentioned in the response, which is directly measured by the *Cover* metric.

The *F1-Gen.* score is the harmonic mean of this precision and recall:

$$F1\text{-}Gen. = 2 \times \frac{(1 - CHAIR) \times Cover}{(1 - CHAIR) + Cover}. \tag{28}$$

This composite metric allows for a fairer comparison between models, as it rewards methods that successfully reduce hallucinations (high precision) without sacrificing the descriptive richness of the output (high recall).

I   DETAILED BENCHMARK RESULTS

This section provides a more detailed breakdown of our model's performance on the AMBER and Object HalBench benchmarks, complementing the main results. These tables offer deeper insight into the consistent and thorough improvements brought by the SCPO framework.

I.1   DETAILED PERFORMANCE ON AMBER

Table 11 presents a detailed performance breakdown on the various semantic categories of the AMBER benchmark. The results show that our SCPO curriculum leads to a broad enhancement of fine-grained capabilities. Across nearly all model scales and sub-tasks (such as *Action*, *Attribute*, and *Existence*), the models trained with our curriculum significantly outperform the original LLaVA baselines. The consistent improvement, particularly in the final stage, confirms our framework's ability to foster a more robust and nuanced semantic understanding.

Table 11: Performance of different variants on the AMBER benchmark across model scales. Results are reported for each difficulty stage (Easy, Middle, Hard) and baseline models. Best and second-best results per metric and model group are **bolded** and underlined, respectively.

Model	Discriminative	Action	Attribution	Existence	Number	Relation	State	Comprehensive_Score
LLaVA-v1.5-7B	85.2	83.1	77.7	94.3	83.4	69.1	74.0	74.94
+Easy	85.6	**85.4**	**78.8**	**97.6**	**84.2**	59.9	75.2	75.98
+Middle	**87.3**	84.5	78.4	97.2	82.2	72.3	76.0	77.78
+Hard	**87.3**	84.2	77.5	96.9	77.0	**76.4**	**76.6**	**78.14**
LLaVA-v1.6-vicuna-7B	87.0	**88.2**	81.5	95.1	83.0	69.6	79.7	80.10
+Easy	87.3	86.9	80.9	**97.8**	83.2	64.5	79.0	80.25
+Middle	88.9	86.9	**82.6**	97.1	**84.9**	73.0	**80.9**	81.37
+Hard	**89.2**	86.6	**82.6**	97.0	84.8	75.3	**80.9**	**81.52**
LLaVA-v1.5-13B	86.6	86.5	82.0	91.3	84.9	**79.4**	79.9	75.72
+Easy	89.2	**89.3**	84.7	95.7	89.5	73.8	81.8	78.03
+Middle	**90.4**	88.1	85.2	96.6	90.2	77.9	82.5	**79.56**
+Hard	**90.4**	88.2	**85.4**	**97.0**	**90.9**	76.4	**82.6**	79.41

I.2   DETAILED PERFORMANCE ON OBJECT HALBENCH

Table 12 provides a direct validation of primary goal of our framework to reduce hallucinations. The results show a clear and steady trend in all models tested. As training progresses through the easy-to-hard curriculum, we observe a consistent and significant decrease in hallucination rates (both

Response Hall. and *Object Hall.*), coupled with a corresponding increase in correctness scores. This provides clear evidence for the effectiveness of our progressive learning strategy in systematically improving the model factuality.

Table 12: Detailed hallucination and correctness metrics on Object HalBench. Best and second-best results per metric and model group are in **bold** and underlined.

Model	Response Hall. (↓)	Response Correct (↑)	Object Hall. (↓)	Object Correct (↑)
LLaVA-v1.5-7B	55.97	73.58	26.49	73.51
+Easy	29.12	70.88	14.98	85.02
+Middle	26.48	73.52	13.33	86.67
+Hard	**21.40**	**78.60**	**11.65**	**88.35**
LLaVA-v1.6-vicuna-7B	12.95	87.05	6.79	93.21
+Easy	9.03	90.97	5.31	94.69
+Middle	9.25	90.75	5.28	94.72
+Hard	**7.61**	**92.39**	**4.42**	**95.58**
LLaVA-v1.5-13B	50.86	49.14	23.86	76.14
+Easy	29.41	70.59	14.44	85.56
+Middle	28.97	71.03	13.58	86.42
+Hard	**24.56**	**75.44**	**11.51**	**88.49**

I.3 ADDITIONAL EVALUATION ON HALLUSIONBENCH

To further demonstrate the comprehensiveness and feasibility of our SCPO framework, we conducted an additional evaluation on HallusionBench (Guan et al., 2024). As presented in Table 13, SCPO achieves consistent improvements across all metrics compared to the baselines, further validating the effectiveness of our approach in mitigating visual illusions and knowledge hallucinations.

Table 13: Evaluation results on HallusionBench metrics across different curriculum stages. Best and second-best results are **bolded** and underlined.

Model	qAcc↑	fAcc↑	aAcc↑
LLaVA-v1.5-7B [§]	11.2	18.5	46.2
+Easy(ours)	**16.5**($\uparrow$5.3)	22.3($\uparrow$3.8)	**52.1**($\uparrow$5.9)
+Medium(ours)	16.3($\uparrow$5.1)	**22.5**($\uparrow$4.0)	50.8($\uparrow$4.6)
+Hard(ours)	15.8($\uparrow$4.6)	21.1($\uparrow$2.6)	50.7($\uparrow$4.5)
LLaVA-v1.6-7B [§]	14.7	19.4	48.4
+Easy(ours)	17.8($\uparrow$3.1)	25.1($\uparrow$5.7)	51.5($\uparrow$3.1)
+Medium(ours)	**18.7**($\uparrow$4.0)	**25.7**($\uparrow$6.3)	**52.7**($\uparrow$4.3)
+Hard(ours)	16.3($\uparrow$1.6)	25.1($\uparrow$5.7)	52.6($\uparrow$4.2)
LLaVA-v1.5-13B [§]	10.8	15.9	46.9
+Easy(ours)	13.8($\uparrow$3.0)	19.9($\uparrow$4.0)	49.4($\uparrow$2.5)
+Medium(ours)	14.5($\uparrow$3.7)	20.5($\uparrow$4.6)	49.5($\uparrow$2.6)
+Hard(ours)	**16.3**($\uparrow$5.5)	**22.8**($\uparrow$6.9)	**50.4**($\uparrow$3.5)

J LLM USAGE

In this work, we used Large Language Models (LLMs) solely for language refinement to improve the clarity and fluency of the text. The LLM played no role in the ideation, analysis, or generation of scientific content.