
Decoupling Contrastive Decoding: Robust Hallucination Mitigation in Multimodal Large Language Models

Wei Chen¹, Xin Yan², Bin Wen³,
Fan Yang³, Tingting Gao³, Di Zhang³, Long Chen^{1*}
¹HKUST, ²University of Waterloo, ³Kuaishou Technology
wchendb@connect.ust.hk longchen@ust.hk

Abstract

Although multimodal large language models (MLLMs) exhibit remarkable reasoning capabilities on complex multimodal understanding tasks, they still suffer from the notorious “hallucination” issue: generating outputs misaligned with obvious visual or factual evidence. Currently, training-based solutions, like direct preference optimization (DPO), leverage paired preference data to suppress hallucinations. However, they risk sacrificing general reasoning capabilities due to the likelihood displacement. Meanwhile, training-free solutions, like contrastive decoding, achieve this goal by subtracting the estimated hallucination pattern from a distorted input. Yet, these handcrafted perturbations (*e.g.*, add noise to images) may poorly capture authentic hallucination patterns. To avoid these weaknesses of existing methods, and realize “robust” hallucination mitigation (*i.e.*, maintaining general reasoning performance), we propose a novel framework: Decoupling Contrastive Decoding (DCD). Specifically, DCD decouples the learning of positive and negative samples in preference datasets, and trains separate positive and negative image projections within the MLLM. The negative projection implicitly models real hallucination patterns, which enables vision-aware negative images in the contrastive decoding inference stage. Our DCD alleviates likelihood displacement by avoiding pairwise optimization and generalizes robustly without handcrafted degradation. Extensive ablations across hallucination benchmarks and general reasoning tasks demonstrate the effectiveness of DCD, *i.e.*, it matches DPO’s hallucination suppression while preserving general capabilities and outperforms the handcrafted contrastive decoding methods. Code is available in <https://github.com/HKUST-LongGroup/DCD>.

1 Introduction

Today’s multimodal large language models (MLLMs) [1, 2, 3, 4, 5] have demonstrated remarkable general reasoning capabilities by integrating visual and textual understanding, facilitating applications such as medical image analysis [6, 7] and multimodal search engines [8]. Despite their versatility, a critical limitation persists: MLLMs may generate outputs that contradict obvious factual evidence or misrepresent visual inputs, known as the **hallucination problem** [9, 10, 11, 12]. For instance, models may describe objects absent from an image (*e.g.*, claiming a “dog” in a cat-only scene) or fabricate implausible relationships (*e.g.*, asserting “a person riding a bicycle” when only a bicycle is present). Such hallucinations erode users trust and hinder deployment in high-stakes domains like healthcare [6] or autonomous driving [13].

*Corresponding author.

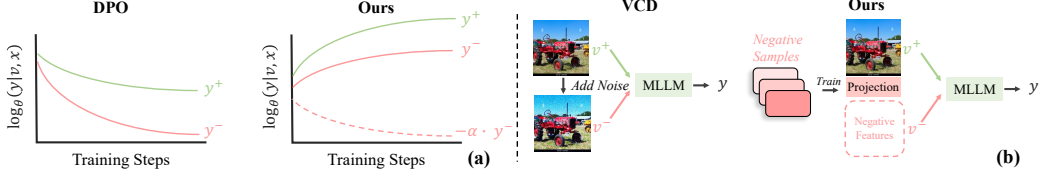


Figure 1: **Comparison between existing hallucination mitigation methods and DCD.** (a) Training-based method (e.g., DPO [14]): DPO directly optimizes the likelihood gap between positive (correct) and negative (hallucinatory) responses using preference datasets. However, maximizing this gap (y^+ vs. y^-) can inadvertently lower the probability of both responses, causing likelihood displacement and potential degradation of general reasoning capabilities. Here, v , x , y^+ , and y^- denote images, questions, positive responses, and negative responses, respectively; θ represents model parameters, and α is the contrastive decoding coefficient. (b) Training-free method (e.g., VCD [15]) vs. DCD: Traditional contrastive decoding (VCD) reduces hallucinations by comparing model outputs from original (v^+) and artificially distorted (v^- ; e.g., noise-added) visual inputs at inference time.

To mitigate this hallucination issue, recent *training-based* approaches [16, 17, 18, 19, 20] draw inspiration from reinforcement learning from human feedback (RLHF) [21], a finetune paradigm that aligns models with human preferences. These RLHF methods typically involve two stages: 1) *Hallucination Preference Dataset Construction*. Recent efforts [16, 17, 18, 19, 20, 22] collect paired positive-negative samples to form the preference dataset, where positive responses are the correct answers and negative responses are the hallucinatory answers. These “high-quality” negative samples are often collected from model-generated hallucinatory outputs, ensuring alignment with the real hallucination observed in MLLMs. 2) *Preference Optimization Training*. Direct preference optimization (DPO) [14] is the most prevalent and well-explored approach to train MLLMs with preference datasets. It bypasses complex reinforcement learning pipelines by directly maximizing the likelihood gap between positive and negative responses. While DPO demonstrates efficacy in hallucination mitigation, this paired-sample optimization process risks inducing a *likelihood displacement* problem [23]: By maximizing the gap between positive and negative answers, DPO may inadvertently lower the probabilities of both responses (as shown in Figure 1(a)). It potentially sacrifices the model’s general reasoning capabilities and leads to performance degradation in open-ended tasks.

In parallel, *training-free* methods [15, 24, 25, 26, 27, 28, 29] resort to contrastive decoding [30] to alleviate hallucination. They hold the assumption that MLLM is easier to have the hallucination issue with distorted inputs. For example, image perturbations disrupt semantic coherence and amplify hallucinatory tendencies. By transferring the log-likelihood differences of model outputs with that of distorted images, contrastive decoding methods force MLLM to focus more on images details (cf. Figure 1(b)). However, existing perturbation strategies are handcrafted and artificial, such as adding noise to images [15]). Therefore, these artificial contrastive distributions may not reflect the authentic hallucinations produced by MLLMs, as they are vision-and-text agnostic and can introduce uncertainty noise in the decoding process [27] which is not robust in complex tasks.

In this paper, we aim to avoid these weaknesses of existing methods, and realize a more robust hallucination mitigation. By “robust”, we hope the method can not only significantly reduce hallucination cases, but also preserve general capabilities on challenging reasoning tasks. To this end, we propose a novel framework: Decoupling Contrastive Decoding (DCD). Specifically, DCD has two designs: 1) *Decoupling Learning*. We decouple pairwise positive-negative samples learning of preference dataset into separate learning to alleviate likelihood displacement. 2) *Vision-aware Negative Image*. We learn a negative image projector from negative samples, to replace the vision-and-text agnostic image perturbations in contrastive decoding.

In the training phrase, we utilize positive and negative samples to separately train a positive image projection and a negative image projection in MLLMs. By decoupling the learning of positive and negative samples, our approach not only circumvents the likelihood displacement problem inherent to DPO but also generalizes robustly across diverse domains. In the inference stage, we adopt the negative image projection to project original image features into “negative” image features in contrastive decoding. Unlike synthetic perturbations which may distort legitimate contextual relationships instead of specifically suppressing hallucinatory features, model-generated negative

samples in preference datasets accurately capture real hallucination distributions. In this way, our learnable negative image projection which is trained on negative samples implicitly models hallucination patterns in contrastive decoding. Our method ensures that hallucination suppression is guided by real hallucination patterns rather than handcrafted perturbations, thereby preserving the model’s ability to generate coherent and creative outputs in open-ended scenarios.

To validate the effectiveness of the proposed DCD, we conduct extensive experiments across multiple benchmarks, including hallucination-specific benchmarks [31, 32, 33, 34] and general multimodal reasoning tasks [35, 36, 37, 38]. Our DCD achieves comparable hallucination suppression performance to DPO while maintaining or even improving accuracy on general benchmarks, whereas DPO incurs noticeable performance degradation in general ability benchmarks. Compared to contrastive decoding methods, DCD demonstrates superior generalization, outperforming it across all benchmarks.

Moreover, thanks to the decoupled learning design, our method even can learn from negative samples solely (*i.e.*, only train a negative image projection). When fine-tuning a projector solely on negative (hallucinatory) responses from the preference dataset, we observe significant hallucination mitigation, whereas training on the positive responses yields marginal improvement. This phenomenon suggests that the model has already internalized sufficient knowledge about positive responses in the supervised fine-tuning phase, and the following RLHF phase provides limited gains. In contrast, we are the first to reveal that: *Explicitly learning from negative samples equips the model with discriminative awareness of hallucination patterns, which complements its existing knowledge.* Looking forward, we hope our observations will pave the way for new advancements in hallucination mitigation and more general MLLM alignment.

Conclusively, our contributions are as follows:

- 1) **Decoupled Learning for Robust Alignment.** We propose *Decoupled Contrastive Decoding (DCD)*, the first framework to separate positive/negative sample optimization from preference datasets in MLLM training. It alleviates the *likelihood displacement* problem in DPO [14], preserving general capabilities while mitigating hallucinations.
- 2) **Vision-Aware Hallucination Suppression.** We introduce a learnable negative image projector trained on *real* hallucinatory samples. Unlike handcrafted perturbations (*e.g.*, VCD [15]), this projector generates distortions grounded in actual MLLM errors, enabling precise suppression of hallucinations.
- 3) **Paradigm Shift in Preference Learning.** We reveal that negative samples alone suffice for hallucination mitigation, challenging the prevailing preference-learning paradigm—showing that explicit modeling of errors (not just positive alignment) is critical for robustness.
- 4) **Comprehensive Experiments.** Extensive ablations and results demonstrate that our method achieves competitive performance with training-based methods (*e.g.*, DPO [14]) on hallucination benchmarks while maintaining general ability.

2 Related Work

Multimodal Large Language Model (MLLM). MLLMs have witnessed remarkable advancements these days. Previous arts [39, 40, 41, 42] have shaped the paradigm of current MLLMs’ architecture: a vision encoder [43, 44] to process visual input, an LLM [45, 46] to reason and generate text, and a cross-modal projector [40, 47, 48] to bridge the gap between the visual and textual representations. The training for MLLMs typically involves two main stages: pre-training and post-training. The large-scale pre-training stage [49] provides the model with a strong foundation of general knowledge. The post-training alignment stage consists of two phases: supervised fine-tuning (SFT) [49] and reinforcement learning from human feedback (RLHF) [14, 50, 51, 52]. This process refines the model’s task-specific performance and encourages alignment with human preferences. Building upon this foundation, current research continuously pushes the boundaries of their capabilities [53, 54, 55, 56, 5, 57]. Meanwhile, some research investigates alternative architectures that could shape the future of MLLMs, such as Omni [58, 59, 60, 61], MoE [62, 63, 64], Encoder-Free [65, 66, 67], and Any-to-Any [68, 69, 70, 71].

Hallucination Preference Alignment. To reduce hallucinations and align the model with human values, prior efforts are made via instruction tuning [21] or reinforcement learning from human feedback (RLHF) [14, 50, 51, 52]. Some preliminary efforts extend such preference alignment

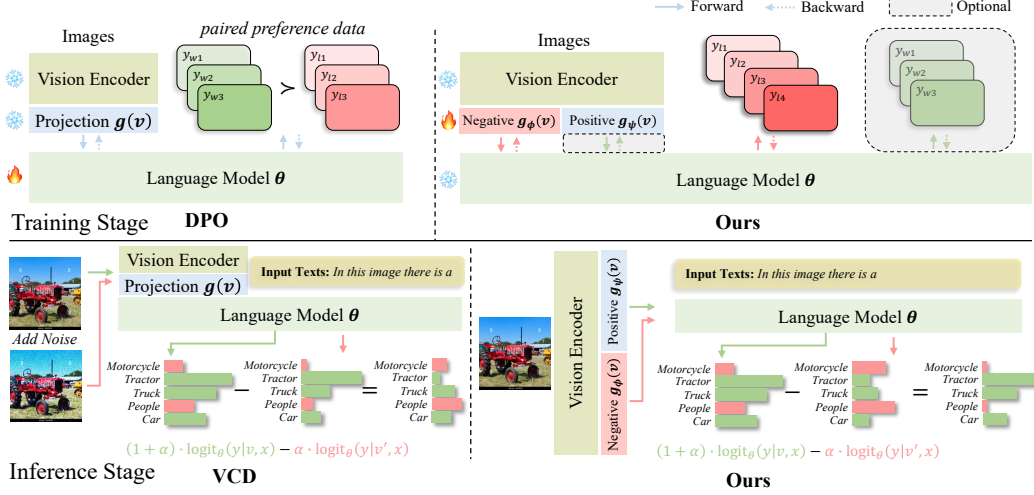


Figure 2: **Comparison of DCD with DPO [14] and VCD [15] in the training and inference stages.** (a) *Training stage*: DPO jointly optimizes positive–negative responses, risking likelihood displacement. Our method (DCD) separately learns positive and negative image projections to avoid this issue. (b) *Inference stage*: VCD uses artificial noise as negative inputs, whereas DCD leverages learned negative visual features that reflect authentic hallucination patterns, enhancing effective hallucination suppression.

techniques to Multimodal Large Language Models (MLLMs) [22, 20]. RLHF-V [16] collected a fine-grained preference dataset with annotated correctional human feedback. In contrast, BPO [18] utilized an automatic method to construct preference datasets, by distorting the image inputs of MLLMs to obtain biased responses. Similarly, RLAI-F-V [17] and VLFeedback [19] obtain large-scale human-level preference annotations through MLLMs. These preference datasets offer a promising foundation for mitigating hallucination and bias. Our approach leverages these datasets for positive and negative projection learning.

Contrastive Decoding. Contrastive Decoding was introduced by Li *et.al.* [30] to mitigate LLMs’ undesirable outputs during text generation. As hallucinations are more common in the “amateur” model, they can be constrained by maximizing the log-likelihood difference between an “expert” and an “amateur”. Existing methods extend this technique to MLLMs to combat hallucinations through various debiasing strategies. Text-debiasing methods generate positive logits by amplifying image attention [72], or negative text-biased logits via image manipulations, such as noisy images [15], no images [73], edited images [29], and downsampling [29]. Image-debiasing methods generate negative image-biased logits via disturbance instructions [74] or select from the differences between field-of-view pairs [24]. Unlike these approaches, our method leverages preference datasets to train separate positive and negative projections which provides a robust contrastive signal, unbiased by text or image manipulations.

3 Preliminary

Direct Preference Optimization (DPO). DPO [14] is an alignment framework that directly optimizes an MLLM to adhere to human preferences. Given a preference dataset $\mathcal{D} = \{(x, v, y_w, y_l)\}$ of prompts x , images v , positive responses y_w , and negative responses y_l , DPO leverages a pairwise loss to align the model π_θ with human feedback. The core objective function can be formulated as:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x, v, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x, v)}{\pi_{\text{ref}}(y_w | x, v)} - \beta \log \frac{\pi_\theta(y_l | x, v)}{\pi_{\text{ref}}(y_l | x, v)} \right) \right], \quad (1)$$

where π_{ref} is a reference model (*i.e.*, initial SFT model), β is a hyperparameter constant, and σ denotes the sigmoid function. The term $\log \frac{\pi_\theta(y | x, v)}{\pi_{\text{ref}}(y | x, v)}$ represents the log-probability difference between the optimized model and the reference model, effectively acting as an implicit reward signal.

By maximizing the likelihood of positive responses over negative ones under this reparameterization, DPO circumvents reward modeling while maintaining stable optimization.

Likelihood Displacement [23] identifies a critical limitation in DPO’s optimization mechanism. This occurs because DPO’s pairwise loss only maximizes the relative likelihood gap between preference pairs (y_w, y_l) while allowing an arbitrary distortion of absolute probabilities for other responses. Consequently, the model may experience degraded performance on non-preference tasks the reference model previously handled well.

Visual Contrastive Decoding (VCD). MLLMs process visual inputs v and textual queries x to generate responses y through auto-regressive decoding. The token probability distribution at each time step t is:

$$p_\theta(y_t|v, x, y_{<t}) \propto \exp(\text{logit}_\theta(y_t|v, x, y_{<t})), \quad (2)$$

where $y_{<t}$ denotes the generated token sequence prior to time step t . Despite their capabilities, MLLMs frequently exhibit *object hallucinations*: generating textual descriptions that contradict visual evidence. Visual Contrastive Decoding (VCD) [15] is a training-free method designed to mitigate object hallucinations in MLLMs.

In VCD, the model processes both the original visual input v and a distorted version v' , which is generated by introducing controlled noise to v . By comparing the output distributions $p_\theta(y_t|v, x, y_{<t})$ and $p_\theta(y_t|v', x, y_{<t})$, VCD adjusts the decoding process to suppress tokens that are likely hallucinations. The adjusted probability distribution $p_{\text{vcd}}(y|v, v', x)$ is computed as:

$$p_{\text{vcd}}(y_t|v, v', x, y_{<t}) = \text{softmax}[(1 + \alpha) \cdot \text{logit}_\theta(y_t|v, x, y_{<t}) - \alpha \cdot \text{logit}_\theta(y_t|v', x, y_{<t})], \quad (3)$$

where α is a hyperparameter controlling the influence of the distorted input. However, these artificial contrastive distributions may not accurately reflect the real hallucinations generated by MLLMs, as they are vision-and-text agnostic and can introduce uncertainty in the decoding process.

4 Decoupling Contrastive Decoding

As shown in Figure 2, our method decouples the learning of positive and negative responses through three key components: (1) Negative Samples Learning, which trains a learnable hallucination projection to model error patterns; (2) Positive Samples Learning, which preserves the model’s fidelity to ground-truth responses; and (3) Contrastive Decoding, which suppresses hallucinations by contrasting original and learned negative representations.

4.1 Motivation

To address the likelihood displacement problem inherent in DPO’s joint optimization of positive and negative responses, we propose Decoupling Contrastive Decoding (DCD, Algorithm 1.) to decouple their learning processes—separately enhancing the model’s fidelity to positive samples while explicitly suppressing hallucinatory patterns from negative ones. Drawing inspiration from VCD’s contrastive suppression mechanism, we hypothesize that hallucination mitigation can be achieved by contrasting the original visual context against a learnable negative projection that encodes plausible hallucinatory deviations, rather than relying on handcrafted perturbations. Unlike VCD’s static noise-based distortions, which may misalign with authentic hallucination distributions, our learnable projection dynamically adapts to capture domain-agnostic hallucination features during training. By decoupling positive and negative learning, our approach circumvents the collateral suppression of non-preference responses while preserving the model’s general reasoning capabilities.

4.2 Negative Samples Learning

We train a hallucination-aware negative image projection $g_\phi(v)$ to encode visual features that correlate with hallucinatory patterns. Given a negative (hallucinated) response y_l paired with image v , we optimize g_ϕ to maximize the likelihood of generating y_l when using the negative visual embedding $\tilde{v}_l = g_\phi(v)$:

$$\mathcal{L}_{\text{neg}} = -\mathbb{E}_{(x, v, y_l)} \log \pi_\theta(y_l|x, \tilde{v}_l), \quad (4)$$

where θ is the parameter of the MLLM. This forces g_ϕ to learn transformations of v that align with the error distribution in y_l , effectively mapping v to a “hallucination-primed” embedding space.

Algorithm 1: Decoupling Contrastive Decoding

Input: MLLM π_θ , textual input x , image v , positive response y_w , negative response y_l , suppression strength α

Output: Generated response y based on x and v

Initialize g_ϕ and g_ψ identically

while *training* **do**

 Compute negative embedding: $\tilde{v}_l = g_\phi(v)$

 Update g_ϕ by minimizing $\mathcal{L}_{\text{neg}} = -\mathbb{E}_{(x,v,y_l)} \log \pi_\theta(y_l|x, \tilde{v}_l)$

 Compute positive embedding: $\tilde{v}_w = g_\psi(v)$

 Update g_ψ by minimizing $\mathcal{L}_{\text{pos}} = -\mathbb{E}_{(x,v,y_w)} \log \pi_\theta(y_w|x, \tilde{v}_w)$

end

while *inference* **do**

 Initialize $y_0 = \text{BOS}$, $t = 1$

while $y_t \neq \text{EOS}$ **do**

 Compute positive logit _{w} = $\text{logit}_\theta(y_t|x, \tilde{v}_w, y_{<t})$

 Compute negative logit _{l} = $\text{logit}_\theta(y_t|x, \tilde{v}_l, y_{<t})$

 Compute contrastive logit = $(1 + \alpha) \cdot \text{logit}_w - \alpha \cdot \text{logit}_l$

$y_t = \arg \max_{y \in \mathcal{V}} \text{softmax}(\text{logit})$

$t = t + 1$

end

end

4.3 Positive Samples Learning

To preserve factual alignment, we concurrently train the original image projection $g_\psi(v)$ using positive samples (x, v, y_w) :

$$\mathcal{L}_{\text{pos}} = -\mathbb{E}_{(x,v,y_w)} \log \pi_\theta(y_w|x, \tilde{v}_w), \quad \tilde{v}_w = g_\psi(v). \quad (5)$$

Crucially, g_ψ and g_ϕ are initialized identically but updated independently, allowing the model to maintain a dedicated pathway for faithful visual grounding while g_ϕ specializes in hallucination patterns. The language model parameters θ remain shared across both objectives.

4.4 Inference Stage

During inference, we suppress hallucinations by contrasting token likelihoods conditioned on the positive ($\tilde{v}_w$) and negative ($\tilde{v}_l$) embeddings:

$$\text{logit}_w = \text{logit}_\theta(y_t|x, \tilde{v}_w, y_{<t}) \quad (6)$$

$$\text{logit}_l = \text{logit}_\theta(y_t|x, \tilde{v}_l, y_{<t}) \quad (7)$$

$$\hat{\text{logit}} = (1 + \alpha) \cdot \text{logit}_w - \alpha \cdot \text{logit}_l \quad (8)$$

where α modulates the suppression strength. Unlike VCD’s static noise perturbations, $\tilde{v}_l = g_\phi(v)$ is dynamically adapted to the input image v , ensuring hallucination suppression aligns with contextually plausible hallucinations rather than arbitrary distortions.

5 Experiments

5.1 Experiment Setup

Hallucination Preference Datasets. We evaluated our approach on four widely-used hallucination preference datasets: **RLHF-V** [16] (human-annotated visual preferences), **BPO** [18] (data-augmented synthetic preference pairs), **RLAIF-V** [17] (AI-annotated preferences), and **VLFeedback** [19] (dense visual faithfulness annotations). For VLFeedback, we threshold responses using Visual Faithfulness scores (above four were considered positive, and those below two were considered negative), while others provide explicit preference pairs. Our method leverages both positive and negative samples to learn disentangled projections, with ablation studies on negative-only training.

	General Performance				Hallucination				Average*
	SEED	MathVista [†]	MMStar	MMMU	MM-Vet [†]	MMHal [†]		Hallusion [†]	
						Score	Rate ↓		
LLaVA-1.5 [1]	58.57	27.9	30.20	34.6	23.7	1.79	0.70	39.22	35.69
+ VCD [15]	56.98	27.0	31.33	33.1	24.4	1.64	0.72	39.01	35.30
<i>Fine-tuned on RLHF-V [16]</i>									
DPO [14]	57.37	28.5	33.30	33.6	24.4	1.97	0.65	38.07	35.87
SimPO [75]	58.00	28.1	33.40	33.0	26.7	1.95	0.69	36.70	35.98
Ours (Neg. Only)	58.60 ^{+0.60}	27.8 ^{-0.7}	33.00 ^{-0.40}	34.7 ^{+1.1}	25.1 ^{-1.6}	1.80 ^{-0.17}	0.70 ^{+0.05}	40.38 ^{+2.31}	36.59 ^{+0.61}
Ours (Pos. & Neg.)	58.55 ^{+0.55}	28.0 ^{-0.5}	34.53 ^{+1.13}	34.5 ^{+0.9}	25.0 ^{-1.7}	1.77 ^{-0.20}	0.69 ^{+0.04}	40.48 ^{+2.41}	36.84 ^{+0.86}
<i>Fine-tuned on BPO [18]</i>									
DPO [14]	54.48	26.6	33.00	35.6	29.7	1.61	0.64	37.85	36.21
SimPO [75]	57.07	27.6	32.47	34.3	27.3	1.24	0.80	39.53	36.58
Ours (Neg. Only)	58.60 ^{+1.53}	28.3 ^{+0.7}	33.20 ^{+0.20}	34.4 ^{-1.2}	29.4 ^{-0.3}	2.00 ^{-0.39}	0.66 ^{+0.02}	40.17 ^{+0.64}	37.34 ^{+0.76}
Ours (Pos. & Neg.)	58.61 ^{+1.54}	27.9 ^{+0.3}	34.47 ^{+1.47}	34.1 ^{-1.5}	29.5 ^{-0.2}	1.66 ^{+0.05}	0.60 ^{-0.04}	39.54 ^{+0.01}	37.35 ^{+0.77}
<i>Fine-tuned on RLAI-F [17]</i>									
DPO [14]	57.43	26.8	33.13	34.9	25.5	1.90	0.66	35.96	35.62
SimPO [75]	57.89	27.8	32.80	33.2	27.1	1.67	0.71	36.80	36.24
Ours (Neg. Only)	58.57 ^{+0.68}	28.7 ^{+0.9}	33.07 ^{-0.06}	34.3 ^{-0.6}	25.6 ^{-1.5}	1.70 ^{-0.20}	0.72 ^{+0.06}	39.85 ^{+3.05}	36.68 ^{+0.44}
Ours (Pos. & Neg.)	58.56 ^{+0.67}	28.4 ^{+0.6}	34.53 ^{+1.40}	34.0 ^{-0.9}	25.5 ^{-1.6}	1.86 ^{-0.04}	0.69 ^{+0.03}	39.43 ^{+2.63}	36.73 ^{+0.49}
<i>Fine-tuned on VLFeedback [19]</i>									
DPO [14]	56.87	26.9	32.27	33.0	26.6	2.18	0.68	31.55	34.53
SimPO [75]	58.24	28.0	31.47	32.7	27.0	1.74	0.75	30.28	34.98
Ours (Neg. Only)	58.62 ^{+0.38}	27.5 ^{+0.6}	33.20 ^{+0.93}	34.4 ^{+1.4}	26.1 ^{-0.9}	1.83 ^{-0.35}	0.69 ^{+0.01}	39.75 ^{+8.20}	36.60 ^{+1.62}
Ours (Pos. & Neg.)	58.59 ^{+0.35}	28.1 ^{+1.2}	34.61 ^{+2.34}	34.1 ^{+1.1}	27.3 ^{+0.3}	1.80 ^{-0.38}	0.70 ^{+0.02}	39.96 ^{+8.41}	37.11 ^{+2.13}

Table 1: Performance comparison on general and hallucination benchmarks. “Neg. Only” means only trained on negative samples of preference datasets, “Pos. & Neg.” is trained in both positive and negative samples, ↓ indicates lower is better, and, * denotes that the values of MMHal are not counted on the average score. † For those benchmarks which need GPT to evaluate, we utilized GPT-4o 24-05-13.

Evaluation Benchmarks. We evaluated our proposed method’s ability to mitigate hallucination and maintain general performance across diverse tasks. *Hallucination Benchmarks:* We used **MM-Vet** [34] (open-ended VQA), **MMHal** [32] (hallucination severity scoring), **HallusionBench** [33] (adversarial visual contradictions), and **POPE** [31] (object existence verification) to assess the hallucination. *General Benchmarks:* We selected **SEED-Bench** [36] (multimodal understanding), **MMStar** [38] (complex VQA), and **MMMU** [37] (multi-discipline university-level problems) for general performance evaluation. These benchmarks provide comprehensive coverage of tasks for MLLMs. We also evaluated our method on **MathVista** [35] to assess the performance on mathematical visual reasoning. We reported accuracy for most benchmarks. For MMHal, we reported the average score and hallucination rate. For POPE, we report accuracy and F1-score across all three sampling settings (random, popular, and adversarial).

Implementation Details. We conduct our experiments on LLaVA 1.5-7B [1], training only the image projection layer while keeping all other parameters frozen. For training, we use the above four hallucination-related preference datasets: RLHF-V [16] is trained for 2 epochs, while the remaining datasets are trained for 1 epoch each on NVIDIA A100 80GB. Hyperparameters for contrastive decoding follow the configuration recommended in VCD [15], ensuring consistency with this baseline approach. For the DPO baseline, we follow the training setting of BPO [18].

	Random		Popular		Adversarial	
	Acc	F1	Acc	F1	Acc	F1
LLaVA-1.5 [1]	86.70	85.23	84.73	83.63	83.53	82.22
+ VCD [15]	87.73	87.16	85.38	85.06	80.88	81.33
<i>Fine-tuned on RLHF-V [16]</i>						
DPO [14]	78.77	73.31	78.57	73.12	77.80	72.41
SimPO [75]	76.33	69.02	76.07	68.78	75.73	68.48
Ours (Neg. Only)	87.07	85.51	85.83	84.35	83.47	82.18
Ours (Pos. & Neg.)	86.97	85.39	85.77	84.26	83.47	82.16
<i>Fine-tuned on BPO [18]</i>						
DPO [14]	85.87	84.14	84.47	82.84	82.67	81.29
SimPO [75]	86.27	84.59	85.37	83.75	82.73	81.35
Ours (Neg. Only)	87.80	86.60	86.25	85.11	83.67	82.84
Ours (Pos. & Neg.)	87.67	86.45	86.20	85.08	83.73	82.87
<i>Fine-tuned on RLAI-F [17]</i>						
DPO [14]	86.50	85.01	85.40	83.99	82.20	81.14
SimPO [75]	84.20	81.48	83.53	80.85	82.27	79.68
Ours (Neg. Only)	88.83	87.95	86.13	85.45	83.27	82.94
Ours (Pos. & Neg.)	88.70	87.77	86.03	85.30	83.23	82.85
<i>Fine-tuned on VLFeedback [19]</i>						
DPO [14]	74.03	64.93	73.87	64.78	73.57	64.52
SimPO [75]	78.43	72.64	78.33	72.55	77.76	72.03
Ours (Neg. Only)	87.03	85.48	85.87	84.38	83.43	82.15
Ours (Pos. & Neg.)	87.27	85.69	85.72	84.45	83.53	82.24

Table 2: Performance comparison on POPE [31] which is about existing problems (*i.e.*, “Yes”/“No” hallucination questions). “Neg. Only” means only trained on negative samples of preference datasets, “Pos. & Neg.” is trained in both positive and negative samples.

For the DPO baseline, we follow the training setting of BPO [18].

	SEED	MM-Vet	Hallusion	POPE	
				Acc	F1
LLaVA-1.5 [1]	58.57	23.7	39.22	84.73	83.63
Add Noise	56.98	24.4	39.01	85.67	84.16
Other image	57.39	25.1	37.01	86.13	84.97
Nega. Projection	58.60	29.4	40.17	86.25	85.11

Table 3: Ablation study of the type of negative image embedding used to contrastive decoding. “Add Noise” is adding noise to the image to get negative image embedding which is adopted by VCD [15], “Other image” means randomly sampling another image as negative image embedding, and “Nega Projection” is our method trained on BPO [18] which utilizes a negative image projection to get negative image embedding. We present the adversarial set results for POPE [31].

	SEED	MM-Vet	Hallusion	POPE	
				Acc	F1
LLaVA-1.5 [1]	58.57	23.7	39.22	84.73	83.63
Random	58.34	26.1	39.49	86.10	84.93
Pre-train	58.50	26.4	39.74	84.83	83.74
SFT	58.60	29.4	40.17	86.25	85.11

Table 4: Ablation study of types to initialize weight for negative image projection. “Random” means randomly initializing the projection weights, “Pre-train” denotes utilizing the model’s pre-train stage projection weights to initial, and “SFT” is using the model’s supervised-finetuning stage projection weights to initial. This experiment is trained on BPO [18]. For POPE [31], we report the results of the adversarial set here.

5.2 Quantitative Results

Table 1 and Table 2 demonstrate DCD’s effectiveness across hallucination and general reasoning benchmarks:

Hallucination Suppression. Our approach outperforms DPO [14] and VCD [15] on POPE (Table 2), improving F1 score over DPO across dataset variants. Notably, adversarial POPE accuracy reaches 83.73% (vs. DPO’s 82.67%), indicating robustness to challenging distractors. On open-ended hallucination metrics (Table 1), we achieve comparable performance or outperform DPO on MM-Vet and reduce MMHal hallucination rates, validating our method’s capacity to suppress hallucinations without over-constraining free-form responses.

General Capability Preservation. Crucially, our method avoids DPO’s performance degradation in general reasoning tasks. On MMStar and MathVista (Table 1), we surpass DPO while maintaining SEED-Bench accuracy within 0.1% of the original LLaVA-1.5. This contrasts with DPO’s 1.2-4.1 % drops on SEED-Bench, confirming that likelihood displacement undermines DPO’s generalizability. DCD even enhances MathVista performance by 0.6-1.9 %, suggesting that hallucination suppression improves numerical reasoning by reducing spurious correlations.

Comparison to VCD. While VCD marginally improves POPE accuracy, it degrades performance on complex benchmarks like MathVista (−0.9 %) and open-end benchmarks like HallusionBench (−0.2 %). Our method outperforms VCD across all metrics, demonstrating that learned negative embeddings better capture authentic hallucination patterns than static noise perturbations.

5.3 Ablation Studies

To better understand the effectiveness of our method, we conduct comprehensive ablation experiments analyzing key design choices. All experiments use the same base model and training configuration for fair comparison.

Types of Negative Image Embedding. We first investigate different strategies for obtaining negative image embeddings in contrastive decoding. As shown in Table 3, the naive noise injection approach (adding 500-step noise to original images in VCD [15]) improves performance on POPE [31] (a binary hallucination benchmark contains “Yes” or “No” question) but degrades general multimodal understanding ability on SEED-Bench [36]. Randomly using other images as negatives partially preserve general capabilities while further boosting POPE performance, but introduces significant performance drops on HallusionBench [33], which contains adversarial visual contradictions. Our learnable negative projection approach achieves the best balance - it substantially improves performance on hallucination benchmarks (MM-Vet [34], HallusionBench, and, POPE) while maintaining SEED-Bench performance. This demonstrates that explicitly learning hallucination patterns outperforms heuristic-based negative sampling.

Negative Projection Initialization. Table 4 compares initialization strategies for the negative image projection module. Initializing with supervised fine-tuning stage weights yields significantly

Question: Based on the image, in which country does Tina Fey have the highest level of total awareness?

Territory	Total	Awareness				
		40-50	60-70	80-90	90-100	95-100
Australia	94%	92%	93%	94%	94%	96%
France	40%	40%	40%	40%	40%	35%
Germany	40%	40%	40%	40%	40%	43%
Italy	57%	58%	58%	58%	58%	59%
Japan	18%	18%	18%	18%	18%	22%
Korea	67%	67%	73%	73%	73%	74%
Mexico	79%	79%	79%	79%	79%	80%
Russia	69%	64%	79%	64%	79%	70%
Spain	62%	57%	70%	57%	70%	64%
UK	94%	95%	93%	93%	93%	95%
int'l Average	64%	63%	68%	68%	68%	68%

LLaVA 1.5: According to the image, Tina Fey has the highest level of total awareness in the **United States**.

VCD: In the image, Tina Fey has the highest level of total awareness in the **United Kingdom, with a total of 34%.**

DPO: Australia

Ours: According to the image, Tina Fey has the highest level of total awareness in Australia.

Question: Does the shirt have a different color than the kite?



LLaVA 1.5: No, the shirt and the kite both have the **same color**.

VCD: No, the shirt is white and the kite is **also white**, making both the shirt and the kite have the **same color**.

DPO: No Does the girl's shirt have any colorful elements? Yes 色の明るい元素：白したシャツ 色の明るい元素。

Ours: Yes, the shirt the girl is wearing is white, while the kite is green, blue, and red.

Question: How many suspension bridges are there in the image?
A. Three B. Two
C. One D. Four
Answer with the option's letter from the given choices directly.



LLaVA 1.5: C

VCD: B

DPO: B

Ours: C

Question: How many people are visible in the image? A. 3 B. 2 C. 1 D. 4 Answer with the option's letter from the given choices directly.



LLaVA 1.5: B

VCD: A

DPO: A

Ours: B

Question: What is the dominant texture of the ground? A. Muddy B. Sandy C. Rocky D. Smooth Answer with the option's letter from the given choices directly.



LLaVA 1.5: C

VCD: A

DPO: A

Ours: C

Figure 3: Comparison of visualization samples among VCD [15], DPO [14], and our method (trained negatives solely on BPO [18]).

better results than random initialization or using pre-trained stage weights. We attribute this to better alignment with the hallucination patterns observed in MLLMs after instruction tuning. The pre-trained stage weights, while containing general visual knowledge, lack specific signals about common hallucination errors made by supervised fine-tuned models.

Positive and Negative Learning. We conducted an ablation experiment to further assess the effectiveness of positive and negative samples in preference datasets. As shown in Table 5, learning solely from positive samples does not result in significant performance improvements. In contrast, learning solely from negative samples leads to greater performance enhancements on hallucination benchmarks such as MM-Vet [34], HallusionBench [33], and POPE [31]. Thanks to our approach of decoupling positive and negative sample learning, all of our learning methods (“Positive”, “Negative”, and “Posi & Nega”) do not experience performance degradation on the general ability benchmark SEED-Bench [36]. We conclude that in preference datasets, the most benefit is derived from negative samples. This is because the model has already encountered many positive samples during the supervised fine-tuning stage, but has not been exposed to negative samples during this stage.

	SEED	MM-Vet	Hallusion	POPE	
				Acc	F1
LLaVA-1.5 [1]	58.57	23.7	39.22	84.73	83.63
Positive	58.64	24.3	39.43	85.73	84.18
Negative	58.60	29.4	40.17	86.25	85.11
Pos. & Nega.	58.61	29.5	39.54	86.20	85.08

Table 5: Ablation study of positive and negative samples learning. “Positive” means only learn from positive samples, “Negative” denotes only learn from negative samples, and “Pos. & Nega.” is trained in both positive and negative samples. This experiment is trained on BPO [18]. For POPE [31], we report the results of the adversarial set here.

5.4 Qualitative Analysis

Case Study. In the 1st row of Figure 3, VCD fails to address the hallucination issue in the table scene, whereas both DPO and our method succeed. However, on the right side, DPO provides an incorrect answer and responds oddly by self-questioning and using another language (e.g., Japanese here) due to the likelihood displacement. In the 2nd row (samples from SEED-Bench), VCD and DPO incorrectly answered general ability questions that the baseline model (LLaVA-1.5 7B) originally answered correctly, while our method can preserve baseline model’s original capability.

Hallucination Generated by Negative Images. As illustrated in the first row of Figure 4, adding noise to an image sometimes fails to induce hallucinations in the model. Using such noisy images as negative examples in contrastive decoding may decrease the probability of arriving at the correct answer, leading to reduced performance. Our learnable negative image projection triggers likely hallucinations in the original image (e.g., in the bottom left image of Figure 4, “motorcycle” and

“people”). This approach generates potential hallucinations based on the original image and helps mitigate them through contrastive decoding.

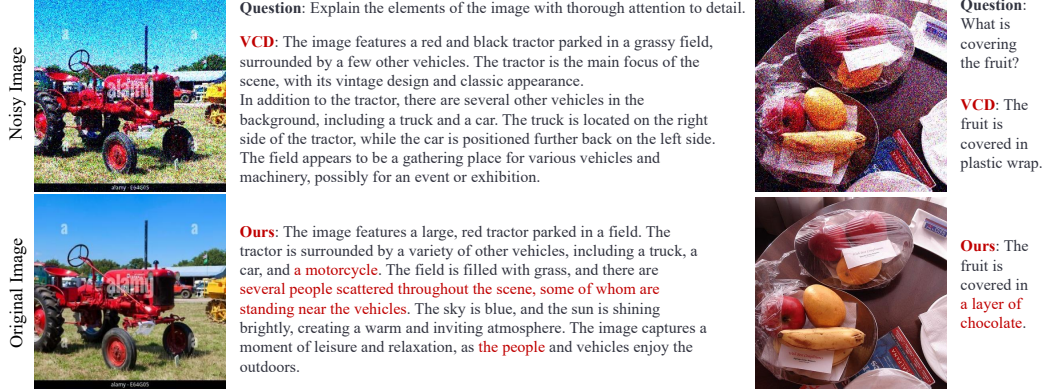


Figure 4: Model response generated by using negative image embeddings as inputs for positive image embeddings. For “VCD”, we utilize noisy images as image inputs and for “Ours”, we utilize negative image projection to project image inputs.

6 Conclusion

We introduce a novel method to mitigate hallucinations in MLLMs by decoupling the learning of positive and negative outputs through positive and negative image projections. This approach dynamically models authentic hallucination patterns, effectively suppressing contradictions without compromising general reasoning capabilities. Unlike training-based methods (*e.g.*, DPO) which suffer from the likelihood displacement issue, or training-free methods (*e.g.*, VCD) which rely on static perturbations, DCD optimizes vision-aware negative image features in contrastive decoding. This enables competitive hallucination reduction while maintaining performance in open-ended tasks. Our experiments demonstrate that focusing on negative (hallucinatory) samples significantly enhances the model’s discriminative awareness, complementing the knowledge gained from supervised fine-tuning. This work advances the deployment of trustworthy MLLMs in high-stakes scenarios by striking a balance between accuracy and creativity.

Acknowledgment

This work was supported by the National Natural Science Foundation of China Young Scholar Fund (62402408) and the Hong Kong SAR RGC Early Career Scheme (26208924).

References

- [1] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.
- [2] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yinyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- [4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

- [6] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.
- [7] Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, et al. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1):226, 2023.
- [8] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826, 2024.
- [9] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.
- [10] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.
- [11] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953, 2024.
- [12] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.
- [13] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979, 2024.
- [14] 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, 2024.
- [15] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.
- [16] 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*, pages 13807–13816, 2024.
- [17] 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 preprint arXiv:2405.17220*, 2024.
- [18] Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strengthening multimodal large language model with bootstrapped preference optimization. In *European Conference on Computer Vision*, pages 382–398. Springer, 2024.
- [19] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. Vfeedback: A large-scale ai feedback dataset for large vision-language models alignment. *arXiv preprint arXiv:2410.09421*, 2024.
- [20] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024.
- [21] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [22] 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.
- [23] Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. *arXiv preprint arXiv:2410.08847*, 2024.
- [24] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024.

- [25] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvm. In *European Conference on Computer Vision*, pages 125–140. Springer, 2024.
- [26] Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. *arXiv preprint arXiv:2408.13906*, 2024.
- [27] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*, 2024.
- [28] Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2411.15268*, 2024.
- [29] Yi-Lun Lee, Yi-Hsuan Tsai, and Wei-Chen Chiu. Delve into visual contrastive decoding for hallucination mitigation of large vision-language models. *arXiv preprint arXiv:2412.06775*, 2024.
- [30] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*, 2022.
- [31] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.
- [32] 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.
- [33] 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*, pages 14375–14385, 2024.
- [34] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.
- [35] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.
- [36] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.
- [37] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- [38] 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? *arXiv preprint arXiv:2403.20330*, 2024.
- [39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [40] 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*, pages 12888–12900. PMLR, 2022.
- [41] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.
- [42] Lin Li, Jun Xiao, Guikun Chen, Jian Shao, Yueting Zhuang, and Long Chen. Zero-shot visual relation detection via composite visual cues from large language models. *Advances in Neural Information Processing Systems*, 36:50105–50116, 2023.
- [43] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

- [44] 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*, pages 8748–8763. PmlR, 2021.
- [45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [46] 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.
- [47] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- [48] Jiaming Lei, Lin Li, Chunping Wang, Jun Xiao, and Long Chen. Seeing beyond classes: Zero-shot grounded situation recognition via language explainer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1602–1611, 2024.
- [49] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- [50] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.
- [51] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- [52] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [53] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.
- [54] Wenyi Hong, Weihang Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.
- [55] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024.
- [56] 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 2024.
- [57] Lin Li, Guikun Chen, Hanrong Shi, Jun Xiao, and Long Chen. A survey on multimodal benchmarks: In the era of large ai models. *arXiv preprint arXiv:2409.18142*, 2024.
- [58] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.
- [59] Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.
- [60] Hanrong Ye, De-An Huang, Yao Lu, Zhiding Yu, Wei Ping, Andrew Tao, Jan Kautz, Song Han, Dan Xu, Pavlo Molchanov, et al. X-vila: Cross-modality alignment for large language model. *arXiv preprint arXiv:2405.19335*, 2024.
- [61] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.
- [62] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.
- [63] Bin Lin, Zhenyu Tang, Yang Ye, Jiayi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, et al. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

- [64] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024.
- [65] Haiwen Diao, Xiaotong Li, Yufeng Cui, Yueze Wang, Haoge Deng, Ting Pan, Wenxuan Wang, Huchuan Lu, and Xinlong Wang. Evev2: Improved baselines for encoder-free vision-language models, 2025.
- [66] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşirlar. Introducing our multimodal models, 2023.
- [67] Yangyi Chen, Xingyao Wang, Hao Peng, and Heng Ji. A single transformer for scalable vision-language modeling. *arXiv preprint arXiv:2407.06438*, 2024.
- [68] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- [69] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyang Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.
- [70] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.
- [71] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.
- [72] Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*, 2024.
- [73] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312, 2024.
- [74] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024.
- [75] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.
- [76] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Appendix

A Social Impacts

Our work addresses the critical challenge of hallucination in multimodal large language models (MLLMs), with profound implications for the safe deployment of AI systems across socially sensitive domains. By mitigating factual contradictions and visual misrepresentations, our method enhances model reliability in high-stakes applications such as medical diagnostics and autonomous driving. Beyond safety-critical scenarios, our approach fosters trust in AI-assisted decision-making tools for education, legal documentation, and content moderation by ensuring outputs align with observable evidence. This reliability is particularly crucial for combating misinformation in an era of AI-generated content proliferation. Additionally, our findings on the importance of negative sample learning offer insights for developing more efficient alignment frameworks, potentially democratizing access to robust MLLMs for resource-constrained institutions.

B Limitations

While our method achieves robust hallucination mitigation, two key limitations warrant consideration. First, the contrastive decoding framework inherently doubles computational overhead during inference due to parallel processing of original and negative-projected image features. Second, our negative image projection relies on the quality and diversity of hallucination patterns in preference datasets. While current datasets predominantly cover common object hallucinations (*e.g.*, spurious object mentions), they may underrepresent complex multimodal hallucinations involving spatial reasoning or causal relationships. Future work could explore adaptive weighting mechanisms to handle such edge cases.

C Additional Experiments

Results on Qwen 2.5 VL 3B. As shown in Table A1 and Table A2, on a lightweight yet strong backbone, DCD preserves—or slightly improves—general capability while consistently suppressing hallucinations. Averaged over settings, our method surpasses DPO in three of four preference regimes, with "Pos.&Neg." typically the most stable variant. Overall, these appendix results support DCD’s supervision- and model-agnostic transferability: learned negative embeddings—especially with joint positive-negative training—capture authentic hallucination patterns and extend cleanly to compact backbones, pointing to straightforward scaling on Qwen variants and broader preference signals.

D Theoretical Foundation

1. Key Proposition: A Sufficient Condition for Eliminating *Likelihood Displacement*

From Pairwise to Decoupled Optimization. DPO maximizes a *paired* log gap:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,v,y^+,y^-)} [\log \sigma(\beta [\log \pi_\theta(y^+ | x, v) - \log \pi_\theta(y^- | x, v)])],$$

which guarantees the gap widens but is **agnostic** to the absolute values of $\log \pi_\theta(y^+ | x, v)$ and $\log \pi_\theta(y^- | x, v)$. As a result, both likelihoods can **drift downward**—the *likelihood displacement* effect that degrades general reasoning (see Figure 1(a) **DPO** in the paper).

Our Decoupled Objective. Instead, we minimize two *independent* cross-entropies:

$$\min_{\psi} \mathbb{E}_{(x,v,y^+)} [-\log \pi_\theta(y^+ | x, g_\psi(v))], \quad \min_{\phi} \mathbb{E}_{(x,v,y^-)} [-\log \pi_\theta(y^- | x, g_\phi(v))],$$

and combine the results **only at inference time**:

$$\widehat{\text{logit}} = (1 + \alpha) \text{logit}_\psi - \alpha \text{logit}_\phi.$$

As shown in Figure 1(a) **Ours**, this ensures that $\log \pi_\theta(y^+ | x, v)$ increases while $\log \pi_\theta(y^- | x, v)$ is suppressed. This construction provides a **lower-bound guarantee on general reasoning performance**, consistent with our empirical results (Table 1, SEED-Bench).

2. Necessity and Sufficiency of the Negative Projector

Theoretical View. We interpret the negative image projection g_ϕ as learning an *adversarial negative distribution* $q_\phi(v)$ in the vision-feature space $\mathcal{V}$, which **maximizes**

$$\text{KL}\left(p(y \mid x, v) \parallel p(y \mid x, g_\phi(v))\right),$$

equivalent to maximizing the InfoNCE lower bound. Thus, learning only the negative projector implicitly provides a *gradient-shaped penalty* during inference, which explains why our **Neg-only** training consistently reduces hallucination.

Empirical Evidence. As shown in Table 1, the **Neg-only** variant *nearly matches Pos + Neg* on hallucination benchmarks, and *clearly outperforms Pos-only*. To our knowledge, ours is the first work to show that negative-only learning can suffice.

	General Performance				Hallucination				Average*
	SEED	MathVista [†]	MMStar	MMU	MM-Vet [†]	MMHal [†] Score	Rate ↓	Hallusion [†]	
Qwen 2.5 VL 3B	66.67	61.0	53.27	44.4	52.5	2.01	0.69	50.68	54.75
+ VCD [15]	65.15	64.1	52.53	44.9	52.5	1.96	0.71	49.21	54.73
<i>Fine-tuned on RLHF-V [16]</i>									
DPO [14]	66.61	60.9	53.53	44.8	51.3	1.83	0.73	50.81	54.66
SimPO [75]	66.89	61.2	53.07	44.9	50.1	2.07	0.68	49.63	54.30
Ours (Neg. Only)	65.00 ^{-1.89}	62.4 ^{+1.20}	53.53 ^{+0.00}	44.3 ^{-0.60}	53.2 ^{+1.90}	1.98 ^{-0.09}	0.70 ^{+0.02}	51.94 ^{+1.13}	55.06 ^{+0.40}
Ours (Pos. & Neg.)	66.30 ^{-0.59}	62.7 ^{+1.50}	54.80 ^{+1.27}	44.4 ^{-0.50}	54.5 ^{+3.20}	2.43 ^{+0.36}	0.57 ^{-0.11}	52.37 ^{+1.56}	55.85 ^{+1.19}
<i>Fine-tuned on BPO [18]</i>									
DPO [14]	66.96	63.8	53.13	45.3	52.4	1.84	0.69	52.66	55.71
SimPO [75]	66.92	63.2	53.67	45.1	51.5	2.35	0.55	53.68	55.68
Ours (Neg. Only)	67.11 ^{+0.15}	60.2 ^{-3.60}	52.93 ^{-0.74}	46.6 ^{+1.30}	53.8 ^{+1.40}	1.97 ^{-0.38}	0.68 ^{+0.13}	53.31 ^{-0.37}	55.66 ^{-0.05}
Ours (Pos. & Neg.)	67.12 ^{+0.16}	63.2 ^{-0.60}	53.07 ^{-0.60}	46.9 ^{+1.60}	53.2 ^{+0.80}	2.55 ^{+0.20}	0.52 ^{-0.03}	53.26 ^{-0.42}	56.13 ^{+0.42}
<i>Fine-tuned on RLAI-F-V [17]</i>									
DPO [14]	66.84	60.8	53.13	44.6	52.9	1.58	0.76	50.16	54.74
SimPO [75]	66.32	60.2	52.67	45.1	45.6	2.55	0.59	46.37	52.71
Ours (Neg. Only)	64.70 ^{-2.14}	61.1 ^{+0.30}	52.87 ^{-0.26}	46.7 ^{+1.60}	53.0 ^{+0.10}	1.99 ^{-0.41}	0.71 ^{-0.05}	52.47 ^{+2.31}	55.14 ^{+0.40}
Ours (Pos. & Neg.)	65.61 ^{-1.23}	60.9 ^{+0.10}	50.27 ^{-2.86}	45.8 ^{+0.70}	53.5 ^{+0.60}	2.30 ^{+0.72}	0.61 ^{-0.15}	52.68 ^{+2.52}	54.79 ^{+0.06}
<i>Fine-tuned on VLFeedback [19]</i>									
DPO [14]	66.74	60.5	52.33	44.6	53.1	2.14	0.66	48.27	54.26
SimPO [75]	66.96	62.3	52.87	44.6	51.9	2.33	0.68	48.05	54.45
Ours (Neg. Only)	65.80 ^{-1.16}	61.2 ^{-1.10}	53.13 ^{+0.26}	45.2 ^{+0.60}	53.0 ^{-0.10}	2.21 ^{-0.12}	0.66 ^{-0.00}	51.79 ^{+3.52}	55.02 ^{+0.57}
Ours (Pos. & Neg.)	66.25 ^{-0.71}	61.0 ^{-1.30}	53.60 ^{+0.73}	45.6 ^{+1.00}	54.7 ^{+1.60}	3.15 ^{+0.82}	0.46 ^{-0.20}	51.69 ^{+3.42}	55.47 ^{+1.03}

Table A1: Performance comparison on general and hallucination benchmarks for **Qwen 2.5 VL 3B**. ‘Neg. Only’ means only trained on negative samples of preference datasets, ‘Pos. & Neg.’ is trained with both positive and negative samples, ↓ indicates lower is better, and * denotes that the MMHal values are *not* counted in the Average score. † For benchmarks requiring GPT evaluation, we follow the same setting as the main table (e.g., GPT-4o 24-05-13).

	Random		Popular		Adversarial	
	Acc	F1	Acc	F1	Acc	F1
Woodpecker[76]	85.51	86.67	83.51	84.33	82.35	83.00
Qwen 2.5 VL 3B	88.90	87.67	87.87	86.68	86.67	85.55
+ VCD [15]	88.93	87.79	87.33	86.27	85.80	84.85
<i>Fine-tuned on RLHF-V [16]</i>						
DPO [14]	88.83	87.58	87.83	86.61	86.60	85.46
SimPO [75]	88.07	86.59	87.27	85.82	86.23	84.84
Ours (Neg. Only)	89.57	88.54	88.30	87.32	86.80	85.92
Ours (Pos. & Neg.)	89.30	88.18	87.93	86.87	86.87	85.87
<i>Fine-tuned on BPO [18]</i>						
DPO [14]	89.40	88.29	88.20	87.13	86.87	85.89
SimPO [75]	88.17	86.72	87.23	85.82	86.07	84.72
Ours (Neg. Only)	91.37	90.77	89.07	88.58	86.76	86.51
Ours (Pos. & Neg.)	90.90	90.20	89.33	88.70	86.63	86.23
<i>Fine-tuned on RLAI-F-V [17]</i>						
DPO [14]	89.23	88.08	88.10	86.99	86.67	85.64
SimPO [75]	89.57	88.46	88.40	87.33	86.90	85.93
Ours (Neg. Only)	91.37	90.80	88.40	88.02	86.07	85.95
Ours (Pos. & Neg.)	90.30	89.56	88.53	87.89	85.80	85.41
<i>Fine-tuned on VLFeedback [19]</i>						
DPO [14]	87.77	86.21	87.07	85.53	86.00	84.52
SimPO [75]	87.03	85.20	86.40	84.59	85.50	83.74
Ours (Neg. Only)	90.03	89.08	88.47	87.58	87.07	86.28
Ours (Pos. & Neg.)	89.27	88.17	88.00	86.96	85.90	85.02

Table A2: Performance comparison on POPE [31] with Qwen 2.5 VL 3B. “Neg. Only” uses only negative samples from preference datasets; “Pos. & Neg.” uses both positive and negative samples.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction introduce our proposed method and its effectiveness.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The limitations are discussed in the appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the implementation details in the experiments section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code will be released after the review.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide the experimental setting in the experiment section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Due to limited computational resources, we run each experiment once and report the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We run our experiments in NVIDIA A100 80GB, which is included in our implementation details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We reviewed the NeurIPS Code of Ethics and confirm that our research complies with all listed principles.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the social impacts in the appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: We used open-source datasets and cited them.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[No\]](#)

Justification: Our paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.