ALIGNMENT-STATE-AWARE PREFERENCE OPTIMIZATION FOR MITIGATING HALLUCINATIONS IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct Preference Optimization (DPO) has emerged as a widely adopted approach to mitigating hallucinations in multimodal large language models (MLLMs), but it relies on response-level preference signals without distinguishing the importance of individual tokens. Recent studies have attempted to achieve fine-grained, token-level preference optimization by leveraging visual relevance; however, these approaches still suffer from two key limitations. First, visual relevance does not necessarily align with the model's optimization needs: initially difficult tokens may become well aligned during training, making continued emphasis less beneficial despite their high visual relevance. Second, fine-grained token weighting alters the aggregated reward scale, causing preference margins to shift uncontrollably and destabilizing the overall optimization strength. To address these issues, we propose Alignment-State-Aware Preference Optimization (AsAPO). Specifically, AsAPO constructs a perceptual boundary from rejected responses to characterize the model's predictive behavior for a given input. Relative to this boundary, it dynamically estimates each token's optimization needs within the preferred response at the current alignment state and assigns weights accordingly, directing supervision toward content that remains insufficiently aligned as training progresses. Furthermore, AsAPO introduces a preference-margin stabilization constraint to suppress excessive margin shifts induced by token weighting, enabling effective fine-grained supervision while maintaining stable preference optimization. Extensive experiments across three MLLMs and four hallucination benchmarks show that AsAPO consistently outperforms standard DPO and existing fine-grained methods, effectively improving visual faithfulness and mitigating hallucinations.

1 INTRODUCTION

In recent years, the rapid development of multimodal large language models (MLLMs) has significantly enhanced their ability to perceive and interact with the real world (Liu et al., 2024; Team, 2025), leading to remarkable progress in tasks such as multimodal understanding and reasoning (Du et al., 2026; Tan et al., 2025; Zhang et al., 2026b). Despite this progress, MLLMs remain prone to hallucinations, where the generated content is inconsistent with the visual evidence provided by the input image, such as mentioning nonexistent objects or mischaracterizing object attributes and relationships (Leng et al., 2024; Liu et al., 2025). Such hallucinations severely undermine the reliability of MLLMs and hinder their safe and trustworthy deployment in real-world applications.

Direct Preference Optimization (DPO) (Rafailov et al., 2023), owing to its effectiveness in model alignment, has been widely adopted to mitigate hallucinations in MLLMs (Yang et al., 2025; Liu et al., 2025). Existing approaches typically treat responses sampled from the model itself as rejected responses y_l, and construct the corresponding preferred responses y_w through correction and refinement. DPO then enlarges the margin between the preference pairs, progressively steering the model away from its original generation distribution toward a more reliable preference distribution that is better aligned with visual evidence. Despite the substantial progress, standard DPO still relies primarily on response-level preference signals without explicitly accounting for the varying importance of individual tokens in hallucination mitigation. Such coarse-grained optimization may dilute the supervision of critical tokens throughout the sequence, limiting the ability of preference

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

learning to provide targeted and effective optimization (Yang et al., 2026). To address this limitation, recent studies have explored fine-grained, token-level preference optimization. For example, TPO (Gu et al., 2025) distinguishes token importance by identifying visual anchors, while CAT-PO (Zheng et al., 2026) integrates visual-relevance rewards at multiple levels to support token-level optimization. These methods incorporate token-level importance into preference learning by assigning differentiated optimization weights based primarily on visual relevance.

However, these methods overlook two critical issues. First, visual relevance does not necessarily reflect the importance of a token in preference optimization. We argue that token importance should instead be determined by its learning needs under the model's current alignment state: tokens that are difficult to align at an early training stage may become sufficiently learned as training progresses, even while remaining highly relevant to the visual content (see Sec. 4.4). In contrast, critical details that remain under-aligned require more targeted supervision. Second, fine-grained weighting not only changes the optimization strength of individual tokens but also alters the aggregated response-level reward scale, causing the preference margin between preferred and rejected responses to be uncontrollably compressed or enlarged, which may in turn lead to excessive optimization of already enlarged margins. This reveals a local–global dilemma in fine-grained preference optimization: token-level supervision should adapt to the model's evolving alignment state, while response-level preference optimization should remain stable. This raises a fundamental challenge: how to achieve adaptive token-level credit assignment without destabilizing the response-level preference margin.

To address this challenge, we present Alignment-State-Aware Preference Optimization (AsAPO) to mitigate hallucinations in MLLMs. DPO can be viewed as progressively strengthening the model's relative preference for generating preferred responses over rejected ones. From this perspective, AsAPO uses the reference model's predictive probability on the rejected response as a baseline, since the rejected response—sampled from the model itself—more faithfully reflects the model's own prediction capability under the given input. Based on this baseline, AsAPO measures the relative prediction difficulty of each token in the preferred response, dynamically identifies tokens that require stronger optimization and assigns weights accordingly. As the model's alignment state evolves, this mechanism dynamically shifts the focus of supervision toward critical content that remains insufficiently learned. Furthermore, to mitigate the preference-margin shifts induced by fine-grained token weighting, we introduce a preference-margin stabilization constraint that prevents excessive deviation from the original response-level margin. In this way, AsAPO enables adaptive token-level supervision while preserving the stability of global preference optimization. Extensive experiments across three MLLMs and four hallucination benchmarks demonstrate that AsAPO achieves superior performance over standard DPO and existing fine-grained optimization methods, effectively improving the visual faithfulness of generated responses and mitigating hallucinations.

The main contributions of this work can be summarized as follows:

- We introduce a new perspective on fine-grained preference optimization for mitigating hallucinations in MLLMs, distinguishing visual relevance from optimization importance and adapting token-level supervision to the model's current alignment needs.
- We propose AsAPO, which dynamically quantifies token importance through the prediction difficulty of preferred responses relative to rejected responses. We further introduce a preference-margin stabilization constraint to curb excessive margin shifts induced by token weighting while enabling differentiated token-level supervision.
- Extensive experiments across multiple models and hallucination benchmarks demonstrate that AsAPO achieves superior overall performance to standard DPO and existing fine-grained DPO methods.

2 PRELIMINARIES

Direct Preference Optimization (Rafailov et al., 2023) directly optimizes a policy model from pairwise preference data without explicitly training a separate reward model, and has become a widely adopted paradigm for model post-training (Yang et al., 2024; Bai et al., 2025). DPO models pairwise human preferences using the Bradley–Terry (BT) formulation. Specifically, given a preference triplet (p, y_w, y_l) sampled from the data distribution $\mathcal{D}$, where p denotes the input prompt and y_w and y_l denote the preferred and rejected responses, respectively, the BT model assumes that the

probability of preferring y_w over y_l is determined by their reward difference:

$$p(y_w \succ y_l \mid p) = \frac{\exp\big(r(p, y_w)\big)}{\exp\big(r(p, y_w)\big) + \exp\big(r(p, y_l)\big)} = \sigma\big(r(p, y_w) - r(p, y_l)\big), \tag{1}$$

where $r(p, y)$ denotes the reward of the response, and $\sigma(\cdot)$ is the sigmoid function.

Model alignment is commonly formulated as a KL-regularized reward maximization problem:

$$\max_{\pi_\theta} \mathbb{E}_{p\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi_\theta(\cdot|p)}\big[r(p, y)\big] - \beta D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid p) \,\|\, \pi_{ref}(\cdot \mid p)\big)\right]. \tag{2}$$

Here, π_θ denotes the policy model, π_{ref} denotes the reference model, and $\beta > 0$ controls the strength of KL regularization. The optimal policy π_θ admits a closed-form solution, which allows the reward function to be expressed as: $r(p, y) = \beta \log \frac{\pi_\theta(y|p)}{\pi_{ref}(y|p)} + \beta \log Z(p)$. Here, $Z(p)$ is the partition function, which depends only on the input p. Substituting this reward parameterization into the BT model cancels the partition-function terms in the reward difference. Using a parameterized policy π_θ to instantiate this implicit reward and substituting it into the BT likelihood yields the DPO objective:

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\mathbb{E}_{(p,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid p)}{\pi_{ref}(y_w \mid p)} - \beta\log\frac{\pi_\theta(y_l \mid p)}{\pi_{ref}(y_l \mid p)}\right)\right]. \tag{3}$$

Nevertheless, standard DPO aggregates these token-level contributions into a single response-level preference margin, without explicitly distinguishing the varying alignment difficulty of individual tokens. This limitation motivates the fine-grained preference optimization developed.

3 METHODOLOGY

To mitigate hallucinations in MLLMs, we propose Alignment-State-Aware Preference Optimization (AsAPO), as illustrated in Fig 1.

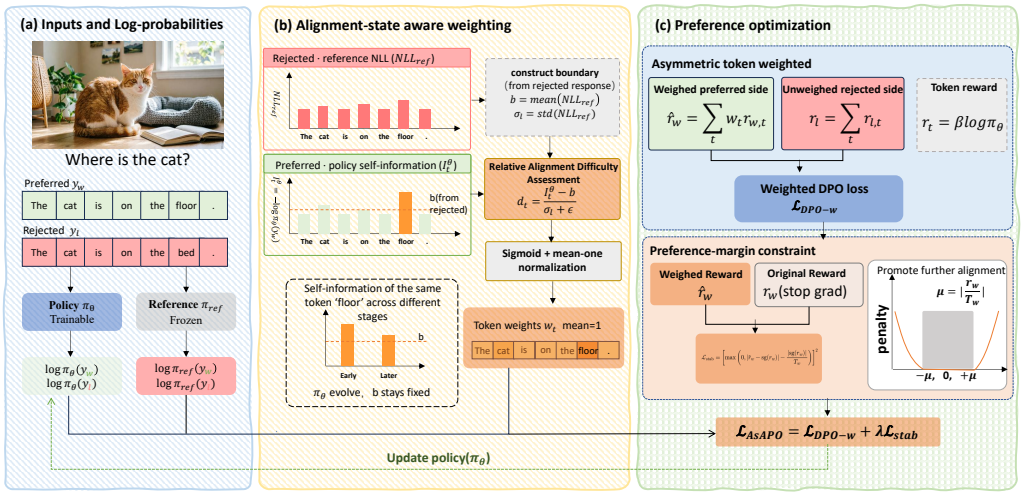


Figure 1: Schematic Illustration of the AsAPO Algorithm.

3.1 ALIGNMENT-STATE-AWARE TOKEN IMPORTANCE ESTIMATION

To estimate token importance from the perspective of the model's current alignment state, AsAPO follows a three-stage process. First, we compute the self-information of each token in the preferred response and construct a perceptual boundary from the rejected response, providing a reference scale for subsequent comparison. Second, we measure the relative alignment difficulty of each preferred token with respect to this boundary, identifying tokens that remain insufficiently aligned

under the current model state. Finally, the resulting difficulty scores are transformed into token-level optimization weights and normalized to control the overall weighting scale.

Self-Information and Perceptual Boundary. Self-information provides a natural measure of how surprising a token is under the current policy. Given a question x, an image m, and a preferred response y_w, we first compute the self-information of each preferred token as:

$$I_{w,t} = -\log \pi_\theta \left(y_{w,t} \mid x, m, y_{w,<t} \right). \tag{4}$$

Here, π_θ denotes the policy being optimized, and $y_{w,<t}$ denotes the response prefix preceding the t-th token. To determine whether a preferred token remains difficult to align, we further introduce a perceptual boundary as a reference anchor. Since rejected responses y_l are closer to model's original generation distribution, their prediction difficulty provides a natural baseline for characterizing the model's predictive level under the given input. Specifically, we define the perceptual boundary as the mean token-level negative log-likelihood (i.e., cross-entropy) of the rejected response under the reference model π_{ref}:

$$b(x, m, y_l) = -\frac{1}{T_l} \sum_{j=1}^{T_l} \log \pi_{\text{ref}} \left(y_{l,j} \mid x, m, y_{l,<j} \right) \tag{5}$$

where T_l denotes the number of valid tokens in the rejected response. The resulting boundary serves as a reference prediction level for the rejected response under the given input. A preferred token with $I_{w,t} > b$ is more difficult for the current policy to predict, indicating a greater need for further alignment. This self-information and perceptual boundary forms the basis for the subsequent estimation of relative alignment difficulty. Note that $I_{w,t}$ is computed under the evolving policy π_θ, while the boundary b is computed under the fixed reference model π_{ref}, serving as a stable anchor against which the model's ongoing alignment progress can be measured.

Relative Alignment Difficulty Assessment. Given the token-level self-information of the preferred response y_w and the perceptual boundary b, we quantify each token's relative alignment difficulty d_t under the current policy through its standardized deviation from the boundary:

$$d_t = \frac{I_{w,t} - b}{\sigma_l + \epsilon} \tag{6}$$

where $\epsilon > 0$ is a small constant for numerical stability, and σ_l is the standard deviation of token-level negative log-likelihoods in the rejected response under the reference model:

$$\begin{cases} \sigma_l = \sqrt{\dfrac{1}{T_l} \sum_{j=1}^{T_l} \left(I_{l,j}^{\text{ref}} - b \right)^2}, \\ I_{l,j}^{\text{ref}} = -\log \pi_{\text{ref}} \left(y_{l,j} \mid x, m, y_{l,<j} \right). \end{cases} \tag{7}$$

This transformation defines a reference coordinate system for prediction difficulty, using the rejected response's mean token-level negative log-likelihood b as its origin and the corresponding standard deviation σ_l as its scale. The self-information of each preferred token under the current policy is mapped into this coordinate system to measure its deviation from the reference prediction level. When $d_t > 0$, the token's prediction difficulty exceeds the reference level, indicating a greater relative alignment need and motivating stronger optimization emphasis. When $d_t \leq 0$, its prediction difficulty is at or below the reference level, indicating a lower relative alignment need for which base supervision is retained. As the policy evolves, $I_{w,t}$ and its standardized deviation d_t change accordingly, providing a dynamic basis for subsequent token weighting.

Token Weight Construction and Scale Normalization. AsAPO maps the relative alignment difficulty d_t to a positive initial score a_t using the sigmoid function g. Then the scores are normalized to obtain the final token-level weights w_t:

$$a_t = g(d_t), \quad w_t = \frac{T_w a_t}{\sum_{j=1}^{T_w} a_j} \tag{8}$$

where T_w denotes the number of valid tokens in y_w. This normalization preserves the relative proportions of the initial weights while ensuring a mean token weight of one: $\frac{1}{T_w} \sum_{t=1}^{T_w} w_t = 1$.

Consequently, tokens with greater relative alignment difficulty receive larger weights, while the total weight remains equal to the response length.

Having obtained the token-level weights w_t, we incorporate them into DPO's sequence-level preference optimization objective. For an autoregressive language model, the sequence-level log-probability ratio for a response y can be decomposed into a sum of token-level contributions:

$$\log \frac{\pi_\theta(y \mid P)}{\pi_{ref}(y \mid P)} = \sum_{t=1}^{T} \left[\log \pi_\theta(y_t \mid P, y_{<t}) - \log \pi_{ref}(y_t \mid P, y_{<t})\right] \tag{9}$$

Based on this decomposition, AsAPO adopts an asymmetric weighting scheme that weights only the token-level reward contributions of the preferred response while retaining the original reward formulation for the rejected response. This design reflects the distinct roles of the two responses: the preferred response provides the target content to be learned, and fine-grained weighting emphasizes critical tokens that remain insufficiently aligned; the rejected response retains its role as the negative comparison in standard DPO. Accordingly, we define the weighted preferred reward and the rejected reward as:

$$\widetilde{r}_w = \beta \sum_{t=1}^{T_w} w_t \log \frac{\pi_\theta(y_{w,t} \mid x, m, y_{w,<t})}{\pi_{\text{ref}}(y_{w,t} \mid x, m, y_{w,<t})}, r_l = \beta \sum_{t=1}^{T_l} \log \frac{\pi_\theta(y_{l,t} \mid x, m, y_{l,<t})}{\pi_{\text{ref}}(y_{l,t} \mid x, m, y_{l,<t})} \tag{10}$$

Substituting these rewards into the Bradley–Terry preference probability formulation yields:

$$p_\theta^w(y_w \succ y_l \mid x, m) = \sigma\big(\widetilde{r}_w - r_l\big) \tag{11}$$

where σ denotes the sigmoid function and β controls the reward scale. Minimizing the negative log-likelihood of the observed preferences gives the token-weighted preference optimization objective:

$$\mathcal{L}_{\text{DPO-w}} = -\mathbb{E}_{(x,m,y_w,y_l)\sim\mathcal{D}}\Big[\log \sigma\big(\widetilde{r}_w - r_l\big)\Big]. \tag{12}$$

This objective allocates optimization weights within the preferred response according to current alignment needs, enabling token-level differentiation under response-level preference supervision.

3.2 Preference Margin Stabilization Constraint

Token-level preference weighting provides fine-grained supervision, but it can also alter the response-level preference margin even when the token weights are normalized. Let r_w and $\tilde{r}_w$ denote the preferred-response rewards before and after token-level weighting, respectively, and let r_l denote the rejected-response reward. The corresponding original and weighted preference margins are $M = r_w - r_l$ and $\widetilde{M} = \widetilde{r}_w - r_l$. Since the rejected-response reward remains unchanged, the margin shift introduced by token weighting can be written as:

$$\Delta M = \widetilde{r}_w - r_w = \sum_{t=1}^{T_w} (w_t - 1) r_{w,t} = T_w \mathrm{Cov}(w_t, r_{w,t}). \tag{13}$$

Where $\mathrm{Cov}(w, r_w) = \frac{1}{T_w}\sum_{t=1}^{T_w}(w_t - 1)\big(r_{w,t} - \bar{r}_w\big)$ denotes the covariance between the token weights and their corresponding reward contributions.

The preceding analysis shows that, the preference-margin shift induced by weighting is determined by the covariance between token weights and their original token-level rewards. Since AsAPO assigns weights based on the model's current alignment state, this covariance is not guaranteed to vanish. Since the margin determines the effective optimization strength of each preference pair in DPO, excessive shifts may over-amplify or suppress supervision and distort the original preference-learning dynamics (Section 4.3). To mitigate this effect, we introduce a preference-margin stabilization constraint. Rather than forcing the weighted margin to exactly match the original one, which would weaken the intended effect of fine-grained weighting, we allow moderate variations and penalize only excessive deviations. Specifically, the absolute mean token reward of the unweighted preferred response is used as a sample-adaptive tolerance threshold, preserving flexible token-level reweighting while stabilizing response-level preference optimization.

$$\mathcal{L}_{\text{stab}} = \left[\max\left(0, |\widetilde{r}_w - \mathrm{sg}(r_w)| - \frac{|\mathrm{sg}(r_w)|}{T_w}\right)\right]^2, \tag{14}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation. It ensures that the original margin serves only as a fixed reference, while the weighted margin is optimized to remain close to it.

3.3 Overall Optimization Objective

We define the complete AsAPO objective by combining alignment-state-guided token-level preference optimization with a preference-margin stabilization constraint:

$$\mathcal{L}_{\mathrm{AsAPO}} = \mathcal{L}_{\mathrm{DPO-w}} + \lambda \mathcal{L}_{\mathrm{stab}} \tag{15}$$

where λ is a hyperparameter controlling the strength of the constraint. Token-level weighting differentiates the contributions of individual tokens according to their current alignment needs, enabling more targeted learning of preferred responses. The margin stabilization constraint curbs excessive preference-margin shifts induced by weighting. Together, these components promote more effective alignment with hallucination-free responses that reflect human preferences, thereby mitigating hallucinations in MLLMs.

4 Experiments

4.1 Experimental Settings

Models and Training Data. We conduct experiments on three widely used MLLMs: LLaVA-7B (Liu et al., 2024), LLaVA-13B, and Qwen3-VL-4B (Team, 2025). Following the data construction strategy of SymMPO (Liu et al., 2025), we randomly sample approximately 7.5K image–question pairs in total from VQA v2 (Goyal et al., 2017), MS COCO (Lin et al., 2014), and TextVQA (Singh et al., 2019), and use them as prompts to construct preference pairs.

Benchmarks and Baselines. We evaluate AsAPO on multiple public hallucination benchmarks. AMBER-G (Wang et al., 2023) comprises 1,004 images with detailed object annotations and assesses hallucinations across multiple dimensions; we report CHAIR, hallucination rate (Hal), and cognitive alignment (Cog). MMHal-Bench (Sun et al., 2024b) contains 96 images spanning 12 categories. Following its official evaluation protocol, we use GPT-4 to score responses on a scale from 0 to 6, with scores below 3 indicating hallucinations. Object HalBench (Rohrbach et al., 2018) is a widely used benchmark for object hallucination evaluation, on which we evaluate 300 samples using sentence-level (CHAIRs) and instance-level (CHAIRi) metrics. AMBER-D (Wang et al., 2023) provides approximately 14K discriminative evaluation samples, which we use to assess hallucinations through questions about object existence.

Baselines. We compare AsAPO with existing preference optimization methods, including POVID (Zhou et al., 2024), RLHF-V (Yu et al., 2024), HALVA (Sarkar et al., 2025), mDPO (Wang et al., 2024), and SymMPO (Liu et al., 2025), as well as fine-grained approaches such as CAT-PO (Zheng et al., 2026), TPO (Gu et al., 2025), and UE-DPO (Zhang et al., 2026a). To isolate the contribution of AsAPO, we additionally train a standard DPO baseline using the same training data, shared hyperparameters, and experimental setup. This controlled comparison allows us to assess the improvements in hallucination mitigation attributable to AsAPO.

Implementation Details. We fine-tune both LLaVA and Qwen3-VL models using LoRA (Hu et al., 2021) for 3 epochs with a learning rate of 1×10^{-5}. We use gradient accumulation to achieve an effective batch size of 64, set the LoRA rank to 128, and use $\beta = 0.1$. The margin constraint coefficient λ is set to 1×10^{-4} for LLaVA and 1×10^{-5} for Qwen3-VL. All experiments are conducted on four NVIDIA A100 GPUs.

4.2 Main Results

Table 1 compares AsAPO with existing methods across multiple hallucination benchmarks on the LLaVA series and Qwen3-VL. Several observations can be made from the results: (1) AsAPO achieves the best or highly competitive performance on most metrics, with particularly clear gains on open-ended generation tasks. For example, on AMBER-G, AsAPO improves CHAIR, Hal, and Cog over LLaVA by 4.8, 22.4, and 3.7 points, respectively. On MMHal-Bench, which relies on GPT-assisted evaluation, AsAPO also achieves a higher overall score and a lower hallucination

Table 1: Main experimental results. The best and second-best results under the same experimental setting (gray region) are highlighted in boldface and underlined, respectively. The results are either taken from prior work or evaluated using the officially released model weights(‡).

Method	Datasize	Object-Hal		MMHal-Bench		AMBER-G			AMBER-D	
		CHAIRs↓	CHAIRi↓	Score↑	HalRate↓	CHAIR↓	HalRate↓	Cog.↓	Acc.↑	F1↑
LLaVA-v1.5-7B (Liu et al., 2024)	✗	55.67	15.96	2.01	0.61	7.7	34.7	4.2	71.7	74.3
+ LLaVA-RLHF (Sun et al., 2024a)	122k	58.00	15.61	1.88	0.71	9.7	46.6	5.3	–	–
+ HALVA (Sarkar et al., 2025)	21.5k	41.40	11.70	2.25	0.54	6.6	32.2	3.4	–	–
+ POVID (Zhou et al., 2024)	17k	50.67	15.28	2.08	0.60	7.4	34.3	3.9	71.9	74.7
+ mDPO (Wang et al., 2024)	10k	35.70	9.80	2.39	0.54	4.4	24.5	2.4	–	–
+ CAT-PO (Zheng et al., 2026)	5.7k	–	–	2.74	0.42	4.8	23.7	2.1	78.0	85.3
+ TPO (Gu et al., 2025)	5.7k	–	–	2.47	0.51	–	–	–	79.3	85.0
+ SymMPO‡ (Liu et al., 2025)	21.4k	27.00	7.34	2.85	0.48	5.7	28.1	3.5	81.5	87.1
+ UE-DPO (Zhang et al., 2026a)	5.7k	13.72	6.69	2.82	0.48	2.9	17.4	1.0	79.1	85.7
+ DPO (Rafailov et al., 2023)	7.5k	24.33	7.96	2.11	0.68	6.6	25.5	1.2	77.4	**85.0**
+ AsAPO(Ours)	7.5k	**13.00**	**4.37**	**2.82**	**0.49**	**2.9**	**12.3**	**0.5**	**79.9**	84.4
LLaVA-v1.5-13B Liu et al. (2024)	✗	51.00	13.71	2.48	0.52	6.8	31.8	3.3	71.3	73.1
+ LLaVA-RLHF (Sun et al., 2024a)	122k	44.67	11.83	2.27	0.64	7.7	38.6	4.0	–	–
+ HALVA (Sarkar et al., 2025)	22.5k	45.40	12.80	2.58	0.45	6.4	30.4	3.2	–	–
+ RLHF-V(HD) (Yu et al., 2024)	1.4k	–	–	2.81	0.49	6.3	25.1	2.1	–	–
+ CAT-PO (Zheng et al., 2026)	5.7k	–	–	2.84	0.42	4.3	22.0	1.6	82.9	88.0
+ TPO (Gu et al., 2025)	5.7k	–	–	2.72	0.46	–	–	–	83.9	88.0
+ SymMPO‡ (Liu et al., 2025)	21.4k	29.00	7.55	2.52	0.54	4.7	27.2	2.4	84.0	88.6
+ UE-DPO (Zhang et al., 2026a)	5.7k	13.70	6.57	2.76	0.45	2.2	14.4	0.9	83.5	88.2
+ DPO (Rafailov et al., 2023)	7.5k	33.00	9.29	2.52	0.55	4.8	24.7	2.1	80.9	87.0
+ AsAPO(Ours)	7.5k	**12.00**	**4.81**	**2.78**	**0.49**	**3.7**	**14.0**	**0.7**	**84.0**	**88.2**
Qwen3-VL-4B (Team, 2025)	✗	62.33	10.8	3.83	0.39	7.8	60.6	3.7	89.9	92.0
+ DPO (Rafailov et al., 2023)	7.5k	41.67	8.78	3.83	0.32	**3.9**	29.1	2.0	88.6	91.9
+ AsAPO(Ours)	7.5k	**30.33**	**6.73**	**4.25**	**0.26**	4.0	**23.0**	**1.0**	**90.0**	**92.5**

rate, indicating improved response quality alongside reduced hallucination. (2) The results further demonstrate the advantage of fine-grained preference supervision. Fine-grained DPO methods generally outperform response-level preference optimization, as illustrated by UE-DPO's gains over SymMPO across multiple metrics on the LLaVA series. More importantly, the gray-shaded rows compare AsAPO with standard DPO under identical training data and settings, where AsAPO consistently performs better, confirming the effectiveness of token-level fine-grained supervision. (3) Compared with existing fine-grained DPO methods such as UE-DPO and CAT-PO, AsAPO still maintains clear advantages on most metrics. In particular, AsAPO achieves substantial improvements on AMBER-G and Object-Hal for both LLaVA-7B and LLaVA-13B, while remaining highly competitive on MMHal-Bench. These results support the core motivation of AsAPO: compared with determining token importance solely based on visual relevance, dynamically adjusting token-level supervision according to the model's current alignment state can provide more targeted fine-grained optimization signals. It is worth noting that the advantage of AsAPO is relatively limited on the discriminative AMBER-D benchmark. A possible explanation is that such tasks typically require only very short responses such as "yes" or "no", where the main objective is closer to calibrating a binary decision boundary rather than correcting localized and sparse hallucinations in open-ended responses. Consequently, there is less room for token-level differential weighting to take effect. Nevertheless, AsAPO still maintains an advantage over standard DPO.

Overall, AsAPO achieves the best or highly competitive results across the evaluated benchmarks, with particularly consistent gains on open-ended generation tasks, demonstrating the effectiveness of alignment-state-aware fine-grained preference optimization for hallucination mitigation.

4.3 Ablation Study and Analysis

Impact of $\mathcal{L}_{stab}$ and $\mathcal{L}_{DPO-w}$. We conduct ablation experiments with two variants, w/o $\mathcal{L}_{stab}$ and w/o $\mathcal{L}_{DPO-w}$, which remove the margin constraint and token-level weighting, respectively. As shown in Table 2, removing the margin constraint consistently degrades performance on both AMBER-G and Object-Hal; for example, CHAIR, Hal, and Cog on AMBER-G decrease by 0.5, 3.5, and 0.6 points, respectively. This suggests that token weighting alone may introduce excessive preference-margin shifts that weaken fine-grained optimization. When token-level weighting is further removed, the margin constraint also becomes ineffective, causing the method to degenerate into standard DPO and resulting in a more pronounced performance drop, particularly on AMBER-G CHAIR and Object-Hal CHAIRs. These results confirm that the two components are complementary: alignment-state-aware weighting provides

Table 2: Ablation Study Results in LLaVA1.5-7B.

Method	Object-Hal		AMBER-G		
	CHAIRs↓	CHAIRi↓	CHAIR↓	HalRate↓	Cog.↓
AsAPO	13.00	4.37	2.9	12.3	0.5
w/o $\mathcal{L}_{stab}$	24.67	6.72	3.4	13.8	1.1
w/o $\mathcal{L}_{DPO-w}$	24.33	7.96	6.6	25.5	1.2
w/o $sg(\cdot)$	21.67	5.81	4.2	18.6	1.2

more targeted token-level supervision, while the margin constraint regularizes the induced reward and margin shifts, together enabling more effective and stable hallucination mitigation.

Impact of $sg(\cdot)$. In the margin constraint, we apply stop-gradient to the original reward so that it serves as a reference. As shown in Table 2, removing stop-gradient (w/o sg) consistently degrades performance on both AMBER-G and Object-Hal. The margin curves in Fig. 2 further show that both settings exhibit similar growth at the early stage of training, while the margin of w/o sg saturates prematurely and remains substantially lower than that of the full AsAPO. Without stop-gradient, the constraint updates both the weighted and original rewards, causing the reference reward to move with the optimization target and weakening its role as a scale reference. This enables AsAPO to stabilize reward scaling while maintaining sufficient preference-margin growth, as supported by both the performance gains and the observed training dynamics.

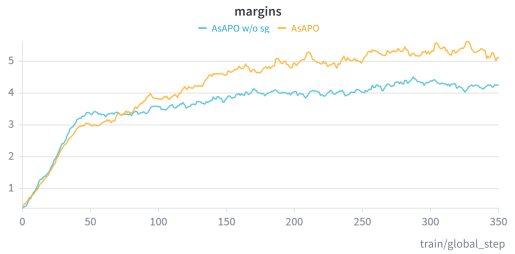
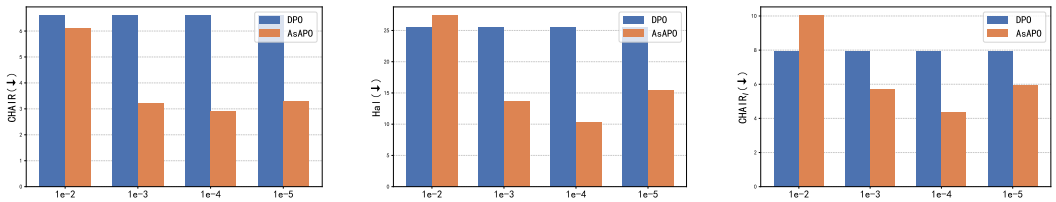


Figure 2: Preference Margin Dynamics of AsAPO and AsAPO w/o sg.



(a) AMBER-G: CHAIR (b) AMBER-G: Hal (c) Object-Hal: CHAIR_I

Figure 3: Performance under Different λ Values on AMBER-G and Object-Hal, (↓) lower is better.

HyperParameter Sensitive Analysis. We conduct a sensitivity analysis of the hyperparameter λ in AsAPO, with the results shown in Fig. 3. When $\lambda = 1e{-}2$, the overly strong margin constraint forces the optimization to place excessive emphasis on limiting reward shifts, thereby weakening the main preference optimization objective. As a result, the performance even falls below that of standard DPO. In contrast, when λ is reduced to $1e{-}3$ or smaller, AsAPO consistently outperforms DPO on both AMBER-G and Object-Hal, indicating a more favorable balance between margin regularization and preference learning. Additional results are provided in the Appendix.

4.4 Visualization of log-probabilities: LLaVA, DPO and AsAPO

To examine whether visual relevance reflects a token's actual learning difficulty, and to illustrate AsAPO's resulting optimization behavior, we visualize the token-level log-probabilities of the preferred response under the original LLaVA, standard DPO, and AsAPO. As shown in Figure 4, tokens such as "truck" and "road" (orange boxes) carry high visual relevance, yet already receive near-zero log-probabilities under the original LLaVA, indicating that high visual relevance does not necessarily entail high learning difficulty. In contrast, tokens such as "right" and "wheels" (purple boxes)



Figure 4: Token-Level Log-Probability Visualization of LLaVA, DPO, and AsAPO.

remain substantially harder to predict. Compared with DPO, AsAPO achieves markedly larger probability gains on these challenging tokens, while making only limited adjustments to tokens such as ”truck” and ”road” that are already well aligned. This indicates that AsAPO does not uniformly reinforce the entire preferred response, but instead allocates stronger optimization signals to tokens that remain under-aligned under the current alignment state. As a result, AsAPO's probability curve exhibits fewer extreme low-probability drops and lower variance across tokens, yielding the highest average log-probability among the three methods (AsAPO: -1.722, LLaVA: -1.804, DPO: -1.835).

5 Related Work

DPO has become an effective paradigm for mitigating hallucinations in MLLMs by aligning model outputs with visual evidence through preference learning. Early methods mainly operate at the response level. HA-DPO (Zhao et al., 2023) constructs hallucination-free and hallucinated preference pairs, while mDPO (Wang et al., 2024), AMP (Zhang et al., 2024), OPA-DPO (Yang et al., 2025), and SymMPO (Liu et al., 2025) improve DPO from the perspectives of visual conditioning, multi-level preferences, policy-data consistency, and objective design. However, response-level supervision overlooking token's different contributions to hallucination mitigation. Recent studies therefore explore token-level fine-grained preference optimization. TPO (Gu et al., 2025), and Cat-PO (Zheng et al., 2026) assign token-level optimization signals based on visual relevance or sensitivity, while UE-DPO (Zhang et al., 2026a) further leverages epistemic uncertainty to identify insufficiently grounded tokens. Despite their effectiveness, existing methods primarily characterize token importance through visual relevance. We instead argue that token importance should reflect the model's current alignment state. From this perspective, AsAPO offers a complementary alignment-state-aware view of fine-grained preference optimization. (A more comprehensive review is provided in the Appendix.)

6 Conclusion

In this work, we propose AsAPO, an alignment-state-aware preference optimization method for mitigating hallucinations in multimodal large language models. AsAPO quantifies token importance according to the model's current alignment state, enabling more targeted fine-grained supervision during preference optimization. Its key contributions include an alignment-state-aware token importance estimation mechanism and a margin constraint that stabilizes the preference optimization process. We evaluate AsAPO on three models with different scales and architectures across four widely used hallucination benchmarks. Extensive experiments demonstrate that AsAPO consistently achieves strong hallucination mitigation performance, while ablation studies and further analyses verify the effectiveness and complementary roles of its key components. Overall, our results highlight the importance of alignment-state-aware fine-grained supervision for more effective preference alignment in multimodal large language models.

AI USE STATEMENT

In this work, we used generative AI tools only for auxiliary writing assistance, including translation, language polishing, and grammar checking. We did not use generative AI tools to develop the research motivation, formulate the core ideas, design the proposed method, conduct experiments, analyze results, or derive scientific conclusions. All AI-assisted content was carefully reviewed and revised by the authors to ensure technical correctness, clarity, and consistency with the intended meaning. The core contributions, methodology, experimental design, and scientific claims were entirely developed by the authors. We take full responsibility for the final content of this work, including any text produced with the aid of generative AI.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan 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. *CoRR*, abs/2502.13923, 2025. doi: 10.48550/ARXIV.2502.13923. URL https://doi.org/10.48550/arXiv.2502.13923.

Henghui Du, Chang Zhou, Xi Chen, and Di Hu. APPO: attention-guided perception policy optimization for video reasoning. *CoRR*, abs/2602.23823, 2026. doi: 10.48550/ARXIV.2602.23823. URL https://doi.org/10.48550/arXiv.2602.23823.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Jihao Gu, Yingyao Wang, Meng Cao, Pi Bu, Jun Song, Yancheng He, Shilong Li, and Bo Zheng. Token preference optimization with self-calibrated visual-anchored rewards for hallucination mitigation, 2025. URL https://arxiv.org/abs/2412.14487.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13872–13882. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01316. URL https://doi.org/10.1109/CVPR52733.2024.01316.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26286–26296. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02484. URL https://doi.org/10.1109/CVPR52733.2024.02484.

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 Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/a1718f361df32ff3a1fc224f8673c556-Abstract-Conference.html.

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. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4035–4045, 2018.

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

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the association for computational linguistics: ACL 2024*, pp. 13088–13110, 2024a.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the association for computational linguistics: ACL 2024*, pp. 13088–13110, 2024b.

Shuhan Tan, Kashyap Chitta, Yuxiao Chen, Ran Tian, Yurong You, Yan Wang, Wenjie Luo, Yulong Cao, Philipp Krahenbuhl, Marco Pavone, et al. Latent chain-of-thought world modeling for end-to-end driving. *arXiv preprint arXiv:2512.10226*, 2025.

Qwen Team. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025. doi: 10.48550/ARXIV.2511.21631. URL https://doi.org/10.48550/arXiv.2511.21631.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088, 2024.

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

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13258–13273, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Ning Yang, Hai Lin, Yibo Liu, Baoliang Tian, Guoqing Liu, and Haijun Zhang. Token-importance guided direct preference optimization. In *International Conference on Learning Representations*, volume 2026, pp. 13888–13910, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10610–10620. IEEE, 2025.

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, 2024.

Huatian Zhang, Zhendong Mao, Lei Zhang, and Yongdong Zhang. Uncertainty-aware exploratory direct preference optimization for multimodal large language models. *arXiv preprint arXiv:2605.04874*, 2026a.

Jialiang Zhang, Junlong Tong, Junyan Lin, Hao Wu, Yirong Sun, Yunpu Ma, and Xiaoyu Shen. Think-as-you-see: Streaming chain-of-thought reasoning for large vision-language models. *arXiv preprint arXiv:2603.02872*, 2026b.

Mengxi Zhang, Wenhao Wu, Yu Lu, Yuxin Song, Kang Rong, Huanjin Yao, Jianbo Zhao, Fanglong Liu, Haocheng Feng, Jingdong Wang, et al. Automated multi-level preference for mllms. *Advances in Neural Information Processing Systems*, 37:26171–26194, 2024.

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

Zhixiao Zheng, Zheren Fu, Zhiyuan Yao, Dongming Zhang, and Zhendong Mao. Cat-po: Cross-modal adaptive token-rewards for preference optimization in truthful multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 113687–113702, 2026.

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.

A Appendix

A.1 Related Work

Despite the remarkable progress of multimodal large language models (MLLMs) in visual understanding and generation tasks (Zhang et al., 2026b; Du et al., 2026; Tan et al., 2025), hallucinations remain a major challenge to their reliability in real-world applications (Leng et al., 2024; Liu et al., 2025). In recent years, Direct Preference Optimization (DPO) has emerged as an important paradigm for mitigating hallucinations in MLLMs by enlarging the reward gap between preferred and rejected responses, thereby encouraging model outputs to better align with visual evidence.

Early studies mainly relied on response-level preference data to align MLLMs. HA-DPO (Zhao et al., 2023) performs direct preference learning by constructing hallucination-free and hallucinated response pairs. mDPO (Wang et al., 2024) observes that MLLMs may underutilize visual information and therefore further incorporates the image condition into DPO, together with a reward anchor to strengthen preference optimization. AMP (Zhang et al., 2024) extends conventional binary preferences to multi-level preferences, enabling the model to capture finer distinctions among responses of different quality. OPA-DPO (Yang et al., 2025) focuses on the consistency between preference data and the policy model, first aligning the model with the data through SFT and then performing preference optimization to alleviate the adverse effects of off-policy training. In addition, SymMPO (Liu et al., 2025) improves upon mDPO from the perspective of objective consistency, aiming to enhance visual understanding and reduce hallucinations. Although these methods substantially improve the effectiveness of DPO, their supervision signals are still mainly applied at the response level, making it difficult to distinguish the different contributions of individual tokens to hallucination mitigation.

To address this limitation, several studies have explored token-level fine-grained DPO. V-DPO (Xie et al., 2024) introduces a vision-guided factor and image-contrast preferences to strengthen the dependence of preference learning on visual context. TPO (Gu et al., 2025) further constructs token-level visual-anchored rewards by comparing token distributions conditioned on the original and

Table 3: Comparison of different λ on the performance of AsAPO

AsAPO	Object-Hal		AMBER-G		
	CHAIRs↓	CHAIRi↓	CHAIR↓	HalRate↓	Cog.↓
λ=1e-2	36.33	10.03	6.1	27.4	2.9
λ=1e-3	20.67	5.70	3.2	13.7	1.0
λ=1e-4	13.00	4.37	2.9	12.3	0.5
λ=1e-5	21.00	5.93	3.3	15.4	0.7

perturbed images, thereby placing greater emphasis on visually relevant tokens. Cat-PO (Zheng et al., 2026) estimates token-level visual relevance from global, local, and semantic perspectives and uses it to construct adaptive token rewards for finer-grained preference supervision. More recently, UE-DPO (Zhang et al., 2026a) points out that visual sensitivity estimated by the model itself may suffer from self-referential bias, and therefore leverages token-level epistemic uncertainty to identify tokens with insufficient visual grounding and adaptively assign optimization strength.

These methods further improve hallucination mitigation over response-level DPO, but token importance is still typically characterized from perspectives such as visual relevance or visual sensitivity. We argue that visual importance and alignment importance are fundamentally different: the former reflects how strongly a token is associated with visual information, whereas the latter characterizes how much optimization the token still requires under the model's current alignment state. Based on this distinction, AsAPO redefines token importance from the perspective of the model's current alignment state, providing a new view of fine-grained preference optimization beyond visual-relevance-based modeling.

A.2 Pseudo-code of AsAPO

Algorithm 1 Alignment-State-Aware Preference Optimization(AsAPO)

Require: Dataset $\mathcal{D}$, policy model π_θ, reference model π_{ref}
Ensure: Aligned policy π_θ

for each epoch **do**
  Sample batch $(x, y_w, y_l) \sim \mathcal{D}$.
  Compute token self-information and boundary: $I_{w,t}, b$(Eq. 4, 5).
  Compute state-aware token scores: $a_t = g(I_t^w, b)$, for $t = 1, \ldots, T_w$(Eq. 8).
  Normalize token weights: $w_t = \frac{T_w a_t}{\sum_{j=1}^{T_w} a_j}$(Eq. 8).
  Compute weighted preferred reward and original rejected reward: $\widetilde{r}_w, r_l$(Eq. 10).
  Compute asymmetric weighted DPO loss:

$$\mathcal{L}_{\text{DPO-w}} = -\mathbb{E}_{(x,m,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\big(\widetilde{r}_w - r_l\big)\Big]$$

  Compute preference margin stabilization constraint loss:

$$\mathcal{L}_{\text{stab}} = \left[\max\left(0, |\widetilde{r}_w - \text{sg}(r_w)| - \tfrac{|\text{sg}(r_w)|}{T_w}\right)\right]^2$$

  Aggregate losses: $\mathcal{L}_{\text{AsAPO}} = \mathcal{L}_{\text{DPO}-\text{w}} + \lambda\,\mathcal{L}_{\text{stab}}$(Eq. 15)
  Update θ using $\nabla_\theta \mathcal{L}_{\text{AsAPO}}$
end for
return π_θ

A.3 HyperParameter Sensitive Analysis

We further analyze the effect of the margin constraint weight λ on AsAPO by evaluating $\lambda \in \{1e-2, 1e-3, 1e-4, 1e-5\}$, with the results reported in Table 3. When $\lambda = 1e-2$, the overly strong margin constraint causes the optimization to place excessive emphasis on limiting the reward shift induced by weighting, thereby weakening the learning signal from the main preference objective and leading to a clear performance drop. In contrast, when λ ranges from $1e-3$ to $1e-5$, the

margin constraint can effectively suppress reward-scale shifts without excessively interfering with preference optimization, resulting in more stable and improved performance. In particular, $\lambda = 1e-4$ achieves the best overall results, suggesting that a moderate constraint strength provides a better balance between fine-grained preference learning and reward-scale stability. It is worth noting that we only conduct a coarse-grained hyperparameter search; therefore, a finer search around $1e-4$ may further improve performance.

A.4 CASE STUDY

To qualitatively illustrate the effectiveness of AsAPO in mitigating MLLM hallucinations, we compare responses from the original LLaVA, the DPO-finetuned model, and AsAPO under identical input conditions, as shown in Figures 5 and 6.

In Figure 5, the original LLaVA describes "a rabbit eating a carrot" in the image, whereas no carrot is actually present. This indicates that the model's output relies more heavily on language priors (the strong co-occurrence association between rabbits and carrots) than on visual evidence from the image. Although the DPO fine-tuned model suppresses this particular hallucination to some extent and produces a richer, more detailed description, it simultaneously introduces new hallucinated content. This suggests that while DPO improves the richness of the generated description, it fails to effectively constrain the consistency between the generated content and visual facts.

Figure 6 further examines model robustness through two challenging cases involving a false-premise question and fine-grained visual localization. In the first case—"How many uncut fruits are in the image?" (the image is actually an anthropomorphic portrait painting of a dog, containing no fruit at all)—the original LLaVA incorrectly answers "there are two uncut fruits in the image", exhibiting a typical object hallucination, while both DPO and AsAPO correctly determine that no fruit is present. However, in the second, more challenging kiwifruit case, when asked "Is the kiwifruit in the front fresh or dried?", the DPO model makes an instruction-comprehension error: rather than answering about the "front" kiwifruit explicitly referenced in the question, it describes "the middle kiwifruit being partially peeled". This response reveals a failure to resolve the question's spatial reference.The original LLaVA, while correctly identifying the target (the front kiwifruit), makes a factual error in its judgment ("judged as fresh due to its skin"). In contrast, AsAPO not only correctly locates the front kiwifruit referenced in the question but also accurately describes its dried characteristics—"yellowish-brown in color with a slightly wrinkled surface"—thereby achieving both instruction-following accuracy and correct visual judgment simultaneously.

Taken together, these cases show that although DPO partially suppresses object-level hallucination present in the original LLaVA (e.g., the fruit misjudgment in Figure 6), its improvements remain largely confined to surface-level factual correction. When a task simultaneously demands rich descriptive detail (Figure 5) or precise spatial referential understanding (Figure 6), the DPO model exhibits, respectively, exacerbated detail-level hallucination and instruction-comprehension drift. This indicates that relying solely on a preference-optimization objective to select hallucinated/non-hallucinated sample pairs is insufficient for the model to genuinely establish dual alignment—between generated content and visual evidence, and between generated content and question semantics. AsAPO, by contrast, maintains consistent content accuracy and instruction-following capability across all cases, demonstrating that its proposed alignment-state-aware optimization mechanism can more finely identify and correct weakly aligned segments of the generation process—thereby suppressing hallucination without sacrificing fine-grained comprehension ability.



What are the animals in the painting and what are they doing?

[LLaVA] The painting features a bear, a cat, and a rabbit sitting around a table, eating food. They are enjoying a meal together, with the bear eating a sandwich, the cat drinking from a cup, and the rabbit eating a carrot.

[LLaVA+DPO] The image features a watercolor-style painting of three animals seated around a small, slightly raised, wooden surface with a white plate containing food items such as berries or fruit slices. The animals are not in full-color, but rather depicted with subtle shading and outlines. The leftmost animal is a dark-colored bear with a visible snout and claws, while the rightmost animal is a light-colored bunny-like creature with erect ears and a pinkish-orange nose. The central animal is a gray-toned cat with green eyes, white whiskers, and a slightly curved body. Each animal has a small, partially filled cup or bowl placed near their respective paws or front legs, suggesting they are not actively consuming the food at the moment.

[LLaVA+AsAPO] The painting features a bear, a cat, and a rabbit sitting around a small wooden table with a white plate containing food items. The bear is on the left side of the table, the cat is in the center, and the rabbit is on the right side. They appear to be enjoying a meal together in a peaceful, natural setting.

Figure 5: Qualitative comparison of responses generated by different methods using greedy decoding, with hallucinated content highlighted in red.



How many uncut fruits are in the image?

[LLaVA] There are two uncut fruits in the image.

[LLaVA+DPO] There are no uncut fruits or other objects in the image. The artwork depicts a canine-like figure with elongated features, wearing a dark robe with intricate folds, and with a subtle landscape or architectural elements in the background.

[LLaVA+AsAPO] There are no uncut fruits in this image. It features a painting or artwork of a dog with a hooded or draped appearance, possibly resembling a hooded figure from a painting or fresco.



Is the kiwi in the front fresh or dried?

[LLaVA] The kiwi in the front is fresh, as it is still attached to its skin.

[LLaVA+DPO] The kiwi in the center is partially peeled, revealing a greenish-yellow color with a slightly rough texture. The other two kiwi-like objects are not peeled and appear to be unripe or underripe, with a slightly darker green color.

[LLaVA+AsAPO] The kiwi in the front is dried, with a yellowish-brown color and a slightly wrinkled surface.

Figure 6: Qualitative comparison of responses generated by different methods using greedy decoding, with hallucinated content highlighted in red.