LEARNING WHAT TO ALIGN: AUTOMATIC TARGET DISCOVERY FOR HALLUCINATION MITIGATION IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models often suffer from visual hallucination, where generated responses are inconsistent with the input image. Existing preference-based alignment methods mainly rely on sequence-level supervision, making it difficult to localize visual evidence and response tokens responsible for hallucinated predictions. Recent target-level approaches improve alignment precision but usually depend on manually annotated regions or token spans. In this paper, we propose **L**earning **W**hat to **A**lign with **T**arget **D**iscovery (**LWA-TD**), which leverages minimal visual contrast to identify visual evidence and response tokens responsible for grounded semantic differences. Given counterfactual image-response pairs, LWA-TD identifies question-relevant visual changes and estimates token sensitivity to counterfactual visual changes. The discovered targets are further incorporated into preference optimization and grounding regularization to improve visual-text consistency. Experiments on multiple hallucination benchmarks demonstrate that LWA-TD effectively reduces hallucination across different model scales and backbones, while maintaining strong general multimodal reasoning performance.

1 INTRODUCTION

Multimodal large language models (MLLMs) have achieved remarkable progress in vision-language understanding by integrating large language models with visual encoders and instruction tuning Touvron et al. (2023); Achiam et al. (2023); Li et al. (2023a); Liu et al. (2024b). They demonstrate strong performance on visual question answering, image captioning, multimodal reasoning, and instruction-following tasks Goyal et al. (2017); Hudson & Manning (2019); Singh et al. (2019); Yu et al. (2023c). Despite these advances, MLLMs still suffer from visual hallucinations, generating objects, attributes, counts, or relations inconsistent with the input image Li et al. (2023b); Zhou et al. (2023); Bai et al. (2024); Liu et al. (2024a), indicating insufficient grounding to visual evidence. Hallucinations are often associated with localized visual evidence, such as small objects, local attributes, or spatial relations. However, most existing alignment methods rely on sequence-level supervision over the entire image and response. As illustrated in Figure 1(a), such global supervision may weaken the learning signal for hallucination-related visual regions and response tokens.

To mitigate hallucinations in MLLMs, recent studies have explored preference-based alignment and contrastive learning. Preference-based methods extend RLHF Ouyang et al. (2022); Bai et al. (2022); Christiano et al. (2017) and DPO Rafailov et al. (2023) to multimodal settings to improve image-conditioned alignment Jiang et al. (2025); Wang et al. (2024b); Xie et al. (2024). However, they mainly rely on sequence-level supervision, making it difficult to localize hallucination-related visual evidence. To improve supervision precision, recent methods introduce token-level or region-level alignment through token-level rewards Zhou et al. (2024) or manually annotated visual regions and response spans Yoon et al. (2025). As illustrated in Figure 1(b), these methods improve alignment but rely heavily on manual annotations, limiting scalability. Contrastive learning methods leverage paired inputs with minimal visual differences to strengthen visual grounding Leng et al. (2024); Wu et al. (2025). However, they still do not explicitly identify the visual regions and response tokens associated with hallucinations.

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.

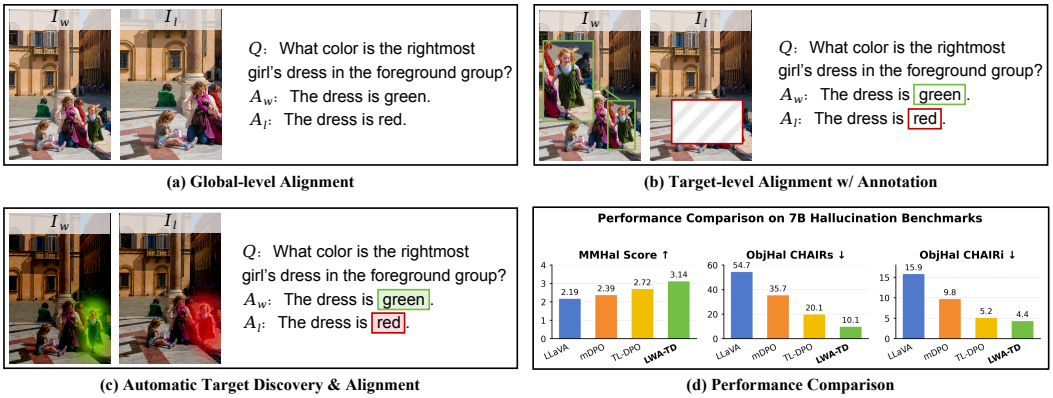


Figure 1: Comparison of alignment strategies for hallucination mitigation in MLLMs. (a) Sequence-level alignment. (b) Target-level alignment with manual annotations. (c) LWA-TD automatically discovers alignment targets from minimal visual differences. (d) Quantitative comparison on hallucination benchmarks.

The above observations suggest that effective hallucination mitigation requires identifying the visual evidence underlying localized response differences. Existing methods either apply sequence-level supervision or rely on manually annotated targets, making it difficult to precisely localize hallucination-related regions and response tokens. Recent contrastive learning studies show that minimal visual differences between paired images naturally induce localized response changes Leng et al. (2024); Wu et al. (2025), providing useful cues for automatic target discovery. However, these cues remain underexplored for alignment. Motivated by this observation, we aim to identify the visual regions and response tokens responsible for response differences and use them to guide preference optimization.

To address these limitations, we propose **L**earning **W**hat to **A**lign with **T**arget **D**iscovery (**LWA-TD**), a framework that automatically discovers visual regions and response tokens from paired images with minimal visual differences, without manual annotations. As illustrated in Figure 1(c), LWA-TD identifies visual regions and response tokens from image differences and incorporates them into target-aware preference optimization to improve grounding between generated responses and visual evidence. Unlike previous target-level methods Yoon et al. (2025), LWA-TD does not require manually annotated regions or tokens. It also differs from contrastive alignment methods Wu et al. (2025) by explicitly discovering alignment targets rather than relying solely on visual contrast. As shown in Figure 1(d), LWA-TD consistently reduces hallucinations across multiple benchmarks.

In summary, our main contributions are:

- We formulate hallucination mitigation in MLLMs as an alignment target discovery problem, aiming to identify hallucination-related visual regions and response tokens from paired images.
- We propose LWA-TD, an annotation-free framework that automatically discovers alignment targets from paired images with minimal visual differences without manual annotations.
- We integrate the discovered targets into target-aware preference optimization and evidence consistency alignment, improving visual grounding and reducing hallucinations in MLLMs.

2 Related Work

2.1 Preference-based Alignment for MLLMs

Preference-based alignment has been widely adopted to improve the reliability and instruction-following ability of large language models, evolving from RLHF Christiano et al. (2017); Ouyang

et al. (2022); Bai et al. (2022) to more efficient preference optimization methods such as DPO Rafailov et al. (2023). Recent studies extend this paradigm to MLLMs to enhance visual grounding and mitigate hallucinations. Representative methods include mDPO, which introduces conditional preference optimization for multimodal inputs Wang et al. (2024b), V-DPO, which incorporates vision-guided preference signals Xie et al. (2024), and MFPO, which considers modality fairness during alignment Jiang et al. (2025). Despite their effectiveness, these methods mainly operate at the sequence level, making it difficult to identify the visual regions and response tokens responsible for hallucinations.

2.2 Alignment with Explicit Targets

To address the limitations of sequence-level supervision, recent studies introduce target-level alignment on specific response tokens or visual regions. TL-DPO focuses preference optimization on annotated hallucination targets Yoon et al. (2025), while T-REG provides token-level reward signals to guide generation Zhou et al. (2024). Other methods incorporate phrase-level or region-level supervision through localized annotations Sarkar et al. (2024); Favero et al. (2024); Xiao et al. (2025). Although these approaches improve alignment precision, they rely on manually annotated tokens or visual regions, which are costly to obtain and difficult to scale.

2.3 Contrastive Learning with Minimal Visual Differences

Recent contrastive approaches improve visual grounding by leveraging paired images with minimal visual differences. S-VCO adopts symmetric contrastive optimization over paired images Wu et al. (2025), while VCD and Octopus mitigate hallucination through contrastive decoding strategies Leng et al. (2024); Suo et al. (2025). More recently, BiDPO combines bidirectional preference optimization with contrastive image pairs to enhance multimodal alignment Liang et al. (2026). Despite their effectiveness, these methods do not explicitly identify the visual regions and response tokens associated with visual changes, limiting their ability to guide target-level alignment.

3 Problem Formulation

3.1 Sequence-level Preference Optimization

Given an image I and a question Q, a policy model π_θ generates a response A. Preference optimization aligns the model with image-conditioned preferences. Given a preferred response A^+ and a dispreferred response A^-, preference learning is commonly formulated using the Bradley–Terry model:

$$P(A^+ \succ A^- \mid I, Q) = \frac{e^{r(I,Q,A^+)}}{e^{r(I,Q,A^+)} + e^{r(I,Q,A^-)}}. \tag{1}$$

where $r(I, Q, A)$ denotes the reward. Following DPO Rafailov et al. (2023), the reward is parameterized by the likelihood ratio between the policy model and a reference model π_{ref}:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(I,Q,A^+,A^-)} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(A^+ \mid I, Q)}{\pi_{\text{ref}}(A^+ \mid I, Q)} - \beta \log \frac{\pi_\theta(A^- \mid I, Q)}{\pi_{\text{ref}}(A^- \mid I, Q)} \right) \right]. \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function and β is the temperature parameter. Standard DPO applies supervision at the sequence level without explicitly modeling hallucination-related visual regions or response tokens.

3.2 Paired Preference Data

Following BiDPO Liang et al. (2026), we adopt paired-image preference data for target discovery. Each training sample is (I_w, I_l, Q, A_w, A_l), where $I_l = \mathcal{T}(I_w; \delta)$, $\mathcal{T}(\cdot)$ denotes a transformation operator, and δ represents a small visual change. The bidirectional preference relation is defined as

$$P(A_w \succ A_l \mid I_w, Q) = \sigma\big(r(I_w, Q, A_w) - r(I_w, Q, A_l)\big), \tag{3}$$

$$P(A_l \succ A_w \mid I_l, Q) = \sigma\big(r(I_l, Q, A_l) - r(I_l, Q, A_w)\big), \tag{4}$$

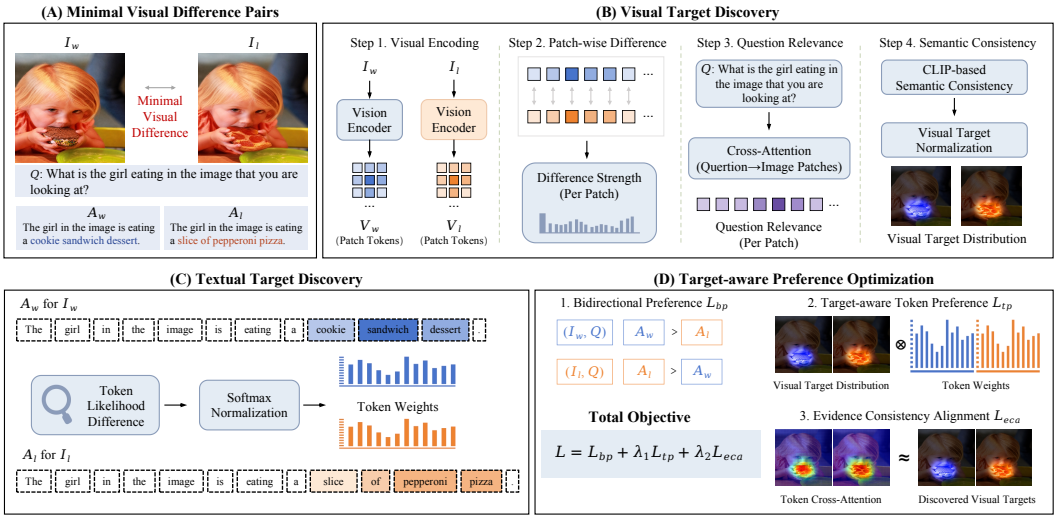


Figure 2: Overview of LWA-TD. The framework discovers visual and textual targets from paired images and incorporates them into target-aware preference optimization.

where $r(\cdot)$ denotes the reward and $\sigma(\cdot)$ is the sigmoid function. Small visual differences between paired images induce localized response differences, providing supervision for visual and textual target discovery.

3.3 Target Discovery

Given paired preference data, we formulate alignment as a target discovery problem, where visual and textual targets are discovered from paired images with minimal visual differences. Let $V_w = \{v_w^{(p)}\}_{p=1}^P$ and $V_l = \{v_l^{(p)}\}_{p=1}^P$ denote the patch-level visual representations. We define a visual target distribution $M = \{M_p\}$ over image patches and a textual target distribution $\omega = \{\omega_t\}$ over response tokens. The target discovery problem is formulated as

$$M = f_v(I_w, I_l, Q), \quad \omega = f_t(I_w, I_l, Q, A), \tag{5}$$

where f_v and f_t denote the visual and textual target functions, respectively.

4 Method

4.1 Overview

Given a preference tuple$\{I_w, I_l, Q, A_w, A_l\}$, where I_w and I_l are paired images, Q is the question, and A_w and A_l are the corresponding responses, LWA-TD aims to discover alignment targets, including visual targets and response tokens, from paired images with minimal visual differences. As illustrated in Figure 2, LWA-TD consists of three stages: visual target discovery, textual target discovery, and target-aware preference optimization. It first discovers question-relevant visual targets, then estimates token importance from paired-image likelihood differences, and finally performs bidirectional preference optimization, target-aware token preference optimization, and evidence consistency alignment.

4.2 Visual Target Discovery

Visual target discovery aims to identify question-relevant image regions associated with visual differences between paired images. Given paired images (I_w, I_l), we first extract patch-level visual representations:

$$V_w = \{v_w^{(p)}\}_{p=1}^P, \quad V_l = \{v_l^{(p)}\}_{p=1}^P, \tag{6}$$

where P is the number of visual patches. We then compute the patch-wise visual difference:

$$D_p = \|v_w^{(p)} - v_l^{(p)}\|_2, \tag{7}$$

which highlights regions affected by visual changes. To focus on question-relevant evidence, we extract question-to-image cross-attention from the MLLM decoder:

$$A_Q = \{A_Q^{(p)}\}_{p=1}^{P}, \tag{8}$$

where $A_Q^{(p)}$ denotes the average attention weight for patch p across decoder layers and attention heads. We further compute a CLIP-based semantic similarity between each visual patch and the question:

$$S_p = \frac{\langle v_p, q \rangle}{\|v_p\|\|q\|}, \tag{9}$$

where $v_p = (v_w^{(p)} + v_l^{(p)})/2$. The final visual target distribution is obtained as

$$M_p = \frac{D_p \cdot A_Q^{(p)} \cdot S_p}{\sum_{j=1}^{P} D_j \cdot A_Q^{(j)} \cdot S_j + \epsilon}. \tag{10}$$

Here, D_p, $A_Q^{(p)}$, and S_p encode visual differences, question relevance, and semantic similarity, respectively.

4.3 Textual Target Discovery

To identify response tokens associated with visual differences, we estimate token weights by comparing token likelihoods under paired image conditions. For response A_w, the token score is

$$s_t^w = \log \pi_\theta(a_t^w \mid a_{<t}^w, I_w, Q) - \log \pi_\theta(a_t^w \mid a_{<t}^w, I_l, Q), \tag{11}$$

where $a_{<t}^w$ denotes the preceding tokens and π_θ is the policy model. Scores for A_l are defined analogously:

$$s_t^l = \log \pi_\theta(a_t^l \mid a_{<t}^l, I_l, Q) - \log \pi_\theta(a_t^l \mid a_{<t}^l, I_w, Q). \tag{12}$$

The scores are normalized into token weights:

$$\omega_t^w = \frac{\exp(s_t^w/\tau)}{\sum_{j=1}^{T_w} \exp(s_j^w/\tau)}, \quad \omega_t^l = \frac{\exp(s_t^l/\tau)}{\sum_{j=1}^{T_l} \exp(s_j^l/\tau)}, \tag{13}$$

where τ controls the concentration of the distribution. The normalized weights are treated as stop-gradient coefficients.

4.4 Target-aware Preference Optimization

Given the discovered visual targets and token weights, we optimize the model with bidirectional preference optimization, target-aware token preference optimization, and evidence consistency alignment.

Bidirectional Preference. To enforce image-conditioned preference alignment across paired images, we optimize the preference objective in both directions. For I_w, the preference loss is

$$\mathcal{L}_{bp}^w = -\log \sigma\left(\beta_1 \log \frac{\pi_\theta(A_w \mid I_w, Q)}{\pi_{\text{ref}}(A_w \mid I_w, Q)} - \beta_1 \log \frac{\pi_\theta(A_l \mid I_w, Q)}{\pi_{\text{ref}}(A_l \mid I_w, Q)}\right). \tag{14}$$

where π_θ and π_{ref} denote the policy and reference models, $\sigma(\cdot)$ is the sigmoid function, and β_1 is the temperature parameter. The loss for image I_l is defined analogously:

$$\mathcal{L}_{bp}^l = -\log \sigma\left(\beta_1 \log \frac{\pi_\theta(A_l \mid I_l, Q)}{\pi_{\text{ref}}(A_l \mid I_l, Q)} - \beta_1 \log \frac{\pi_\theta(A_w \mid I_l, Q)}{\pi_{\text{ref}}(A_w \mid I_l, Q)}\right). \tag{15}$$

The final objective is

$$\mathcal{L}_{bp} = \mathcal{L}_{bp}^w + \mathcal{L}_{bp}^l. \tag{16}$$

Target-aware Token Preference. To emphasize hallucination-related tokens, we reweight token importance using the discovered visual targets:

$$\tilde{\omega}_t^w = \omega_t^w \cdot \sum_{p=1}^{P} \mathrm{Attn}_{t,p}^w M_p, \quad \tilde{\omega}_t^l = \omega_t^l \cdot \sum_{p=1}^{P} \mathrm{Attn}_{t,p}^l M_p, \tag{17}$$

where ω_t denotes the textual target weight, $\mathrm{Attn}_{t,p}$ is the averaged decoder cross-attention between token t and patch p, and M_p is the visual target weight. The reweighted token weights are treated as stop-gradient coefficients. The token-level preference margins for A_w and A_l are defined as

$$\Delta_t^w = \beta_2 \left(\log \frac{\pi_\theta(a_t^w | a_{<t}^w, I_w, Q)}{\pi_{\mathrm{ref}}(a_t^w | a_{<t}^w, I_w, Q)} - \log \frac{\pi_\theta(a_t^w | a_{<t}^w, I_l, Q)}{\pi_{\mathrm{ref}}(a_t^w | a_{<t}^w, I_l, Q)} \right). \tag{18}$$

$$\Delta_t^l = \beta_2 \left(\log \frac{\pi_\theta(a_t^l | a_{<t}^l, I_l, Q)}{\pi_{\mathrm{ref}}(a_t^l | a_{<t}^l, I_l, Q)} - \log \frac{\pi_\theta(a_t^l | a_{<t}^l, I_w, Q)}{\pi_{\mathrm{ref}}(a_t^l | a_{<t}^l, I_w, Q)} \right). \tag{19}$$

where β_2 is the temperature parameter. The token-level preference loss is

$$\mathcal{L}_{tp} = -\sum_{t=1}^{T_w} \tilde{\omega}_t^w \log \sigma(\Delta_t^w) - \sum_{t=1}^{T_l} \tilde{\omega}_t^l \log \sigma(\Delta_t^l), \tag{20}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

Evidence Consistency Alignment. We align token cross-attention with the discovered visual targets:

$$\mathcal{L}_{eca} = \sum_{t=1}^{T_w} \omega_t^w (1 - \cos(\mathrm{Attn}_t^w, M)) + \sum_{t=1}^{T_l} \omega_t^l \Big(1 - \cos(\mathrm{Attn}_t^l, M) \Big). \tag{21}$$

where M denotes the visual target distribution, Attn_t is the token cross-attention distribution over image patches, and $\cos(\cdot, \cdot)$ denotes cosine similarity.

Total Loss. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{bp} + \lambda_1 \mathcal{L}_{tp} + \lambda_2 \mathcal{L}_{eca}, \tag{22}$$

where λ_1 and λ_2 balance the contributions of $\mathcal{L}_{tp}$ and $\mathcal{L}_{eca}$, respectively.

5 Experiment

5.1 Experimental Setup

Comparison Methods. We compare LWA-TD with representative preference optimization, target-level alignment, contrastive grounding, and hallucination mitigation methods, including mDPO Wang et al. (2024b), V-DPO Xie et al. (2024), MFPO Jiang et al. (2025), TL-DPO Yoon et al. (2025), BiDPO Liang et al. (2026), HALVA Sarkar et al. (2024), HSA-DPO Xiao et al. (2025), S-VCO Wu et al. (2025), CHiP Fu et al. (2025), Octopus Suo et al. (2025), VCD Leng et al. (2023), RoVRM Wang et al. (2024a), and NaPO Zhang et al. (2025). Following prior work, experiments are conducted on both 7B and 13B models. We implement LWA-TD on LLaVA-1.5-7B/13B Liu et al. (2024b), and Muffin-13B Yu et al. (2023a), and further evaluate its generalization on Qwen2.5-VL-7B Bai et al. (2025) and LLaVA-Next-7B Liu et al. (2024c).

Evaluation Metrics. We evaluate hallucination mitigation on MMHalBench Sun et al. (2023), Object HalBench Rohrbach et al. (2018), and AMBER Wang et al. (2023), using MMHal scores, CHAIR metrics, coverage, and hallucination rates. We further assess general multimodal capability on GQA Hudson & Manning (2019), SQA Lu et al. (2022), TextVQA Singh et al. (2019), MMVet Yu et al. (2023c), and POPE Li et al. (2023b).

Model	Model Size	MMHalBench		Object HalBench		AMBER	
		Score↑	HalRate↓	CHAIRs↓	CHAIRi↓	Cover.↑	HalRate↓
mDPO Wang et al. (2024b) (EMNLP'24)	7B	2.39	54.0	35.7	9.8	52.4	24.5
S-VCO Wu et al. (2025) (ACL'25)	7B	2.75	46.0	-	-	-	-
VCD Leng et al. (2023) (CVPR'24)	7B	2.12	54.2	48.8	24.3	51.5	39.0
RLAIF-V Yu et al. (2024) (CVPR'25)	7B	2.95	32.3	10.5	5.2	50.4	16.2
V-DPO Xie et al. (2024) (EMNLP'24)	7B	2.16	56.0	-	-	49.7	27.3
POVID Zhou et al. (arXiv'24)	7B	2.08	56.2	48.1	24.4	51.3	34.3
Octopus Suo et al. (2025) (arXiv'25)	7B	2.61	50.0	20.8	6.6	49.2	23.4
RoVRM Wang et al. (2024a) (AAAI'25)	7B	2.36	48.9	27.0	16.3	48.2	23.4
NaPO Zhang et al. (2025) (arXiv'25)	7B	**3.31**	35.0	25.7	6.2	54.1	20.7
OPA-DPO Yang et al. (2025) (CVPR'25)	7B	2.83	45.0	13.0	4.3	47.9	**11.6**
MFPO Jiang et al. (2025) (arXiv'24)	7B	2.69	49.0	13.4	6.6	55.7	22.5
HALVA Sarkar et al. (2024) (ICLR'25)	7B	2.25	54.0	41.4	11.7	53.0	32.2
TL-DPO Yoon et al. (2025) (CVPR'25)	7B	2.72	-	20.1	5.2	-	-
BiDPO Liang et al. (2026) (TechRxiv'26)	7B	3.09	31.2	10.4	4.9	56.3	13.6
CHiP Fu et al. (2025) (ICLR'25)	7B	2.90	39.6	**4.9**	**3.2**	**57.8**	24.5
LLaVA-1.5 Liu et al. (2024b)	7B	2.19	57.0	54.7	15.9	51.8	34.7
+LWA-TD (Ours)	7B	3.14	**30.7**	10.1	4.4	57.3	11.9
RoVRM Wang et al. (2024a) (AAAI'25)	13B	2.57	43.8	25.0	13.2	47.7	19.5
OPA-DPO Yang et al. (2025) (CVPR'25)	13B	3.07	39.0	16.3	5.5	48.3	12.8
RLHF-V Yu et al. (2023b) (CVPR'24)	13B	2.81	49.0	12.2	7.5	46.1	25.1
MFPO Jiang et al. (2025) (arXiv'24)	13B	2.94	42.0	11.4	4.6	56.1	19.4
NaPO Zhang et al. (2025) (arXiv'25)	13B	3.72	30.0	23.7	5.9	55.7	19.0
HSA-DPO Xiao et al. (2025) (AAAI'25)	13B	2.61	48.0	-	-	47.3	13.4
BiDPO Liang et al. (2026) (TechRxiv'26)	13B	3.74	28.9	11.2	4.1	57.8	11.9
CHiP Fu et al. (2025) (ICLR'25)	13B	2.60	49.0	**6.2**	**3.9**	45.3	17.6
HALVA Sarkar et al. (2024) (ICLR'25)	14B	2.58	45.0	45.4	12.8	52.6	30.4
Muffin Yu et al. (2023a)	13B	2.40	60.4	21.5	11.6	48.3	32.1
+LWA-TD (Ours)	13B	3.31	**27.9**	12.0	4.6	57.9	12.1
LLaVA-1.5 Liu et al. (2024b)	13B	3.03	47.0	53.3	14.5	50.9	37.6
+LWA-TD (Ours)	13B	**3.86**	29.3	10.7	4.0	**58.5**	**11.2**

Table 1: Hallucination evaluation results on MMHalBench, Object HalBench, and AMBER. LWA-TD improves hallucination mitigation performance over its base models across different model scales and backbones. Best results are shown in **bold** and second-best results are underlined.

Implementation Details. Unless otherwise specified, all models are fine-tuned for 2 epochs using AdamW with a learning rate of 1×10^{-6}, weight decay of 0.05, and cosine decay scheduling. We use a global batch size of 32 on 8 NVIDIA A100 80GB GPUs. Images are resized to 336×336, and the maximum text length is set to 128. The hyperparameters are set to $\tau = 0.1$, $\beta_1 = 0.1$, $\beta_2 = 0.5$, $\lambda_1 = 0.5$, and $\lambda_2 = 0.3$. The discovered visual targets and token weights are treated as fixed coefficients during optimization. Additional implementation details, training efficiency analysis, and parameter sensitivity results are provided in Appendix A.1 and Appendix A.4.

Training Data. We train LWA-TD on BiDPO-data-12k Liang et al. (2026), which is constructed from LLaVA-Instruct-150K Liu et al. (2024b). The dataset contains 12,000 preference tuples $\{I_w, I_l, Q, A_w, A_l\}$, where I_l is obtained by applying localized visual modifications to I_w. Modified images and responses are generated with GPT-4o and filtered using CLIP Radford et al. (2021) similarity to preserve global semantics while retaining sufficient local differences. The dataset covers object, attribute, count, and position changes, providing supervision for visual and textual target discovery.

Model	Size	GQA	SQA	TextVQA	MMVet	POPE
mDPO Wang et al. (2024b)	7B	-	66.5	41.7	31.8	-
POVID Zhou et al.	7B	64.1	68.8	64.6	31.8	86.9
RLHF-V Yu et al. (2023b)	7B	62.9	67.1	64.8	30.9	86.2
BiDPO Liang et al. (2026)	7B	64.9	74.3	68.0	45.2	87.4
S-VCO Wu et al. (2025)	7B	-	73.9	60.3	44.0	-
LLaVA-1.5 Liu et al. (2024b)	7B	62.0	66.8	58.2	31.1	85.9
+LWA-TD (Ours)	7B	**65.3**	**75.6**	**69.4**	**45.7**	**87.6**

Table 2: Results on general multimodal benchmarks. Best results are shown in **bold** and second-best results are underlined.

5.2 Main Results

Table 1 reports the results on MMHalBench, Object HalBench, and AMBER. LWA-TD consistently improves hallucination mitigation across different model scales and backbones. On LLaVA-1.5-7B, it improves the MMHal score from 2.19 to 3.14 while reducing the hallucination rate from 57.0 to 30.7. Similar gains are observed on Object HalBench and AMBER. Compared with recent methods such as TL-DPO, S-VCO, and BiDPO, LWA-TD achieves competitive or superior performance across most metrics. The improvements generalize to larger models, achieving a MMHal score of 3.86 on LLaVA-1.5-13B and reducing the hallucination rate of Muffin-13B from 60.4 to 27.9. Table 2 shows that LWA-TD preserves strong general multimodal capability. On LLaVA-1.5-7B, it consistently improves GQA, SQA, TextVQA, MMVet, and POPE over the base model, outperforming recent preference optimization methods on most benchmarks. Additional backbone generalization and controlled comparison results are provided in Appendix A.2 and Appendix A.3, respectively.

Variant	MMHalBench		AMBER	
	Score↑	HalRate↓	Cover.↑	HalRate↓
w/o Visual Target Discovery	2.96	35.4	55.9	15.1
w/o Textual Target Discovery	2.99	34.1	56.2	14.5
w/o Evidence Consistency Alignment	3.05	32.9	56.7	13.6
w/o Bidirectional Preference Optimization	2.87	37.6	55.4	16.8
w/o Attention-guided Token Weighting	3.06	33.5	56.3	14.2
LWA-TD (Ours)	**3.14**	**30.7**	**57.3**	**11.9**

Table 3: The effect of different components in LWA-TD on LLaVA-1.5-7B.

5.3 Ablation Study

Key Components. Table 3 presents the ablation results on LLaVA-1.5-7B. Removing any component degrades performance on both MMHalBench and AMBER. Bidirectional preference optimization has the largest impact, reducing the MMHal score from 3.14 to 2.87 and increasing the AMBER hallucination rate from 11.9 to 16.8 when removed. Visual target discovery is also important, lowering the MMHal score to 2.96 without explicit region supervision. Similar degradation is observed when removing textual target discovery. Attention-guided token weighting and evidence consistency alignment further improve the consistency between generated tokens and visual evidence.

Visual Target Components			MMHalBench		AMBER	
Visual Difference	Question Relevance	Semantic Similarity	Score↑	HalRate↓	Cover.↑	HalRate↓
✓	×	×	2.98	34.2	56.0	14.8
✓	✓	×	3.07	32.5	56.6	13.2
✓	×	✓	3.04	33.1	56.4	13.7
✓	✓	✓	**3.14**	**30.7**	**57.3**	**11.9**

Table 4: The effect of different visual target discovery components.

Visual Target Discovery. Table 4 compares different components in visual target discovery. Using only visual differences achieves a MMHal score of 2.98. Incorporating question relevance further improves the score to 3.07 and reduces the hallucination rate from 34.2 to 32.5, highlighting the importance of question-aware grounding. Semantic similarity based on CLIP provides additional improvements by filtering noisy visual changes. Combining all three components yields the best performance, achieving a MMHal score of 3.14 and reducing the AMBER hallucination rate to 11.9.

Visual Target Faithfulness. For each image, visual patches are ranked by their target weights. High-target Mask removes the top 20% patches, Low-target Mask removes the bottom 20%, and Random Mask removes the same number of randomly selected patches. Margin Drop denotes the reduction in the response-level preference margin after masking. Table 5 shows that masking low-target regions only slightly affects performance, reducing the MMHal score from 3.14 to 3.04. In

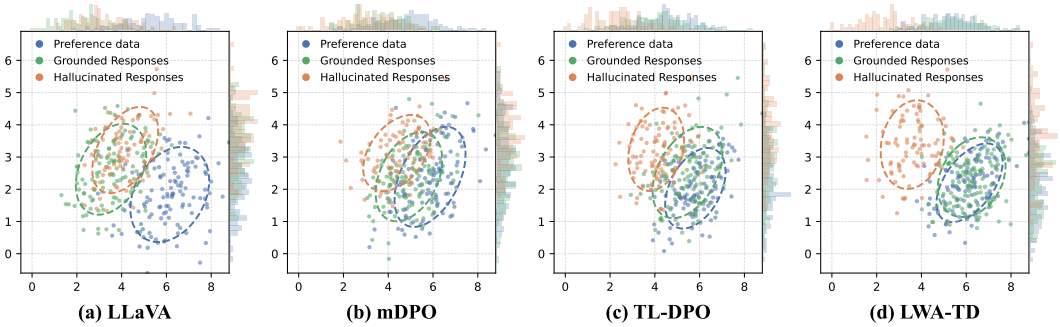


Figure 3: t-SNE visualization of preference data, grounded responses, and hallucinated responses.

Mask Strategy	MMHalBench		AMBER		Margin Drop↑
	Score↑	HalRate↓	Cover.↑	HalRate↓	
Original	3.14	30.7	57.3	11.9	0.00
Low-target Mask	3.04	32.1	56.7	12.8	0.19
Random Mask	2.91	36.4	55.0	15.6	0.52
High-target Mask	2.57	43.9	52.4	19.3	**1.24**

Table 5: The effect of masking visual regions with different target weights.

contrast, masking high-target regions causes a substantial drop, decreasing the MMHal score to 2.57 and increasing the AMBER hallucination rate to 19.3, while random masking yields intermediate results. A similar trend is observed for preference margins: the margin drop increases from 0.19 under Low-target Mask to 1.24 under High-target Mask. Additional analyses on different token weighting strategies are presented in Appendix A.5.

5.4 Representation Distribution Analysis

We visualize mean-pooled final-layer representations of preference data, grounded responses, and hallucinated responses using t-SNE. As shown in Figure 3, LLaVA exhibits substantial overlap between grounded and hallucinated responses. mDPO and TL-DPO improve the separation, while LWA-TD achieves the clearest separation and keeps grounded responses aligned with preference data.

5.5 Case Study

Figure 4 compares the responses generated by LLaVA, mDPO, TL-DPO, and LWA-TD. Existing methods tend to infer unsupported events from crowded scenes, producing descriptions such as "music festival," "concert tickets," and "outdoor festival." In contrast, LWA-TD describes only observable activities, including people standing, walking, and gathering near an outdoor booth area.

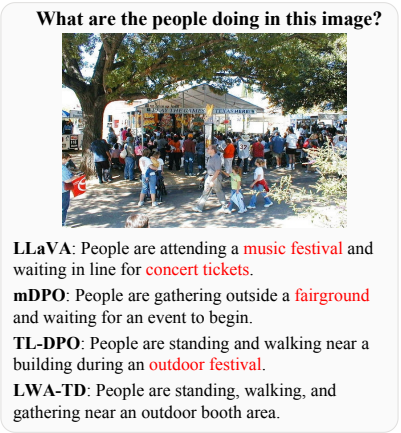


Figure 4: Qualitative comparison of scene-level hallucinations.

6 Conclusion

We propose LWA-TD, which automatically discovers visual regions and response tokens from paired images with minimal visual differences. The discovered targets guide target-aware preference optimization and evidence consistency alignment, improving grounding and reducing hallucinations. Experiments demonstrate consistent improvements across multiple benchmarks and model backbones.

AI USE STATEMENT

Generative AI tools were used to assist with editing the manuscript for readability, including language polishing, grammatical correction, and refinement of scientific writing. They were also used for limited assistance with LaTeX formatting and presentation. Generative AI tools were not used to develop the proposed methodology, formulate mathematical claims, design the experiments, implement the methods, generate or modify the experimental data, conduct data analysis, interpret the experimental results, or determine the scientific conclusions. All AI-assisted text and formatting were reviewed and verified by the authors. The authors take full responsibility for the content, claims, results, and conclusions presented in this work.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Sam Fort, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Alessio Favero, Luca Zancato, Matthew Trager, Sudhanshu Choudhary, Priya 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 (CVPR)*, 2024.

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

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.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Songtao Jiang, Yan Zhang, Ruizhe Chen, Tianxiang Hu, Yeying Jin, Qinglin He, Yang Feng, Jian Wu, and Zuozhu Liu. Modality-fair preference optimization for trustworthy mllm alignment. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 403–411, 2025. doi: 10.24963/ijcai.2025/46.

Shan Leng, Hao Zhang, Guotian Chen, Xiaoxue Li, Shihan Lu, Chang Miao, and Li 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 (CVPR)*, 2024.

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. *IEEE*, 2023.

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

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*, 2023b.

Jing Liang, Rui Wang, Long Hu, Xianzhi Li, Yixue Hao, Tao Guan, and Min Chen. Bidpo: Bidirectional preference optimization with token-level supervision for vision-language models. *TechRxiv*, 2026(0212), 2026. doi: 10.36227/techrxiv.177091884.42596378/v1.

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*, 2024a.

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*, pp. 26296–26306, 2024b.

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 2024c. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Pan Lu, Swaroop Mishra, Tianming Xia, Lin Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, 2022.

Long Ouyang, Jeff 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. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

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

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

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

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan. Ark, and Tomas Pfister. Data-augmented phrase-level alignment for mitigating object hallucination. 2024.

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

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang Yan Gui, Yu Xiong Wang, and Yiming Yang. Aligning large multimodal models with factually augmented rlhf. 2023.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29904–29914, 2025.

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.

Chenglong Wang, Yang Gan, Yifu Huo, Yongyu Mu, Murun Yang, Qiaozhi He, Tong Xiao, Chunliang Zhang, Tongran Liu, and Quan Du. Rovrm: A robust visual reward model optimized via auxiliary textual preference data. 2024a.

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. *arXiv preprint arXiv:2406.11839*, 2024b.

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.

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

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback, 2025. URL https://arxiv.org/abs/2404.14233.

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. *arXiv preprint arXiv:2411.02712*, 2024.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. 2025.

Dokyoon Yoon, Youngsook Song, and Woomyong Park. Stop learning it all to mitigate visual hallucination, focus on the hallucination target. *arXiv preprint arXiv:2506.11417*, 2025.

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

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai Tao Zheng, and Maosong Sun. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *IEEE*, 2023b.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, and Taiwen He. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. 2024.

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*, 2023c.

Zefeng Zhang, Hengzhu Tang, Jiawei Sheng, Zhenyu Zhang, Yiming Ren, Zhenyang Li, Dawei Yin, Duohe Ma, and Tingwen Liu. Debiasing multimodal large language models via noise-aware preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9423–9433. IEEE, 2025.

Wenxuan Zhou, Shujian Zhang, Lingxiao Zhao, and Tao Meng. T-reg: Preference optimization with token-level reward regularization. *arXiv preprint arXiv:2412.02685*, 2024.

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

Yuan Zhou, Chen Cui, Jinsung Yoon, Lei Zhang, Zhewei Deng, Chelsea Finn, Mohit Bansal, and Hongyang Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2311.11199*, 2023.

A Appendix

A.1 Implementation Details

All models are fine-tuned for 2 epochs using 8 NVIDIA A100 80GB GPUs with a global batch size of 32. We use the AdamW optimizer with a learning rate of 1×10^{-6}, weight decay of 0.05, cosine learning rate scheduling, and a warm-up ratio of 5%. Following the default settings of the base models, images are resized to 336×336, and the maximum text sequence length is set to 128. For visual target discovery, the semantic consistency score is computed using CLIP cosine similarity between visual patches and question embeddings. For textual target discovery, the temperature parameter is set to $\tau = 0.1$. The temperature parameters are set to $\beta_1 = 0.1$ and $\beta_2 = 0.5$ for bidirectional preference optimization and token-level preference optimization, respectively. The weighting coefficients for the token-level preference loss and evidence consistency alignment loss are set to $\lambda_1 = 0.5$ and $\lambda_2 = 0.3$, respectively. The discovered visual targets and token weights are treated as fixed targets during optimization. All other hyperparameters follow the default settings of the corresponding open-source implementations. Figure 5 presents the training time and peak GPU memory across different MLLM backbones under the same training configuration using the BiDPO-data-12k dataset.

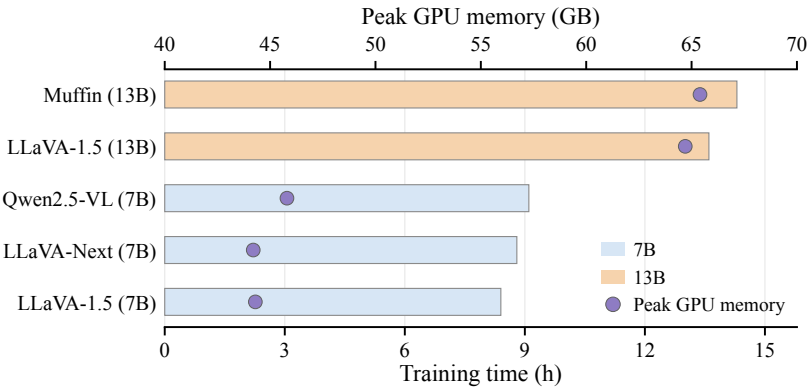


Figure 5: Training time and peak GPU memory of LWA-TD across different MLLM backbones under the same training configuration.

Model	Size	MMHal Score	MM-Vet	POPE	SQA
Qwen2.5-VL	7B	3.30	67.1	80.2	87.5
+ LWA-TD (Ours)	7B	**3.39**	**68.8**	**84.3**	**88.4**
LLaVA-Next	7B	2.56	43.9	86.5	70.1
+ LWA-TD (Ours)	7B	**3.18**	**48.7**	**88.2**	**73.7**

Table 6: Generalization results of LWA-TD across different MLLM backbones.

A.2 Generalization Across Different MLLMs

Table 6 reports the generalization results of LWA-TD on different MLLM backbones, including Qwen2.5-VL-7B and LLaVA-Next-7B. LWA-TD consistently improves both hallucination mitigation and general multimodal reasoning performance across different backbones. On Qwen2.5-VL-7B, LWA-TD improves the MMHal score from 3.30 to 3.39 while also achieving consistent gains on MM-Vet, POPE, and SQA. The POPE score improves from 80.2 to 84.3, suggesting enhanced object-level grounding ability. On LLaVA-Next-7B, LWA-TD also achieves consistent improvements across all evaluated benchmarks. LWA-TD improves the MMHal score from 2.56 to 3.18 and increases MM-Vet from 43.9 to 48.7. The method also improves POPE and SQA performance, demonstrating good generalization across different MLLM backbones.

Method	Training Data	MMHalBench		AMBER	
		Score↑	HalRate↓	Cover.↑	HalRate↓
DPO	BiDPO-data-12k	2.61	45.8	54.1	21.7
BiDPO	BiDPO-data-12k	3.09	31.2	56.3	13.6
w/o Target Discovery	BiDPO-data-12k	3.02	33.5	56.3	14.2
LWA-TD (Ours)	BiDPO-data-12k	**3.14**	**30.7**	**57.3**	**11.9**

Table 7: Controlled comparison of different optimization strategies under identical training data.

A.3 Controlled Comparison of Optimization Strategies

Table 7 presents a controlled comparison where all methods are trained on the same BiDPO-data-12k dataset under identical training settings. This experiment isolates the contribution of the optimization strategy from that of the training data. Compared with standard DPO, BiDPO achieves better hallucination mitigation performance, indicating that bidirectional preference optimization can better utilize counterfactual preference supervision. Specifically, the MMHal score improves from 2.61 to 3.09, while the MMHal hallucination rate decreases from 45.8 to 31.2. The AMBER hallucination rate is also reduced from 21.7 to 13.6. Although BiDPO already introduces bidirectional preference optimization, LWA-TD further improves performance by explicitly identifying and utilizing relevant visual and textual targets. Compared with BiDPO, LWA-TD improves the MMHal score from 3.09 to 3.14 and further reduces the AMBER hallucination rate from 13.6 to 11.9. In contrast, removing target discovery leads to consistent performance degradation across all metrics. For example, the MMHal hallucination rate increases from 30.7 to 33.5 after removing target discovery. These results indicate that the performance gains mainly come from the proposed target-aware optimization strategy rather than the training data alone.

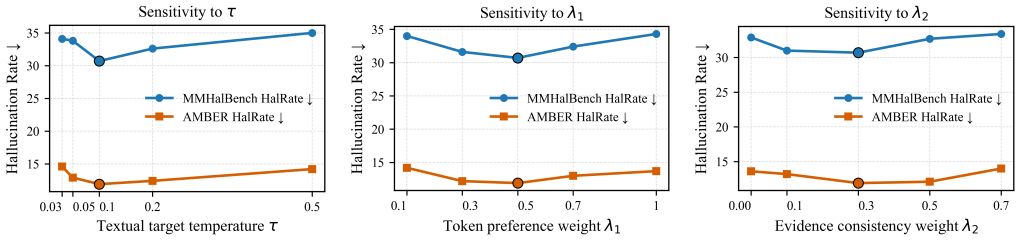


Figure 6: Parameter sensitivity analysis of LWA-TD on MMHalBench and AMBER. Each hyperparameter is varied independently while keeping the others fixed.

A.4 Parameter Sensitivity Analysis

We analyze the sensitivity of LWA-TD to three key hyperparameters, including the token weighting temperature τ, the token preference weight λ_1, and the evidence consistency weight λ_2. The results are shown in Figure 6. For the token weighting temperature τ, performance first improves and then gradually decreases as τ increases. When τ is too small, the token weight distribution becomes highly concentrated and reduces the contribution of contextual tokens. As τ further increases, the token weights become overly smooth and weaken the distinction between important and less relevant tokens. The best performance is achieved around $\tau = 0.1$, where both MMHalBench and AMBER hallucination rates reach their lowest values. For the token preference weight λ_1, the model achieves the best performance at an intermediate value. A smaller λ_1 weakens the contribution of token-level preference optimization, while a larger value overemphasizes token-level supervision. The lowest hallucination rates are obtained near $\lambda_1 = 0.5$, indicating that balancing sequence-level and token-level supervision is important. For the evidence consistency weight λ_2, introducing evidence consistency alignment consistently improves performance. When $\lambda_2 = 0$, the alignment term is removed, corresponding to the ablation setting without evidence consistency alignment. Performance improves as λ_2 increases to a moderate range and achieves the best result around $\lambda_2 = 0.3$. Further increasing λ_2 slightly degrades performance, suggesting that overly large alignment weights may constrain token-to-region alignment.

Token Weighting Strategy	MMHalBench		AMBER	
	Score↑	HalRate↓	Cover.↑	HalRate↓
Uniform Weighting	2.99	34.1	56.2	14.5
Top-k Weighting	3.03	33.2	56.5	13.9
Threshold Weighting	3.05	32.8	56.6	13.6
Soft Weighting (Ours)	**3.14**	**30.7**	**57.3**	**11.9**

Table 8: The effect of different token weighting strategies.

A.5 Token Weighting Strategies.

Table 8 compares different token weighting strategies. For hard weighting, Top-k selects the top 20% tokens with the highest scores, while Threshold Weighting assigns non-zero weights to tokens with scores above 0.5. Uniform weighting achieves a MMHal score of 2.99, whereas the proposed Soft Weighting improves it to 3.14 and reduces the AMBER hallucination rate from 14.5 to 11.9.