This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Stop learning it all to mitigate visual hallucination, Focus on the hallucination target.

Dokyoon Yoon, Youngsook Song, Woomyong Park*
SIONIC AI
dokyoon, song, max@sionic.ai

Abstract

Multimodal Large Language Models (MLLMs) frequently suffer from hallucination issues, generating information about objects that are not present in input images during vision-language tasks. These hallucinations particularly undermine model reliability in practical applications requiring accurate object identification. To address this challenge, we propose TL-DPO, a preference learning approach that mitigates hallucinations by focusing on targeted areas where they occur. To implement this, we build a dataset containing hallucinated responses, correct responses, and target information (i.e., objects present in the images and the corresponding chunk positions in responses affected by hallucinations). By applying a preference learning method restricted to these specific targets, the model can filter out irrelevant signals and focus on correcting hallucinations. This allows the model to produce more factual responses by concentrating solely on relevant information. Experimental results demonstrate that TL-DPO effectively reduces hallucinations across multiple vision hallucination tasks, improving the reliability and performance of MLLMs without diminishing overall performance.

1. Introduction

Multimodal Large Language Models (MLLMs) have achieved remarkable performance in multimodal tasks by leveraging extensive training and fine-tuning on large-scale image-text pairs, which has significantly enhanced their contextual understanding and instruction-following abilities [2, 4, 8, 13]. However, during the integration of text and vision modalities, a new type of hallucination issue has emerged, where the model generates object information that does not exist in the images or creates inaccurate spatial representations [1, 3, 6, 7, 17].

To address these issues, recent studies have introduced preference-based learning methods, most notably Reinforcement Learning from Human Feedback (RLHF), to improve alignment across modalities and correct hallucinations [12, 14–16, 18]. However, the problem of object-level hallucinations within images remains unresolved [11, 17]. Although preference-based learning has contributed to reducing distribution gaps across modalities and enhancing factuality, its global feature-centric learning poses limitations in addressing object-level hallucinations.

While humans tend to correct only specific areas when they misdraw a region in an illustration or use an incorrect word in writing, previous approaches to addressing halluci-

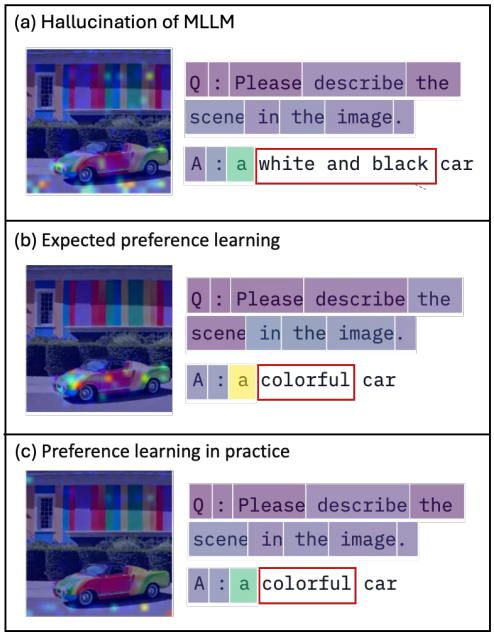


Figure 1. **Visualization of attention maps for hallucination mitigation** shows how preference learning influences model's attention. (a) illustrates hallucinations caused by focusing on incorrect objects or missing relevant ones. Ideally, as in (b), preference learning would guide attention to the correct objects, but in reality, there is a risk of overfitting to text rather than object-based information, as shown in (c). Additional analysis is in Appendix A.

*Corresponding author

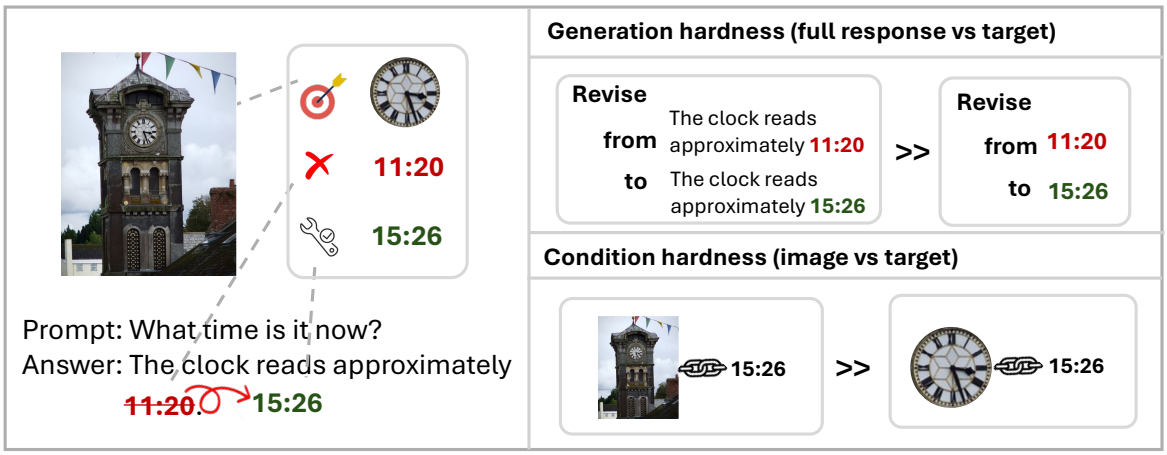


Figure 2. Target learning paradigm. (Left) Correcting hallucinations in image-based QA requires identifying the hallucinated object, the hallucination, and the correct (e.g., identifying incorrect and correct times for a clock-related hallucination). (Right) Training on specific hallucinated parts and their corrections is easier than learning from whole responses, and focusing on relevant image regions simplifies learning compared to using the entire image.

nation issues in the literature have often suggested "redrawing everything." In this study, we examine the necessity of directly addressing the image hallucination problem, as humans would, by investigating whether existing preference-based learning aligns with our expectations for hallucination correction. As shown in 1, existing preference learning methods have often learned signals unrelated to hallucinations (e.g., irrelevant chunks or regions in the image), leading to unintended side effects.

Based on these observations, we hypothesize that the reason for the lack of effective hallucination correction in preference learning is due to learning from irrelevant signals. Consequently, we propose a novel approach, termed "target learning", which focuses exclusively on the image regions where hallucinations occur and the specific chunks in the responses that contain hallucinations, separating these as the primary targets for learning. Additionally, we provide a logical explanation showing how target learning enhances both the efficiency and effectiveness of the learning process.

To correct hallucinations in MLLMs, this study introduces **TL-DPO** (**T**arget-**L**earning **D**irect **P**reference **O**ptimization), a learning method that identifies image regions and response chunks associated with hallucinations as targets, while categorizing unrelated parts as irrelevant signals, thereby focusing solely on the target elements during training. In TL-DPO, a new loss function is employed to focus on targets, enabling selective learning by distinguishing them from irrelevant signals. This approach not only guides multimodal models to generate outputs without hallucinations but also mitigates other incorrect behaviors.

To sum up, our contributions are as follows:

1. We identify the learning of irrelevant signals as the primary reason for the failure of conventional preference learning in correcting hallucinations. To address this, we propose **target learning**, which explicitly excludes irrelevant signals, and provide a solid theoretical foundation for this approach.
2. We introduce TL-DPO a novel method for implementing Target Learning. To support this method, we expand the existing dataset by constructing a preference-based image dataset that includes the target
3. We validate the effectiveness of the proposed method through experiments on various benchmark models, demonstrating reduced hallucinations.

2. Preliminaries

In language generation, a language model (LM) generates a response y based on a prompt containing a question x, where both x and y consist of sequences of tokens. Direct Preference Optimization (DPO) [10] builds upon the reinforcement learning (RL) objective used in RLHF (Reinforcement Learning from Human Feedback) [9].

$$\max_{\pi_\theta} \mathbb{E}_{x\sim D, y\sim\pi_\theta(\cdot|x)}[r(x,y) - \beta D_{\mathrm{KL}}(\pi_\theta(\cdot \mid x)|\pi_{\mathrm{ref}}(\cdot \mid x))], \tag{2.1}$$

where D represents the human preference dataset, $r(x,y)$ denotes the reward function, and $\pi_{\mathrm{ref}}(\cdot \mid x)$ is the reference model, typically chosen as the language model after supervised fine-tuning. Here, π_θ denotes the model being adjusted through reinforcement learning. This objective function incorporates Kullback-Leibler (KL) divergence constraints at the token level, improving the regulation of KL divergence.

DPO derives a connection between the reward model and the optimal policy by applying reverse KL divergence, allowing the reward function to be expressed in relation to the model's policy. This is captured as:

$$r(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x), \tag{2.2}$$

where $Z(x)$ is a normalization term called the partition function. To effectively capture human preferences, DPO uses the Bradley-Terry model to compare pairs of possible responses, assigning a probability to each preferred response y_1 over another response y_2 given the prompt x:

$$P_{\text{BT}}(y_1 \succ y_2 \mid x) = \frac{\exp(r(x, y_1))}{\exp(r(x, y_1)) + \exp(r(x, y_2))}. \tag{2.3}$$

This formulation is refined by substituting the reward expression into the comparison model, leading to a loss function based on the negative log-likelihood:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim D}[\log \sigma(u(x, y_w, y_l))], \tag{2.4}$$

where $\sigma(\cdot)$ is the sigmoid function and $u(x, y_w, y_l)$ quantifies the relative preference between the more favorable response y_w and the less favorable response y_l. Specifically, $u(x, y_w, y_l)$ is defined as:

$$u(x, y_w, y_l) = r(x, y_w) - r(x, y_l). \tag{2.5}$$

The gradient of this loss, critical for model updates, is:

$$\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim D}\left[\sigma(-u(x, y_w, y_l)) \nabla_\theta u\right], \tag{2.6}$$

where u acts as a shorthand for $u(x, y_w, y_l)$. This formulation enables DPO to iteratively refine the policy by focusing on minimizing the divergence while aligning with human preference through straightforward pairwise comparison, enhancing the overall quality of the model's responses without complex reward model sampling.

3. Target Learning Problem

Although conventional preference learning has been effective in improving the overall output quality of models, as shown in 2, it may also result in learning elements that are not designated as targets in preference learning. This is due to a limitation, as represented in 1, where the optimization process prioritizes global features, often overlooking chunks where hallucinations occur in responses or localized object information that has caused hallucinations.

In 3.1, we redefine conventional preference learning as a problem focused solely on target learning. We formalize this as a target learning approach and logically demonstrate in 3.2 that the target learning approach achieves greater efficiency than conventional preference learning by excluding irrelevant signals.

This approach rejects hallucinated values in target chunks where hallucinations occur, guiding the model to favor corrected values. In the human preference learning stage, a reward model $\hat{r}$ is trained on target chunks, enabling the model to output hallucination-free chunks and reducing the probability of generating hallucinated chunks. This reward model can assign rewards across various chunks, reflecting the presence of hallucinations as human preference. Once the reward model is established, it provides feedback in an additional fine-tuning stage, guiding a policy model π_θ based on human preferences regarding the presence or absence of hallucinations.

3.1. Target optimization

In this section, we first reformulate the RLHF [9] objective (constrained reward maximization problem) in terms of the target. The objective function of DPO in Eq. 3.1 operates at the sentence level. In contrast, we propose a new objective function designed to work on alternative target generation in the context of image goals:

$$\max_{\pi_\theta} \mathbb{E}_{x \sim D, y^t \sim \pi_\theta(\cdot|x)}[r(x, y^t) - \beta D_{\text{KL}}(\pi_\theta(\cdot \mid x) | \pi_{\text{ref}}(\cdot \mid x))], \tag{3.1}$$

where y^t represents the subset of targets corresponding to the full target set T. This objective function is designed to enable the model to learn more effectively toward specified targets in image generation, while regulating the KL divergence with the reference model's policy. This balance enhances the model's stability and maintains the quality of generated outputs.

In this work, we design the MLLMs to enable preference learning by comparing chunks within the response where hallucinations occur and the localized objects that cause them, to correct hallucinations. For this, we set the assumption that human preference differences regarding hallucinations occur only in signals related to hallucinations, with no differences in unrelated signals like below:

Assumption

Assumption 3.1 For hallucination-related response y, the reward for response satisfies $\sum r_t = \sum r$, where r_t represents the reward funtion for the target in response y^t.

Following, in the preference learning problem dealing with hallucinations, we can see that the value of the Bradley-Terry model remains the same when unrelated signals are removed:

Lemma

Lemma 3.2 Given a reward function $r(x, y)$, assuming Assumption 3.1, we can establish the equivalence between the Bradley-Terry model to the 2.2.

Lemma 3.2 ensures that, under Assumption 3.1, the preference reward for a response remains the same even when non-target tokens are excluded, preserving the overall reward $r(x, y)$. Here, $r(x, y)$ represents the overall reward for the response y given the prompt x. Therefore, when considering text generation as a target-level generation problem, $r(x, y^t)$ can be viewed as the reward for the generated text aligned with the target, rather than the entire sentence.

Using Lemma 3.2, we can limit the calculation scope to the target itself. This strategy can also be applied in DPO by restricting the comparison of text generations to the target. This concept is summarized in the following theorem.

Theorem 3.3 Under Assumption 3.1, preference learning remains equivalent when non-target tokens are excluded.

That is, the reward function and policy updates for target-restricted DPO align with those of the original DPO. Formally, let π_θ denote the policy in DPO and $r(x, y)$ the reward function. When Assumption 3.1 holds, for any prompt x and response y:

$$\mathbb{E}_{y\sim\pi_\theta}\left[r(x, y)\right] = \mathbb{E}_{y^t\sim\pi_\theta}\left[r(x, y^t)\right], \tag{3.2}$$

Thus, the policy gradient update for target-restricted DPO follows the same form as for the original DPO:

$$\nabla_\theta \mathbb{E}_{y^t\sim\pi_\theta}\left[r(x, y^t) - \beta D_{\text{KL}}(\pi_\theta|\pi_{\text{ref}})\right] \tag{3.3}$$

The proof of theorem 3.3 is provided in the Appendix B. Theorem **??** provides a structured approach for modeling human preference probability within a KL-constrained setting by using the Bradley-Terry model. In this framework, the optimal policy π_θ is designed to align closely with human preferences by directly modeling the relative preference between two target outputs, y_h^t and y_r^t , conditioned on the prompt x .

3.2. Efficiency in Target Learning

In 3.1, we were able to modify the existing Bradley-Terry model to the target level, transforming the preference objective (eq.3.1) into a target-level problem. Intuitively, as shown in 2, focusing on correcting the target rather than the entire sentence is more efficient. This approach enables the model to learn the relationship between the specific region in the image where the fact is located and the fact itself, rather than the relationship between the entire image and the fact. This section emphasizes that the proposed target learning approach is theoretically more efficient than traditional preference learning methods that focus on the overall output. Traditional preference learning can indeed enhance the general quality of model outputs; however, it also tends to include elements outside of the designated target. This inefficiency arises as non-target elements in the image are often learned, resulting in a need for more extensive data during the learning process.

Proposition 1. ***Efficiency comparison in target learning*** *Let $\mathcal{H}_{pl}$ and $\mathcal{H}_{tl}$ be the hypothesis spaces to learning methods without target(pl) and with target(tl), respectively. The number of samples required to achieve the same generalization error ϵ and confidence level $1-\delta$ satisfies $m_{tl} < m_{pl}$.*

By proving the above proposition in the Appendix B, it is shown that the target learning approach can achieve the same level of generalization performance with fewer samples compared to conventional holistic learning methods. This improvement in efficiency is expected to yield a more effective hallucination mitigation effect in real.

4. Methods

To address the tendency of conventional Preference Learning to ignore images and learn from undesired signals, this study proposes a new learning approach in 4.1 and 4.2 to mitigate these issues. The primary objective is to resolve the object hallucination problem in MLLMs through targeted learning. Considering the complexity of data construction and model training, we have chosen Direct Preference Optimization (DPO) over traditional RLHF (RL with Human Feedback) methods to restrict model preferences. DPO, a simpler and more efficient approach that does not require reinforcement learning, serves as our fundamental strategy. We extend this strategy to the multimodal domain to eliminate hallucinations and enhance the reliability and accuracy of multimodal model outputs. We construct a TL dataset that includes incorrect responses, correct responses, and the positional information of targets with bounding boxes (Details in Appendix C). This dataset is subsequently used to train the TL-DPO model. Models trained using this method show a substantial reduction in hallucinations in tasks involving detailed descriptions and dialogue.

4.1. Generation of Target-Object-Inclusive Preference Datasets

In this study, we utilize the Visual Genome (VG) dataset as the main source to create hallucinated and corrected samples. For each given image and related question, we use a baseline model to generate an initial response, then identify instances where hallucinations occur to produce the hallucinated response y_h and its corrected version y_r. Throughout this process, we record the positions of hallucinated elements in the original response and the corresponding corrected elements, allowing us to define hallucinated segments y_h^t and their corrected counterparts y_r^t. Additionally, we incorporate bounding box information from

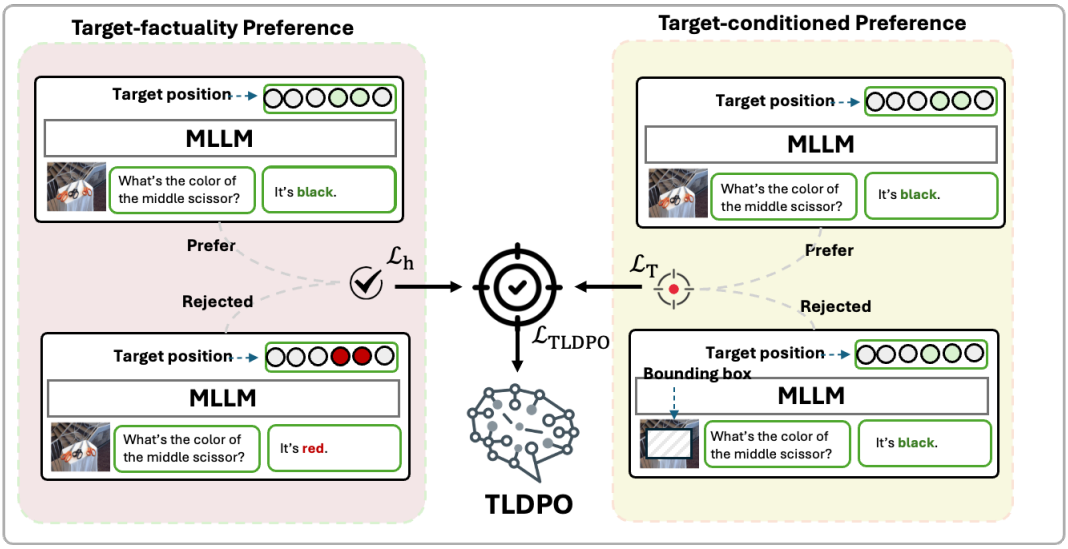


Figure 3. Overview of TL-DPO. To correct hallucinations, TL-DPO (right) applies learning only to hallucinated target chunks, correcting only erroneous information and learning factual information. (left) It applies a masked condition to the target object where hallucination occurs utilizing bounding box information from the image dataset, training the model to extract accurate information from the desired part of the image. TL-DPO combines these two learning objectives to correct erroneous responses extracted from the desired image regions and transform them into correct responses.(Details in Appendix D)

the VG dataset to identify the specific objects within the image that contribute to the hallucination, resulting in the hallucination-related object information m_i^t. Consequently, the structured dataset contains the hallucinated response y_h along with its hallucinated segment y_h^t, the corrected response y_r with its corrected segment y_r^t, the full image m_i, and the specific hallucination-triggering object within the image m_i^t. This comprehensive setup facilitates detailed analysis and targeted correction of hallucinations in generated responses. Detailed explanations are provided in Appendix C.

4.2. Proposed Loss in TL-DPO

Target Generation Loss When learning from hallucinations, the model may unintentionally learn irrelevant information, which we address by applying target loss. Given a pretrained MLLM that is prone to hallucinations, our goal is to minimize the likelihood of generating hallucinated responses by using pairs of erroneous and corrected responses, y_h^t, y_r^t, obtained through generative data augmentation. To this end, we define a preference optimization expression for erroneous and corrected response pairs.

For example, as shown in 2, suppose the erroneous response y_h is "The clock reads approximately 11:20", and the corrected response y_r is "The clock reads approximately 15:26". Here, y_h^t represents the hallucinated information in y_h , and y_r^t denotes the corrected information in y_r that replaces y_h^t with the correct answer. In this example, the erroneous information y_h^t is "11:20", and the correct information y_r^t is "15:26".

Assuming that irrelevant information does not affect the Reward difference, the DPO formula3.1 can be formulated as follows :

$$\mathcal{L}_{\text{t}} = -\mathbb{E}_{(x,y_r^t,y_h^t)\sim D}[\log \sigma(u(x, y_r^t, y_h^t))], \tag{4.1}$$

Here, t denotes a target, y_r^t is a target from the corrected response, and y_h^t is a target from the hallucinated response. T represents the set of targets and u is defined in eq.2.5. In summary, while the standard DPO objective maximizes $\sigma(r(x, y_r) - r(x, y_h))$, TL-DPO maximizes the target-specific reward $\sigma(r(x, y_r^t) - r(x, y_h^t))$, leveraging the fact that not all positions reflect the reward in TL-DPO.

Target Condition Loss In this study, we propose target-conditional preference optimization to guide the model to utilize target object information. The key idea is to extend the approach proposed in mDPO[14] by leaving the target object in the image as a variable, allowing the model to determine preference labels based solely on target object information. For example, as shown in 2, let's assume that the given image m_i is of a clock tower, and m_i^t represents the specific part of the image that caused the hallucination y_h^t. In m_i^t, the object in the image that caused the hallucination is masked. In this example, y_r^t is "11:20", m_i^t is the clock face showing the time, and $\tilde{m}_i^t$ is the image where the clock

Method	Hallucination Benchmark				Comprehensive Benchmark			
	CHAIR_s (↓)	CHAIR_i (↓)	POPE (↑)	MMHal (↑)	SciQA-IMG (↑)	MM-Vet (↑)	MMBench (↑)	LLaVA-Bench(↑)
LLaVA-1.5	66.8	12.7	85.9	2.42	66.8	30.5	63.0	63.4
+VLfeedback	56.3	11.4	83.72	2.62	66.2	31.2	63.9	62.1
+Human-Preference	54.0	9.3	81.50	2.53	65.8	31.1	60.4	63.7
+RLHF-V	44.6	7.9	86.20	2.59	67.1	30.9	63.6	65.4
+POVID	35.2	8.3	86.9	2.08	67.5	30.5	63.2	64.8
+HA-DPO	37.2	10.0	86.9	1.97	70.3	30.7	64.0	66.2
+HALVA	46.6	23.1	**87.0**	2.25	**70.5**	32.1	66.1	67.2
+ TL-DPO	**20.1**	**5.2**	86.95	**2.72**	70.3	**32.2**	**67.8**	**71.2**

Table 1. **Main results on the Hallucination Benchmark and Comprehensive Benchmark.** We report CHAIR-s, CHAIR-i, POPE, and MMHal scores under the Hallucination Benchmark, and SciQA-IMG, MM-Vet, MMBench, and LLaVA-Bench scores under the Comprehensive Benchmark. The best results for each metric in each group are presented in bold. Additionally, we provide additional results using various preference data and learning objectives for reference.

Method	Hallucination Benchmark				Comprehensive Benchmark			
	CHAIR_s (↓)	CHAIR_i (↓)	POPE (↑)	MMHal (↑)	SciQA-IMG (↑)	MM-Vet (↑)	MMBench (↑)	LLaVA-Bench(↑)
LLaVA-1.5	66.8	12.7	85.90	2.42	66.8	30.5	63.0	63.4
+ TL-DPO	**20.1**	**5.2**	**86.95**	**2.72**	**70.3**	**32.2**	**67.8**	**71.2**
Qwen VL Chat	48.2	9.1	87.07	2.89	**68.2**	**41.2**	**60.6**	**67.6**
+ TL-DPO	**31.4**	**7.8**	**85.28**	**2.89**	64.8	38.5	59.7	64.0
LLaVA-Next	29.08	8.08	84.8	2.56	70.3	**40.2**	63	**72.7**
+ TL-DPO	**25.1**	**6.72**	**87.1**	**3.12**	**72.1**	40.1	**63.1**	72.6
Llama3	**5.5**	10.1	82.8	3.19	83.9	**57.6**	85.8	**65.8**
+ TL-DPO	7.1	**9.71**	**87.1**	**3.86**	**87.2**	53.8	**87.3**	58.5
InternVL-2.5(8B)	18.4	8.7	86.5	3.3	96.3	62.8	68.6	82.5
+ TL-DPO	**7.6**	**4.1**	**87.0**	**3.6**	**97.1**	**68.1**	**80.0**	**87.2**

Table 2. **Comparative results of different MLLMs with and without TL-DPO.** Results are provided for baseline models (LLaVA-1.5, DocOwl, Qwen VL Chat, LLaVA-Next, Llama3, InternLV-2.5) both with and without TL-DPO for comparison, with the improved scores highlighted in bold.

face showing the time is masked. Therefore, given preference data pairs (m_i, q, y_r) and $(\tilde{m}_i^t, q, y_r)$, if the unmasked target m_t is more appropriate for q and y_r than the masked target $\tilde{m}_i^t$, we reconstruct the loss function in equation 4.1 by adding the following conditions.

$$\mathcal{L}_{\text{c}} = -\mathbb{E}_{(m_i, \tilde{m}_i^t, x, y_r) \sim D}[\log \sigma(u^*(m_i, \tilde{m}_i^t, x, y_r)], \quad (4.2)$$

With additional target conditions, $u^*(m_i, \tilde{m}_i^t, x, y_r)$ is defined as an extension of 2.5 as follows:

$$u^*(m_i, \tilde{m}_i^t, x, y_r) = r(m_i, x, y_r) - r(\tilde{m}_i^t, x, y_r). \quad (4.3)$$

The challenge here is to ensure that responses based on masked information from hallucination-inducing object parts are less preferred than responses that use all available information to provide a correct answer. To achieve this, our study applies masking using the bounding box of the hallucination-inducing object. This setup encourages the MLLM to utilize the object's image in order to respond accurately.

The objective of TL-DPO is a combination of the target generation loss (eq.4.1) and target condition loss (eq.4.2):

$$\mathcal{L}_{TL-DPO} = \mathcal{L}_t + \mathcal{L}_c \quad (4.4)$$

5. Experiment Results

5.1. Experiment Settings

5.1.1. Implementation Details

We applied TL-DPO, to LLaVA-v1.5-7B to validate the model's performance. Model training was conducted with a batch size of 32 for 3 epochs, using a learning rate of 0.00001, a cosine learning rate scheduler, and a warm-up ratio of 0.1. The β value for preference optimization was set to 0.1. Following previous research, we tuned the model using LoRA, setting the α value to 128 and the rank to 64. TL-DPO, and the standard TL-DPO, share the same configuration settings.

Method	Hallucination Benchmark				Comprehensive Benchmark			
	CHAIR_s (↓)	CHAIR_i (↓)	POPE (↑)	MMHal (↑)	SciQA-IMG (↑)	MM-Vet (↑)	MMBench (↑)	LLaVA-Bench(↑)
LLaVA-1.5	66.8	12.7	85.90	2.42	66.8	30.5	63.0	63.4
+ target condition	32.4	8.6	84.40	2.58	67.0	31.3	61.2	66.5
+ target generation	**14.6**	6.1	**89.64**	2.70	68.2	32.1	62.4	68.7
TL-DPO (both)	20.1	**5.2**	86.95	**2.72**	**70.3**	**32.2**	**67.8**	**71.2**

Table 3. **Ablation Results on TL-DPO** with target generation or target condition removed. Both components contribute to correcting hallucinations, but the combined approach achieves the best overall performance (highlighted in bold).

5.1.2. Training Data

Training Data: Based on the VG dataset ([5]), responses to given queries were generated using a baseline model. GPT-4 was then used to classify the responses as correct or incorrect, with the incorrect responses forming a hallucination dataset. Subsequently, incorrect responses were paired with the corrected answers from GPT-4 to create a preference dataset. Detailed generation methods are provided in the Appendix C.

5.1.3. Evaluation Benchmark

To validate performance, we used eight multimodal LLM benchmarks, focusing on hallucination issues. MMHal-Bench and Object HalBench assess response and object-level hallucinations, with Object HalBench measuring rates at both $CHAIR_s$ (response) and $CHAIR_i$ (object) levels. POPE evaluates preference alignment based on human feedback, while SciQA-IMG and MM-Vet assess scientific and veterinary-specific Q&A accuracy, respectively. MMBench offers a general multimodal evaluation of factual accuracy, and LLaVA-Bench focuses on mitigating visual hallucinations in LLaVA models. Together, these benchmarks provide a comprehensive framework for evaluating MLLM performance in diverse multimodal contexts.



Figure 4. **Qualitative Results from MMHalBench.** An MLLM trained with standard DPO often rephrases image content into approximate meanings, leading to hallucinations resembling the correct answer (e.g., responding with "Come Back" instead of "Return" in OCR tasks). In contrast, an MLLM trained with TL-DPO gives accurate answers grounded in the actual objects in the image. Additional results are in the Appendix E.

5.2. Main results

Table 1 summarizes the experimental results on two benchmark categories: the Hallucination Benchmark and the Comprehensive Benchmark. In the Hallucination Benchmark, our method, TL-DPO, consistently achieves the lowest hallucination scores, showing significant reductions across key metrics. Specifically, TL-DPO, records scores of 20.1 in $CHAIR_s$, 5.2 in $CHAIR_i$, and 2.72 in MMHal, significantly outperforming all other methods. These reductions indicate that TL-DPO, effectively mitigates hallucinations related to target objects, addressing a major limitation of conventional MLLM-based models. By selectively masking or generating content centered on target object information to enhance factual consistency, TL-DPO, demonstrates outstanding performance in effectively controlling hallucination-prone elements in model outputs.

Similarly, in the Comprehensive Benchmark, TL-DPO, leads in overall scores, achieving 70.3 in SciQA-IMG, and the highest values in MM-Vet (32.2), MMBench (67.8), and LLaVA-Bench (71.2). These high scores confirm TL-DPO's ability to maintain balanced performance across a wide range of vision-language tasks while reducing hallucinations. By combining target learning and preference-aligned data, TL-DPO, achieves these results without sacrificing performance on general metrics, demonstrating robustness and adaptability across different task settings. This contrasts with methods such as VLfeedback, Human-Preference, RLHF-V, and POVID, which struggle to achieve the same balance between accuracy and hallucination reduction. Additionally, results from HALVA and HA-DPO show further improvements in hallucination reduction but at the cost of comprehensive performance metrics.

5.3. Generalization Across Models

Table 2 shows the performance improvements across both the Hallucination Benchmark and Comprehensive Benchmark for the LLaVA-1.5, Qwen VL Chat, LLaVA-Next, Llama3, and InternVL-2.5 models enhanced with TL-DPO. For LLaVA-1.5, TL-DPO, significantly reduces the $CHAIR_s$ score from 66.8 to 20.1 and the $CHAIR_i$ score from 12.7 to 5.2, while also improving metrics such as POPE and MMHal. Almost all baseline models also benefit from TL-DPO, integration, showing substantial reductions in $CHAIR_s$ and $CHAIR_i$ scores and consistently improved or maintained performance on comprehensive metrics such as MM-Vet, SciQA-IMG, and MMBench. These results confirm that TL-DPO, enhances robustness and generalization across a range of MLLM models, not only LLaVA, effectively reducing hallucinations while maintaining strong overall performance. However, while improvements are evident in the Hallucination Benchmark, certain metrics in the Comprehensive Benchmark show slight declines, suggesting room for further optimization in future work.

5.3.1. Ablation study

In this study, we propose a target learning approach that focuses on learning specific targets, adding a masking method to the target's location on the image to enable the model to learn the target while utilizing information from the image. Table 3 presents the ablation study results for TL-DPO 's two methods: target condition and target generation, both individually and combined. Applying only the target condition already shows substantial improvements in hallucination metrics (e.g., reducing $CHAIR_s$ from 66.8 to 32.4) while maintaining stable performance on comprehensive benchmarks. Applying target generation also enhances the results, achieving lower hallucination scores ($CHAIR_s$ 14.6) and higher comprehensive metrics (POPE 89.64). The combined TL-DPO achieves the best performance across most benchmarks, demonstrating the effectiveness of integrating target condition and target generation for target-focused learning.

6. Related Works

Factuality and Hallucinations With recent advances in MLLMs, research on errors and hallucinations in model-generated sentences has been actively pursued. MLLMs, trained on large datasets, can produce linguistically coherent sentences across various topics, but they often present information that does not align with factual accuracy. Such hallucinations are considered particularly problematic in applications where factual accuracy is crucial, such as information retrieval, medicine, and law. Key causes of hallucinations include data bias in the training process, noise within the training data, and information gaps between pre-training and fine-tuning stages. As a result, various approaches have been proposed for factual verification in sentence generation. These approaches include evaluating model outputs through reliability scores and reducing hallucination potential by integrating with external knowledge bases.

Preference Alignment in MLLMs Aligning large models with human preferences has become a critical issue due to safety and ethical considerations in real-world applications. Preference alignment can be broadly categorized into two approaches: feedback-based alignment, which includes feedback generated by both humans and AI, and alignment through prompt guidance. Initial research on preference alignment for MLLMs has recently been conducted. Specifically, methods have been proposed to distill preferences to enhance relevance and accuracy based on MLLMs' visual context. These methods collect human preferences to adjust model behavior by modifying hallucinated content at the segment level. While these initial results are promising, they heavily rely on traditional preference data generation processes for LLMs, where both preferred and non-preferred responses are generated, but neither is necessarily guaranteed to be accurate. In contrast, our proposed method directly corrects hallucination issues through a preference learning approach that leverages target information. This approach specifically targets hallucinations by utilizing information from the image objects that previously caused hallucinations, guiding the model to provide correct responses and effectively overcoming the limitations of existing methods.

7. Conclusion

In conventional MLLM preference learning, a significant amount of information had to be acquired to correct hallucinations in image-related responses. To address this issue and mitigate object hallucinations in MLLMs, we propose TL-DPO, a learning method that utilizes a dataset containing target object information. This approach uses generative data augmentation to selectively alter the correct answer phrases, generating pairs of hallucinated and correct responses along with target information for the target objects. These pairs are used to train MLLMs through the proposed TL-DPO, enabling the model to accurately respond with information about the target objects and to effectively locate the correct target objects. Our extensive research demonstrates the effectiveness of TL-DPO in mitigating various forms of object hallucinations, including hallucinations related to existence and attributes, visual illusions, and hallucinations caused by complex charts. Unlike traditional fine-tuning-based solutions, TL-DPO effectively mitigates hallucinations across diverse vision-language tasks while maintaining or even enhancing performance in general vision-language tasks.

References

[1] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey, 2024. 1

[2] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1

[3] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models, 2024. 1

[4] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding, 2024. 1

[5] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 7

[6] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023. 1

[7] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning, 2024. 1

[8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. 1

[9] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. 2, 3

[10] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. 2

[11] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning, 2019. 1

[12] Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö. Arık, and Tomas Pfister. Data-augmented phrase-level alignment for mitigating object hallucination, 2024. 1

[13] OpenAI Team. Gpt-4o system card, 2024. 1

[14] 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, 2024. 1, 5

[15] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback, 2024.

[16] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness, 2024. 1

[17] Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, Chunyuan Li, and Manling Li. Halle-control: Controlling object hallucination in large multimodal models, 2024. 1

[18] Mengxi Zhang, Wenhao Wu, Yu Lu, Yuxin Song, Kang Rong, Huanjin Yao, Jianbo Zhao, Fanglong Liu, Yifan Sun, Haocheng Feng, and Jingdong Wang. Automated multi-level preference for mllms, 2024. 1