Token Preference Optimization with Self-Calibrated Visual-Anchored Rewards for Hallucination Mitigation

Anonymous ACL submission

Abstract

Direct Preference Optimization (DPO) has been demonstrated to be highly effective in mitigating hallucinations in Large Vision Language Models (LVLMs) by aligning their outputs more closely with human preferences. Despite the recent progress, existing methods suffer from two drawbacks: 1) Lack of scalable token-level rewards; and 2) Neglect of visual-anchored tokens. To this end, we propose a novel Token Preference Optimization model with self-calibrated rewards (dubbed as TPO), which adaptively attends to visual-correlated tokens without fine-grained annotations. Specifically, we introduce a token-level *visual-anchored reward* as the difference of the logistic distributions of generated tokens conditioned on the raw image and the corrupted one. In addition, to highlight the informative visual-anchored tokens, a visual-aware training objective is proposed to enhance more accurate token-level optimization. Extensive experimental results have manifested the state-of-the-art performance of the proposed TPO. For example, by building on top of LLaVA and Qwen, our TPO boosts the performance absolute improvement for hallucination benchmarks.

1 Introduction

Recently, Large Vision Language Models (LVLMs) have showcased their remarkable capabilities in handling multimodal information, excelling in tasks such as image captioning, visual question-answering, and complex visual reasoning (Team et al., 2023; Bai et al., 2023; Hurst et al., 2024; Yang et al., 2023). Specifically, by integrating pre-trained language models with meticulously designed visual encoders, LVLMs are capable of effectively capturing the semantic correlations between visual and textual data. This integration supports more accurate and contextually relevant tasks of visual understanding and generation.

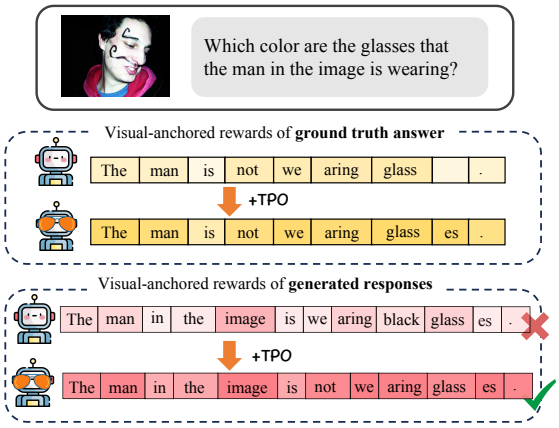

Figure 1: An example of visual Q&A. The upper box contains the ground truth answer, while the lower box shows the LVLM responses before and after training with our method. In each box, we visualize the rewards for each token which can reflect the degree of visual anchoring, with the top representing scores before training and the bottom after. Scoring is detailed in Equation 4, and we've applied sigmoid normalization in this score.

Despite the advancements, the issue of "hallucination", where the generated responses are not grounded in the input visual contexts, greatly impedes the reliability and practical deployment of LVLMs (Liu et al., 2024a; Bai et al., 2024). To alleviate this, various methods have been proposed from the perspectives of data quality (Liu et al., 2023; Zhai et al., 2023) and inference-time strategies (Yin et al., 2023; Zhou et al., 2023; Huang et al., 2024). Recently, direct preference optimization (DPO) (Rafailov et al., 2024) is introduced to align outputs with human preferences, therefore reducing the risk of generating hallucinatory or nonsensical responses.

Existing DPO-like methods, however, still suffer from two drawbacks: 1) *Lack of scalable token-level rewards*. The fine-grained token-level rewards enable precise adjustments to individual parts of generated responses. Existing methods, however, either provide global sentence rewards or rely on

manual efforts for fine-grained segment-level annotations (Yu et al., 2024b). Therefore, designing a scalable token-level reward generation strategy has become a clearly defined necessity (*c.f*. Table 1); 2) *Neglect of visual-anchored tokens:* By "visual-anchored tokens", we refer to response tokens that are essential and highly correlated with the input visual embeddings. RLHF-V assigns all the hallucinated segments with a fixed reward value. Recent studies (Guan et al., 2024) attribute the hallucination issue to an inherent imbalance between the visual and textual modalities. Specifically, due to the large-scale pre-trained textual corpus, LVLMs tend to prioritize language-based information even at the costs of overriding the provided visual content. Therefore, we argue that not all the tokens are equal, *i.e*., visual-anchored tokens (*e.g*., `glass` in Figure 1) are more prone to hallucination and deserve great emphasis. As shown in Table 1, the concurrent pre-print V-DPO (Xie et al., 2024) also focuses on visual-anchored tokens; however, it requires the additional construction of a synthetic dataset, whereas our method eliminates the need for any extra annotations.

To alleviate these aforementioned problems, we propose a novel **T**oken **P**reference **O**ptimization with self-calibrated rewards (dubbed as **TPO**), which rectifies the fine-grained token-level hallucinations and attends to visual-anchored tokens without the need of fine-grained annotations. Specifically, to mine the visual-anchored tokens, we compute the differences between the logits distributions of generated tokens conditioned on the raw image and the corrupted one. We regard this distribution difference as token-wise rewards. In Figure 1, we apply this visual-anchored score mining strategy on both golden truth and the generated responses. As shown, this strategy effectively helps highlight visual-anchored tokens. Then, we propose a token preference optimization loss by integrating the self-calibrated rewards into the vanilla DPO. In particular, we multiply the like-hood distribution with token-wise rewards to generate our desired visual-correlated ones.

Overall, the main contributions of this work are:

- We propose TPO for hallucination mitigation in LVLMs, which implements token-level distribution rectification without the reliance of fine-grained manual annotations.
- We mine visual-anchored tokens by comparing the response distributions conditioned on the raw image and the corrupted one.
- Extensive experiments on the popular hallucination benchmarks demonstrate the state-of-the-art performance of the proposed TPO.

Methods	Visual-Anchored	Token-level	Non Fine-grained Annotations
DPO	✗	✗	✓
POVID	✗	✗	✓
CSR	✓	✗	✓
MDPO	✓	✗	✓
V-DPO	✓	✓	✗
RLHF-V	✗	✓	✗
TPO (Ours)	✓	✓	✓

Table 1: Comparisons with hallucination mitigation methods from the perspective of whether attending to vision-anchored tokens, whether generating token-level rewards and whether requiring fine-grained annotations. The compared methods include DPO (Rafailov et al., 2024), POVID (Zhou et al., 2024a), CSR (Zhou et al., 2024b), MDPO (Wang et al., 2024a), V-DPO (Xie et al., 2024), RLHF-V (Yu et al., 2024b).

2 Related Works

2.1 LVLMs' Hallucination

Leveraging the rich knowledge in large language models and the vision understanding capabilities of vision encoders, LVLMs have shown exceptional performance in image understanding and generation tasks (Li et al., 2023b; Zhu et al., 2023). However, imbalances in parameters and data scale during pre-training can lead to LVLMs being overly influenced by biases in the language model, resulting in inadequate attention to visual information and potential hallucination issues (Zhou et al., 2023; Zhang et al., 2024). Consequently, addressing the issue of hallucinations in LVLMs has become one of the key research focuses in this field.

Previous studies have mitigated hallucinations by enhancing training data quality, refining decoding strategies, and post-processing generated responses (Huang et al., 2024; Leng et al., 2024; Yu et al., 2024a; Han et al., 2024; Chen et al., 2024; Zhou et al., 2023; Yin et al., 2023; Lee et al., 2023; Shao et al., 2024a; Jiang et al., 2024; Yue et al., 2024; Xiao et al., 2025; Sarkar et al., 2024; Zhao et al., 2023). While these methods can lead to more accurate responses, they do not fundamentally resolve the issue of inadequate visual information association in LVLMs.

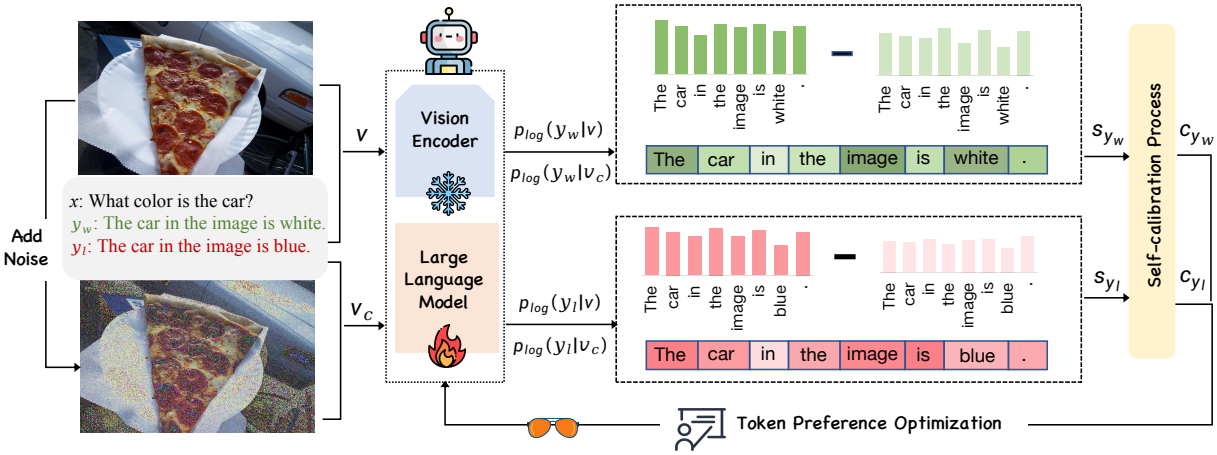

Figure 2: Outline of our TPO pipline. The process is divided into three parts for each data at every training step. First, 1) add noise to the image, then, 2) calculate Self-Calibrated Visual-Anchored Rewards, and finally 3) perform Token Preference Optimization. At the end of each training step, we calibrate the model and calculate new Visual-Anchored Rewards for the next step.

2.2 Preference Learning Methods

More recently, reinforcement learning from human feedback (RLHF) (Sun et al., 2023) is gradually becoming a prevalent approach to mitigate the hallucination. As a more direct and effective method, DPO (Rafailov et al., 2024) and its variants are more widely utilized for preference alignment.

Several studies based on DPO focus on developing more robustly constructed preference data. For example, the POVID (Zhou et al., 2024a) method constructs negative samples for preferred data by adding noise to the image and providing hallucinated patterns to guide the model to generate hallucinated responses. The MDPO incorporate optimization for image preference, training with the images before and after alteration as positive and negative samples. Apart from these works, RLAIF (Yu et al., 2024c) and CSR (Zhou et al., 2024b) methods, which are built upon on-policy DPO strategy, construct preference pairs by iteratively performing self-rewarding to select preference pairs. R1-Onevision (Yang et al., 2025) enhances the visual reasoning capabilities by employing Group Relative Policy Optimization (GRPO). However, assigning response-level rewards for each generated sequence is insufficient for effectively aligning with genuinely hallucination-prone contents.

Other studies, RLHF-V (Yu et al., 2024b) and V-DPO (Xie et al., 2024), investigated this issue and achieved more fine-grained alignment of preference data. Nevertheless, this approach depends on resource-intensive annotations or data constructions and applies a fixed reward to all hallucinated segments, thus failing to account for the differing levels of relevance these segments may have to visual information. It is worth mentioning that CSR also considered this problem and introduced CLIP (Radford et al., 2021) to calculate the relevance score between generated text and vision information as an additional reward, and TLDR (Fu et al., 2024) score each token by training a scoring model. However, these methods requires the introduction of an additional model, which reduces the training efficiency.

In this paper, we propose a token-level preference optimization method with self-calibrated visual-anchored rewards (TPO), aimed at addressing the aforementioned challenges. TPO facilitates finer-grained alignment in LVLMs, enhancing accuracy in visual information correlation and reducing hallucinations during response generation.

3 Methodology

The schematic illustration of the proposed TPO is demonstrated in Figure 2. In Sec. 3.1, we present the preliminaries including the definition and off-policy optimization of DPO. Then we detail the visual-anchored rewards and token preference optimization loss in Sec. 3.2 and Sec. 3.3, respectively.

3.1 Preliminaries

DPO (Rafailov et al., 2024) is designed to directly maximize the reward margin between positive and negative responses to align human preferences. Given a textual input x, a visual input v, a negative response y_l, and a preferred positive response y_w, the reward function $r(x, v, y_l/y_w)$ is defined

as follows.

$$r(x, v, y) = \beta \log \frac{\pi_\theta(y|x, v)}{\pi_{\text{ref}}(y|x, v)}, \tag{1}$$

where $\pi_{\text{ref}}(y|x, v)$ and $\pi_\theta(y|x, v)$ respectively represent the reference model and current policy model. On this basis, the formulation of a maximum likelihood objective is defined as:

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,v,y_w,y_l)\sim D}\Big[\log \sigma \Big(\beta \log \frac{\pi_\theta(y_w|x, v)}{\pi_{\text{ref}}(y_w|x, v)} - \beta \log \frac{\pi_\theta(y_l|x, v)}{\pi_{\text{ref}}(y_l|x, v)}\Big)\Big], \tag{2}$$

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

3.2 Visual-Anchored Rewards

Different to the equal confidence for each token in DPO, we propose a visual-anchored by measuring the token-wise visual reliance. Specifically, we firstly add noise into the embedding of the input image v in a total k steps to obtain the corrupted image v_c:

$$v_c(k) = \sqrt{\bar{\xi}_k} \cdot v + \sqrt{1 - \bar{\xi}_k} \cdot \epsilon, \tag{3}$$

where ξ is a predefined noise parameter derived from a list with 1,000 equally spaced elements[1]. $\bar{\xi}_k$ is a cumulant, *i.e.*, $\bar{\xi}_k = \prod_{i=0}^{k} \xi_i$.

Subsequently, the difference of generated token distribution is computed as follow:

$$s_{y_i} = p_{log}(y_i|x, v, y_{<i}) - p_{log}(y_i|x, v_c, y_{<i}), \tag{4}$$

where s_{y_i} denotes the distribution difference of the token y_i of the response y. p_{log} refers to the raw logits output of the model, before applying `softmax` normalization. One example case is demonstrated in Figure 1, which demonstrates that s reflects the visual relevance of each token y_i.

Then, a self-calibration process is proposed to generate the final visual-anchored rewards c_{y_i}.

$$c_{y_i} = \begin{cases} a + \sigma(s_{y_i}) & \text{if } y_i \in y_w \\ a + 1 - \sigma(s_{y_i}) & \text{if } y_i \in y_l \end{cases} \tag{5}$$

where a is a margin value. We set $a = 0.5$ in Equation (5), so that when $s = 0$, $c = 1$, the rewards will not take effect. This process aims to ensure that positive samples receive higher rewards than negative samples while optimizing the visual relevance of visual-anchored tokens in all responses.

[1]More details can be found in Appendix A, and experimental analysis can be found in Appendix E

3.3 Token Preference Optimization

After obtaining the reward c_{y_i} to y_i, the output cumulative distribution can be calculated:

$$\pi^v(y|x, v) = \prod_{y_i \in \mathcal{Y}} c_{y_i} \tag{6}$$

Especially, when $c_{y_i} = 1$, the probability of y_i will not be accumulated. By multiplying the probability distribution with the visual-anchored rewards, we obtain a novel KL-constrained reward maximization objective:

$$\max_\pi E_{(x,v,y)}\Big[r'(x, v, y) - \beta D_{KL}\Big(\pi_\theta(y|x, v) \cdot \pi^v_\theta(y|x, v), \pi_{\text{ref}}(y|x, v) \cdot \pi^v_{\text{ref}}(y|x, v)\Big)\Big], \tag{7}$$

where $D_{KL}(\cdot, \cdot)$ denotes the KL divergency computation. $\pi^v_\theta(y|x, v)$ and $\pi^v_{\text{ref}}(y|x, v)$ are calculated using the policy model and the reference model, respectively. Thus, the optimal solution formula for the maximization objective of the KL-constrained reward is as follows:

$$\pi_\theta(y|x, v) \cdot \pi^v_\theta(y|x, v) = \frac{1}{Z(x, v)} \pi_{\text{ref}}(y|x, v) \cdot \pi^v_{\text{ref}}(y|x, v) \exp\big(\frac{1}{\beta} r'(x, v, y)\big). \tag{8}$$

The partition function of Eq (8) is as follows.

$$Z(x, v) = \sum_y \pi_{\text{ref}}(y|v, x) \cdot \pi^v_{\text{ref}}(y|x, v) \cdot \exp\big(\frac{1}{\beta} r'(x, v, y)\big) \tag{9}$$

Rearranging Eq (8), we obtain the reward function:

$$\begin{aligned} & r'(x, v, y) \\ &= \beta \log \frac{\pi_\theta(y|x, v) \cdot \pi^v_\theta(y|x, v)}{\pi_{\text{ref}}(y|x, v) \cdot \pi^v_{\text{ref}}(y|x, v)} + \beta Z(x, v) \\ &= \beta \sum_{y_i \in y} \Big[\log \Big(p_\theta(y_i|x, v, y_{<i}) \cdot c^\theta_{y_i}\Big) \\ &\quad - \log \big(p_{\text{ref}}(y_i|x, v, y_{<i}) \cdot c^{\text{ref}}_{y_i}\big)\Big] + \beta Z(x, v) \\ &= \beta \sum_{y_i \in y} \Big[\log p_\theta(y_i|x, v, y_{<i}) - \log p_{\text{ref}}(y_i|x, v, y_{<i}) \\ &\quad + \log \frac{c^\theta_{y_i}}{c^{\text{ref}}_{y_i}}\Big] + \beta Z(x, v), \end{aligned} \tag{10}$$

where $c^\theta_{y_i}$ and $c^{\text{ref}}_{y_i}$ represent the token reward calculated using the policy model and the reference model, respectively.

Compared to the original reward function in DPO (Eq (1)), we multiply each $p(y_i|x, v, y_{<i})$ by the generated visual-anchored rewards c_{y_i} at the token level. $c^{\theta}_{y_i}$ is continuously updated at each step during training as the model changes. To calculate each token in the entire reward function, we add a term $log\frac{c^{\theta}_{y_i}}{c^{\text{ref}}_{y_i}} \in (-log3, log3)$ (as we set $a = 0.5$ in Equation (5)), which has a reasonable upper and lower bound. For positive samples, this term is expected to increase, while for negative samples, it is expected to decrease. Due to the different methods of calculating c_{y_i} that we set in Eq (5), this will encourage the increase of s_{y_i} during the training process, making the token generation focus more on visual information.

Thus, following the Bradley-Terry model, when given the positive and negative samples $\mathcal{D} = \{x^{(k)}, v^{(k)}, y_w^{(k)}, y_l^{(k)}\}_{k=1}^{N}$, we obtain our maximum likelihood objective:

$$\begin{aligned}
&\mathcal{L}_{TPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,v,y_w,y_l)\sim D}\Big[\log\sigma \\
&\Big(\beta\log\frac{\pi_\theta(y_w|x,v)\cdot\pi_\theta^v(y_w|x,v)}{\pi_{\text{ref}}(y_w|x,v)\cdot\pi_{\text{ref}}^v(y_w|x,v)} - \\
&\beta\log\frac{\pi_\theta(y_l|x,v)\cdot\pi_\theta^v(y_l|x,v)}{\pi_{\text{ref}}(y_l|x,v)\cdot\pi_{\text{ref}}^v(y_l|x,v)}\Big)\Big] \\
&= \mathcal{L}_{DPO}(\pi_\theta; \pi_{\text{ref}}) + \mathbb{E}_{(x,v,y_w,y_l)\sim D}\Big[\log\sigma \\
&\Big(\beta\log\frac{\pi_\theta^v(y_w|x,v)}{\pi_{\text{ref}}^v(y_w|x,v)} - \beta\log\frac{\pi_\theta^v(y_l|x,v)}{\pi_{\text{ref}}^v(y_l|x,v)}\Big)\Big]
\end{aligned} \tag{11}$$

According to Eq (10), we can deduce as follows.

$$\begin{aligned}
&\mathcal{L}_{TPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,v,y_w,y_l)\sim D}\Big[\log\sigma \\
&\Big(\beta\sum_{y_{w_i}\in y_w}\Big[\log\left(p_\theta(y_{w_i}|x,v,y_{w_{<i}})\right) \\
&-\log p_{\text{ref}}(y_{w_i}|x,v,y_{w_{<i}}) + \log\frac{c^{\theta}_{y_{w_i}}}{c^{\text{ref}}_{y_{w_i}}}\Big] \\
&+\sum_{y_{l_i}\in y_l}\Big[\log\left(p_\theta(y_{l_i}|x,v,y_{l_{<i}})\right) \\
&-\log p_{\text{ref}}(y_{l_i}|x,v,y_{l_{<i}}) + \log\frac{c^{\theta}_{y_{l_i}}}{c^{\text{ref}}_{y_{l_i}}}\Big]\Big)\Big]
\end{aligned} \tag{12}$$

where $c^{\theta}_{y_{w_i}}$ and $c^{\text{ref}}_{y_{w_i}}$ represent the token reward calculated for y_w using the policy model and the reference model, respectively. The same applies to $c^{\theta}_{y_{w_i}}$, $c^{\text{ref}}_{y_{w_i}}$ and y_l.

4 Experiment

4.1 Setup

Aligning with previous DPO-based approaches on hallucination mitigation, we mainly adopt the popular LVLM, LLaVA-1.5 (Liu et al., 2024b), as the backbone model to validate the effectiveness of our TPO. Furthermore, to evaluate the effectiveness of TPO on more advanced and powerful model, we implement TPO training based on Qwen2-VL (Wang et al., 2024b), and compare it with the DPO method. For the dataset, we directly utilize the preference pairs provided by RLHF-V (5K) without their fine-grained human annotations.

Benchmarks We primarily conduct the experiments on three hallucination benchmarks: AMBER (Wang et al., 2023), MMHal-Bench (Sun et al., 2023), and HallusionBench (Guan et al., 2024). In this section, we mainly focus on AMBER's discriminative task and report the accuracy and F1 metrics referencing (Yu et al., 2024c). In addition, we provide the results of its Chair metric in Appendix D. Moreover, we also evaluate the performance of TPO on four general benchmarks: SEED Bench (Li et al., 2023a), MMBench (Liu et al., 2025), LLaVA Bench (Liu et al., 2024c) and MM-Vet (Yu et al., 2023). These benchmarks are used to evaluate the performance of the models on general tasks after hallucination alignment.

Baselines We mainly compare TPO with the R1-Onevision (Yang et al., 2025), LLaVA-1.5-7B SFT model, as well as with the DPO and V-DPO (Xie et al., 2024) methods trained using RLHF-V (Yu et al., 2024b) data, along with two improved methods, CSR (Zhou et al., 2024b) and POVID (Zhou et al., 2024a). Moreover, to evaluate the effectiveness and robustness of TPO as the model size increases, we further evaluate the performance of TPO on the LLaVA-1.5-13B model and compared it with DPO. Additionally, to demonstrate the advantages of TPO, we reproduced the strong baseline method, RLHF-V, on LLaVA-1.5-13B and conducted a comparison. Furthermore, we additionally employ Qwen2-VL-7B (Wang et al., 2024b) as the baseline mode and compare our TPO with DPO.

4.2 Main Results

In Table 2, we present the main results of our TPO and baselines. In hallucination benchmarks, our method shows significant improvements over all

Method	AMBER		MMHal		HallusionBench			General Benchmarks			
	Acc	F1	Score	Hal ↓	Easy	Hard	aAcc	SEED	MMB	LLaVA	MMVet
R1-Onevision	80.2	85.7	3.85	36.46	63.74	50.47	62.80	35.2	–	83.7	67.8
LLaVA-1.5-7B	71.7	74.3	2.01	61.46	42.64	41.16	47.21	66.1	73.3	65.6	31.6
+ DPO	77.5	82.1	2.14	58.33	37.36	37.21	43.84	66.4	73.3	69.1	31.6
+ CSR	73.2	76.1	2.05	60.42	**43.08**	41.16	47.48	65.9	73.0	68.9	31.0
+ POVID	71.9	74.7	2.26	55.21	42.86	41.63	47.56	66.1	73.2	68.2	31.7
+ RLHF-V	74.8	78.5	2.02	60.42	42.20	43.72	48.27	66.1	73.1	68.0	32.3
+ MDPO	–	–	2.39	54.00	–	–	–	–	–	–	–
+ V-DPO	–	81.6	2.16	56.00	–	–	**51.63**	–	–	–	–
+ TPO (Ours)	**79.3**	**85.0**	**2.47**	**51.04**	41.76	**48.37**	50.22	**66.6**	**73.6**	**70.2**	**33.0**
LLaVA-1.5-13B	71.3	73.1	2.38	53.13	44.40	36.51	46.94	68.2	76.7	73.1	36.1
+ DPO	83.2	86.9	2.47	51.04	**45.49**	43.49	50.22	68.6	76.6	72.8	37.5
+ RLHF-V	79.2	82.3	2.50	52.08	43.96	40.00	48.27	68.2	76.7	**76.7**	**38.5**
+ TPO (Ours)	**83.9**	**88.0**	**2.72**	**45.83**	44.40	**46.05**	**50.93**	**68.7**	**76.8**	72.8	36.2
Qwen2-VL-7B	86.5	**90.0**	3.5	29.0	67.0	48.8	64.0	**45.0**	**79.0**	82.4	61.4
+ DPO	86.5	**90.0**	3.7	28.1	67.3	49.3	64.5	**45.0**	**79.0**	81.9	60.2
+ TPO (Ours)	86.4	89.9	**4.2**	**18.8**	**67.9**	**50.0**	**65.2**	**45.0**	**79.0**	**82.9**	**61.4**

Table 2: Performence of LLaVA-1.5 on hallucination and general benchmarks. Score and Hall refer to the overall GPT-4 (Achiam et al., 2023) score and hallucination rate, respectively. Easy represents the accuracy of with original images, hard represents the accuracy with manually edited challenging images, and aAcc is the average accuracy for each question. The results for POVID (Zhou et al., 2024a) and CSR (Zhou et al., 2024b) are based on our testing of their open-source model weights, while the results for V-DPO (Xie et al., 2024), MDPO (Wang et al., 2024a) are taken from previous work

.

previous methods for both the 7B and 13B models, surpassing even the GRPO-based (Shao et al., 2024b) R1-Onevision model. Specifically, compared to the original LLaVA model, we achieve improvements of 20.4 % on AMBER F1, 22.8% on the MMHAL score, and 8.5% on HallusionBench aAcc at most. This validates the effectiveness of our method in helping the model mitigate hallucination issues and improve the performance of visual question answering.

Notably, on the HallusionBench evaluation metrics, "Easy" represents the accuracy of original image-based questions, which tend to rely on prior knowledge, while "Hard" represents the accuracy of questions based on manually edited images, which tend to rely on visual information. Our method leads to the most significant improvement for the original model on hard questions. This indicates that our approach enables the model to focus more on visual information rather than textual prior knowledge to provide accurate answers.

In general benchmarks, our approach remains stable against the backbone models and achieves improvement on most benchmarks. We attribute it to our method helping the model associate with more visual information when answering questions. This shows that our approach can improve hallucination issues while maintaining good performance in general evaluation tasks.

4.3 Results on Qwen2-VL-7B

As Table 2 shown, we report the results on the key metrics of three hallucination benchmarks. The results indicate that our TPO outperforms DPO on most benchmarks. On Qwen2-VL-7B, which has strong inherent capabilities, using 5K RLHF-V data for DPO alignment barely improves the performance. However, introducing TPO leads to a significant further enhancement. This demonstrates that TPO can capture and learn more subtle preferences from the data and brings higher data utilization efficiency. In addition, the results on the chair metric in Figure 6 further demonstrate that TPO can also significantly solve the object hallucination problem of Qwen2-VL-7B.

4.4 Ablation Studies

Visual-Anchored Rewards Table 3 demonstrates that TPO can enhance model performance when rewards are assigned separately to positive and negative samples, achieving results comparable to those obtained by rewarding both simultaneously. However, by providing opposite rewards to positive and negative samples, where rewards are negatively correlated with the visual relevance of

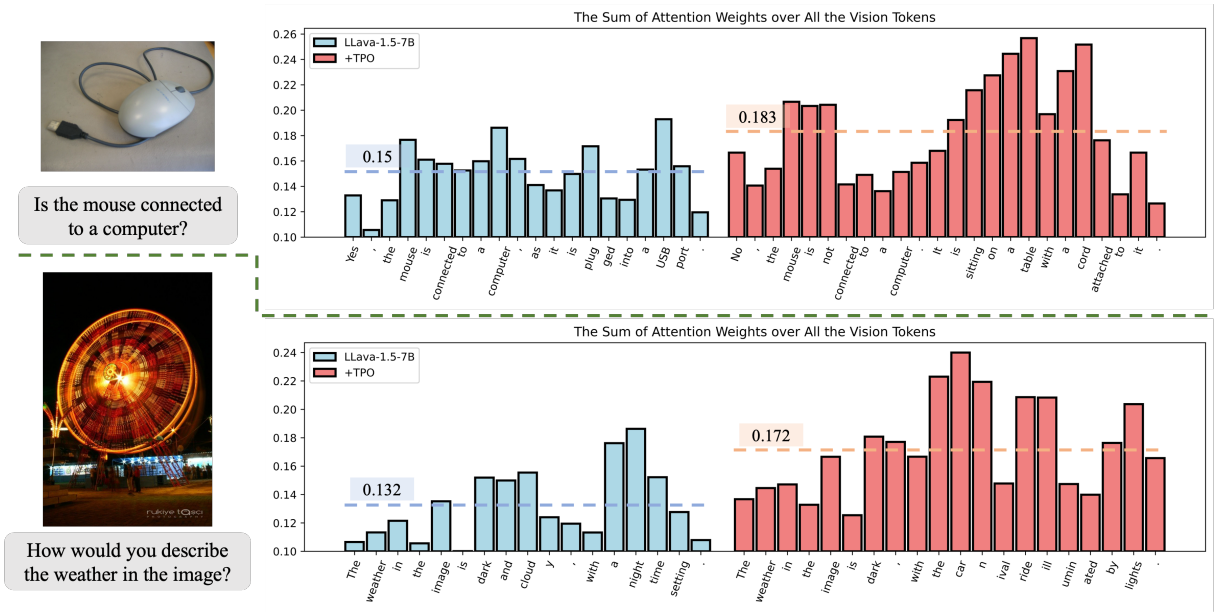

Figure 3: Comparison of attention weights for LLaVA before and after TPO training. Each horizontal line represents the mean of that data. The blue section response incorrectly, with many 'visual-anchored tokens' tokens having high attention weights but resulting in hallucinated responses (*e.g.* USB). The red section on the right answered correctly.

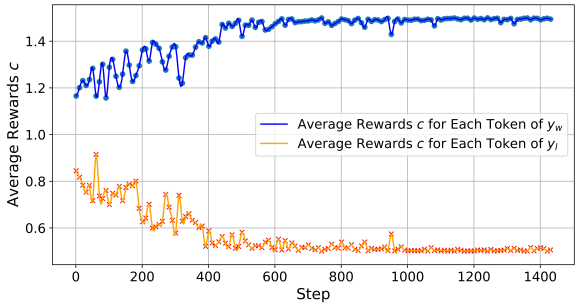

Figure 4: The curve of changes in self-calibrated rewards for positive and negative samples over training steps, with a sample point taken every 10 steps.

positive samples and positively correlated with that of negative samples, TPO's performance significantly deteriorates. In some metrics, this approach yields even poorer results than the original LLaVA-1.5 model. This further underscores the validity of the designation of visual-anchored rewards.

Hyperparameters To optimize the hyperparameters in TPO, we perform comparative experiments on the noise steps (Section 3.2) and parameter a in Equation 5. As the Figure 5 shown, the model performs best with 500 noise steps. Testing a across the range $a = [0, 0.5, 1]$, we find that a=0.5 produces the best outcome. This supports our hypothesis that setting $s = 0$ and $c = 1$, without adding extra reward signals, leads to superior performance[2].

[2]The more detail results and discussions are provided in Appendix.

4.5 Analysis

Visual-Anchored Rewards As Figure 1 shown, the proposed visual-anchored rewards can reflect the degree to which a token depends on visual information. To further prove this statement, we construct the analysis experiment on the MMhal dataset as shown in Table 4. Intuitively, nouns and adjectives in responses are thought to most associate the content of an image. Therefore, we first perform part-of-speech (POS) tagging on the model responses and count the average number of noun/adjective tokens and other types of tokens. Specifically, in the ground-truth responses, 39.6% of the tokens are nouns or adjectives. In the responses from LLAVA-1.5-7B, the proportion of noun and adjective tokens remains nearly constant at 39.2%, both before and after TPO.

Afterwards, we count the average score from Equation 4 of noun/adjective tokens and other types of tokens. The results show that noun and adjective tokens have significantly higher scores than other types, indicating higher relevance to images. After applying TPO, these scores of all the tokens increased notably. The results supports our conclusions: 1) The visual-anchored rewards reflects token-image relevance. 2) TPO enhances the alignment of generated tokens with image content.

Attentions To further validate TPO's effectiveness in enhancing visual alignment, we measure the relevance using the sum of attention weights

Method	AMBER		MMHal		HallusionBench			General Benchmarks			
	Acc	F1	Score	Hal ↓	Easy	Hard	aAcc	SEED	MMB	LLaVA	MM-Vet
LLaVA-1.5-7B	71.70	74.3	2.01	61.46	42.64	41.16	47.21	66.1	73.3	65.6	31.6
Only Win	79.10	84.5	2.24	56.25	**44.62**	46.05	**50.40**	66.6	73.6	69.8	31.7
Only Loss	79.20	84.8	2.33	53.13	42.20	47.91	49.87	66.6	73.5	**70.7**	32.0
Opposite	75.30	80.7	1.91	64.58	42.42	45.58	48.63	65.6	73.1	68.9	32.1
TPO (Ours)	**79.30**	**85.0**	**2.47**	**51.04**	41.76	**48.37**	50.22	**66.6**	**73.6**	70.2	**33.0**

Table 3: Ablation Studies. Performence of LLaVA-1.5 on hallucination and general benchmarks.

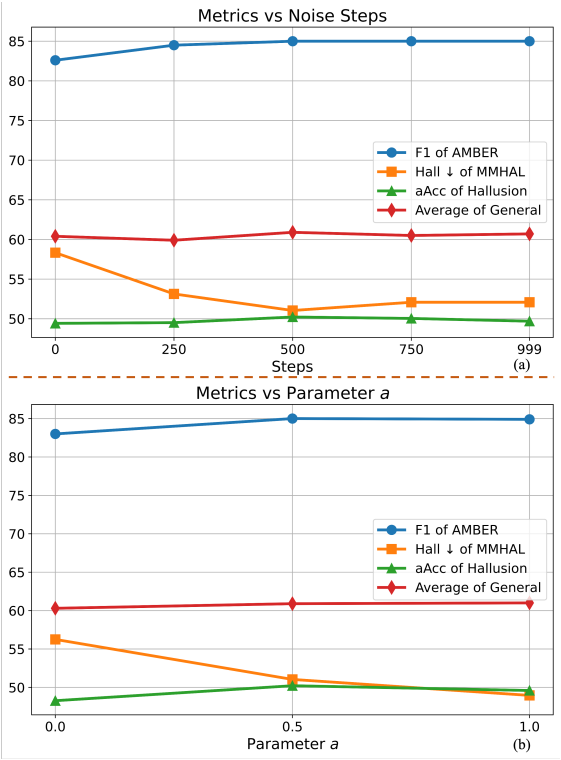

Figure 5: Performance curves with the change of the noise steps-(a) and the change of parameter a-(b), We separately present the F1 of AMBER, the hallucination rate of MMHAL, the aACC of HallusionBench, and the average value of the general benchmarks. More detailed metrics can be found in the Appendix B.

between responses and images. On the MMHal dataset, the overall image attention weights for LLaVA-1.5-7B increased from **0.14** before TPO training to **0.17** afterward. Additionally, Figure 3 visualizes the cases, showing a significant increase in image attention weights for response tokens, especially for visual-anchored tokens (e.g., `table`, `cord`). This highlights our method's success in improving the model's integration of visual information, thus reducing hallucinations.

Self-Calibration To illustrate that our method enables the model to progressively enhance its focus on visual information through continuous self-calibration during training, we present the evolution of scores for positive and negative samples, as calculated by Equation (5), across various training steps. With $a = 0.5$, it follows that $c_{y_i} \in (0.5, 1.5)$. As shown in Figure 4, the scores for positive samples gradually approach their maximum values, while those for negative samples approach their minimum values, indicating convergence. This trend illustrates the self-calibrating effect of our method, which ultimately enhances the model's ability to focus on visual information.

Average score	Noun/Adj	Others
Ground Truth	1.83	0.90
Ground Truth (TPO)	**5.72**	4.87
Response of LLaVA	1.48	0.83
Response of LLaVA+TPO (TPO)	**5.67**	4.59

Table 4: Average score from Equation 4 of Noun/Adj tokens and other tokens. Here, Ground Truth and Ground Truth (TPO) represent the scores calculated for the ground truth answer using LLaVA-1.5-7B and LLaVA-1.5-7B+TPO. Response of LLaVA and LLaVA+TPO (TPO) correspond to the outputs before and after TPO training and the scores calculated by LLaVA-1.5-7B and LLaVA-1.5-7B+TPO, corresponding to Fiure 1.

5 Conclusion

In this study, we propose a novel pereference alignment method, TPO, to mitigation hallucinations in LVLMs. TPO incorporates a self-calibrated visual-anchored reward mechanism that automatically identifies "vision-anchored tokens" and adaptively assigns appropriate rewards to them. By adding noise to the visual input and capturing changes in the generation probability of each token, TPO computes a score indicating each token's relevance to visual information. Based on the self-calibrated visual-anchored reward, TPO can perform more efficient token-level preference alignment optimization for LVLMs. Experimental results have proved that TPO not only alleviates the hallucination problem but also strengthens the model's attention to visual input when generating responses.

6 Limitation

Although our method has achieved outstanding performance in addressing the hallucination problem, the self-calibrated visual-anchored rewards approach we used in this paper can be extended to even broader areas. By altering the way noise is added to images, we can shift from adding noise to the entire image to adding noise to specific key objects. It can enable the model to specifically improve its focus on image information in certain domains, thus having extensive industrial applications. Besides, we believe that the core part of the TPO method, the visual-anchored reward scoring method, possesses strong extensibility. For example, these token-level rewards can also be used to weight the probability distributions in the calculations for other RLHF methods, enhancing the visual attention of multimodal models.

We will continue to expand in this direction, and we believe that the technology we have proposed in this paper has a vast space for further development and application.

7 Ethic Statement

The main purpose of this article is to alleviate the hallucination problem in LVLM using reinforcement learning method. By employing a self-calibrated visual-anchored reward approach, we propose the TPO method, which significantly addresses the hallucination issue and helps the model connect with more visual information. All the models and datasets we used are open source, so we believe that the work in this paper does not pose any potential threats.

References

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*.

Deqing Fu, Tong Xiao, Rui Wang, Wang Zhu, Pengchuan Zhang, Guan Pang, Robin Jia, and Lawrence Chen. 2024. Tldr: Token-level detective reward model for large vision language models. *arXiv preprint arXiv:2410.04734*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Zongbo Han, Zechen Bai, Haiyang Mei, Qianli Xu, Changqing Zhang, and Mike Zheng Shou. 2024. Skip\n: A simple method to reduce hallucination in large vision-language models. *arXiv preprint arXiv:2402.01345*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2023. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. 2024. Mitigating object hallucination via data augmented contrastive tuning. *arXiv e-prints*, pages arXiv–2405.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024a. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. 2025. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25543–25551.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. *arXiv preprint arXiv:2411.02712*.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2024a. Hallucidoctor: Mitigating

hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024b. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024c. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Zihao Yue, Liang Zhang, and Qin Jin. 2024. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, and Manling Li. 2023. Halle-switch: Controlling object hallucination in large vision language models. *arXiv e-prints*, pages arXiv–2310.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. 2024. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*.

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

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024a. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. 2024b. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Implement Details

A.1 Setup

In our experiments, we trained the LLaVA-v1.5 model. For our TPO method and the vanilla DPO method, we set the maximum learning rate to 5e-8 on the 7B version and trained for 4 epochs. We set the maximum learning rate to 2e-7 on the 13B version and trained for 4 epochs. The RLHF-V training was set according to the paper (Yu et al., 2024b). All parts requiring GPT-4 evaluation use the GPT-4-0613 8K version, and the MM-Vet testing is conducted on the official evaluation website.

During the training process, we froze the vision encoder and only trained the LLM.

We also trained the Qwen2-VL model. For our TPO method and the vanilla DPO method, we set the maximum learning rate to 5e-9 for 7B model, 1e-9 for 2B model and trained for 4 epochs.

For a fair comparison, we set the seed to 42 during training and greedy decoding was used during inference.

Our experiments were all conducted on a server equipped with 8 Nivdia A100 GPUs; in specific cases (such as the 13B model), we utilized 32 Nivdia A100 GPUs. For the hyperparameter settings, all hyperparameters are consistent with those of our main experiment. Moreover, the level of diffusion noise in our model is represented by a formula $\xi = Sigmoid(l_t) \times (0.5 \times 10^{-2} - 10^{-5}) + 10^{-5}$, where l_t is a list of 1,000 numbers taken at equal intervals over the interval $[-6, 6]$, and $\epsilon \in N(0, 1)$.

The cases in Figure 1 and Figure 3 come from benchmarks (Sun et al., 2023), while the cases in Figure 2 come from the RLHF-V training set (Yu et al., 2024b).

A.2 Benchmarks

The three hallucination benchmarks: (1) AMBER : a multi-dimensional hallucination benchmark with more than 15K samples, including discriminative and description tasks. (2) MMHal-Bench : it measures the hallucination rate and informativeness of responses. (3) HallusionBench : it evaluates visual illusions and knowledge hallucinations through systematically structured discriminative tasks.

Method	AMBER		MMHal		HallusionBench			General Benchmarks			
	Acc	F1	Score	Hal ↓	Easy	Hard	aAcc	SEED	MMB	LLaVA	MM-Vet
LLAVA-1.5-7B	71.7	74.3	2.01	61.46	42.64	41.16	47.21	66.1	73.3	65.6	31.6
0 setp	77.6	82.6	2.10	58.33	**44.40**	45.35	49.42	66.2	73.2	69.9	32.1
250 steps	79.0	84.5	2.33	53.13	43.52	46.05	49.51	66.6	73.4	68.5	31.3
750 steps	79.30	85.0	2.40	52.08	41.76	48.14	50.04	**66.7**	73.5	69.2	32.8
999 steps	79.20	85.0	2.41	52.08	41.76	47.67	49.69	**66.7**	73.5	69.2	**33.3**
500 steps (Ours)	**79.30**	**85.0**	**2.47**	**51.04**	41.76	**48.37**	**50.22**	66.6	**73.6**	**70.2**	33.0

Table 5: Detail of Figure 5 (a).

Method	AMBER		MMHal		HallusionBench			General Benchmarks			
	Acc	F1	Score	Hal ↓	Easy	Hard	aAcc	SEED	MMB	LLaVA	MM-Vet
LLAVA-1.5-7B	71.7	74.3	2.01	61.46	42.64	41.16	47.21	66.1	73.3	65.6	31.6
$a = 0$	79.2	83.0	2.24	56.25	**42.20**	43.72	48.27	66.6	73.5	68.4	32.8
$a = 1$	79.2	84.9	2.44	**48.96**	41.54	47.44	49.60	**66.7**	**73.6**	**70.8**	**33.1**
$a = 0.5$ **(Ours)**	**79.3**	**85.0**	**2.47**	51.04	41.76	**48.37**	**50.22**	66.6	**73.6**	70.2	33.0

Table 6: Detail of Figure 5 (b).

The four general benchmarks: (1) SEED Bench : a benchmark for LVLMs on generative comprehension. (2) MMBench: a comprehensive benchmark designed to evaluate the capabilities across various tasks and modalities. (3) LLaVA Bench: a benchmark for evaluating multi-modal conversation, detailed description, and complex reasoning. (4) MM-Vet: a benchmark to assess integrated capabilities.

A.3 Training Efficiency

In TPO, generating corrupted images at each step incurs almost no time cost, as it is done during the initial data preparation. The main time consumption comes from calculating logits $p_{log}(y_i|x, v_c, y_{<i})$ for the noisy images.

We have also conducted a careful analysis of the time consumption for LLava-1.5-7B under the settings in Section A.1, the training durations for DPO and TPO were 1 hour 24 minutes and 1 hour 57 minutes, respectively, indicating about a 40% increase in time. Nevertheless, all training methods aimed at eliminating hallucinations inevitably incur additional time costs, compared to other methods requiring fine-grained annotations, our self-calibrated approach with 40% time increase proves to be sufficiently efficient.

It has also shown superior outcomes on 5K training data training to CSR training on 13K data and POVID training on 17K data. This highlights the efficacy of our method in guiding the model to pay more attention to image details and in reducing hallucinations. We promise we will further elaborate on our efficiency in detail in the final version.

B Ablation Analysis

Noise Step We ablate on the noise steps in Figure 5 (a). As shown, the optimal performance is achieved at the step of 500. This medium corruption enables the model to grasp the general outline of the image while missing the detailed contents, which is prone to generate hallucinations of the visual-anchored tokens.

The Figure also shows when step=0, TPO still effective and significantly better than DPO. This confusion is a code-implementation issue. In implementation as shown in Listing 1, we first convert the image into a tensor, add noise, and then convert it back into an image. This encode-decode process introduces some losses. Our method of setting the noise step to 0 serves as an ablation experiment to test the impact of this loss on our method, and it allows our experiment to more comprehensively demonstrate the advantages of TPO. The following portion of code may help you better understand our encode-decode process for adding noise. We will also open source all the code once the paper is accepted.

Parameter a We conduct experiments by varying the parameter a introduced in Equation (5) with the results shown in Figure 5 (b). By setting $a = [0, 0.5, 1]$, we observed consistently good performance across all configurations. This suggests

that effective performance is achieved as long as the reward mechanism successfully highlights token differences and identifies visually anchored tokens. Notably, the best overall results are obtained with $a = 0.5$, validating our proposed method and hypothesis. This indicates that when the visual-anchored score $s = 0$, setting $c = 1$, not introducing additional reward signals can yield better outcomes.

```
pil_to_tensor = transforms.ToTensor()
tensor_to_pil = transforms.ToPILImage()
image = Image.open(default_image_path).
    convert("RGB")
image_tensor = pil_to_tensor(image)
image_noisy = add_diffusion_noise(
    image_tensor, 500)
image_noisy = tensor_to_pil(image_noisy)
```

Listing 1: Example Python Code for Noise Addition

C Comparison with Related Methods

To more comprehensively highlight the advantages of the TPO method, we conducted comparisons with other related works (Jiang et al., 2024; Yue et al., 2024; Xiao et al., 2025; Sarkar et al., 2024; Zhao et al., 2023; Leng et al., 2024; Huang et al., 2024; Zhou et al., 2023) aimed at addressing the hallucination problem. The results show that TPO achieves more significant hallucination reduction.

Preference alignment and decoding strategies are two important and parallel categories of methods for hallucination mitigation. We believe that training with preference alignment offers several advantages: 1) Direct Optimization of Output Preferences: This approach directly optimizes the model's output to align with desired preferences without requiring changes to the decoding strategy. 2) Higher Inference Efficiency: Preference alignment typically results in more efficient inference, as it does not introduce additional complexity during the decoding process.

One key advantage of decoding methods is that they do not require retraining the model, making them highly efficient for deployment. However, this does not preclude the benefits of preference alignment. In fact, we believe combining these two approaches can yield even better results.

D Results on Object Hallucination

In the AMBER benchmark, there is a subset for evaluating object hallucinations in image description tasks. Since this paper focuses on visual question answering, this part of the experiment is included in this section. To assess the proportion of object hallucinations in image descriptions, AMBER uses **Chair** as the metric.

Method	AMBER		MMHal	
	Acc	F1	Score	Hal↓
LLaVA-1.5-7B	71.7	74.3	2.01	61.46
VCD	71.8	74.9	2.12	54.20
LURE	73.5	77.7	1.64	60.40
OPERA	75.2	78.3	2.15	54.20
HACL	2.13	50	-	-
EOS	2.03	59	-	-
HA-DPO	1.97	60	75.2	79.9
HALVA	2.25	54	-	83.4
DPO	77.5	82.1	2.14	58.33
TPO	**79.3**	**85**	**2.47**	**51.04**
LLAVA-1.5-13B	2.38	53	71.3	73.1
HSA-DPO	2.61	48	-	-
HALVA	2.58	**45**	-	86.5
DPO	2.47	51	83.2	86.9
TPO	**2.72**	46	**83.9**	**88**

Table 7: Comparison of Results

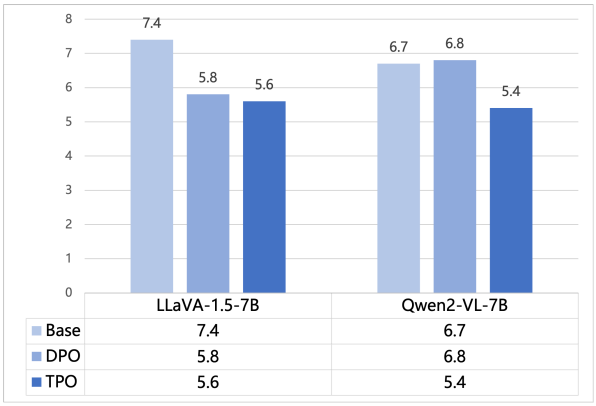

Figure 6: Chair Performance Comparison.

The results are shown in Figure 6. Note that 'Chair' represents the hallucination ratio, where a smaller value indicates better model performance. To more clearly illustrate the comparison between methods in the figure, we use $10 - chair$ as the indicator. The results show that TPO can not only mitigate the hallucination in visual question answering, but also eliminate the object hallucination in image descriptions to a certain extent.

E Comparison of Different Noise Adding Methods.

To evaluate the impact of different methods of adding noise to images on our approach, we test a scheme where noise images were replaced with white images under the same experimental conditions. The results, shown in Table 8, demonstrate the superior performance of our method. We be-

Method	AMBER		MMHal		HallusionBench		
	Acc	F1	Score	Hal↓	Easy	Hard	aAcc
LLaVA-1.5-7B	71.7	74.3	2.01	61.5	42.6	41.2	47.2
+TPO (white)	78.0	82.7	2.26	55.2	**44.2**	45.4	49.3
+TPO	**79.3**	**85.0**	**2.5**	**51.0**	41.8	**48.4**	**50.2**

Table 8: Comparison of different noise adding method. “white" indicates that blank images are used in place of noisy images.

lieve that the noise addition method used in our paper can control noise levels to create images that are more likely to induce hallucinations in the model, thereby achieving better results.