LPOI: Listwise Preference Optimization for Vision Language Models

Fatemeh Pesaran zadeh[1] **Yoojin Oh**[1] **Gunhee Kim**[1]*
[1]Seoul National University,
fatemehpesaran@vision.snu.ac.kr, snujin20@snu.ac.kr
gunhee@snu.ac.kr

Abstract

Aligning large VLMs with human preferences is a challenging task, as methods like RLHF and DPO often overfit to textual information or exacerbate hallucinations. Although augmenting negative image samples partially addresses these pitfalls, no prior work has employed listwise preference optimization for VLMs, due to the complexity and cost of constructing listwise image samples. In this work, we propose LPOI, the first object-aware listwise preference optimization developed for reducing hallucinations in VLMs. LPOI identifies and masks a critical object in the image, and then interpolates the masked region between the positive and negative images to form a sequence of incrementally more complete images. The model is trained to rank these images in ascending order of object visibility, effectively reducing hallucinations while retaining visual fidelity. LPOI requires no extra annotations beyond standard pairwise preference data, as it automatically constructs the ranked lists through object masking and interpolation. Comprehensive experiments on MMHalBench, AMBER, and Object HalBench confirm that LPOI outperforms existing preference optimization methods in reducing hallucinations and enhancing VLM performance. We make the code available at https://github.com/fatemehpesaran310/lpoi.

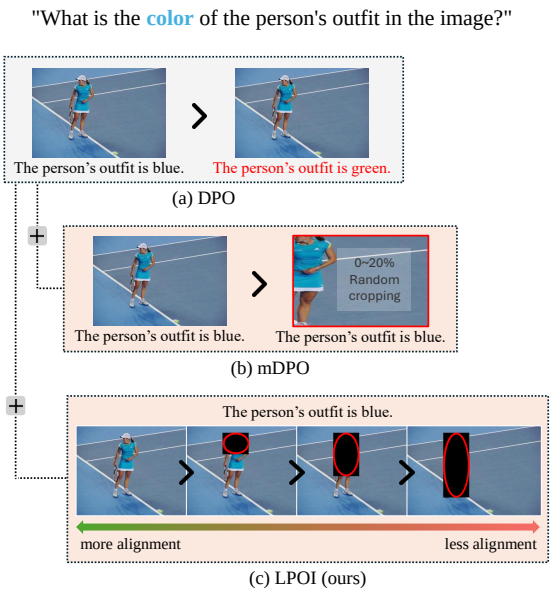

Figure 1: Comparison of preference optimization (PO) strategies for VLMs, with text and image negatives backgrounded in gray and orange, respectively. (a) DPO (Rafailov et al., 2024): PO with text negatives. (b) mDPO (Wang et al., 2024a): DPO + PO using randomly cropped images as binary image negatives. (c) The proposed LPOI method: DPO + listwise PO with ranked image negatives, consisting of four samples: (1) the full image, (2) an image with the partial outfit, (3) an image with no outfit but some parts of person, and (4) an image with neither outfit nor person.

1 Introduction

Aligning large language models (LLMs) or vision language models (VLMs) with human preferences has been an emergent challenge in the field. Approaches like Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022; Glaese et al., 2022; Bai et al., 2022; Stiennon et al., 2022) and Direct Preference Optimization (DPO) (Rafailov et al., 2024; Li et al., 2023a) have increasingly tackled this problem in the text domain. However, adapting these methods to multimodal settings introduces substantial challenges; simply substituting textual preference data with multimodal ones often leads to unreliable results and can even amplify critical issues like hallucinations (Zhao et al., 2024; Yue et al., 2024).

In this regard, a line of research has revealed that multimodal models often overfit to textual information in the preference data, overlooking the necessary information in the image (Wang et al., 2024a; Xie et al., 2024). They propose augment-

* Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 26830–26844

ing the negative samples in the preference data via randomly cropping the image or editing the image using diffusion models.

Meanwhile, recent studies have demonstrated that the methods employing listwise samples for preference optimization often surpass the ones based on pairwise samples by directly optimizing the entire ranking order in the list (Cao et al., 2007a; Wu et al., 2019; Li et al., 2023b). This approach can capture interdependencies among items, unlike pairwise ranking that only compares two items at a time. Although efforts have been made to adapt DPO to listwise ranking in the text domain (Bansal et al., 2024; Liu et al., 2024c; Song et al., 2024; Yuan et al., 2023), applying this to images remains unexplored due to the complexity of ranking visual data and high cost of collecting listwise image samples.

To address this, we propose LPOI (Listwise Preference Optimization via Interpolating between Images), an object-aware listwise preference optimization framework for reducing hallucinations in VLMs. LPOI begins by identifying the critical object in an image based on textual context and creating *hard negative* images by masking this object while keeping overall context. Next, LPOI interpolates the masking ratios between the positive and hard negative images, automatically generating a preference list to be optimized (Figure 2). Finally, the model is trained to rank these interpolated images using a listwise preference loss.

LPOI ranks images by how much of a critical object mentioned in the associated text they reveal (Figure 1). Thus, the model's likelihood of generating positive text about the object increases with its visibility. By aligning the model's output with the object's actual presence, LPOI can lower hallucination rates compared to the state-of-the-art VLM preference optimization approaches. We also employ visual prompting (Shtedritski et al., 2023) to highlight the masked region in each negative example, redirecting the model's focus to the missing object (Figure 4). By efficiently generating diverse image lists without costly annotations or diffusion models, LPOI helps the model learn subtle distinctions between factual and hallucinating text, learning more robust and nuanced representation.

To empirically evaluate LPOI's reduction of hallucination, we fine-tune three VLM models, Idefics-8B (Laurençon et al., 2024), LLaVA-v1.5-7B, and LLaVA-v1.5-13B (Liu et al., 2024a), and assess their performance on the MMHalBench (Sun et al., 2023), AMBER (Wang et al., 2024b), and Object HalBench (Rohrbach et al., 2019). Our experiments demonstrate that preference learning in multimodal model benefits from the use of incrementally ranked listwise negatives, in reducing hallucinations and improving overall model performance.

Our contributions can be outlined as follows.

- We present LPOI, the first approach to apply listwise ranking for VLM preference optimization to reduce hallucinations without requiring additional annotation beyond standard pairwise preference data. This is achieved by masking the image's critical object, and then interpolating the mask ratios between positive and negative images to generate the preference list automatically.
- We evaluate LPOI with three VLM models across three hallucination benchmarks. The results show that LPOI consistently achieves a lower hallucination rate compared to state-of-the-art VLM preference learning methods. Furthermore, LPOI outperforms existing methods in various scenarios, including when trained on datasets of different sizes or compared under a fixed budget of GPU hours.

2 Related Work

Preference Learning. Aligning LLMs or VLMs with human preferences and values, known as preference learning, is an emerging challenge. Reinforcement learning with human feedback (RLHF) typically involves a multi-phase pipeline, including supervised fine-tuning of the policy model, training a reward model, and optimizing the policy based on the reward model (Christiano et al., 2023; Ouyang et al., 2022; Ziegler et al., 2020; Gao et al., 2022; Zadeh et al., 2024). Direct Preference Optimization (DPO) (Rafailov et al., 2024) has emerged as a promising alternative, demonstrating remarkable performance while simplifying the process by eliminating the need for reward model training. Following the DPO, numerous works have been proposed to enhance preference alignment for LLMs (Hong et al., 2024; Xu et al., 2024b; Meng et al., 2024; Xu et al., 2024a).

Preference Learning for VLMs. Several studies have focused on adapting DPO to VLMs, primarily by constructing preference datasets (Xiao et al., 2024; Zhou et al., 2024; Pi et al., 2024; Deng et al.,

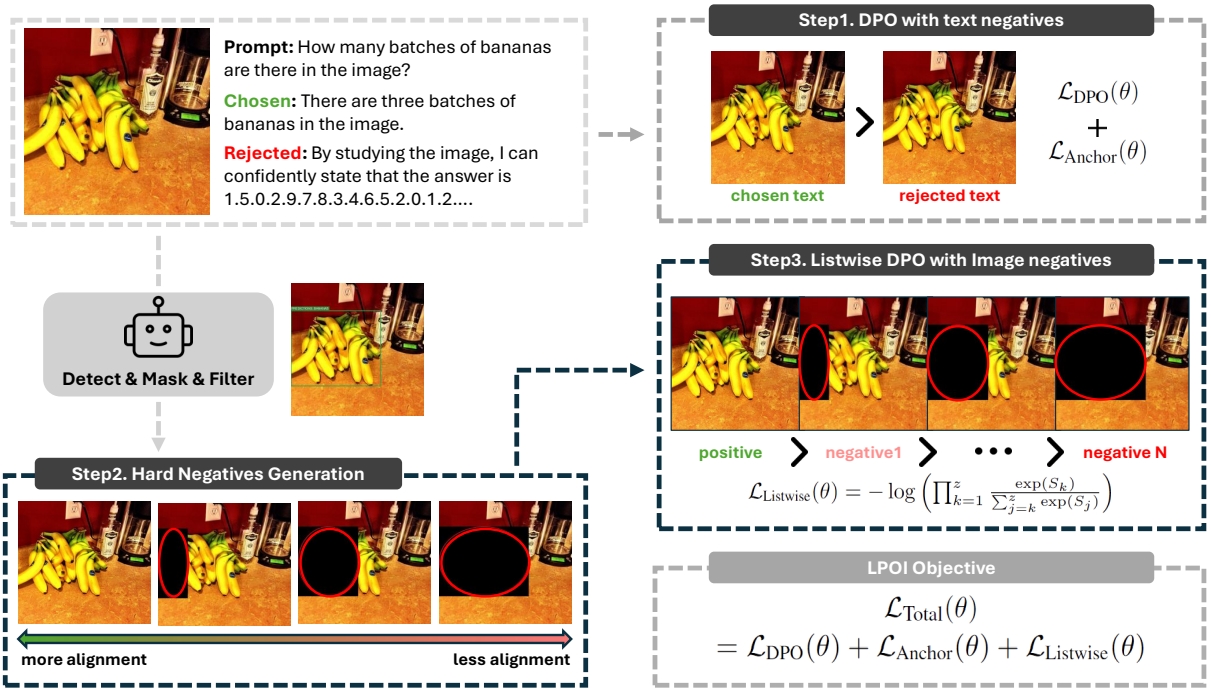

Figure 2: Overview of the LPOI framework. (1) Given an input image, prompt and corresponding set of chosen and rejected responses, we first compute L_{DPO} and L_{Anchor} using the response pairs similar to traditional DPO. (2) An object detection model and a VLM are employed to identify the most important object in the image. These objects are progressively masked in a sequence, with more visual clues being masked as the image deviates further from the positive example. (3) We optimize our model using this sequence of progressively masked images, which allows it to better differentiate between varying levels of hallucination, thereby improving its ability to discern subtle changes in visual context and generate responses more accurately grounded in the image.

2024). Other approaches have explored generating negative images and using them in preference learning, either through random cropping (Wang et al., 2024a) or using computationally expensive diffusion models (Xie et al., 2024). In this work, we propose automatically generating hard negative samples by identifying the critical objects in the image using an object detection module and textual information, and then masking these objects out of the original image.

Hard Negative Mining. Hard negative mining has been extensively explored in deep metric learning and contrastive learning, with techniques like contrastive loss (Hadsell et al., 2006), triplet loss (Schroff et al., 2015), and adaptive sampling (Robinson et al., 2021). They aim to enhance representation learning by identifying challenging negatives that are semantically close to positive samples. In our work, we adapt this principle to create hard negative images by preserving the overall semantic context of the image while masking out the critical object.

Listwise Ranking. Empirical and mathematical studies have shown that listwise ranking is more effective than pairwise ranking (Cao et al., 2007a; Li et al., 2023b; Wu et al., 2019), since it optimizes the entire ranked list simultaneously, considering the relative positions of all items within the list. While prior work has focused on adapting DPO for listwise ranking in text-based applications (Bansal et al., 2024; Liu et al., 2024c; Song et al., 2024; Yuan et al., 2023), adapting listwise ranking in the VLM domain remains underexplored due to the high costs associated with collecting listwise image preference data. Our approach is the first to effectively leverage listwise ranking for VLM preference optimization to reduce hallucinations, without incurring additional annotation costs.

3 Approach

A major challenge in preference learning for VLMs is that models often overfit to textual patterns and overlook the image information (Wang et al., 2024a). This issue can lead to object hallucination (Rohrbach et al., 2019), where the model erroneously describes objects or attributes that do

Algorithm 1 Listwise Preference Optimization via Interpolating between Images (LPOI)

Require: Policy network π_θ, reference policy network π_{ref}, dataset $\mathcal{D}$, parameters N, list size L

1: **for** $i = 1$ to N **do**
2: Sample $(x, q, w, l) \sim \mathcal{D}$ ▷ x: input image, q: question, w: chosen answer, l: rejected answer
3: Calculate $\mathcal{L}_{\text{DPO}}(\theta) = -\log \sigma \left(\beta \log \frac{\pi_\theta(w \mid x,q)}{\pi_{\text{ref}}(w \mid x,q)} - \beta \log \frac{\pi_\theta(l \mid x,q)}{\pi_{\text{ref}}(l \mid x,q)} \right)$
4: Calculate $\mathcal{L}_{\text{Anchor}}(\theta) = -\log \sigma \left(\beta \log \frac{\pi_\theta(w \mid x,q)}{\pi_{\text{ref}}(w \mid x,q)} - \delta \right)$
5: Extract bounding box of the main object b from x, prompt q and chosen answer w.
6: **for** $k = 1$ to L **do** ▷ Create k-th negative sample in the list
7: Define m_k as the mask obtained by masking $\left(\frac{k-1}{L-1}\right) \times 100\%$ of the bounding box b.
8: $x_k = \text{Highlight}(\text{Mask}(x, m_k))$ ▷ Apply masking and visual prompting
9: **end for**
10: **if** filtering model answers (x_L, q, w) to be positive answer **then**
11: Go to Line 5 with different object b
12: **end if**
13: Calculate $\mathcal{L}_{\text{Listwise}}(\theta) = -\log \left(\prod_{k=1}^{z} \frac{\exp(S_k)}{\sum_{j=k}^{z} \exp(S_j)} \right)$ where $S_k = \beta \log \frac{\pi_\theta(w \mid x_k,q)}{\pi_{\text{ref}}(w \mid x_k,q)}$
14: Minimize $\mathcal{L}_{\text{Total}}(\theta) = \mathcal{L}_{\text{DPO}}(\theta) + \mathcal{L}_{\text{Anchor}}(\theta) + \mathcal{L}_{\text{Listwise}}(\theta)$ ▷ Optimize towards $S_1 > S_2 > \cdots > S_L$
15: **end for**

not actually appear in the visual scene; particularly when there are no proper negative image samples during training. In this work, we propose to reduce object hallucination by addressing two key objectives: (1) Generating hard negative image samples, in which the critical object mentioned in the text is missing but the overall context is preserved (Section 3.1). (2) Creating listwise samples without any additional costly annotations, where the images are aligned with the object's actual presence. (Section 3.2).

3.1 Hard Negative Sample Generation

We generate hard negative image samples—images that turn the originally preferred answer into the hallucinated one while preserving the overall semantic context—through two steps of detecting the object to be masked, and applying the mask (Figure 2). First, we run the zero-shot object detection module, Grounding-DINO-Tiny with 172M parameters (Liu et al., 2024b), through the input image. We select the object to be masked in the following orders: objects in the first sentence of the chosen answer, then those in the query, and finally any remaining objects in the answer. We also randomly select a detected object that are not in the text. For the selected object, we mask its bounding box and highlight it using a visual prompting technique (e.g., a red circle) (Shtedritski et al., 2023), directing the model's attention to the masked area.

We then verify that the masked image is indeed a hard negative sample by making sure that Idefics2-8B (Laurençon et al., 2024) hallucinates. If it does not hallucinate, another object is selected, and the process is repeated (Algorithm 1, Lines 5–12).

3.2 Listwise Optimization

We automatically create listwise samples with no annotation by interpolating the masking ratios between the positive image and the hard negative image. Specifically, when generating k-th image in the list, we progressively mask $\frac{k-1}{L-1} \times 100\%$ of the bounding box starting from the side closest to the image edge, where L denotes the list size. As a result, we obtain a list of samples aligned by the visibility, where images with less masking are more positive and those with more masking are more negative.

Once the listwise samples are created, we optimize the model to have higher likelihood of generating positive response according to the order of the list. This is achieved by using a listwise ranking loss, which can be interpreted as the negative log-likelihood of a given permutation (Cao et al., 2007b; Rafailov et al., 2024; Liu et al., 2024c):

$$\mathcal{L}_{\text{Listwise}}(\theta) = -\log \left(\prod_{k=1}^{z} \frac{\exp(S_k)}{\sum_{j=k}^{z} \exp(S_j)} \right), \tag{1}$$

where $S_k = \beta \log \frac{\pi_\theta(w \mid x_k,q)}{\pi_{\text{ref}}(w \mid x_k,q)}$. Here, π_θ and π_{ref} denote the fine-tuned model and the base model, respectively. S_k is the normalized log-likelihood of the model π_θ describing the relevant object given the image x_k. x_1 is the original image, x_L is the hard negative image, and x_k is the interpolated image with the masking ratio of $\frac{k-1}{L-1} \times 100\%$.

By minimizing the listwise loss in eq. (1), we optimize the values of S_k to be $S_1 > S_2 > \cdots > S_L$, which implies that the model's likelihood of

Method	Object HalBench		MMHalBench		AMBER			
	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Score ↑	HalRate ↓	$CHAIR_s \downarrow$	Cover. ↑	HalRate ↓	Cog. ↓
LLaVA-v1.5-7B (Liu et al., 2024a)	49.7	26.1	2.02	0.65	7.7	49.8	31.9	3.7
+ DPO (Rafailov et al., 2024)	42.3	23.2	2.00	0.69	6.7	**53.2**	33.7	3.3
+ HALVA (Sarkar et al., 2024)	-	-	-	-	6.6	53.0	32.2	3.4
+ HA-DPO (Zhao et al., 2024)	39.9	19.9	-	-	6.7	49.8	30.9	3.3
+ V-DPO (Xie et al., 2024)	-	-	-	-	6.6	49.1	30.8	3.1
+ mDPO (Wang et al., 2024a)	30.7	16.0	**2.40**	**0.59**	5.0	52.5	27.5	2.4
+ LPOI (Ours)	**24.3**	**14.6**	**2.40**	**0.59**	**4.3**	51.9	**26.4**	**2.0**
LLaVA-v1.5-13B (Liu et al., 2024a)	44.3	21.2	2.09	0.64	6.3	51.0	30.2	3.0
+ DPO (Rafailov et al., 2024)	38.3	19.4	2.36	0.61	6.2	**54.3**	31.8	2.6
+ mDPO (Wang et al., 2024a)	33.3	16.6	2.50	**0.57**	4.6	52.6	25.0	2.0
+ LPOI (Ours)	**24.3**	**11.7**	**2.54**	**0.57**	**3.9**	52.9	**22.3**	**1.8**
Idefics2-8B (Laurençon et al., 2024)	6.3	4.2	2.62	0.43	3.4	36.5	7.6	0.4
+ DPO (Rafailov et al., 2024)	6.0	4.2	2.48	0.45	3.5	37.4	8.1	**0.2**
+ mDPO (Wang et al., 2024a)	7.3	5.4	2.80	0.40	2.7	**37.7**	6.2	**0.2**
+ LPOI (Ours)	**5.3**	**3.6**	**2.88**	**0.36**	**2.6**	36.4	**5.7**	**0.2**

Table 1: Performance comparison between various preference learning methods on Object HalBench, MMHalBench, and AMBER benchmarks. We use three base VLM models: Llava-v1.5-7B/13B and Idefics2-8B. The results of DPO and mDPO are reproduced under a fair setting with LPOI. HALVA, HA-DPO, and V-DPO are taken from their respective papers; they are included for reference.

generating positive text about the object increases as its visibility in the image grows (Figure 2). This approach helps the model reduce hallucinations, as it encourages the model to mention the object in proportion to its visibility.

In addition to the listwise loss, we also use the standard DPO loss and the anchor loss:

$$\mathcal{L}_{\text{Anchor}} = -\log \sigma \left(\beta \log \frac{\pi_\theta(w \mid x, q)}{\pi_{\text{ref}}(w \mid x, q)} - \delta \right),$$

which is proposed in mDPO (Wang et al., 2024a). Minimizing the anchor loss further increases the likelihood that the model generates postive responses when given the original image. In total, our objective becomes

$$\mathcal{L}_{\text{Total}}(\theta) = \mathcal{L}_{\text{DPO}}(\theta) + \mathcal{L}_{\text{Anchor}}(\theta) + \mathcal{L}_{\text{Listwise}}(\theta).$$

Algorithm 1 summarizes the overall procedure of the proposed LPOI method.

4 Experiment

4.1 Experimental Setup

Baselines. We compare our LPOI approach against established methods, including DPO (Rafailov et al., 2024), mDPO (Wang et al., 2024a), HALVA (Sarkar et al., 2024), HA-DPO (Zhao et al., 2024), and V-DPO (Xie et al., 2024). We evaluate each method using three VLMs including the LLaVA-v1.5-7B, LLaVA-v1.5-13B (Liu et al., 2024a), and Idefics2-8B (Laurençon et al., 2024). For DPO and mDPO, we report reproduced results using the same training dataset as our LPOI method. For HALVA, HA-DPO, and V-DPO, we report the originally published performance for reference.

Evaluation. We evaluate both the base and fine-tuned versions of VLMs using MMHalBench (Sun et al., 2023), Object HalBench (Rohrbach et al., 2019), and AMBER (Wang et al., 2024b), which are standard benchmarks for assessing hallucination and the quality of generated text of VLMs. We report the CHAIR metric (Rohrbach et al., 2019) to measure object hallucination and the MMHalBench score (computed via GPT-4o (OpenAI, 2024)) to quantify the quality of generated outputs.

Training setup. We conduct the preference learning via LoRA fine-tuning (Hu et al., 2021). For training sets, we randomly sample 10K preference data from Silkie (Li et al., 2023a) and instruction datafrom LLaVA-Instruct-150K (Liu et al., 2023), following the setup of mDPO (Wang et al., 2024a). Idefics2-8B is trained for 3 epochs with a learning rate of 5e-6, and LLaVA-v1.5 (7B and 13B) for 1 epoch with a learning rate of 1e-6. We employ 1 RTX A6000 GPU for fine-tuning Idefics2-8B and LLaVA-v1.5-7B, and employ 2 RTX A6000 GPU for LLaVA-v1.5-13B. Refer to Appendix A for details on hyperparameters.

Method	Object HalBench		MMHalBench		AMBER			
	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Score ↑	HalRate ↓	$CHAIR_s \downarrow$	Cover. ↑	HalRate ↓	Cog. ↓
Idefics2-8B (Laurençon et al., 2024)	6.3	4.2	2.62	0.43	3.4	36.5	7.6	0.4
+ DPO (Rafailov et al., 2024)	6.0	4.3	2.29	0.51	3.1	36.4	6.8	**0.3**
+ mDPO (Wang et al., 2024a)	8.7	5.6	2.71	0.42	**2.8**	**37.2**	6.5	**0.3**
+ LPOI (Ours)	**5.3**	**4.0**	**2.81**	**0.38**	**2.8**	36.2	**6.2**	**0.3**

Table 2: Performance comparison under the same training cost (20 hours on a single RTX A6000 GPU) for Idefics2-8B model on Object HalBench, MMHalBench, and AMBER benchmarks.

Method	Object HalBench		MMHalBench		AMBER	
	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Score ↑	HalRate ↓	$CHAIR_s \downarrow$	HalRate ↓
without V.P.	5.3	4.0	2.74	0.40	2.7	6.0
with V.P.	**5.0**	**3.4**	**2.91**	**0.35**	**2.6**	**5.8**

Table 3: Performance comparison with and without visual prompting for the Idefics2-8B model on Object HalBench, MMHalBench, and AMBER benchmarks.

Method	Object HalBench		MMHalBench		AMBER	
	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Score ↑	HalRate ↓	$CHAIR_s \downarrow$	HalRate ↓
List size 3	7.3	5.1	2.86	**0.36**	2.9	6.6
List size 4	6.7	4.5	2.86	**0.36**	**2.5**	**5.6**
List size 5	**5.3**	**3.6**	**2.88**	**0.36**	2.6	5.7

Table 4: Performance comparison across different list sizes for the Idefics2-8B model on Object HalBench, MMHalBench, and AMBER benchmarks.

4.2 Results

We present the results in Table 1. Our proposed LPOI consistently improves performance of different VLMs across most benchmarks. Notably, it excels at hallucination related metrics, including the HalRate in MMHalBench, the CHAIR metric in Object HalBench, and the CHAIR and cognition metric in AMBER. Specifically, our method achieves 24.3 in $CHAIR_s$ and 14.6 in $CHAIR_i$ for LLaVA-v1.5-7B on Object HalBench, which is superior than state-of-the-art mDPO with 30.7 in $CHAIR_s$ and 16.0 in $CHAIR_i$ in the same setting. It is also worth noting that although our coverage performance is on par with other methods, this metric often grows at the expense of increased hallucination since it measures how much ratio of correct objects are detected by the model. Thus, models that generate more mentions, even if some are erroneous, can inflate their coverage score.

We further note that Object HalBench is generally more challenging than AMBER with respect to the CHAIR score, and models tend to exhibit a higher hallucination rate on this benchmark. Our method yields a notably larger performance gain on Object HalBench compared to AMBER, where models already maintain a low hallucination rate and the scores are largely saturated.

4.3 Human Evaluation

To further assess the quality of responses, we conduct a human evaluation using 80 randomly selected image-question pairs, 40 from the AMBER benchmark and 40 from the Object HalBench. We present the results in Figure 3. Each pair is presented to three crowd workers recruited via Amazon Mechanical Turk from English-speaking countries, with a maximum payment of $0.50 per HIT. The annotators are provided with two responses generated by the Idefics2-8B, one fine-tuned using our LPOI and the other using mDPO, which is the strongest baseline in Table 1. Workers are instructed to select the response that is more accurate and reliable, considering the visual information in the image.

We also compare with DPO under the same conditions. Annotators consistently prefer responses from our fine-tuned model over those from mDPO and DPO. Inter-annotator agreement is measured using Krippendorff's α, which yields a value of 0.735 for DPO and 0.671 for mDPO on the AMBER benchmark, and a value of 0.823 for DPO and 0.627 for mDPO on the Object HalBench. These values reflect the level of agreement among annotators regarding the relative quality of the responses, with three possible choices: A is better, B is better, or a tie. More details can be found in Appendix I.

4.4 Analysis

Comparison Under Equal Training Budget. We present the results of evaluating DPO, mDPO and LPOI (ours) under the same training budget (GPU hours). Since the listwise objective inherently incurs a higher training cost compared to the pairwise objective, we further present the results of training LPOI, DPO, and mDPO for 20 hours on a single RTX A6000 GPU using a 5K subsample of the preference dataset. Table 2 demonstrates that, even under the same training budget, our method

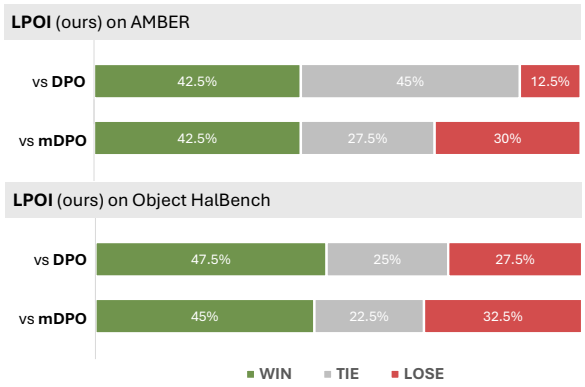

Figure 3: Human evaluation results on a subset of the AMBER and Object HalBench benchmark. We compare responses generated by the Idefics-2B model fine-tuned using LPOI (ours), DPO, and mDPO.

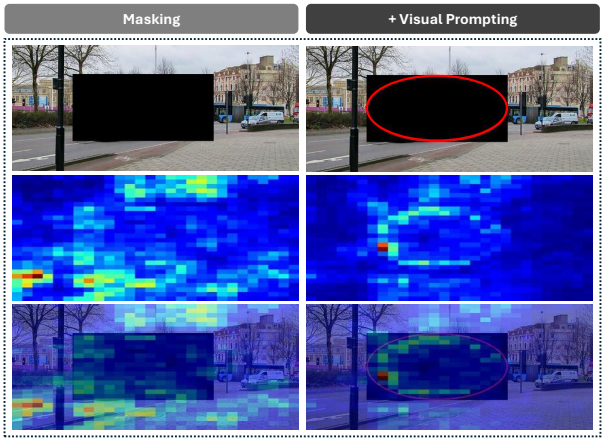

Figure 4: Comparison of saliency maps with or without visual prompting (highlighted in red circle). Visual prompting shifts the model's attention towards the masked area, guiding it to focus more on the region of interest. In the saliency maps, blue indicates low saliency, while red indicates high saliency.

consistently outperforms DPO and mDPO, particularly in terms of hallucination scores and the overall quality of the generated outputs.

Advantages of Visual Prompting. Masking the critical object in an image may not always turn the original preferred answer into a negative one, when VLMs can still infer the correct answer by using surrounding context. Thus, we apply visual prompting (Shtedritski et al., 2023; Wu et al., 2024; Lin et al., 2024; Cai et al., 2024) to highlight more the masked region and guide the model's attention there. We validate that visual prompting directs the model's focus and increase the performance.

Figure 4 shows the saliency maps of the masked

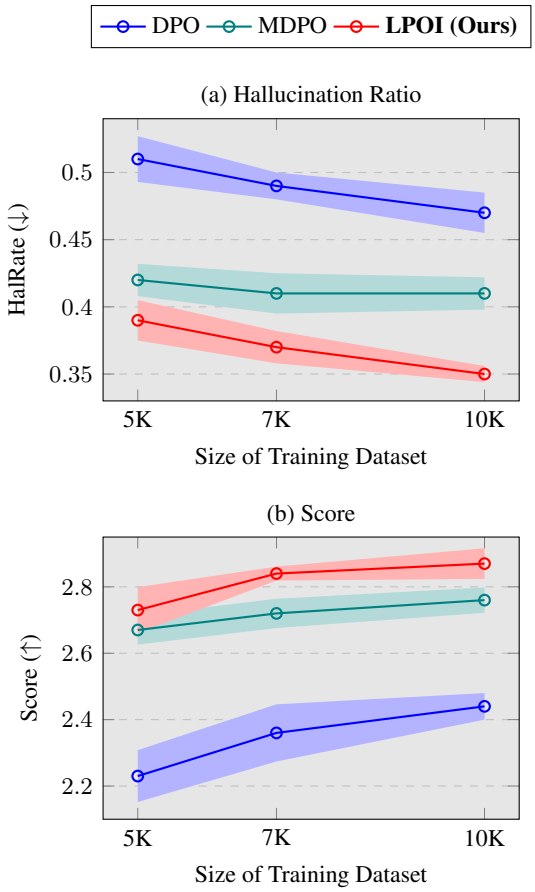

Figure 5: MMHalBench results for different preference optimization methods trained on three different sizes of training sets.

image with and without visual prompting. The saliency maps are computed using a gradient-based method with ResNet-50 (He et al., 2015), where we aggregate the absolute gradient of the prediction loss with respect to the input image across RGB channels. After the masked area is highlighted with a red circle, the saliency score in that region increases, showing that the model focuses more on the masked area. Please refer to Appendix G for more examples. Table 3 shows that using visual prompting improves the quality of generated outputs and in reducing hallucinations, thanks to generating higher-quality negative images.

Impact of List Sizes. We present the results of LPOI with list sizes of 3, 4, and 5, assessing the impact of the list size on the performance. Table 4 shows that larger list sizes result in improved overall performance, particularly for the Idefics2-8B on the Object HalBench dataset. We speculate that this improvement comes from the model being able to learn from a wider range of preference samples, i.e., larger list sizes provide more opportunities for the

Q. Compare the tomatoes in the photo.

(base) In the photo, **there are three tomatoes** hanging from a vine. The first tomato is located on the left side of the image, the second tomato is in the middle, and the third tomato is on the right side. **They all appear to be ripe and ready for harvest.** The tomatoes are hanging from the vine, which is a common sight in gardening and agriculture.

(DPO) In the photo, **there are four tomatoes** hanging from a vine, each at a different stage of ripeness. **The first tomato on the left is the most unripe,** with a greenish hue and a firm texture. **The second tomato, in the middle, appears to be slightly riper,** with a more yellowish color and a slightly softer texture. **The third tomato, on the right, is the ripest,** with a deep yellow color and a noticeably softer texture. **The fourth tomato,** at the far right, is the most unripe of the group, with a greenish hue and a firm texture. This progression of ripeness from left to right indicates that the tomatoes are at different stages of maturity, with the ripest tomato being ready for consumption and the others still needing more time to ripen.

(mDPO) In the photo**, there are three tomatoes** hanging from a vine. They appear to be at different stages of ripeness. **The tomato on the left seems to be the least ripe, with a greenish tint, indicating that it is still unripe.** The tomato in the middle is more ripe, with a yellowish hue, suggesting that it is at an intermediate stage of ripeness. **The tomato on the right is the most ripe, with a deep orange color, indicating that it is fully ripe and ready to be picked.** The differences in color and texture suggest that the tomatoes are at different stages of maturity, **with the left one being the least ripe and the right one being the most ripe.**

(OURS) The photo features **three tomatoes** hanging from a vine, with each tomato appearing to be at a different stage of ripeness. **The first tomato is the most ripe,** with a bright orange color and a slightly yellowish tint. **The second tomato is slightly less ripe,** with a more orange hue and a slightly greenish tint. **The third tomato is the least ripe,** with a more greenish-yellow color and a noticeable stem. This progression of ripeness suggests that the tomatoes are at different stages of maturity, **with the first tomato being ready for immediate consumption, while the other two tomatoes may require more time to ripen further before they are at their peak flavor and nutritional value.**

Q. Please provide a detailed description.

(base) The image features a large, freshly baked pizza with a generous amount of sauce and cheese. The pizza is placed on a metal tray, which is sitting on a dining table. **The table is surrounded by chairs, with one on the left side and another on the right side. In the background, there is a person partially visible, likely enjoying the meal or preparing to eat the pizza.** The overall atmosphere of the image is casual and inviting, with the focus on the delicious pizza and the comfortable dining setup.

(DPO) The image features a large, freshly baked pizza with a generous amount of tomato sauce and melted cheese. The pizza is placed on a metal tray, which is sitting on a dining table. **The table appears to be in a restaurant setting, as there are chairs positioned around it.** The pizza is cut into several slices, with some of the slices appearing to be missing, indicating that it has been partially eaten. The sauce is spread evenly across the pizza, with some areas appearing to be more saucy than others. The cheese is melted and bubbly, creating a mouth-watering appearance. The overall mood of the image is one of indulgence and enjoyment, as the pizza looks delicious and inviting. The restaurant setting suggests a casual dining atmosphere, where people can gather to share a meal and socialize.

(mDPO) The image features a close-up view of a partially eaten pizza with a generous amount of sauce, sitting on a metal tray. The pizza appears to be a cheese and tomato-based variety, with a slightly burnt crust. The sauce is predominantly red, and the cheese is white, creating a visually appealing contrast. **In the background, there is a person partially visible, likely enjoying the pizza or preparing to eat it.** The dining table is also present, but it is not the main focus of the image. The overall mood is casual and inviting, as the pizza is a popular and comforting food choice.

(OURS) The image features a close-up view of a large, partially eaten pizza with a generous amount of red sauce, likely tomato-based. The pizza is placed on a metal tray, which is the main focus of the image. The crust appears to be slightly burnt, giving the pizza a slightly charred appearance. The pizza is sliced into several pieces, with a few visible in the foreground and others in the background. The sauce is spread across the pizza, with some areas appearing more saturated than others. The close-up view emphasizes the details of the pizza, including the cheese and the sauce, creating a sense of depth and texture. **The image does not include any people or objects, which allows the viewer to focus solely on the pizza and its details.** The overall mood or atmosphere of the image is one of indulgence and satisfaction, as the pizza appears to be a delicious and satisfying meal.

Figure 6: Qualitative results of the base model (LLaVA-v1.5-7B) its finetuned versions with DPO, mDPO, and LPOI (Ours). **Correct answers** and **hallucinations** are highlighted.

model to capture fine-grained differences between candidates, leading to a better model performance.

Ablating the DPO Loss. The listwise preference loss only utilizes positive text (with multiple masked images). Without the text DPO loss, negative text samples in the dataset are not used, meaning that the model would not learn from any textual preference information (i.e., learn from only image preference information). To demonstrate the effect of incorporating the text DPO loss, we conducted ablation experiments by training the Idefics2-8B model on a 5K training dataset with LPOI (a list size of 3) for 3 epochs. We compared three scenarios: (1) LPOI without the text DPO loss, (2) LPOI with neither the text DPO loss nor the anchor loss, and (3) the full LPOI loss as proposed in this paper. The results, presented in Table 5, show that excluding the DPO loss leads to suboptimal performance compared to using the complete LPOI loss.

	Object HalBench		MMHalBench		AMBER	
Method	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Score ↑	HalRate ↓	$CHAIR_s \downarrow$	HalRate ↓
Idefics2-8B	6.3	4.2	2.62	0.43	3.4	7.6
+LPOI (without DPO loss)	7.7	4.6	2.56	0.44	3.3	7.4
+LPOI (without DPO, anchor loss)	6.0	4.1	2.50	0.45	3.5	7.5
+LPOI	**5.7**	**3.6**	**2.74**	**0.40**	**2.8**	**6.4**

Table 5: Ablation experiments comparing (1) LPOI without text DPO loss, (2) LPOI without both text DPO and anchor loss, and (3) the full LPOI loss, using the Idefics2-8B model trained for 3 epochs on 5K dataset, using list size of 3.

Results on Different Training Sets. We quantitatively compare between DPO, mDPO, and LPOI when training on smaller datasets for Idefics2-8B in Figure 5. We train them on the subsets with sizes of 5K, 7K, and 10K, repeating the process three times for each subset, and report the average and standard deviation of the GPT-score and hallucination ratio on the MMHalBench benchmark. Our experiments demonstrate a consistent advantage of LPOI over the other methods, both in terms of

output quality and hallucination reduction, across preference datasets of varying sizes.

Qualitative Examples Figure 6 presents a comparative analysis of outputs from the LLaVA-v1.5-7B base model and its fine-tuned variants using DPO, mDPO, and LPOI. For instance, in the first example, where the main factor in hallucination is determining which tomato is the ripest, our model accurately selects the leftmost tomato while other models erroneously choose the rightmost one. The baselines' explanations often contradict what is clearly observable in the image. These results highlight the importance of guiding the model to focus on subtle incremental visual changes. By doing so, our LPOI enables the model to ground its responses more reliably in the image, improving the recognition of fine details and reducing the likelihood of hallucinating common yet irrelevant objects.

5 Conclusion

In this work, we addressed the challenge of aligning VLMs with human preferences by proposing LPOI, a novel framework that combines hard negative sampling with listwise ranking. By generating object-aware hard negatives through masking key objects in images and interpolating between them and positive samples, we provide an efficient method for creating listwise preference data without additional annotation cost. Extensive evaluations on Object HalBench, MMHalBench, and AMBER benchmarks demonstrate that LPOI significantly improves performance by mitigating hallucinations and enhancing multimodal alignment.

Ethics Statement

We have used open source models, libraries, datasets, and closed source models for their intended use and license, and not use other than research purposes.

Limitations

A potential limitation of our approach is that while we focus on listwise sample generation for the vision and language domain, we do not address other modalities, such as the audio domain. Future work could explore further optimization strategies and extend listwise preference learning to additional modalities, including audio, by adapting similar interpolation strategies to reduce hallucinations in those domains. Additionally, the prompts provided are exclusively in English but it can be expanded to include multiple languages in future iterations.

Acknowledgements

We would like to thank the anonymous reviewers and Professor Chenglin Fan for their valuable feedback. This work was financially supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II191082, SW StarLab, No. RS-2022-II220156, Fundamental research on continual meta-learning for quality enhancement of casual videos and their 3D metaverse transformation, and No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2023R1A2C2005573), and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2023-00274280).

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Hritik Bansal, Ashima Suvarna, Gantavya Bhatt, Nanyun Peng, Kai-Wei Chang, and Aditya Grover. 2024. Comparing bad apples to good oranges: Aligning large language models via joint preference optimization. *Preprint*, arXiv:2404.00530.

Mu Cai, Haotian Liu, Dennis Park, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, and Yong Jae Lee. 2024. Vip-llava: Making large multimodal models understand arbitrary visual prompts. *Preprint*, arXiv:2312.00784.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007a. Learning to rank: From pairwise approach to listwise approach. volume 227, pages 129–136.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007b. Learning to rank: from pairwise

approach to listwise approach. In *International Conference on Machine Learning*.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. 2024. Yolo-world: Real-time open-vocabulary object detection. *Preprint*, arXiv:2401.17270.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. *Preprint*, arXiv:1706.03741.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang. 2024. Enhancing large vision language models with self-training on image comprehension. *Preprint*, arXiv:2405.19716.

Leo Gao, John Schulman, and Jacob Hilton. 2022. Scaling laws for reward model overoptimization. *Preprint*, arXiv:2210.10760.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. *Preprint*, arXiv:2209.14375.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *Preprint*, arXiv:2310.14566.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *Preprint*, arXiv:1512.03385.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA. Association for Computational Linguistics.

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

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023a. Silkie: Preference distillation for large visual language models. *Preprint*, arXiv:2312.10665.

Zheng Li, Caili Guo, Xin Wang, Zerun Feng, and Yanjun Wang. 2023b. Integrating listwise ranking into pairwise-based image-text retrieval. *Preprint*, arXiv:2305.16566.

Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. 2024. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *Preprint*, arXiv:2403.20271.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *Preprint*, arXiv:2303.05499.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, Peter J. Liu, and Xuanhui Wang. 2024c. Lipo: Listwise preference optimization through learning-to-rank. *Preprint*, arXiv:2402.01878.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Preprint*, arXiv:2405.14734.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. 2024. Scaling open-vocabulary object detection. *Preprint*, arXiv:2306.09683.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. 2024. Strengthening multimodal large language model with bootstrapped preference optimization. *Preprint*, arXiv:2403.08730.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. *Preprint*, arXiv:2010.04592.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object hallucination in image captioning. *Preprint*, arXiv:1809.02156.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö. Arık, and Tomas Pfister. 2024. Data-augmented phrase-level alignment for mitigating object hallucination. *Preprint*, arXiv:2405.18654.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 815–823. IEEE.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does clip know about a red circle? visual prompt engineering for vlms. *Preprint*, arXiv:2304.06712.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. *Preprint*, arXiv:2306.17492.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *Preprint*, arXiv:2009.01325.

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

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. *Preprint*, arXiv:2406.11839.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2024b. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *Preprint*, arXiv:2311.07397.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A. Rossi, Ruiyi Zhang, Subrata Mitra, Dimitris N. Metaxas, Lina Yao, Jingbo Shang, and Julian McAuley. 2024. Visual prompting in multimodal large language models: A survey. *Preprint*, arXiv:2409.15310.

Liwei Wu, Cho-Jui Hsieh, and James Sharpnack. 2019. Sql-rank: A listwise approach to collaborative ranking. *Preprint*, arXiv:1803.00114.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Hao Jiang, Fei Wu, and Linchao Zhu. 2024. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. *Preprint*, arXiv:2404.14233.

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. *Preprint*, arXiv:2411.02712.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024a. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *Preprint*, arXiv:2401.08417.

Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. 2024b. Some things are more cringe than others: Iterative preference optimization with the pairwise cringe loss. *Preprint*, arXiv:2312.16682.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears. *Preprint*, arXiv:2304.05302.

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

Fatemeh Pesaran Zadeh, Juyeon Kim, Jin-Hwa Kim, and Gunhee Kim. 2024. Text2chart31: Instruction tuning for chart generation with automatic feedback. *Preprint*, arXiv:2410.04064.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2024. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *Preprint*, arXiv:2311.16839.

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

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

A Experimental Details

Training setup and hyperparameters We report the hyperparameters for training LPOI in Table 6. We fine-tune base models with LoRA adapter with the configuration in Table 6.

Model	LLaVA-v1.5-7B	LLaVA-v1.5-13B	Idefics2-8B
Training epochs	1	1	3
Training set size	10K	10K	10K
Batch size	64	64	64
Optimizer	AdamW	AdamW	AdamW
Learning rate	1e-6	1e-6	5e-6
Learning rate scheduling	Linear	Linear	Linear
Mixed precision	FP16	FP16	BF16
LoRA rank	8	8	8
LoRA alpha	8	8	8
LoRA dropout	0.0	0.0	0.0

Table 6: Training hyperparameters for fine-tuning LLaVA-v1.5-7B, LLaVA-v1.5-13B, and Idefics2-8B models.

B Computational Overhead and Performance Analysis

We present the training time of DPO, mDPO, and LPOI with the list sizes of 3, 4, and 5 on 5K examples for 1 epoch, measured on an RTX A6000 GPU in Table 7. Additionally, we include the number of epochs and scores on the MMHalBench benchmark when trained with the same GPU budget (20 GPU hours), also in Table 7. As the list size increases, LPOI introduces computational overhead, but it provides richer signals that help reduce hallucinations, leading to a lower hallucination ratio (See Table 4). Moreover, with sufficient optimization time, LPOI outperforms both mDPO and DPO within the same GPU training budget, benefiting from these richer signals.

Methods	Time per epoch	Epochs under 20 GPU hours	MMHalBench GPT-Score (↑)	MMHalBench HalRate (↓)
DPO	2.2 hrs	9 epochs	2.29	0.51
mDPO	4.0 hrs	5 epochs	2.71	0.42
LPOI (list size 3)	4.5 hrs	4.5 epochs	–	–
LPOI (list size 4)	5.3 hrs	3.8 epochs	–	–
LPOI (list size 5)	6.2 hrs	3 epochs	**2.81**	**0.38**

Table 7: Training time per epoch on 5K examples for DPO, mDPO, and LPOI (list sizes 3, 4, 5), using an RTX A6000 GPU, along with the number of epochs and MMHalBench results under the same GPU budget.

C Extended Benchmark Comparison

We further evaluated our method (LPOI) and the baselines (DPO and mDPO) on the HallusionBench benchmark (Guan et al., 2024) using Idefics2-8B model, and presented the results in Table 8. LPOI consistently outperforms or matches the baseline methods across most of the metrics.

D Additional Results with Increased Training Data

We chose to use a 10K subset of Silkie and LLaVA-Instruct-150K for preference fine-tuning by following the experiment setup in mDPO. Furthermore we conducted additional experiments by fine-tuning the Idefics2-8B model on 15K data for 1 epoch, using our method (LPOI) and baselines (DPO, mDPO). The results, presented in Table 9, demonstrate that our method consistently outperforms the baselines across most metrics.

Model	Question Pair Acc	Figure Acc	Easy Acc	Hard Acc	All Acc
Idefics2-8B (Laurençon et al., 2024)	8.35	14.16	32.53	30.93	35.08
+DPO (Rafailov et al., 2024)	15.82	22.54	49.45	33.72	46.68
+mDPO (Wang et al., 2024a)	16.48	**24.28**	50.33	36.05	48.45
+LPOI (Ours)	**17.80**	23.70	**51.65**	**36.98**	**49.78**

Table 8: Performance comparison between various preference learning methods on HallusionBench benchmark.

	Object HalBench		MMHalBench		AMBER			
Method	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Score ↑	HalRate ↓	$CHAIR_s \downarrow$	Cover. ↑	HalRate ↓	Cog. ↓
Idefics2-8B (Laurençon et al., 2024)	6.3	4.2	2.62	0.43	3.4	36.5	7.6	0.4
+ DPO (Rafailov et al., 2024)	6.3	4.4	2.57	0.44	3.3	36.4	7.3	**0.3**
+ mDPO (Wang et al., 2024a)	7.7	5.0	2.74	0.41	**3.0**	**37.6**	**6.8**	**0.3**
+ LPOI (Ours)	**5.0**	**3.7**	**2.75**	**0.38**	**3.0**	36.8	**6.8**	**0.3**

Table 9: Performance comparison between various preference learning methods with larger dataset (15K).

E Analysis and Ablation of the Verification Module

For the full 10K dataset with a list size of 5, object detection takes 1,298 seconds (21 minutes), and the verification module takes 18,938 seconds (5.26 hours), averaging 0.166 seconds and 2.43 seconds per data point, respectively. While we reported the version with verification to achieve the best performance, we note that our method performs well even without the verification step, outperforming all baseline methods in this case. To further illustrate this, we conducted an additional experiment using only the object detection module, focusing on a single salient object per image and excluding the verification step, and presented the results in Table 10. Despite this simplification, the LPOI still enables the model to outperform baseline methods like DPO and mDPO across most metrics—especially on hallucination scores, as shown in Table 10. This demonstrates that our approach can maintain strong performance while significantly reducing preprocessing time.

F Details on Object Detection Model

For the object detection component in Section 3.1, we utilize the Grounding-DINO-Tiny model. Since generating accurate hard negative samples is vital for our pipeline, and precise object detection plays a key role in this process, we evaluate various object detection models to find the most suitable one for our task. Specifically, we compare different versions of Grounding-DINO (Liu et al., 2024b), OwlV2 (Minderer et al., 2024), and YOLO-World (Cheng et al., 2024) on a 1k subset of our dataset. The chosen model, with 172 million parameters, effectively detects around 80% of the key noun objects present in the image.

G Details on Visual Prompting

Figure 7 illustrates 3 more examples of the impact of incorporating an additional visual prompting represented by a red circle in the image, to guide the model's attention toward the region of interest. In each group, the left column displays an image from our dataset with only the applied mask, its corresponding saliency map, and an overlap visualization of the two. The right column shows the same image, but with the visual prompt added by circling the masked area.

H Qualitative Analysis

We provide additional examples generated by the fine-tuned models using DPO, mDPO, and LPOI (ours). In the first example, shown at the left, all models except ours mistakenly claim that the kiwi in the foreground, which is dried into chips, is fresh. In the third example at the right, the image shows a motorcycle without a rider. When asked to determine the gender of the person riding the motorcycle, our model correctly states that no person is visible, while the other models erroneously identify a woman as the rider. These examples highlight how our method reduces common hallucinations in vision-language

Method	Object HalBench		MMHalBench		AMBER			
	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Score ↑	HalRate ↓	$CHAIR_s \downarrow$	Cover. ↑	HalRate ↓	Cog. ↓
Idefics2-8B (Laurençon et al., 2024)	6.3	4.2	2.62	0.43	3.4	36.5	7.6	0.4
+ DPO (Rafailov et al., 2024)	**6.0**	4.2	2.48	0.45	3.5	37.4	8.1	**0.2**
+ mDPO (Wang et al., 2024a)	7.3	5.4	2.80	0.40	**2.7**	37.7	6.2	**0.2**
+ LPOI (without verification)	**6.0**	**4.1**	**2.86**	**0.35**	**2.7**	36.1	**5.9**	**0.2**
+ LPOI (with verification)	5.3	3.6	2.88	0.36	2.6	36.4	5.7	0.2

Table 10: Performance of DPO, mDPO, and LPOI on the Idefics2-8B model trained for 3 epochs. LPOI preserves its superiority over the baselines even without verification module. LPOI with verification is included for reference.

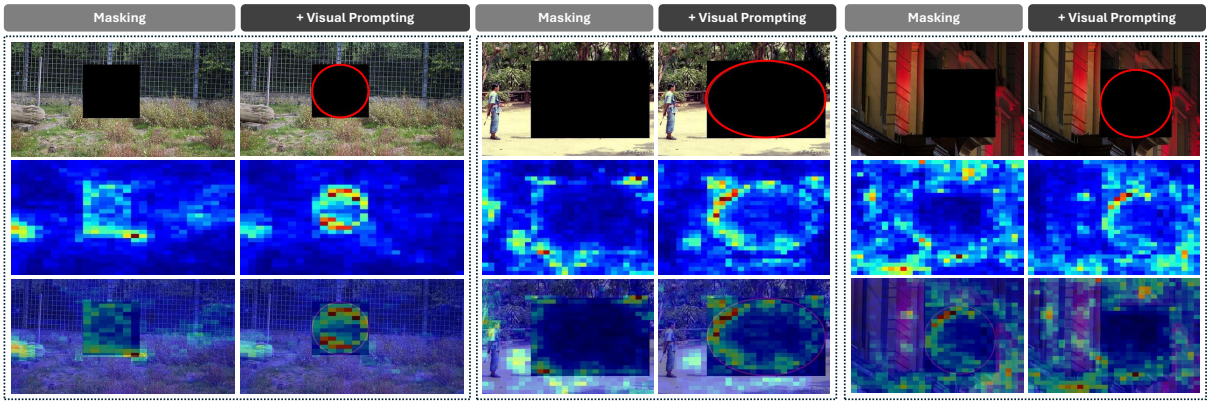

Figure 7: Comparison of saliency maps with or without visual prompting (highlighted in red circle).

models, such as the false assumption of co-occurring objects, the failure to recognize subtle object features or the provision of answers to questions that cannot be derived from the image alone.

I Details on Human Evaluation

Figure 9 shows the user interface where annotators select the less hallucinatory response between two answers generated by mDPO and LPOI (ours). Each worker is presented with two responses generated by the Idefics2-8B model: one fine-tuned using mDPO or DPO, and the other using the LPOI method. Workers are instructed to select the response they consider more accurate and reliable based on the visual information in the image. If the responses are identical or both factually incorrect, workers are asked to choose the 'tie' option. The workers' answers are then aggregated using a majority vote. To prevent bias, the order of the responses (Response A and Response B) is shuffled for each datapoint, and workers must also provide justifications for their selections. These justifications are reviewed to ensure the reliability and consistency of the answers, which are then used to validate the integrity of the evaluation process.

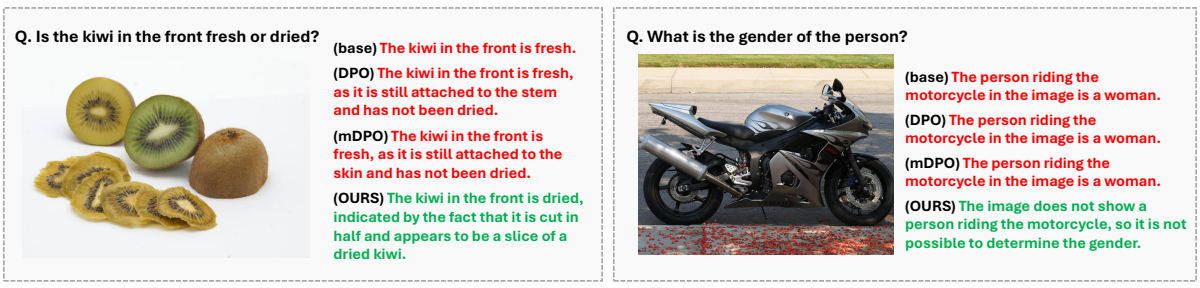

Figure 8: Qualitative results of the base model (LLaVA-v1.5-7B) its variants optimized with DPO, mDPO, and LPOI(Ours). **Correct answers** and **hallucinations** are highlighted.

Choose a Response with Less Hallucination

Hello! We are looking for interested annotators who can contribute to comparing different language model responses. Your job is to look at the provided image and determine which response about that image contains fewer hallucinations. Here, hallucination refers to when an AI model generates outputs that are factually incorrect, misleading, or entirely fabricated despite appearing plausible. [For a brief intro to AI Hallucination]. We greatly appreciate your expertise and dedication in helping us improve AI language models.

Additional Details

This is the **main test** for annotators who demonstrated the best performance in our previous tests. We greatly appreciate your hard work and insight, and we're excited to continue this final task with you. This test is paid at a rate of **$0.25 per question, and it is much less challenging than before**. If possible, **we encourage you to work through the entire batch**, as it should be more manageable and quicker than previous rounds!

As always, please determine which of the two responses is more accurate and reliable. **If both answers are identical or both answers are factually incorrect, you should choose "TIE".**

Disclaimer: All responses provided in this task will be used solely for research and improvement of AI systems. Your participation is voluntary, and you may stop at any time.

Now, please take a look at the provided image, and carefully determine which of the two responses is more accurate and reliable. The chosen answer does not have to be perfect, but it has to be better than the rejected answer.

Image:

Question:

Describe this image.

Response A:

A mans name is written in stones with a back pack sitting next to it.

Response B:

A man sits on a stone ledge with the name Moose RVS written on it.

Which of the two responses is more reliable and accurate?

- ○ Response A
- ○ Response B
- ○ Tie

Please share a brief reason supporting your choice:

I answered that response (A) is better because (B) wrongly states that the person's hat is green while it is actually blue...

Figure 9: User interface and instruction for human evaluation.