MDPO: Conditional Preference Optimization for Multimodal Large Language Models

Fei Wang[1]   Wenxuan Zhou[1]   James Y. Huang[1]   Nan Xu[1]
Sheng Zhang[2]   Hoifung Poon[2]   Muhao Chen[3]

[1]University of Southern California   [2]Microsoft Research   [3]University of California, Davis

https://feiwang96.github.io/mDPO

{fwang598,zhouwenx}@usc.edu   shezhan@microsoft.com   muhchen@ucdavis.edu

Abstract

Direct preference optimization (DPO) has shown to be an effective method for large language model (LLM) alignment. Recent works have attempted to apply DPO to multimodal scenarios but have found it challenging to achieve consistent improvement. Through a comparative experiment, we identify the *unconditional preference* problem in multimodal preference optimization, where the model overlooks the image condition. To address this problem, we propose MDPO, a multimodal DPO objective that prevents the over-prioritization of language-only preferences by also optimizing image preference. Moreover, we introduce a reward anchor that forces the reward to be positive for chosen responses, thereby avoiding the decrease in their likelihood—an intrinsic problem of *relative preference* optimization. Experiments on two multimodal LLMs of different sizes and three widely used benchmarks demonstrate that MDPO effectively addresses the unconditional preference problem in multimodal preference optimization and significantly improves model performance, particularly in reducing hallucination.

1 Introduction

Direct preference optimization (DPO) has emerged as the predominating method for aligning large language models (LLMs) with human preferences (Rafailov et al., 2023; Zhao et al., 2024). Building on its success in the language modality, recent studies have extended DPO to multimodal scenarios (Li et al., 2023; Yu et al., 2024a; Zhou et al., 2024b; Zhao et al., 2023). However, transferring this approach across modalities presents significant challenges. Merely substituting textual preference data with multimodal preference data does not consistently yield positive outcomes and can exacerbate issues such as hallucinations (Li et al., 2023; Sarkar et al., 2024).

While recent efforts in multimodal preference learning focus on improving performance through
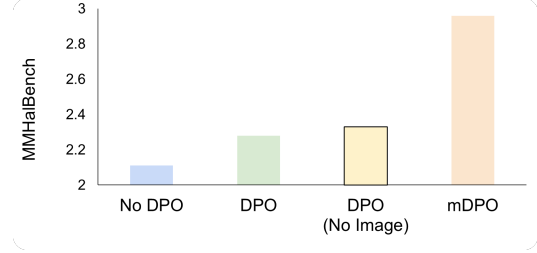

Figure 1: We train Bunny-v1.0-3B (He et al., 2024) on 10K multimodal preference data from Silkie (Li et al., 2023) with different variants of DPO. We perform DPO (No Image) where all images are removed from the preference data. Counterintuitively, the overall score on the MMHalBench (Sun et al., 2023) for DPO (No Image) is similar to that of DPO with images. This finding suggests that DPO may suffer from unconditional preferences, neglecting the visual modality during optimization. Our proposed method, MDPO, effectively addresses this issue and improves model performance.

enhanced multimodal preference data (Li et al., 2023; Zhao et al., 2023; Xiao et al., 2024; Zhou et al., 2024b; Pi et al., 2024; Sarkar et al., 2024; Yu et al., 2024b; Deng et al., 2024), we investigate the pitfalls of multimodal DPO from a different perspective. Through controlled comparisons, we discover that multimodal LLMs can achieve similar performance even when all images are *removed* from the multimodal preference data during DPO (see Fig. 1). This counterintuitive finding suggests that the failure of DPO in multimodal scenarios may not be solely attributed to data quality. We attribute this to a systematic gap between the theoretical expectations and practical implementations of the DPO objective in multimodal settings (refer to Fig. 2). While DPO aims to compute implicit rewards conditioned on all input modalities, it may prioritize language-only preferences and overlook the image condition (*i.e.*, *unconditional preference*), leading to suboptimal model performance and increased hallucination. After applying DPO, the model may show an increased tendency to ig-
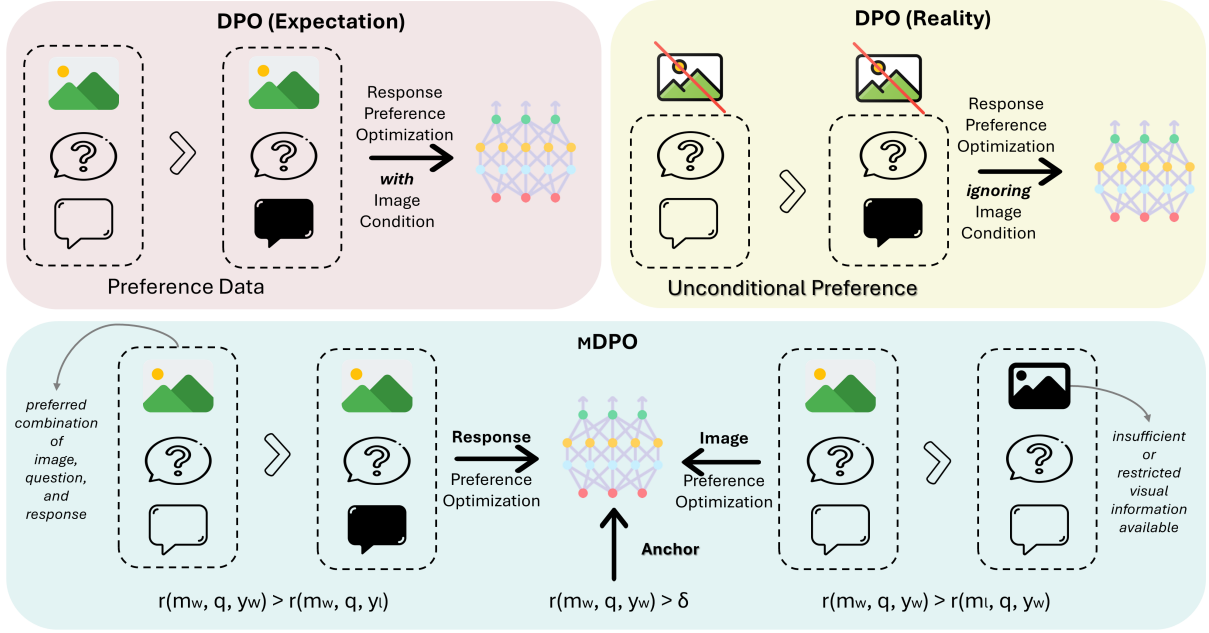
Figure 2: **Overview of MDPO**. *Top Left*: Standard DPO expects the multimodal LLM to learn response preferences conditioned on both the image and the question. *Top Right*: However, in practice, the learning process often disregards the image condition. *Bottom*: To address this issue, MDPO introduces an additional image preference learning objective to emphasize the relationship between the image and the response. Furthermore, MDPO incorporates a reward anchor to ensure that the probability of the chosen response does not decrease.

nore the provided image and generate responses based solely on the question (illustrated in Fig. 3).

In this paper, we propose MDPO, a multimodal DPO objective that utilizes conditional preference optimization on images to prevent the overly prioritization of language-only preferences. As depicted in Fig. 2, in addition to the original preference pairs contrasting responses, MDPO introduces new preference pairs contrasting images. The rejected image is derived from the original (*i.e.*, chosen) image by reducing effective visual information. This approach, combined with the standard DPO objective, encourages the multimodal LLM to simultaneously emphasize both visual and language features. Furthermore, we observe that DPO often experiences a decrease in the likelihood of chosen responses in multimodal scenarios despite an increase in the implicit reward. To address this, MDPO incorporates a reward anchor that maintains the likelihood of chosen responses by regularizing the reward to be positive.

To validate the effectiveness of MDPO, we conduct experiments using Bunny-v1.0-3B (He et al., 2024) and LLaVA-v1.5-7B (Liu et al., 2024a). Both automatic and human evaluations on MMHal-bench (Sun et al., 2023), Object HalBench (Yu et al., 2024a), and AMBER (Wang et al., 2023)

consistently demonstrate that MDPO outperforms standard DPO in multimodal scenarios, effectively reducing hallucinations across varying model and data scales. Detailed analyses reveal that conditional preference plays a crucial role in enhancing the effectiveness of DPO for multimodal LLMs. Fine-grained and qualitative studies further illustrate that MDPO significantly improves the model's ability to comprehend images and mitigates language biases in model responses.

Our contributions are three-fold. First, we identify unconditional preference towards the visual modality as a primary reason for the pitfalls of DPO in multimodal LLMs. Second, we propose MDPO, a multimodal DPO objective that incorporates conditional preference optimization and anchored preference optimization to mitigate these pitfalls. Third, we verify the effectiveness of MDPO across different model and data scales. Using MDPO, we achieve the best-performing 3B multimodal LLM in terms of reducing hallucinations.

2 The Pitfall of Preference Optimization

In this section, we first introduce the background of DPO (§2.1) and then delve into the issue of unconditional preference in multimodal DPO (§2.2).

Adversarial

What is a vicuna standing in the sand looking at?

MDPO: The image features a **drawing of a toucan, not a vicuna**. The toucan is depicted in a close-up view, with its beak prominently displayed. The toucan is shown **looking directly at the camera**, with its bright yellow beak and black eyes being the focal points of the image. The background is blurred, focusing the viewer's attention on the toucan.

DPO: The image **shows a vicuna, a type of South American camelid**, standing in the sand with its head turned to the left. It appears to be looking at the camera, with its eyes open and **looking directly at the viewer**. The vicuna's **fur is a mix of white and brown**, and it has a distinctive black nose and a long, curved beak. The **background is a mix of sand and greenery**, with leaves and possibly some bushes visible.

Bunny: The **vicuna** is **looking at the camera**.

Relation

Is the mouse connected to a computer?

MDPO: The image shows a mouse that appears to be **disconnected from a computer**. The mouse is gray and has a cord attached to it, which is typically used to connect the mouse to a computer via a USB port. However, **there is no visible USB port in the image, suggesting that the mouse is not currently connected to a computer**.

DPO: The mouse appears to be **connected to a computer** via a USB cable. **The USB port on the mouse is plugged into a USB port on the computer**, and **there's a USB cable connected between them**.

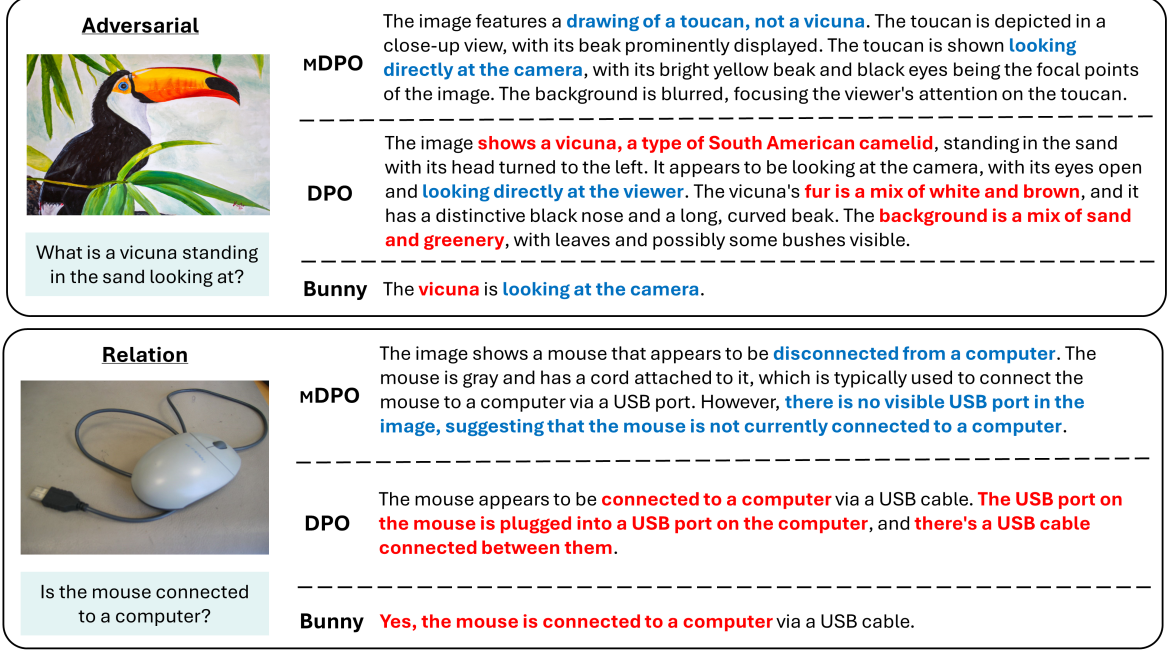**Bunny**: **Yes, the mouse is connected to a computer** via a USB cable.

Figure 3: **Qualitative Results** from MMHalBench. *Top*: When trained with standard DPO, Bunny often assumes the image description in the question is correct, responding accordingly, even if the question contains an adversarial premise regarding the image. In contrast, MDPO identifies the false premise in the question by referencing the image. *Bottom*: Bunny trained with standard DPO may disregard the image and provide an educated guess for the answer. Conversely, MDPO delivers a correct answer that is conditioned on the image.

2.1 Background: Preference Optimization

Preference optimization seeks to align LLMs with human preferences, thereby enhancing their capabilities to respond to human needs. In the context of LLMs, it aims to encourage the model to learn that, for a given question q, the response y_w chosen by the evaluator is preferred over the rejected one y_l. DPO is the predominant method for this purpose. Derived from reward modeling in RLHF (Ouyang et al., 2022), it seeks to maximize the difference between the reward for the chosen response $r(q, y_w)$ and that for the rejected response $r(q, y_l)$. Specifically, given a model to be optimized π_θ and a reference model π_{ref}, which is typically initialized from a supervised finetuning model, DPO formulates the reward as follows:

$$r(q, y) = \beta \log \frac{\pi_\theta(y|q)}{\pi_{\text{ref}}(y|q)} + Z(q),$$

where $Z(q)$ is a partition function, β is a hyperparameter that controls the deviation from the reference model. Then, based on the Bradley-Terry model (Bradley and Terry, 1952), the preference optimization objective becomes:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|q)}{\pi_{\text{ref}}(y_w|q)} - \beta \log \frac{\pi_\theta(y_l|q)}{\pi_{\text{ref}}(y_l|q)} \right),$$

which is essentially maximizing

$$\sigma(r(q, y_w) - r(q, y_l)).$$

In the multimodal scenario, each instance in the preference data contains an image m, in addition to q, y_w, and y_l, and the preference label is decided based on both the image and the question. DPO expects the multimodal LLM to learn to maximize

$$\sigma(r(m, q, y_w) - r(m, q, y_l)),$$

and the objective becomes:

$$\mathcal{L}_{\text{DPO}_m} = -\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|m,q)}{\pi_{\text{ref}}(y_w|m,q)} - \beta \log \frac{\pi_\theta(y_l|m,q)}{\pi_{\text{ref}}(y_l|m,q)} \right).$$

2.2 Problem: Unconditional Preference

Recent studies have found inconsistent improvements in model capabilities when applying DPO to multimodal LLMs, often attributing this issue to the quality of preference data (Li et al., 2023; Sarkar et al., 2024). However, our controlled experiments suggest that the problem arises because DPO does not effectively utilize the visual modality in the preference dataset. To explore this, we introduce a variant called DPO (No Image), which is trained on the preference dataset with the visual signal removed, forcing the model to maximize

$\sigma(r(q, y_w) - r(q, y_l))$ without visual cues. We apply both DPO and DPO (No Image) to the Bunny-v1.0-3B model and 10K preference instances from the LLaVA-Instruct (Liu et al., 2024b) subset of Silkie (Li et al., 2023). Results shown in Fig. 1 indicate that DPO (No Image) performs similarly, or even slightly better, than DPO on MMHalBench. This finding underscores the issue of *unconditional preference* learned by multimodal LLMs during DPO, where the model may disregard image information, as illustrated in Fig. 2.

3   MDPO

In this section, we introduce MDPO, an improved DPO approached dedicated to multimodal preference alignment. As depicted in Fig. 2, MDPO introduces two additional preference optimization objectives to DPO: *conditional preference optimization* to address the issue of ignoring visual information (see §3.1), and *anchored preference optimization* to prevent a decrease in the likelihood of the chosen response (see §3.2).

3.1   Conditional Preference Optimization

We propose a conditional preference optimization objective to address the issue of ignoring visual information in preference data. The core idea is to construct preference data where the image is the only variable, forcing the model to determine the preference label based on visual information. Specifically, given a pair of tuples (m_w, q, y_w) and (m_l, q, y_w), where m_w is more compitible with q and y_w than m_l, the conditional preference optimization objective is formulated as:

$$\mathcal{L}_{\text{CoPO}} = -\log \sigma\Big(\beta \log \frac{\pi_\theta(y_w|m_w,q)}{\pi_{\text{ref}}(y_w|m_w,q)} - \beta \log \frac{\pi_\theta(y_w|m_l,q)}{\pi_{\text{ref}}(y_w|m_l,q)}\Big).$$

The challenge then lies in constructing appropriate pairs of m_w and m_l. On the one hand, m_l should contain different visual information from m_w to make it less compatible. On the other hand, it should also share some common features with m_w to serve as a hard negative. We find that a straightforward strategy, using the original image as m and creating m_l by randomly cropping less than 20% of the original image, yields the best performance, as shown in §4.4.

To summarize, the standard DPO objective maximizes $\sigma(r(m_w, q, y_w) - r(m_w, q, y_l))$, while the conditional preference optimization objective maximizes $\sigma(r(m_w, q, y_w) - r(m_l, q, y_w))$. The two objectives work in collaboration to ensure that the

multimodal LLM captures preferences based on both visual and language cues. Notably, while we focus on the multimodal setting, this conditional preference optimization could be beneficial to other preference optimization scenarios involving multiple input components. In such settings, DPO may also ignore specific input components and encourage the model to learn unconditional preferences.

3.2   Anchored Preference Optimization

We also observe that the likelihood of the chosen response often decreases during the optimization process of DPO. This occurs because the standard DPO objective only encourages the model to learn a relative preference. Without further regularization, the model may reduce the likelihood of the chosen response to enlarge the likelihood gap between the chosen and rejected responses. This can harm model performance, as the chosen responses are often of high quality. To address this problem, we add an anchor to the preference optimization, forcing the reward of the chosen response to be higher than a specific value: $\sigma(r(m_w, q, y_w) - \delta)$ with δ as the anchor value. The corresponding objective is

$$\mathcal{L}_{\text{AncPO}} = -\log \sigma\Big(\beta \log \frac{\pi_\theta(y_w|m_w,q)}{\pi_{\text{ref}}(y_w|m_w,q)} - \delta\Big).$$

In this way, we introduce absolute reward regularization to the preference optimization process, effectively avoiding the likelihood decrease of the chosen response. The anchor is decided based on the data properties and expected model behavior. While we keep it simple in the default setting of MDPO, one can always change the anchor values and set multiple anchors for different purposes. The anchor can also be added to the rejected response in the opposite direction, forcing its reward to be lower than a specific value. We compare other anchors in §4.4.

The objective of MDPO is a combination of the standard DPO, conditional preference optimization, and anchored preference optimization:

$$\mathcal{L}_{\text{MDPO}} = \mathcal{L}_{\text{DPO}_m} + \mathcal{L}_{\text{CoPO}} + \mathcal{L}_{\text{AncPO}}.$$

4   Experiment

In this section, we begin with the experimental setup (§4.1). Then we present the main results on three benchmarks (§4.2) and the human evaluation

	MMHalBench		Object HalBench		AMBER			
	Score ↑	HalRate ↓	CHAIR$_s$ ↓	CHAIR$_i$ ↓	CHAIR$_s$ ↓	Cover. ↑	HalRate ↓	Cog. ↓
Referenced Results (Not Directly Comparable)								
GPT-4V (Achiam et al., 2023)[†♯]	3.49	0.28	13.6	7.3	4.6	67.1	30.7	2.6
LLaVA-v1.5-7B (Liu et al., 2024a)[‡♯]	2.11	0.54	53.6	25.2	7.8	51.0	36.4	4.2
+ HACL (Jiang et al., 2024)[‡]	2.13	0.50	-	-	-	-	-	-
+ POVID (Zhou et al., 2024b)[♯]	2.08	0.56	48.1	24.4	-	-	-	-
+ OPERA (Huang et al., 2024)[♯]	2.15	0.54	45.1	22.3	-	-	-	-
+ VCD (Leng et al., 2024)[♯]	2.12	0.54	48.8	24.3	-	-	-	-
+ EOS (Yue et al., 2024)[‡♯]	2.03	0.59	40.3	17.8	5.1	49.1	22.7	2.0
+ HA-DPO (Zhao et al., 2023)[‡♯]	1.97	0.60	39.9	19.9	6.7	49.8	30.9	3.3
+ HALVA (Sarkar et al., 2024)[‡]	2.25	0.54	-	-	6.6	53.0	32.2	3.4
LLaVA-v1.5-13B (Liu et al., 2024a)[†]	2.42	-	46.3	22.6	7.8	51.0	36.4	4.2
+ RLHF-V (Yu et al., 2024a)[†]	2.81	0.49	12.2	7.5	6.3	46.1	25.1	2.1
+ HSA-DPO (Xiao et al., 2024)[†]	2.61	0.48	5.2	3.2	2.1	47.3	13.4	1.2
+ HALVA (Sarkar et al., 2024)[‡]	2.58	0.45	-	-	6.4	52.6	30.4	3.2
Qwen-VL-Chat (Bai et al., 2023)[†]	2.89	0.43	36.0	21.3	6.6	53.2	31.0	2.9
+ Silkie-80K (Li et al., 2023)[†]	3.01	0.41	25.3	13.9	5.4	55.8	29.0	2.0
3B Multimodal LLMs								
Bunny-v1.0-3B (He et al., 2024)	2.11	0.58	43.0	8.9	9.8	**75.6**	64.9	6.0
+ DPO	2.28	0.56	44.3	7.6	7.9	74.1	58.9	4.8
+ MDPO	**2.96**	**0.42**	**27.0**	**4.6**	**4.9**	67.4	**37.7**	**2.4**
7B Multimodal LLMs								
LLaVA-v1.5-7B (Liu et al., 2024a)	2.19	0.57	54.7	15.9	7.4	51.8	34.7	4.1
+ DPO	2.14	0.65	49.0	13.0	6.5	**55.1**	34.5	**2.3**
+ MDPO	**2.39**	**0.54**	**35.7**	**9.8**	**4.4**	52.4	**24.5**	2.4

Table 1: **Main results** of Bunny-v1.0-3B and LLaVA-v1.5-7B trained with different preference optimization objectives. We report overall score and hallucination rate (HalRate) on MMHalBench, CHAIR scores at both response and object levels on Object HalBench, along with CHAIR scores, object coverage (cover.), hallucination rate (HalRate), and cognition (Cog.) on AMBER. The best result for each metric in each group is in bold. For reference, we also provide additional results using various multimodal LLMs, preference data, and learning objectives, although these are not directly comparable. Results from contemporary work focusing on multimodal preference data: [†]Xiao et al. (2024), [‡]Sarkar et al. (2024), and [♯]Yu et al. (2024b).

(§4.3) of MDPO. We further provide in-depth analysis (§4.4) and fine-grained results (§4.5). Finally, we conduct a qualitative study (§4.6).

4.1 Experimental Setup

Models. We apply MDPO on two multimodal LLMs in different sizes. Bunny-v1.0-3B (He et al., 2024) is a 3B model building upon SigLIP (Zhai et al., 2023) and Phi-2 (Javaheripi et al., 2023). It is pretrained on 2M image-text pairs and finetuned on 695K instruction tuning data. LLAVA-v1.5-7B (Liu et al., 2024a) is a 7B model based on CLIP (Radford et al., 2021) and Vincuna (Chiang et al., 2023). It is pretrained on 558K image-text pairs and finetuned on 665K instruction tuning data.

Preference Data. We sample 10K preference data from Silkie (Li et al., 2023) with instructions from LLaVA-Instruct-150K (Liu et al., 2024a) for training. The original Silkie dataset contains 80K preference data collected on 12 multimodal LLMs. While the original Silkie paper explores the effect of extreme data size, we follow the majority of prior works using around 10K data for preference optimization (Sun et al., 2023; Zhao et al., 2023).

Evaluation Benchmarks. We evaluate the performance of MDPO on three widely used benchmarks for multimodal LLMs with a special focus on hallucination. MMHalBench (Sun et al., 2023) is a practical question answering benchmark containing eight question categories and 12 object topics. Following the official setting, we use GPT4 (Achiam et al., 2023) to assess the overall quality of responses with a score betwen zero and six, and the hallucination rate. Object HalBench (Rohrbach et al., 2018) is a widely adopted benchmark to as-
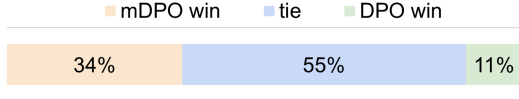
Figure 4: **Human evaluation** on MMHalBench.

Figure 5: **Impact of data scale** on the performance of standard DPO and MDPO, using Bunny as the base model. We assess the overall score and hallucination rate on MMHallBench. MDPO is effective across different scales, whereas standard DPO does not exhibit a scaling effect in multimodal scenarios.

sess object hallucination. We follow the setting of Yu et al. (2024a) to augment the benchmark with eight diverse prompts and evaluating on 300 instances. We report the CHAIR scores (Rohrbach et al., 2018) assessing hallucination rate of response level (CHAIR_s) and object level (CHAIR_i). AMBER (Wang et al., 2023) is a multimodal LLM hallucination benchmark with fine-grained object annotation. We focus on the generative task consisting of 1K images. Using the official evaluation tool, we report a variant of CHAIR score, object coverage, rate of hallucinated responses, and hallucination rate overlapping with human cognition.

Baselines. We primarily compare MDPO with standard DPO. The standard DPO baseline shares the same training process, data, and hyperparameters, despite different learning objectives. We further provide the results of other multimodal LLMs for reference, although they are not directly comparable due to different base models, preference data, and alignment methods. This group contains GPT-4V (Achiam et al., 2023), LLaVA-v1.5-13B (Liu et al., 2024a), Qwen-VL-Chat (Bai et al., 2023), POVID (Zhou et al., 2024b), HACL (Jiang et al., 2024), OPERA (Huang et al., 2024), VCD (Leng et al., 2024), EOS (Yue et al., 2024), HA-DPO (Zhao et al., 2023), HALVA (Sarkar et al., 2024), RLHF-V (Yu et al., 2024a), HSA-DPO (Xiao et al., 2024), and Silkie (Li et al., 2023). Some of them are contemporary works to ours.

Implementation Details. We train all the models for 3 epochs with a batch size of 32. We use a learning rate of 0.00001, a cosine learning rate scheduler, and a warmup ratio of 0.1. We set the β of preference optimization to 0.1. Following prior work (Zhao et al., 2023; Li et al., 2023), we use LoRA (Hu et al., 2021) to tune the model. Specifically, we set the α to 128 and rank to 64 for LoRA. MDPO and standard DPO share the same configuration above. For MDPO, we set $\delta = 0$ by default.

4.2 Main Results

Tab. 1 presents the main results. On all three benchmarks, MDPO consistently performs better than

DPO for Bunny and LLaVA. Notably, MDPO enhances the 3B model (Bunny) with 10K preference data to be comparable to a stronger 7B base model (Qwen-VL-Chat) trained with DPO on 80K data. The former preference data is only a subset of the latter. This result highlights that a proper objective can be more important than data scale and diversity in multimodal preference optimization. Moreover, MDPO is specifically effective in reducing hallucination, which aligns with the objective of conditional preference optimization. While MDPO may lead to a decrease in object coverage, this decrease is minor given the significant improvement in overall quality and reduction in hallucination.

4.3 Human Evaluation

To further verify the effectiveness of MDPO, we conduct human evaluation on MMHalBench, in which we ask domain experts to pick the better response generated by Bunny trained with either DPO or MDPO. The results are presented in Fig. 4. Overall, responses from MDPO are of better or same quality on 89% instances compared to DPO. In contrast, DPO only achieves better performance on 11% instances.

4.4 Analysis

MDPO is effective across different scales of preference data. We assess the overall score and hallucination rate on MMHallBench, as shown in Fig. 5. We find that MDPO is effective and consistently outperforms DPO across different data scales, demonstrating that our conditional preference method enhances multimodal preference opti-

	MMHalBench		Object HalBench	
	Score	HalRate	CHAIR$_s$	CHAIR$_i$
mDPO	**2.96**	**0.42**	**27.0**	**4.6**
- conditional	2.36	0.53	40.3	7.1
- anchored	2.50	0.48	34.3	5.7
- both (i.e., DPO)	2.28	0.56	44.3	7.6

Table 2: **Ablation results on MDPO with conditional preference or/and anchored preference removed.** While both components are essential in MDPO, anchored preference alone brings only slight improvement over DPO. This indicates that conditional preference is crucial in multimodal scenarios.

	MMHalBench		Object HalBench	
	Score	HalRate	CHAIR$_s$	CHAIR$_i$
Random image	2.81	0.46	40.7	6.6
Crop 0-20%	**2.96**	**0.42**	**27.0**	**4.6**
Crop 20%-50%	2.92	0.42	33.7	5.4
MoCo v2	2.82	0.44	32.3	5.9

Table 3: **Comparison of strategies to create rejected images in MDPO.** Among all the strategies, the default strategy in MDPO, Cropping 0-20% of the chosen images, retains some similarities with the original images but contains insufficient visual information, thereby providing effective preference optimization signals. In contrast, random images are too easy to identify, while MoCo v2's data augmentation (Chen et al., 2020) may not produce clearly worse images.

mization. Additionally, we observe that MDPO's performance increases with the scale of data, while DPO does not exhibit a scaling effect. This indicates that MDPO better utilizes multimodal preference data compared to DPO. Specifically, DPO struggles to fully leverage multimodal preference data, and its neglect of the visual modality cannot be mitigated by merely increasing the size of the preference data.

Both designs in MDPO are effective, with conditional preference being more crucial. We conduct an ablation study to evaluate the contributions of each component in MDPO, as shown in Tab. 2. While both anchored preference and conditional preference enhance the overall performance of MDPO, the results indicate that conditional preference leads to greater improvements than anchored preference. This suggests that conditional preference is the key factor in enhancing the effectiveness of DPO in multimodal scenarios, ensuring that the model better utilizes the visual modality.

Using hard negative images for rejection im-

Anchor	MMHalBench		Object HalBench	
	Score	HalRate	CHAIR$_s$	CHAIR$_i$
$\mathbf{y_w}$	2.96	0.42	**27.0**	**4.6**
y_w & y_l	**2.98**	**0.39**	29.3	5.0
y_w & y_l & m_l	2.85	0.4	34.7	6.1

Table 4: **Comparison of anchors used in MDPO.** MDPO adds an anchor to regularize the $r(m_w, q, y_w)$ to be positive by default. Adding additional anchors to regularize the rewards of instances with rejected responses (y_l) or images (m_l) to be negative does not show an obvious improvement.

proves preference optimization. We evaluate MDPO with different methods for constructing the rejected image, as shown in Tab. 3. Among all the strategies, the default strategy in MDPO, which involves cropping 0-20% of the chosen images, consistently outperforms the others. This indicates that using hard negative images, which retain some similarities with the original images but also have parts erased, provides effective preference optimization signals. In contrast, random images are too easy to identify, while MoCo v2's data augmentation (Chen et al., 2020) is for creating similar images.

Adding anchors to rejected responses or images brings litter improvement. In MDPO, we introduce an anchor to regularize $r(m_w, q, y_w)$ to be positive by default. We also experimented with adding additional anchors to regularize $r(m_w, q, y_l)$ and $r(m_l, q, y_w)$ to be negative. However, the additional anchors do not yield significant improvements. The results, shown in Tab. 4, indicate that only using the anchor on $r(m_w, q, y_w)$ is sufficient. Adding anchors to rejected responses or images may complicate the training process without providing clear advantages.

4.5 Fine-grained Results

We further compare the fine-grained results of DPO and MDPO on MMHalBench. As shown in Tab. 5, among the eight question categories, MDPO outperforms standard DPO on six of them. MDPO shows significant improvement particularly on adversarial questions with false premises about images. MDPO can identify the incorrect information in the question according to the image, while DPO fail to do so. These results also show the advantage of MDPO under various practical scenarios.

	overall	attribute	adversarial	comparison	counting	relation	environment	holistic	other
Bunny	2.11	**3.92**	0.83	**2.17**	2.33	2.67	2.25	1.75	1.00
+ DPO	2.28	3.25	1.50	1.42	2.50	2.67	**4.25**	1.75	0.92
+ MDPO	**2.96**	3.08	**4.17**	2.00	**3.50**	**3.25**	4.08	**2.17**	**1.42**

Table 5: **Fine-grained results on MMHalBench** with a maximum score of six. MDPO outperforms standard DPO on six out of eight types of questions, showing significant improvement particularly on adversarial questions with false premises about images.

4.6 Qualitative Study

In Fig. 3, we compare the When trained with standard DPO, Bunny often assumes the image description in the question is correct, responding accordingly, even if the question contains an adversarial premise regarding the image. In contrast, MDPO identifies the false premise in the question by referencing the image. Moreover, Bunny trained with standard DPO may disregard the image and provide an educated guess for the answer. Conversely, MDPO delivers a correct answer that is conditioned on the image.

5 Related Work

Reinforcement learning from human feedback (RLHF; Christiano et al. 2017; Ouyang et al. 2022) has proven to be an effective approach for aligning LLMs with human values. Direct preference optimization (DPO; Rafailov et al. 2023), which involves directly optimizing LLMs based on human preferences, has been widely adopted in RLHF due to its strong performance and the elimination of the need for a separate reward model. Significant efforts have been made to further enhance the efficacy and efficiency of DPO, which can be categorized into algorithmic and data-related advancements. On the algorithmic side, various approaches aim to improve the efficiency of DPO. For example, ORPO (Hong et al., 2024) models preferences using an odds ratio and combines instruction fine-tuning and preference optimization into a unified training process. Methods such as CPO (Xu et al., 2024a), TPO (Saeidi et al., 2024), and SimPO (Meng et al., 2024) simplify DPO by eliminating the use of a reference model, thereby reducing computational and memory overhead. Additionally, IPO (Azar et al., 2024) addresses the issue of reward overfitting in DPO. On the data side, approaches such as KTO (Ethayarajh et al., 2024) and NCA (Chen et al., 2024) seek to overcome DPO's requirement for paired preference data by designing optimization goals that can also utilize

unpaired data. Iterative DPO (Xu et al., 2023; Yuan et al., 2024; Xiong et al., 2024) and SPPO (Wu et al., 2024) propose sampling preference data in an on-policy manner, achieving better results than off-policy DPO. WPO (Zhou et al., 2024a) adapts off-policy data to resemble on-policy data by reweighting preference pairs. More closely related to our work, studies by Park et al. (2024) and Dong et al. (2024) address the reward hacking problem in textual preference optimization, where human preference may be biased towards longer outputs. They propose to calibrate the rewards in DPO with respect to output length. In this work, we discover a novel challenge in multimodal DPO, where preference optimization often neglects images. We then propose a solution to this problem through conditional preference optimization.

In multimodal scenarios, recent works mainly focus on creating multimodal preference data (Li et al., 2023; Zhao et al., 2023; Xiao et al., 2024; Zhou et al., 2024b; Pi et al., 2024; Sarkar et al., 2024; Yu et al., 2024b; Deng et al., 2024). These efforts include collecting human preference (Sun et al., 2023; Yu et al., 2024a), preference from advanced multimodal LLMs (Li et al., 2023; Yu et al., 2024b), and preference from the model to align itself (Deng et al., 2024). In terms of learning objectives, recent works mainly follows DPO for LLMs (Li et al., 2023; Zhao et al., 2023; Zhou et al., 2024b). Some also apply reinforcement learning (Sun et al., 2023; Jing and Du, 2024) and contrastive learning (Sarkar et al., 2024; Jiang et al., 2024). Our work studies an overlooked but crucial problem in the multimodal DPO objective.

6 Conclusion

We propose MDPO, a preference optimization method dedicated to multimodal scenarios. MDPO leverages conditional preference optimization to encourage multimodal LLMs to capture preference labels based on both visual and language cues. It further introduces anchored preference optimization to prevent the likelihood of preferred

responses from decreasing. Experiments show that MDPO consistently enhances multimodal LLM performance and reduces hallucination across different model sizes on three widely used benchmarks. Our method could be extended to reduce hallucinations and enhance trustworthiness across broader multimodal data and scenarios, including multiple images (Wang et al., 2024), videos (Li et al., 2024), in-context learning (Xu et al., 2024b), and risk-sensitive domains (Chaves et al., 2024).

Acknowledgement

Limitation

While we have conducted comprehensive experiments to show the effectiveness of MDPO, there are still several limitations. First, experiments on more multimodal LLMs will provide further evidence on the advantages and disadvantages of MDPO, specifically on models across various sizes and different architectures. Second, we focus on the unconditional preference problem in multimodal preference optimization. However, many contemporary studies have explored enhancing DPO from other perspectives, which may be complementary to ours. We leave the analysis of combining methods for future work. Third, while we have evaluated MDPO on three benchmarks, they still represent a limited range of tasks and settings compared with the numerous scenarios in the real world. Further evaluation on more benchmarks can deepen our understanding of the proposed method.

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*.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. 2024. Training small multimodal models to bridge biomedical competency gap: A case study in radiology imaging. *arXiv preprint arXiv:2403.08002*.

Huayu Chen, Guande He, Hang Su, and Jun Zhu. 2024. Noise contrastive alignment of language models with explicit rewards. *arXiv preprint arXiv:2402.05369*.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

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

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. 2024. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Reference-free monolithic preference optimization with odds ratio. *arXiv preprint arXiv:2403.07691*.

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. *arXiv preprint arXiv:2106.09685*.

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.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.

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.

Liqiang Jing and Xinya Du. 2024. Fgaif: Aligning large vision-language models with fine-grained ai feedback. *arXiv preprint arXiv:2404.05046*.

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.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. 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. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*.

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. *arXiv preprint arXiv:2403.08730*.

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. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Amir Saeidi, Shivanshu Verma, Aswin RRV, and Chitta Baral. 2024. Triple preference optimization: Achieving better alignment with less data in a single step optimization. *arXiv preprint arXiv:2405.16681*.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. 2024. Mitigating object hallucination via data augmented contrastive tuning. *arXiv preprint arXiv:2405.18654*.

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*.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. 2024. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.

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*.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. 2024. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*.

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. *arXiv preprint arXiv:2404.14233*.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. 2024. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*.

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. *arXiv preprint arXiv:2401.08417*.

Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. 2023. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*.

Nan Xu, Fei Wang, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024b. From introspection to best practices: Principled analysis of demonstrations in multimodal in-context learning. *arXiv preprint arXiv:2407.00902*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024a. 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. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

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

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models.

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*.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. 2024a. Wpo: Enhancing rlhf with weighted preference optimization.

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