Modality-Fair Preference Optimization for Trustworthy MLLM Alignment

Songtao Jiang[1] , **Yan Zhang**[2] , **Ruizhe Chen**[1] , **Tianxiang Hu**[1]
Yeying Jin[3] , **Qinglin He**[1] , **Yang Feng**[4] , **Jian Wu**[1] , **Zuozhu Liu**[1†]
[1]Zhejiang University
[2]ByteDance
[3]National University of Singapore
[4]Angelalign Inc., China
{songtao.22, zuozhuliu}@intl.zju.edu.cn

Abstract

Multimodal large language models (MLLMs) have achieved remarkable success across various tasks. However, separate training of visual and textual encoders often results in a misalignment of the modality. Such misalignment may lead models to generate content that is absent from the input image, a phenomenon referred to as hallucination. These inaccuracies severely undermine the trustworthiness of MLLMs in real-world applications. Despite attempts to optimize text preferences to mitigate this issue, our initial investigation indicates that the trustworthiness of MLLMs remains inadequate. Specifically, these models tend to provide preferred answers even when the input image is heavily distorted. Analysis of visual token attention also indicates that the model focuses primarily on the surrounding context rather than the key object referenced in the question. These findings highlight a misalignment between the modalities, where answers inadequately leverage input images. Motivated by our findings, we propose Modality-Fair Preference Optimization (MFPO), which comprises three components: the construction of a multimodal preference dataset in which dispreferred images differ from originals solely in key regions; an image reward loss function encouraging the model to generate answers better aligned with the input images; and an easy-to-hard iterative alignment strategy to stabilize joint modality training. Extensive experiments on three trustworthiness benchmarks demonstrate that MFPO significantly enhances the trustworthiness of MLLMs. In particular, it enables the 7B models to attain trustworthiness levels on par with, or even surpass, those of the 13B, 34B, and larger models.

1 Introduction

Recent advances in multimodal large language models (MLLMs) [Liu *et al.*, 2024b; Wang *et al.*, 2023b; Jiang *et al.*, 2024b] have achieved remarkable performance on diverse multimodal tasks. However, separate training of visual and textual encoders often causes a misalignment of the modality [Cui *et al.*, 2023; Guan *et al.*, 2024] of MLLMs, which can cause the model to generate content that does not exist in the visual input—a phenomenon known as hallucination [Bai *et al.*, 2024]. This issue compromises the reliability of MLLMs in real-world applications [Shah *et al.*, 2019].

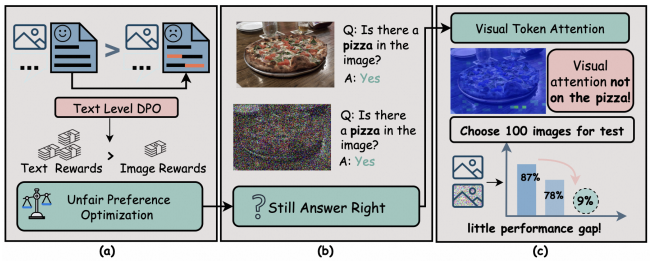

Figure 1: (a) Current preference optimization (PO) in LLaVA-v1.5 results in imbalanced modality rewards. (b) After PO, LLaVA-v1.5 exhibits a persistent bias towards preferred responses, even with distorted image inputs. (c) Visual token attention analysis, supported by experimental evidence, indicates that MLLMs tend to memorize text preferences rather than achieving modality alignment.

To address this issue, recent work has applied preference optimization to improve the alignment of modalities in MLLM [Rafailov *et al.*, 2024; Naveed *et al.*, 2023]. These methods typically train models to favor accurate responses over those that contain erroneous image descriptions. They achieve this by using inaccurate outputs as rejected options to enhance learning from correct responses, thus improving the alignment of modality [Wang *et al.*, 2024; Zhang and Rong, 2024].

Although current preference optimization methods show success on trustworthiness benchmarks, they mainly rely on text preference optimization. This raises a critical question: *What does the model learn during preference optimization when only text preferences are used? Is it merely memorizing text preferences or is it learning to align image and text data?* Our experiments shown in Figure 1 reveal that even with severe noise-degrading image content [Leng *et al.*, 2024], the model consistently provides accurate answers. Across a sample of 100 images, the disparity in performance between the original and degraded images is only 9%. Furthermore, the

† Corresponding author: zuozhuliu@intl.zju.edu.cn

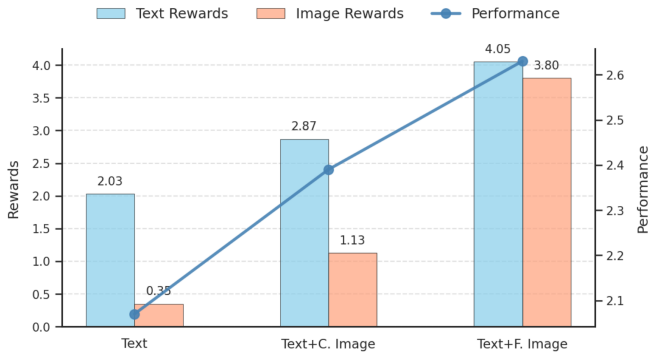

Figure 2: "Text" refers to methods utilizing manually annotated text preference datasets (RLHF-V [Yu *et al.*, 2024a]). "Text + C. Image" involves image preference datasets generated by randomly cropping images [Wang *et al.*, 2024], while "Text + F. Image" uses fine-grained image preference datasets constructed with SAM and manual annotations. As the quality of image preference datasets improves, rewards across modalities become more balanced, leading to higher overall rewards and enhanced trustworthiness.

visualization of visual token attention shows that the model mainly focuses on the surrounding contexts rather than the key object mentioned in the question. These findings suggest that the model seems to lean more towards memorizing text preferences rather than achieving modality alignment after preference optimization.

We explore implementing more modality balanced preference optimization to encourage MLLMs to focus more on modality alignment for improved trustworthiness. As shown in Figure 2, we evaluated adding image-related rewards with a simple image preference dataset and a more fine-grained image preference dataset. As we progressively update the approach, the image and text rewards become higher and also more balanced, and the model's trustworthy performance consistently improves. The limited impact of simple image degradation may be due to the model quickly learning to distinguish these variations after only a few training steps, which leads it to continue prioritizing text preference optimization. These findings illustrate the effectiveness of more balanced optimization in achieving trustworthiness. Motivated by our investigation, we propose Modality-Fair Preference Optimization (MFPO) for trustworthy MLLM alignment. To circumvent mere memorization of text responses, we propose to jointly optimize over both text and image preference in a modality-fair manner. In particular, we first construct a multimodal preference dataset encompassing both text and image preference data. Beginning with a text preference dataset, we characterize the relationship between words using a multipartite graph [Jin and Zhang, 2014; Boudin, 2018] to identify the top-K keywords. These keywords are then linked to relevant image regions using a modified Segment Anything Model (SAM) [Kirillov *et al.*, 2023]. By introducing diffusion noise, we generate dispreferred image data that remains largely faithful to the originals while introducing meaningful differences [Jin and Zhang, 2014; Dawande *et al.*, 2001], which are critical for preference optimization. Afterwards, to encourage fair alignment over both modalities, we devise a novel loss function consisting of text preference alignment, image preference alignment and a margin loss for stablized training. Furthermore, to tackle the instability in joint modality optimization, we propose an iterative alignment approach [Paas *et al.*, 2003]. By categorizing training data into varying complexities using semantic entropy, we initially train the model on simpler data and gradually advance to more challenging samples.

Extensive experiments demonstrate the effectiveness of MFPO in improving model trustworthiness. Remarkably, MFPO empowers 7B LLaVA-v1.5 to achieve trustworthiness levels that are comparable to, or even exceed, those of significantly larger 13B, 34B, and larger-scale models. Specifically, when applied to models such as LLaVA-v1.5-13B, MFPO outperforms GPT-4V, achieving a substantial 40% improvement on Object HalBench and setting new state-of-the-art results on both Object HalBench and AMBER benchmarks. Furthermore, LLaVA-v1.5-7B+MFPO and LLaVA-v1.5-13B+MFPO demonstrate superior performance over GPT-4V across five out of eight evaluation metrics. Comprehensive ablation studies further confirm the efficacy of the three core components in MFPO. Our work underscores the importance of balanced preference optimization over different modalities for MLLMs.

2 Preliminaries

Reinforcement Learning from Human Feedback (RLHF) aligns models with human preferences using a reward model r_ϕ [Schulman *et al.*, 2017], trained on pairwise preference data [Christiano *et al.*, 2017; Ouyang *et al.*, 2022; Casper *et al.*, 2023]. The model assigns higher rewards to preferred outputs, with the cross-entropy loss:

$$L_{\text{RM}} = -\log\left(\sigma\left(r_\phi(x, y_w) - r_\phi(x, y_l)\right)\right), \tag{1}$$

where $\sigma(\cdot)$ is the logistic sigmoid, and $r_\phi(x, y_w)$ and $r_\phi(x, y_l)$ are the rewards for the preferred and less preferred outputs.

After training, the policy π_θ is optimized by maximizing the expected reward while regularizing it to prevent divergence from the reference policy π_{ref}, using a KL divergence penalty:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim D, y\sim\pi_\theta(y|x)}\left[r_\phi(x, y) - \beta D_{\text{KL}}(\pi_\theta(y|x) \parallel \pi_{\text{ref}}(y|x))\right], \tag{2}$$

where β balances reward maximization and regularization. The KL divergence ensures policy stability [Peng *et al.*, 2023].

Direct Preference Optimization (DPO) directly optimizes a policy π_θ using preference data D [Rafailov *et al.*, 2024], without relying on a reward model. Concretely, DPO derives a mapping between the reward $r(x, y)$ and the policy π_θ:

$$r(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x), \tag{3}$$

where $Z(x)$ is the partition function, defined as $Z(x) = \sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)$. Substitute Eq. 3 into Eq. 1, we can derive the following loss:

$$L_{\text{DPO}}(\pi_\theta; \pi_{\text{sft}}) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log\sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{sft}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{sft}}(y_l|x)}\right)\right], \tag{4}$$

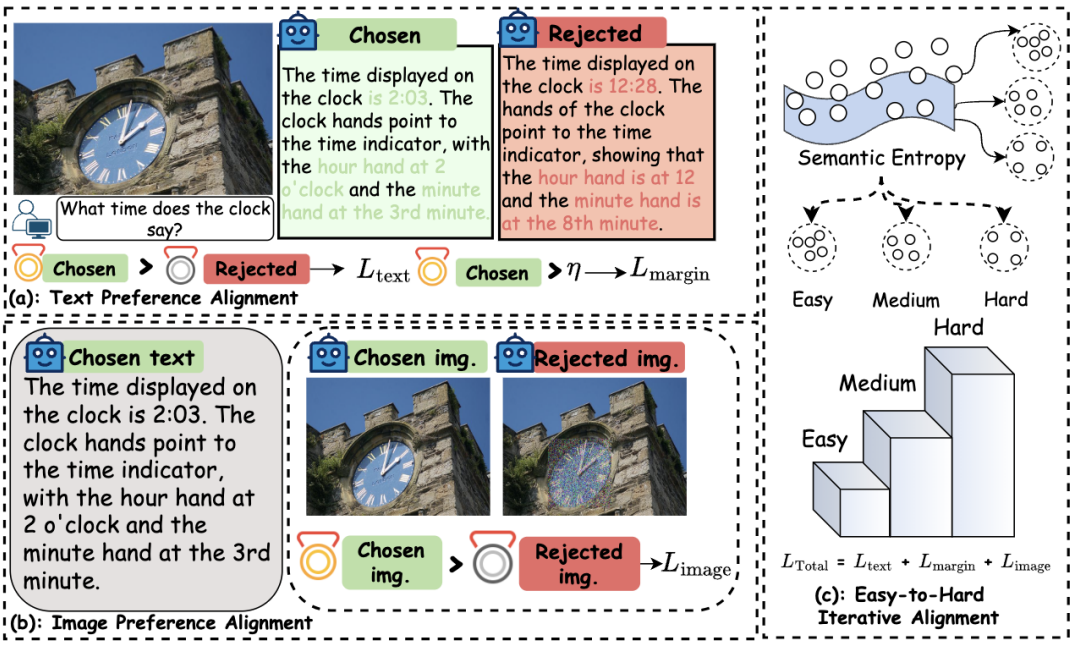

Figure 3: Overview of the MFPO framework. (a) Chosen and rejected text responses are compared to compute text loss L_{text}, with margin loss L_{margin} for reward stability. (b) Chosen and perturbed rejected images are compared, optimizing image preferences with loss L_{image}. (c) Samples are categorized by difficulty using semantic entropy, facilitating progressive training. The total loss L_{Total} combines L_{text}, L_{margin}, and L_{image}.

where y_w and y_l denote the preferred and less preferred outputs, respectively. This approach bypasses the complexity of RLHF by allowing the policy gradient to be computed analytically, ensuring efficient alignment with human preferences [Xu *et al.*, 2024; Feng *et al.*, 2024].

The common approach in DPO for MLLMs [Yu *et al.*, 2024a; Yu *et al.*, 2024b; Sun *et al.*, 2023; Zhou *et al.*, 2024] involves concatenating the image m and question t into a single input x, followed by optimization using Eq. 4. Chosen responses y_w typically provide accurate, hallucination-free information, while rejected responses y_l often contain hallucinated details, making them less preferred.

3 Modality-Fair Preference Optimization

3.1 Image Preference Data Generation

We first introduce the construction of our multimodal preference data as shown in Figure 4. A key question is: *what defines effective image preference data in DPO for MLLMs?* Previous works on text preference data [Chen *et al.*, 2024] suggest that optimal preference data consists of chosen and rejected responses with moderate differences, as this balance enhances gradient effectiveness during training and helps the model capture subtle content distinctions. Inspired by this, our approach focuses on generating image preference data with fine-grained variation, which involves introducing perturbations only to the key regions of the original image most relevant to the chosen responses. The details of the hyperparameters in this section can be found in Supplementary Section 2.

Keyword Selection. To achieve balanced optimization of image and text preferences, we refine the construction of image preference data at the region level to better capture image preferences. We first construct a multipartite graph [Boudin, 2018], $G = (V, E)$, where nodes V represent words and edges E encode both positional and semantic relationships. The positional influence between two words k_i and k_j is defined as $\theta_{ij} = \sum_{l_i \in \mathcal{L}(k_i)} \sum_{l_j \in \mathcal{L}(k_j)} \frac{1}{1+|l_i - l_j|^{\varphi}}$, where φ controls the decay with distance. Semantic similarity is calculated using cosine similarity: $S(k_i, k_j) = \frac{\mathbf{v}(k_i) \cdot \mathbf{v}(k_j)}{\|\mathbf{v}(k_i)\| \|\mathbf{v}(k_j)\|}$, and the edge weight is a combination of positional and semantic factors: $\mu_{ij} = \theta_{ij} + \sigma_{ij}$, with $\sigma_{ij} = \gamma \cdot S(k_i, k_j)$. We further refine these relationships by adjusting edge weights contextually using related words and apply positional decay $\kappa_{ij} = e^{-\lambda l_i}$ to emphasize proximity. The final edge weight $\omega_{ij} = \tau_{ij} \cdot \kappa_{ij}$ integrates positional, semantic, and contextual factors.

Once the graph is constructed, we compute a ranking score for each word based on its centrality using a weighted PageRank-inspired algorithm [Bianchini *et al.*, 2005; Mihalcea and Tarau, 2004]. The initial score for each word k_i is $r_0(k_i) = (1 - \alpha_{\text{page}})$, where α_{page} is a damping factor. The score of each word is influenced by its neighbors, with the contribution from a neighboring word k_j calculated as $r_c(k_i, k_j) = \frac{\omega_{ij} \cdot r(k_j)}{\sum_{k_m \in \mathcal{N}(k_j)} \omega_{jm}}$, where ω_{ij} is the edge weight, $r(k_j)$ is the neighbor's score, and $\mathcal{N}(k_j)$ represents the neighbors of k_j. The score for each word k_i is updated iteratively using $r(k_i) = (1 - \alpha_{\text{page}}) + \alpha_{\text{page}} \cdot \sum_{k_j \in \mathcal{N}(k_i)} r_c(k_i, k_j)$, with α_{page} controlling the influence of neighbors. This process continues until convergence, where all words are ranked by their final scores $r(k_i)$. The top-K keywords, corresponding to the most significant regions in the image, are then selected.

Region Mapping and Perturbation. The top-ranked keywords guide a modified Segment Anything Model (SAM) [Kir-

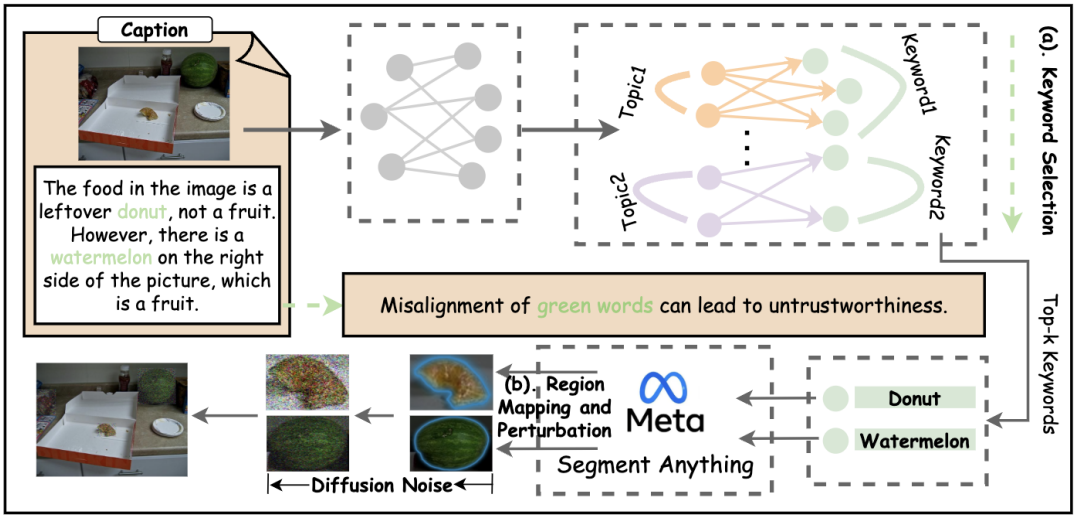

Figure 4: Our data generation pipeline consists of two stages: (a) Keyword Selection, where top-K keywords are identified using a multipartite graph; and (b) Region Mapping and Perturbation, where selected regions are perturbed with diffusion noise to generate rejected images.

illov *et al.*, 2023], mapping each keyword k_i to its corresponding visual regions. We then apply diffusion noise to perturb these regions [Ho *et al.*, 2020]. For each region R_i of the image m, the perturbation follows the diffusion process $m' = \sqrt{\alpha_{\text{diff},t}} \cdot R_i + \sqrt{1-\alpha_{\text{diff},t}} \cdot \epsilon$, where $\alpha_{\text{diff},t} = \prod_{j=0}^{t} \beta_{\text{diff},j}$ (with $\beta_{\text{diff},t} \in (0,1)$ as a hyperparameter controlling the noise schedule), and ϵ is random noise sampled from a standard normal distribution. This perturbation selectively distorts key regions, encouraging the model to focus on critical areas. For images whose keywords fail to locate any visual region, we add diffusion noise across all image areas to generate rejected image data. These perturbed images m' serve as rejected preference images in DPO, refining the model's alignment with fine-grained visual details.

3.2 MFPO Training Process

We leverage the generated image preference data for a unified optimization of the policy π_θ across both text and image.

Text Preference Alignment. The DPO policy objective for text preference optimization is:

$$L_{\text{text}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_D \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|t,m)}{\pi_{\text{ref}}(y_w|t,m)} - \beta \log \frac{\pi_\theta(y_l|t,m)}{\pi_{\text{ref}}(y_l|t,m)} \right) \right], \tag{5}$$

where t and m represents the original text and image inputs. y_w and y_l are the preferred and less preferred outputs. π_{ref} is the reference policy.

Image Preference Alignment. For image preference optimization, we use the original image m and the perturbed image m' generated in earlier steps. The image preference loss is defined as:

$$L_{\text{image}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_D \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|t,m)}{\pi_{\text{ref}}(y_w|t,m)} - \beta \log \frac{\pi_\theta(y_w|t,m')}{\pi_{\text{ref}}(y_w|t,m')} \right) \right]. \tag{6}$$

Note that we only use the preferred response y_w for image preference alignment.

Margin Loss for Stability. To address potential instability during the joint optimization of text and image preferences, we introduce a margin loss to penalize situations where both chosen and rejected responses experience a reduction in reward, drawing inspiration from previous research [Meng *et al.*, 2024; Chen *et al.*, 2024; Wang *et al.*, 2024]. The margin loss is defined as:

$$L_{\text{margin}} = -\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|t,m)}{\pi_{\text{ref}}(y_w|t,m)} - \eta \right), \tag{7}$$

where η is the margin parameter that enforces a greater separation between positive (chosen) and negative (rejected) responses. The total loss for the training phase is defined as $L_{\text{total}} = L_{\text{text}} + L_{\text{image}} + L_{\text{margin}}$, ensuring that both text and image preferences are jointly optimized while maintaining stability in their alignment.

3.3 Easy-to-Hard Iterative Alignment

We introduce an Easy-to-Hard Iterative Alignment algorithm to stabilize the training of MFPO.

Entropy Calculation. Leveraging semantic entropy [Venhuizen *et al.*, 2019; Farquhar *et al.*, 2024], we estimate the uncertainty of the model's responses. Given a probability distribution of predicted answers $P = \{p_1, p_2, \ldots, p_n\}$, where p_i represents the probability of the i-th predicted token, the entropy H is calculated as $H(P) = -\sum_{i=1}^{n} p_i \log(p_i)$. This measure quantifies the uncertainty inherent in the model.

Sorting by Difficulty. After calculating entropy for all training samples, we rank the training dataset according to their entropy scores, where higher values denoting more challenging inputs. We then divide the dataset into three distinct difficulty levels: "easy", "medium", and "hard".

Iterative Alignment. By progressively moving from easy to hard examples, the feedback distribution is iteratively updated. This ensures that the model first learns simpler patterns to build a foundation before tackling more complex cases. As the training progresses, the alignment becomes more refined, allowing the model to effectively handle harder examples.

Methods	MMHalBench		Object HalBench		AMBER			
	Score ↑	HalRate ↓	$CHAIR_s$ ↓	$CHAIR_i$ ↓	$CHAIR_s$ ↓	Cover. ↑	HalRate ↓	Cog. ↓
7B MLLMs								
LLaVA-v1.6-7B [Liu *et al.*, 2024b]	2.46	0.52	16.4	9.4	9.1	**61.7**	50.2	4.7
LLaVA-v1.5-7B [Liu *et al.*, 2024b]	2.07	0.59	53.6	25.2	7.8	51.0	36.4	4.2
+ HACL [Jiang *et al.*, 2024a]	2.13	0.50	–	–	–	–	–	–
+ POVID [Zhou *et al.*, 2024]	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]	2.04	0.58	48.0	22.3	–	–	–	–
+ DPO [Rafailov *et al.*, 2024]	2.14	0.65	49.0	13.0	6.5	55.5	34.5	2.3
+ mDPO [Wang *et al.*, 2024]	2.39	0.54	35.7	9.8	4.4	52.4	24.5	2.4
+ 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	0.59	39.9	19.9	6.7	49.8	30.9	3.3
+ HALVA [Sarkar *et al.*, 2024]	2.08	0.60	46.6	53.0	6.6	53.0	33.2	3.4
LLaVA-v1.5-7B + MFPO	2.69 (↑0.62)	0.49 (↓0.1)	13.4 (↓40.2)	6.6 (↓18.6)	4.1 (↓3.7)	55.7 (↑4.7)	22.5 (↓13.9)	1.9 (↓2.3)
LLaVA-v1.6-7B + MFPO	**2.89** (↑0.43)	**0.45** (↓0.07)	**10.6** (↓5.8)	**5.1** (↓4.3)	**3.1** (↓6.0)	58.8 (↓2.9)	**18.7** (↓31.5)	**1.1** (↓3.6)
¿=13B MLLMs								
GPT-4V [Achiam *et al.*, 2023]	3.49	0.28	13.6	7.3	4.6	67.1	30.7	2.6
MiniGemini-34B [Li *et al.*, 2024]	3.08	0.38	14.5	8.0	–	–	–	–
Qwen-VL-Chat [Bai *et al.*, 2023]	2.89	0.43	36.0	21.3	6.6	53.2	31.0	2.9
LLaVA-v1.5-13B [Liu *et al.*, 2024a]	2.42	0.53	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
+ HALVA [Sarkar *et al.*, 2024]	2.84	0.48	–	–	6.4	52.6	30.4	3.2
LLaVA-v1.5-13B + MFPO	**2.94** (↑0.52)	**0.42** (↓0.11)	**11.4** (↓34.9)	**4.6** (↓18.0)	**3.4** (↓4.4)	**56.1** (↑5.1)	**19.4** (↓17.0)	**1.4** (↓2.8)

Table 1: Results for MMHalBench, Object HalBench, and AMBER benchmarks. Text in red highlights comparisons before and after incorporating MFPO.

	overall	attribute	adversarial	comparison	counting	relation	environment	holistic	other
LLaVA-RLHF-7B	2.05	2.92	1.83	2.42	1.92	2.25	2.25	1.75	1.08
LLaVA-RLHF-13B	2.53	3.33	2.67	1.75	2.25	2.33	3.25	**2.25**	**2.42**
LLaVA-v1.5-7B	2.07	3.08	1.08	**2.58**	2.25	2.0	3.0	1.42	1.33
+ MFPO	**2.69**	**3.33**	**3.67**	2.42	**2.25**	**2.75**	**3.42**	2.00	1.83

Table 2: Performance comparison across different dimensions in MMHalBench.

4 Experiements

4.1 Experimental Setup

Implementation Details. We use LLaVA-v1.5 as the backbone for all experiments and include the latest top-performing LLaVA-v1.6 to validate the effectiveness of our method. The training consists of three stages: the first two stages follow standard LLaVA training, while MFPO is introduced in the third stage. Here, we construct image preference data based on Section 3.1, using text preference data from RLHF-V [Yu *et al.*, 2024a], and apply MFPO optimization. Details are in Supplementary Section 4.

Evaluation. We evaluate the trustworthiness and general model capabilities with comprehensive experiments. For trustworthiness, we employ three widely used benchmarks to evaluate trustworthiness reflecting the degree of hallucination. Object HalBench [Rohrbach *et al.*, 2018] was used to assess common object hallucinations in image descriptions, with CHAIR scores for both response-level (CHAIRs) and object-level (CHAIRi) hallucination rates. MMHal-Bench [Sun *et al.*, 2023] measured response quality and hallucination rates by comparing model outputs with human responses and object labels. AMBER [Wang *et al.*, 2023a] evaluated generative tasks, providing metrics on CHAIR variants, object coverage, and cognitive hallucination rates. For general capabilities, we employ the LLaVA-Bench [Liu *et al.*, 2024b] for systematic comprehension, encompassing two categories: conversation, detailed description. See details in Supplementary Section 5.

Baselines. We compare our model with state-of-the-art baselines under different settings. For general MLLMs, we include LLaVA1.5 [Liu *et al.*, 2024a], Qwen-VL-Chat [Bai *et al.*, 2023], LLaVA1.6, and MiniGemini [Li *et al.*, 2024]. We also compare with preference-feedback models such as POVID [Zhou *et al.*, 2024], RLHF-V [Yu *et al.*, 2024a], Silkie [Sun *et al.*, 2023] and DPO [Rafailov *et al.*, 2024], as well as feedback-independent methods such as VCD [Leng *et al.*, 2024]. We also benchmark against GPT-4V [Achiam *et al.*, 2023]. See details in Supplementary Section 6.

4.2 Main Results

Comparison to Existing Methods. The main results are presented in Table 1. With MFPO integration, both LLaVA-v1.5 and LLaVA-v1.6 show consistent improvements across all three trustworthiness evaluation datasets. On MMHalBench and Object HalBench, our approach achieves state-of-the-art performance among all 7B-parameter models. As shown in Table 2, MFPO enhances performance across all metrics, demonstrating universal improvements in MLLM trustworthiness. These results highlight MFPO's ability to balance preference optimization across text and image modalities, leading to superior trustworthiness.

Notably, LLaVA-v1.5-7B+MFPO achieves trustworthiness levels comparable to or exceeding those of larger models, such as LLaVA-v1.5-13B and MiniGemini-34B. Furthermore, while baseline LLaVA-v1.5-7B and LLaVA-v1.5-13B underperform GPT-4V across all eight metrics, MFPO integration enables both models to outperform GPT-4V on five out of eight metrics. This underscores MFPO's capability to significantly

Method	Conversation	Captioning
LLaVA-1.5	53.3	53.4
+ VIfeedback	51.3	49.3
+ Human-Preference	49.6	43.3
+ RLHF-V	55.8	56.1
Ours	**65.8**	**60.0**

Table 3: Comparison of different methods in LLaVA-Bench.

L_{text}	L_{image}	L_{margin}	MMHalBench		Object HalBench	
			Score ↑	HalRate↓	CHAIR_s ↓	CHAIR_i ↓
✓	x	✓	2.46	0.53	21.9	10.9
x	✓	✓	2.34	0.56	24.4	13.1
✓	✓	x	2.61	0.50	16.9	7.9
✓	✓	✓	**2.69**	**0.49**	**13.4**	**6.6**

Table 4: Ablation study of different loss compositions.

Margin	MMHalBench		Object HalBench	
	Score ↑	HalRate ↓	CHAIR_s ↓	CHAIR_i ↓
0 (Ours)	**2.69**	**0.49**	**13.4**	**6.6**
0.2	2.58	0.50	15.7	7.1
0.4	2.63	0.52	16.0	7.9

Table 5: Comparison of different margins.

Ratio	MMHalbench		Object Halbench	
	Score↑	HalRate↓	CHAIRs↓	CHAIRi↓
1:1:1	**2.69**	**0.49**	**13.4**	**6.6**
1:5:1	2.66	0.50	13.9	6.9
5:1:1	2.59	0.55	15.6	8.2
1:1:5	2.62	0.52	14.5	7.0

Table 6: Ablation of Ratio $L_{\text{text}} : L_{\text{image}} : L_{\text{margin}}$

Training Method	MMHalBench		Object HalBench	
	Score ↑	HalRate ↓	CHAIR_s ↓	CHAIR_i ↓
Easy-to-hard	**2.69**	**0.49**	**13.4**	**6.6**
End-to-end	2.53	0.53	16.0	8.0

Table 7: Comparison of different training methods.

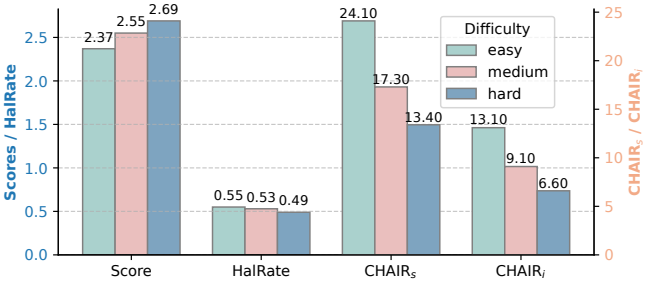

Figure 5: Each stage's performance.

Construction Method	MMHalBench		Object HalBench	
	Score ↑	HalRate ↓	CHAIR_s ↓	CHAIR_i ↓
Global	2.56	0.56	21.7	10.3
Random 20%	2.45	0.52	16.5	8.5
Ours	**2.69**	**0.49**	**13.4**	**6.6**

Table 8: Comparison of different image preference data construction.

enhance the trustworthiness of smaller MLLMs, bridging the performance gap with much larger models.

General Perception Evaluation. To verify whether MFPO can enhance general perception capability of MLLM, we analyze different preference collection strategies on LLaVA-Bench benchmark. We focus on the performance related to conversation and detailed descriptions, which represent general perception abilities. As shown in Table 3, our method outperforms both LLaVA-v1.5 and other alignment methods, indicating that our approach not only enhances trustworthiness but also improves overall perception capability of the model.

4.3 Ablation Study and Analysis

Ablation on Joint Text-Visual Reward Loss. We conduct experiments to validate the loss composition in MFPO. As shown in Table 4, optimizing only the text modality (L_{text}) improves performance but overfits to text preferences, while optimizing only the image modality (L_{image}) addresses the text-image preference imbalance but underperforms due to lack of text optimization. This highlights the necessity of joint text-visual optimization in multimodal tasks.

Ablation on Margin Loss. Adding margin loss (L_{margin}) further improves performance. Figure 6 shows that without L_{margin}, rewards for both chosen and rejected responses decrease simultaneously, while L_{margin} stabilizes rewards by penalizing reductions for chosen responses. Table 5 demonstrates consistent performance gains across margin values (0, 0.2, 0.4), with the best results at 0.

Ablation on Loss Ratio. In Table 6, we present an ablation study on the weighting of different loss components. A balanced ratio of 1:1:1 yields the best results.

Ablation of Easy-to-Hard Alignment Scheme. We evaluate the effectiveness of the easy-to-hard alignment scheme against simultaneous training on all data. Table 7 shows that our approach consistently outperforms the traditional method. Figure 7 further reveals that the easy-to-hard strategy achieves better convergence, with losses steadily decreasing across easy, medium, and hard phases, indicating more efficient alignment learning. The loss reduction is fastest in the easy phase and gradually slows in later phases. Moreover, the easy-to-hard training curve exhibits greater stability compared to full training, demonstrating its stability and efficiency.

How Does Noise Proportion Affect Preference Optimization Balance? We study how noise granularity and proportion affect preference optimization balance by comparing three methods: applying global diffusion noise to the entire image, adding noise to 20% of random image regions, and our fine-grained noise addition strategy. Table 8 demonstrates that image preference data construction enhances image-based optimization during alignment. However, as noise granularity becomes coarser, optimization increasingly favors text preferences, leading to performance degradation. Coarse-grained methods oversimplify image feature optimization, impairing image-text alignment. In contrast, fine-grained construction shifts optimization toward the image modality, achieving a more balanced multimodal alignment.

How Does Keyword Selection Perform? To evaluate the accuracy of extracting key visual keyphrases from the original

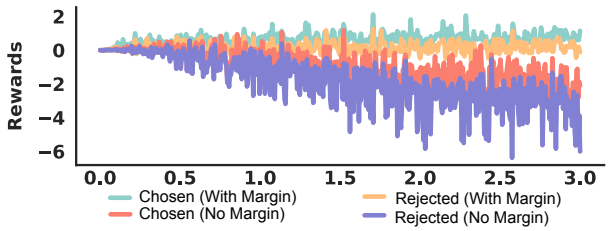

Figure 6: Rewards with and without margin loss in 3 epochs.

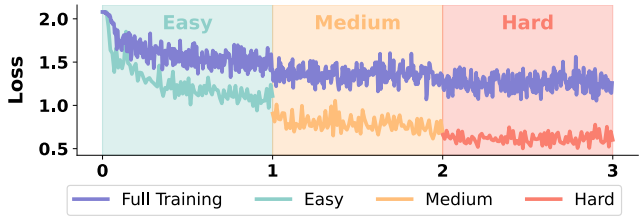

Figure 7: Loss comparison in 3 epochs.

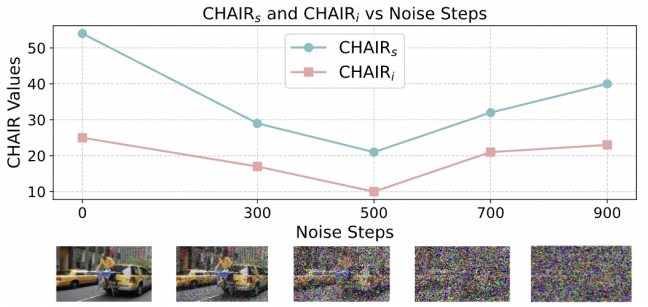

Figure 8: Performance of LLaVA-v1.5 under different noise levels.

Methods	MMHalbench		Object Halbench	
	Score↑	HalRate↓	CHAIRs↓	CHAIRi↓
mDPO (RLHF-V)	2.42	0.56	29.3	9.2
MFPO (RLHF-V)	**2.69**	0.49	**13.4**	**6.6**
mDPO (Silkie)	2.39	0.54	35.7	9.8
MFPO (Silkie)	2.67	**0.46**	14.7	7.1

Table 9: Further Comparision of mDPO and MFPO.

text preference data using a multipartite graph, we randomly sample 100 examples and manually verify their correctness. We find that 86% of the extracted keyphrases correspond to important objects in the associated preference data. In the remaining 14% of cases, the selected keywords are not the most critical objects; however, using SAM to add diffusion noise to these regions still generates finely dispreferred images. We attribute the effectiveness of the multipartite graph to the relatively short text length and simple object relationships in the preference datasets, which enable robust keyword selection.

How Does SAM Perform on Region Mapping? To validate SAM's accuracy in region mapping, we randomly sample 100 images for manual evaluation. SAM achieves over 90% accuracy, with the remaining 10% of failures occurring when keywords refer to objects requiring contextual understanding beyond SAM's scope. In such cases, we apply global diffusion noise to the entire image. While this approach is suboptimal, the resulting dispreferred images still improve model performance. Furthermore, we experiment with more advanced segmentation models, such as LISA [Lai *et al.*, 2024], which achieves over 95% accuracy in region mapping. Models with reasoning-based segmentation capabilities offer greater robustness to diverse keywords, enabling the generation of more refined dispreferred images.

How Do Different Noise Levels Affect Performance? We investigate the impact of varying noise levels on preference optimization. As shown in Figure 8, adding diffusion noise to the entire image improves performance, with the best results achieved at 500 steps. At this level, the image undergoes an appropriate level of corruption: it avoids excessive noise that would erase critical details and make dispreferred images too easily distinguishable, while also preventing insufficient noise that would result in minimal image changes, failing to establish meaningful rewards for image optimization.

How to Achieve Modality-Fair Optimization? We investigate modality-fair optimization from both data and loss perspectives. On the data side, we analyze the impact of noisy region proportion and noise intensity in images. Our results show that region-level noise is more effective than global noise in guiding preference toward the visual modality. Moreover, moderate noise levels yield a better optimization balance than extremely high or low noise. On the loss side, jointly optimizing text and image rewards facilitates balanced modality alignment while preserving overall performance.

More Comparisions to mDPO. While mDPO demonstrated its effectiveness by only training on Silkie, Table 9 reports MFPO's performance when trained on Silkie and mDPO's performance when trained on RLHF-V. This ensures a fair comparison under equivalent dataset conditions. The results show that dataset differences have minimal impact on MFPO's performance, and MFPO consistently outperforms mDPO across both datasets, further validating MFPO's effectiveness.

5 Conclusion

In this paper, we investigate the limitation of text preference optimization and propose Modality-Fair Preference Optimization (MFPO) for trustworhty alignment in MLLMs. We constructed a new dataset with fine-grained image preference, and implemented balanced preference optimization between text and image using a multi-stage alignment strategy alongside novel loss functions. Extensive experiments and analysis demonstrate that MFPO notebly diminishes hallucinations and attains state-of-the-art performance across trustworthiness and general capability benchmarks.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 12326612, 62476241), the Natural Science Foundation of Zhejiang Province, China (Grant No. LZ23F020008), and the Zhejiang University-Angelalign Inc. R&D Center for Intelligent Healthcare.

References

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

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

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

[Bianchini *et al.*, 2005] Monica Bianchini, Marco Gori, and Franco Scarselli. Inside pagerank. *ACM Transactions on Internet Technology (TOIT)*, 5(1):92–128, 2005.

[Boudin, 2018] Florian Boudin. Unsupervised keyphrase extraction with multipartite graphs. *arXiv preprint arXiv:1803.08721*, 2018.

[Casper *et al.*, 2023] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

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

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

[Cui *et al.*, 2023] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023.

[Dawande *et al.*, 2001] Milind Dawande, Pinar Keskinocak, Jayashankar M Swaminathan, and Sridhar Tayur. On bipartite and multipartite clique problems. *Journal of Algorithms*, 41(2):388–403, 2001.

[Farquhar *et al.*, 2024] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[Feng *et al.*, 2024] Duanyu Feng, Bowen Qin, Chen Huang, Zheng Zhang, and Wenqiang Lei. Towards analyzing and understanding the limitations of dpo: A theoretical perspective. *arXiv preprint arXiv:2404.04626*, 2024.

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

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Huang *et al.*, 2024] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 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, 2024.

[Jiang *et al.*, 2024a] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 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, 2024.

[Jiang *et al.*, 2024b] Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. *arXiv preprint arXiv:2404.10237*, 2024.

[Jin and Zhang, 2014] Ya-Lei Jin and Xiao-Dong Zhang. Complete multipartite graphs are determined by their distance spectra. *Linear Algebra and its Applications*, 448:285–291, 2014.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[Lai *et al.*, 2024] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589, June 2024.

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

[Li *et al.*, 2024] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.

[Liu *et al.*, 2024a] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 26296–26306, 2024.

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

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

[Mihalcea and Tarau, 2004] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[Naveed *et al.*, 2023] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

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

[Paas *et al.*, 2003] Fred Paas, Alexander Renkl, and John Sweller. Cognitive load theory and instructional design: Recent developments. *Educational psychologist*, 38(1):1–4, 2003.

[Peng *et al.*, 2023] Baolin Peng, Linfeng Song, Ye Tian, Lifeng Jin, Haitao Mi, and Dong Yu. Stabilizing rlhf through advantage model and selective rehearsal. *arXiv preprint arXiv:2309.10202*, 2023.

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

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

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

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[Shah *et al.*, 2019] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658, 2019.

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

[Venhuizen *et al.*, 2019] Noortje J Venhuizen, Matthew W Crocker, and Harm Brouwer. Semantic entropy in language comprehension. *Entropy*, 21(12):1159, 2019.

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

[Wang *et al.*, 2023b] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

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

[Xu *et al.*, 2024] Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*, 2024.

[Yu *et al.*, 2024a] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 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, 2024.

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

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

[Zhang and Rong, 2024] Mengxi Zhang and Kang Rong. Automated multi-level preference for mllms. *arXiv preprint arXiv:2405.11165*, 2024.

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

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