ASPO: Adaptive Sentence-Level Preference Optimization for Fine-Grained Multimodal Reasoning

Yeyuan Wang[1,*], **Dehong Gao**[2,3,*], **Rujiao Long**[4], **Lei Yi**[4],
Linbo Jin[4], **Libin Yang**[2,†], **Xiaoyan Cai**[1,†]
[1]Northwestern Polytechnical University, School of Automation, Xi'an, China
[2]Northwestern Polytechnical University, School of Cybersecurity, Xi'an, China
[3]Binjiang Institute of Artificial Intelligence, ZJUT, Hangzhou, China
[4]Alibaba Group, Hangzhou, China
{wangyeyuan, dehong.gdh, libiny, xiaoyanc}@nwpu.edu.cn
{rujiao.lrj, yilei.yi, yuyi.jlb}@alibaba-inc.com

Abstract

Direct Preference Optimization (DPO) has gained significant attention for its simplicity and computational efficiency in aligning large language models (LLMs). Recent advancements have extended DPO to multimodal scenarios, achieving strong performance. However, traditional DPO relies on binary preference optimization, rewarding or penalizing entire responses without considering fine-grained segment correctness, leading to suboptimal solutions. The root of this issue lies in the absence of fine-grained supervision during the optimization process. To address this, we propose Adaptive Sentence-level Preference Optimization (ASPO), which evaluates individual sentences for more precise preference optimization. By dynamically calculating adaptive rewards at the sentence level based on model predictions, ASPO enhances response content assessment without additional models or parameters. This significantly improves the alignment of multimodal features. Extensive experiments show that ASPO substantially enhances the overall performance of multimodal models.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in Natural Language Processing (NLP) (Touvron et al., 2023; Zhang et al., 2022; Achiam et al., 2023), which greatly accelerates the development of Multimodal Large Language Models (MLLMs) (Kirillov et al., 2023; Li et al., 2024; Anil et al., 2023). The training of MLLMs typically involves a pre-training stage followed by the Supervised Fine-Tuning (SFT) stage (Liu et al., 2024b; Dai et al., 2023). SFT enhances the model's ability to understand and execute multimodal instructions (Liu et al., 2024a; Yu et al., 2023a; Ye et al., 2024). Models trained in this manner exhibit impressive multimodal conversational abilities (Yu et al., 2023a; Huang et al., 2023).

Despite these advancements, SFT often leads to hallucinated outputs, causing performance saturation (Zhao et al., 2023; Jiang et al., 2024). As noted in (Hong et al., 2024), increasing the probability of preferred outputs can inadvertently raise the likelihood of dis-preferred outputs, making models more prone to errors in complex reasoning tasks (Wang et al., 2024b; Zhang et al., 2024). Additionally, as response length grows, models become increasingly susceptible to generating hallucinations (Favero et al., 2024). This issue is exacerbated by machine-generated data containing noisy sentences, where segments may be **partially** correct or incorrect (Yu et al., 2024; Lai et al., 2024). Training on such data risks convergence to suboptimal solutions, emphasizing the need for strategies to suppress undesirable outputs (Liao et al., 2024).

In response, researchers have developed various preference fine-tuning methods to better align MLLMs with human preferences (Sun et al., 2024; Yuan et al., 2024a; Xie et al., 2024). Among these, DPO-based methods (Rafailov et al., 2024) have gained prominence due to their simplicity and low computational cost (Wang et al., 2024a; Zhu et al., 2024). However, traditional DPO methods rely on binary preference data and a coarse reward mechanism, which lacks the fine-grained preference granularity needed to identify specific errors in responses (Liao et al., 2024). This limitation hampers the model's ability to refine its reasoning capabilities (Lai et al., 2024; Liao et al., 2024). To this end, we introduce a fine-grained preference mechanism that assigns variable reward values, enabling more fine-grained reasoning optimization.

As shown in Figure 1, we present an Adaptive Sentence-level Preference Optimization (ASPO) approach, where each sentence in a response serves as the fundamental unit of preference optimization. Unlike traditional DPO, which evaluates prefer-

---

[*] Equal Contribution.
[†] Corresponding Authors.

Findings of the Association for Computational Linguistics: ACL 2025, pages 5149–5160
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

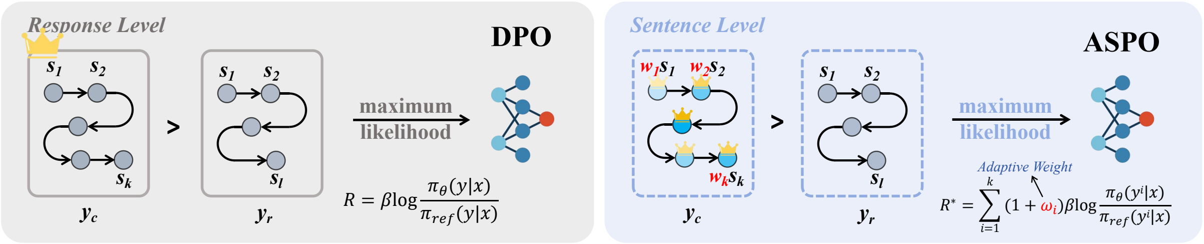


Figure 1: Comparison between DPO and our ASPO. Unlike the traditional DPO algorithm that calculates implicit rewards at the response level, ASPO calculates adaptive rewards using sentences as the smallest units to achieve a fine-grained self-supervised preference optimization method.

ences at the response level (Li et al., 2023c; Wang et al., 2024a), ASPO systematically evaluates the correctness and importance of each sentence and assigns appropriate rewards. By prioritizing accurate reasoning and rejecting errors based on fine-grained criteria, our method efficiently identifies erroneous sentences for targeted optimization. This reduces the impact of noisy data and substantially improves the model's comprehensive capability.

Our contributions are summarized as follows:

- We propose a fine-grained ASPO approach that uses sentences as the fundamental units for preference optimization, surpassing traditional response-level approaches.
- We design an adaptive reward mechanism that dynamically adjusts rewards at sentence-level, enabling precise preference optimization.
- We demonstrate significant improvements through extensive experiments and establish the broad applicability of our approach across various multimodal reasoning tasks.

2 Related Work

2.1 Multimodal Large Language Models

With the unprecedented success of LLMs (Touvron et al., 2023; Zhang et al., 2022), recent research focused on integrating visual encoders with advanced LLMs to construct MLLMs (Liu et al., 2024b; Dai et al., 2023). Experimental evidence indicates that MLLM performs effectively across various multimodal tasks (Anil et al., 2023; Achiam et al., 2023).

Despite significant advancements in understanding high-level semantics, MLLMs continue to encounter challenges in addressing fine-grained details (Yuan et al., 2024b; Guo et al., 2024). In complex visual scenarios, these models often struggle to capture critical aspects such as object relationships, attribute alignment, and recognition of local features (Yuan et al., 2024b; Guo et al., 2024; Ouali et al., 2025). Moreover, in tasks like visual question answering or image captioning, models frequently generate hallucinated outputs (Chen et al., 2024; Jiang et al., 2024). Existing hallucination problems in models often introduce noise into the generated data, increasing the risk that models trained on these data will converge to suboptimal solutions.

To address these challenges, we propose ASPO that incorporates an adaptive reward mechanism to mitigates the impact of noisy data. This mechanism dynamically adjusts reward weights based on the accuracy of response segments, facilitating precise fine-grained optimization.

2.2 Preference Alignment

Preference alignment (Chen et al., 2023a; Ethayarajh et al., 2024) is widely used to enhance model's instruction-following capabilities (Wang et al., 2024c). Early work primarily relied on Reinforcement Learning from Human Feedback (RLHF) (Lee et al., 2024; Wang et al., 2023). However, RL-based methods face challenges in stability and efficiency (Rafailov et al., 2024). DPO (Rafailov et al., 2024) addresses these issues by redefining the training objective as a classification loss-based optimization problem. By mapping reward functions to optimal policies, the constrained reward maximization problem could be precisely optimized through single-stage policy training (Rafailov et al., 2024; Wang et al., 2024c).

In the multimodal domain, early DPO-based methods relied on response-level rewards (Li et al., 2023c; Zhou et al., 2024a; Wang et al., 2024c). For example, SILKIE (Li et al., 2023c) constructs a vision-language feedback dataset by gathering annotations from multiple MLLMs and employs GPT-4V to evaluate outputs across various dimensions. MDPO (Wang et al., 2024a) addresses the unconditional preference problem by prioritizing image preferences over language preferences. However, response-level rewards have limitations in granularity and specificity (Yoon et al., 2024), potentially leading to less precise model optimization due to the inclusion of noisy segments in the training data. This highlights the need for more fine-grained feedback mechanisms.

Subsequent work has explored much more fine-grained feedback mechanisms at the sentence level (Yu et al., 2024; Lee et al., 2024; Ouali et al., 2025; Zhou et al., 2024b). RLHF-V (Yu et al., 2024) collects human preferences in the form of segment-level corrections on hallucinations and performs dense DPO over corrective human feedback, enhancing the trustworthiness of MLLMs. CSR (Zhou et al., 2024b) iteratively generates candidate responses, evaluates their rewards, and curates preference data for fine-tuning.

Recent studies have investigated token-level reward mechanisms with notable success (Yoon et al., 2024; Cui et al., 2024; Fu et al., 2024). FiSAO (Cui et al., 2024) leverages the model's own visual encoder as a fine-grained verifier to improve vision-language alignment by exploiting token-level feedback from the visual encoder. TLDR (Fu et al., 2024) introduces a token-level detection reward model to provide fine-grained annotations for MLLMs, offering a foundation for post-training methods such as DPO and Proximal Policy Optimization (PPO) (Schulman et al., 2017).

Despite these advancements, most approaches rely on model-generated datasets, which tend to be noisy due to hallucinations (Yu et al., 2024; Li et al., 2023c). Traditional DPO treats all segments of a response equally regardless of their correctness, introducing noise and hindering model optimization. In addition, most previous methods rely on paid APIs or deploy other external MLLMs, which is not cost-friendly.

To address these issues, we propose ASPO, which dynamically adjusts reward weights of response segments without relying on paid APIs, additional training data or deploying other external MLLMs. This approach enables more fine-grained preference optimization, alleviating the impact of noise and providing a promising direction for advancing preference alignment research.

3 Methodology

In this section, we introduce ASPO for fine-grained multimodal reasoning. ASPO incorporates two adaptive reward weights into DPO, i.e., the image-text similarity weight, which optimizes multimodal alignment to mitigate hallucination, and the textual perplexity weight, which enhances text confidence to improve the model's reasoning capabilities.

3.1 Preliminaries

RLHF leverages human feedback to train models in scenarios where direct supervision, such as ground-truth labels, is unavailable or insufficient. It is widely used to optimize LLMs and MLLMs, aligning their outputs with human preferences.

In RLHF, a Bradley-Terry (BT) reward model is commonly used, defining the human preference distribution as:

$$p^*(y_c \succ y_r|x) = \frac{\exp\left(r^*(x, y_c)\right)}{\exp\left(r^*(x, y_c)\right) + \exp\left(r^*(x, y_r)\right)} \tag{1}$$

where y_c and y_r denote the *chosen* and *rejected* responses conditioned on the model's input prompt x, and $r^*(x, y)$ represents the latent reward model.

The reward model $r_\phi(\cdot)$, parameterized by ϕ, is trained via maximum likelihood estimation using a preference database $\mathcal{D}$, which is annotated by human evaluators (Yu et al., 2024) or advanced AI models (Lee et al., 2024):

$$\mathcal{D} = \{x^i, y_c^i, y_r^i\}_{i=1}^N \tag{2}$$

The optimization objective maximizes the preference policy:

$$\max_\theta \mathbb{E}_{x,y} \left\{ r_\phi(x, y) - \beta \mathbb{D}_{\text{KL}}[\pi_\theta(y|x) | \pi_{\text{ref}}(y|x)] \right\} \tag{3}$$

where the reference model π_{ref} is typically initialized using the supervised fine-tuned model π_{SFT} to ensure that the learned parameter θ does not deviate excessively. The resulting learned policy $\pi_\theta(\cdot)$ aligns more effectively with human intentions.

DPO simplifies RLHF by deriving a closed-form solution (Rafailov et al., 2024) from Eq. 3, representing the optimal reward model r^* in terms of the learned optimal preference model π^* as:

$$r^*(x, y) = \beta \log \frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} + \Delta \tag{4}$$

where Δ is a constant factor. Substituting r^* in Eq. 1 with Eq. 4, the final optimized loss function for DPO is expressed as (Rafailov et al., 2024):

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_c|x)}{\pi_{\text{ref}}(y_c|x)} - \beta \log \frac{\pi_\theta(y_r|x)}{\pi_{\text{ref}}(y_r|x)} \right) \right] \tag{5}$$

where θ denotes the trainable parameters in the policy model, σ is the logistic function and the $\beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ item can be regarded as an "implicit reward". The objective of DPO is to align the "implicit reward" directly with human preference data.

3.2 ASPO

ASPO is designed to address the limitations of traditional DPO approache, which optimizes entire responses using binary rewards and fail to account for variations among individual response segments. ASPO introduces a sentence-level adaptive reward mechanism that utilizes **image-text similarity** and **textual perplexity** metrics to compute fine-grained rewards for each response segment. This enables targeted and differentiated optimization, significantly improving the model's performance. The overall process is shown in Algorithm 1.

3.2.1 Rewards of ASPO

Instead of treating a response as an indivisible unit, ASPO decomposes it into sentences for more precise analysis. Two critical features are computed for each sentence as follows.

(1) Image-Text Similarity. To ensure the model's output is consistently aligned with the input image, reducing hallucinations and improving model credibility, we measure the semantic relevance of each sentence to the input image as part of the adaptive reward weight.

Let s_i denotes the i-th sentence in a response, and let x represents the corresponding image. The similarity function between s_i and x is denoted as:

$$S_i = cosine(s_i, x), \quad i = 1, 2, \ldots, n \tag{6}$$

where n is the total number of sentences in the response and the similarity score is calculated by CLIP (Radford et al., 2021).

Next, we apply min-max normalization to scale each similarity score into the range [0,1]:

$$S_i' = \frac{S_i - S_{\min}}{S_{\max} - S_{\min}}, \quad i = 1, 2, \ldots, n \tag{7}$$

where S_i' represents the normalized similarity score for the i-th sentence with respect to the image x. $S_{\min}$ and $S_{\max}$ represent the minimum and maximum similarity scores across all sentences, respectively.

(2) Textual Perplexity. Perplexity is used to evaluate the quality of a language model, which reflects the uncertainty of the model about a given text. The lower the value, the better the model's prediction of the text, that is, the model believes that the text is more likely to appear. Prior research has shown that outputs with higher model confidence are more likely to be correct (Kuhn et al., 2023; Chen and Mueller, 2024). Thus, we incorporate the model's predicted output probabilities as part of the adaptive reward weight computation. Fluency, coherence, and confidence are assessed using metrics such as log-likelihood, perplexity, or information entropy. Unless otherwise specified, we use textual perplexity as the default reward weight.

Algorithm 1 ASPO Training Process

```
Input: Preference dataset $\mathcal{D} = \{(x^i, y_c^i, y_r^i)\}_{i=1}^N$, Reference model $\pi_{\text{ref}}$, Policy model $\pi_\theta$
Output: Optimized policy model $\pi_{\theta'}$
repeat
    for each $(x, y_c, y_r)$ in $\mathcal{D}$ do
        Split $y_c$ into sentences $\{s_1, s_2, \ldots, s_n\}$
        for each sentence $s_i$ in $\{s_1, s_2, \ldots, s_n\}$ do
            Compute Image-Text Similarity $S_i'$
            Compute Textual Perplexity $PPL_i'$
            Compute Sentence Weight $w_i$
        end for
        Calculate adaptive implicit reward margin $\mathcal{M}^*$ by Equation (12)
    end for
    Update policy parameters $\theta'$ using gradient descent
until Convergence or maximum iterations reached
```

The perplexity of a text sequence is defined as the geometric mean of the inverse probabilities of each word in the sequence, given the preceding context. For a sequence of N words $w_1, w_2, \ldots, w_N$, the perplexity can be expressed as:

$$PPL = \left(\prod_{j=1}^{N} \frac{1}{P(w_j | w_{<j})} \right)^{\frac{1}{N}} \tag{8}$$

Since the reward coefficient is calculated in units of sentences in our algorithm, assuming that the i-th sentence contains N tokens and the prior $i-1$ sentences contain a total of M tokens, using logarithms for computational stability, the perplexity of the i-th sentence can be written as:

$$PPL_i = \exp\left(-\frac{1}{N} \sum_{j=M+1}^{M+N} \log P(w_j | x, w_{<j}) \right) \tag{9}$$

where $P(w_j | x, w_{<j})$ is the conditional probability of the j-th word given the preceding image x and context $w_{<j} = w_1, w_2, \ldots, w_{j-1}$, and exp denotes the exponential function.

Since a higher perplexity indicates a higher uncertainty in the model about a given text, we first negate the perplexity at the sentence level and then perform min-max normalization to get the normalized perplexity PPL'_i for the i-th sentence follows Equation (7).

(3) Final Adaptive Weight. To fully leverage the advantages of different metrics, we integrate them as follows:

$$w_i = \alpha S'_i + (1-\alpha) PPL'_i \tag{10}$$

where α is the metric-weighting factor introduced to balance the contribution of different metrics and w_i is used to adjust the reward of each sentence.

Sentence-Level Reward Aggregation. In the standard DPO algorithm, the "implicit reward" is shared across all steps, and the output of the model at each step receives the same reward weight. Specifically, for the triplet (x, y_c, y_r), the "implicit reward" margin is mathematically defined as:

$$\begin{aligned} \mathcal{M} &= \beta \log \frac{\pi_\theta(y_c \mid x)}{\pi_{\text{ref}}(y_c \mid x)} - \beta \log \frac{\pi_\theta(y_r \mid x)}{\pi_{\text{ref}}(y_r \mid x)} \\ &= \sum_{i=1}^{K} \beta \log \frac{\pi_\theta(s_i^c \mid x)}{\pi_{\text{ref}}(s_i^c \mid x)} - \sum_{i=1}^{L} \beta \log \frac{\pi_\theta(s_i^r \mid x)}{\pi_{\text{ref}}(s_i^r \mid x)} \end{aligned} \tag{11}$$

To address the limitation of the shared reward weight, which fails to differentiate the importance of individual sentences, ASPO introduces an adaptive reward mechanism to adjust the rewards for each sentence. Specifically, ASPO uses the adaptive weight w_i to modulate the reward strength of each sentence. The adaptive M is mathematically reformulated as:

$$\begin{aligned} \mathcal{M}^* = &\frac{R_c}{R_c^*} \sum_{i=1}^{K} \beta(1+w_i) \log \frac{\pi_\theta(s_i^c \mid x)}{\pi_{\text{ref}}(s_i^c \mid x)} \\ &- \beta \log \frac{\pi_\theta(y_r \mid x)}{\pi_{\text{ref}}(y_r \mid x)} \end{aligned} \tag{12}$$

where R_c and R_c^* represent the original and the reweighted sum of the sentence-level implicit rewards of chosen response, respectively. We introduce the $\frac{R_c}{R_c^*}$ item to normalize the influence of sentence-level rewards across responses of varying lengths, which prevents longer responses from disproportionately benefiting from a total weight increase, aligning better with our objective of implementing fine-grained rewards.

Note that when a response contains only one sentence, the adaptive weight w_i is normalized to 0 and our method degenerates into the standard DPO.

3.2.2 Optimization Objective of ASPO

The ASPO loss function extends the traditional DPO objective by incorporating sentence-level granularity. For a preference pair (y_c, y_r), the optimization objective becomes:

$$\mathcal{L}_{\text{ASPO}} = -\mathbb{E}_{\mathcal{D}} \left(\log \sigma \mathcal{M}^* \right) \tag{13}$$

By focusing on individual sentences, ASPO allows the model to refine specific response segments, leading to improved fine-grained understanding. Adaptive weighting enables the model to prioritize high-quality or critical sentences while minimizing the influence of less relevant ones. ASPO can be seamlessly integrated into existing DPO-based training pipelines with minimal adjustments.

4 Experiment

In this section, we conduct comprehensive experiments to evaluate the effectiveness of the proposed ASPO. We detail the experimental setups, results, and analysis in the following sections.

4.1 Experimental Setups

Pretrained Models. To assess ASPO's generalization, we test it on the popular open source MLLMs in different architectures and sizes. LLaVA-v1.5-7B and LLaVA-v1.5-13B (Liu et al., 2024a) uses MLP as an alignment module, while InstructBLIP-13B (Dai et al., 2023) uses Q-Former (Li et al., 2023b), a query-based alignment module to align multimodal features. We use the models after SFT as base models and conduct DPO training on them.

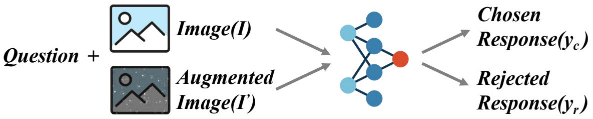


Figure 2: Data collection pipeline.

Preference data. We sample about 20K instructions from LLaVA-Instruct-150K (Liu et al., 2024b) and go through SeVa pipline (Zhu et al., 2024) to produce the preference dataset. The data collection pipeline is shown in Figure 2. Specifically, diffusion noise is added to the images in the instruction data, and the original and augmented images with the instruction are then fed into the MLLM to be train to obtain the chosen and rejected responses, respectively. The identical paired preference data is filtered out and we finally get about 16K preference data pairs. For all experiments, the noise

Method	MMVet	MMB^{D}	MMB^{T}	MMB^{C}	SEED^{I}	LLaVA^{W}	SQA^{I}	GQA	POPE	SHR (↓)	Avg.
BLIP-2-7B	22.4	–	–	–	46.4	38.1	61.0	41.0	85.3	–	–
MiniGPT-4-7B	22.1	23.0	–	–	42.8	–	–	–	74.5	–	–
Shikra-13B	–	58.8	–	–	–	–	–	–	–	–	–
LLaVA-7B	26.7	34.1	–	14.1	25.5	63.0	38.5	–	–	–	–
IDEFICS-7B	–	48.2	–	25.2	–	–	–	38.4	–	–	–
IDEFICS-65B	–	54.5	–	38.1	–	–	–	45.2	–	–	–
mPLUG-Owl2-7B	36.2	64.5	–	–	–	–	68.7	56.1	85.8	–	–
Qwen-VL-7B	–	38.2	–	7.4	56.3	–	67.1	59.3	–	–	–
Qwen-VL-chat-7B	–	60.6	–	56.7	58.2	–	68.2	57.5	–	–	–
InstructBLIP-13B	25.6	33.5	34.0	**26.3**	45.4	58.2	44.7	48.1	78.9	51.2	43.86
+DPO	26.1	33.5	34.3	26.1	45.8	62.7	44.3	47.5	80.2	53.8	44.50
+ASPO	**27.0**	**33.7**	**35.1**	26.0	**46.3**	**67.4**	**45.1**	**48.2**	**83.8**	**50.5**	**45.84**
LLaVA-1.5-7B	30.5	64.3	66.4	58.3	65.7	63.4	66.8	**62.0**	85.9	36.7	62.59
+DPO	33.3	64.7	66.4	58.3	66.1	65.7	66.4	61.0	86.2	40.1	63.12
+ASPO	**35.3**	**65.6**	**67.7**	**59.5**	**66.3**	**75.7**	**67.7**	**62.0**	**86.6**	**33.9**	**65.16**
LLaVA-1.5-13B	35.4	67.7	67.0	63.6	68.2	70.7	71.6	63.3	85.9	37.2	65.93
+DPO	38.8	69.6	69.9	64.1	68.2	74.7	70.2	62.3	85.5	47.4	67.03
+ASPO	**41.2**	**70.4**	**70.7**	**64.7**	**68.5**	**82.0**	**71.8**	**63.4**	**86.5**	**34.8**	**68.80**

Table 1: Comparison results with state-of-the-art methods. ASPO consistently improves base model's performance on multiple benchmarks. The last column shows the average values of all metrics except for SHR.

Method	Level	MMVet	MMB^{D}	LLaVA^{W}	SQA^{I}	POPE	Avg.
LLaVA-1.5-7B	–	30.5	64.3	63.4	66.8	85.9	62.18
+ Vlfeedback	Response	31.2	64.0	62.1	66.2	83.7	61.44
+ Human-Prefer	Response	31.1	63.4	63.7	65.8	81.5	61.10
+ POVID	Response	31.8	64.9	68.7	68.8	**86.9**	64.22
+ SIMA	Response	31.6	64.9	66.1	69.1	86.5	63.64
+ HA-DPO	Response	30.5	64.4	64.2	68.5	85.8	62.68
+ RLHF-V	Sentence	30.9	63.6	65.4	67.1	86.2	62.64
+ RLAIF-V	Sentence	30.5	63.4	–	68.4	81.5	–
+ CLIP-DPO	Sentence	–	64.9	–	67.6	85.8	–
+ CSR iter-1	Sentence	32.2	64.7	69.7	70.3	85.8	64.54
+ CSR iter-3	Sentence	33.9	65.4	71.1	**70.7**	85.9	65.40
+ FiSAO	Token	30.7	64.8	–	69.3	85.7	–
+ ASPO	**Sentence**	**35.3**	**65.6**	**75.7**	67.7	86.6	**66.18**

Table 2: Comparison results of different preference optimization-based methods. Most baseline results are from (Zhou et al., 2024a). Benchmarks like GQA in Table 1 are not included since most baseline methods have not been tested on them.

steps are set to 500. In this way, preference data is obtained with the same style and distribution as previous studies (Zhao et al., 2023) suggesting that unifying factors such as style and distribution in preference data could enhance model optimization.

Baselines. We compare ASPO with several previous State-Of-The-Art (SOTA) open-source MLLMs and preference optimization-based methods. The **open-source MLLMs** include BLIP-2 (Li et al., 2023b), MiniGPT-4 (Zhu et al., 2023), InstructBLIP (Dai et al., 2023), LLaVA (Liu et al., 2024b), LLaVA-v1.5 (Liu et al., 2024a), Shikra (Chen et al., 2023b), Qwen-VL (Bai et al., 2023), Qwen-VL-Chat (Bai et al., 2023), mPLUG-Owl2 (Ye et al., 2024) and IDEFICS (Laurençon et al., 2024). For **preference optimization-**

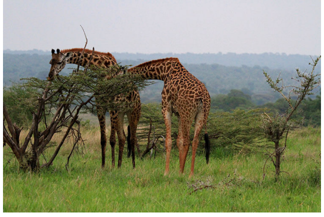
Sample of Visual Qustion Answering:	
	
User	Are the trees taller than the giraffes?
LLaVA-1.5	Yes, the trees are taller than the giraffes, as they are reaching up to eat leaves from the trees.
ASPO (Ours)	No, the trees are not taller than the giraffes. The giraffes are eating leaves from the trees, which are at a height that is accessible to them.

Table 3: Our approach is more faithful to the image input and demonstrates stronger reasoning capabilities. Hallucination outputs are shown in red.

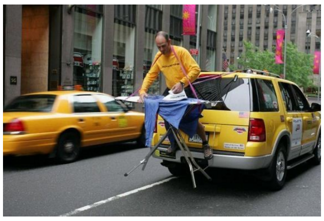
Sample of Image Captioning:	
	
User	Describe this photo in detail.
LLaVA-1.5	The image features a man standing on the back of a yellow SUV, ironing a shirt. He is using an ironing board placed on the back of the vehicle, which is an unusual and creative way to iron clothes while on the go. In the background, there are two other people visible, one on the left side and another on the right side of the scene. A potted plant can be seen in the middle of the scene, adding a touch of greenery to the urban environment. Additionally, there is a traffic light in the background, indicating that the scene takes place in a busy city area.
ASPO (Ours)	The image features a man ironing a shirt while standing on the back of a yellow SUV. It is driving down a city street, and the man appears to be multitasking by attending to his clothes and navigating the busy street. There is also a yellow taxi on the street. Additionally, some trees can be seen on the sidewalk, adding a touch of greenery to the urban environment.

Table 4: ASPO provides more reliable and contextually accurate descriptions. Hallucination outputs are shown in red.

based methods, we compare ASPO with SILKIE (Vlfeedback) (Li et al., 2023c), LLaVA-RLHF (Human-prefer) (Sun et al., 2023), POVID (Zhou et al., 2024a), SIMA (Wang et al., 2024c), HA-DPO (Zhao et al., 2023), RLHF-V (Yu et al., 2024), RLAIF-V (Lee et al., 2024), CLIP-DPO (Ouali et al., 2025), CSR (Zhou et al., 2024b) and FiSAO (Cui et al., 2024).

Evaluation Benchmarks We evaluate the performance of ASPO on several widely used multimodal benchmarks across different dimensions, including comprehensive benchmarks (MMVet (Yu et al., 2023b)), MMBench-Dev (Liu et al., 2025), MMBench-Test (Liu et al., 2025), MMBench-Chinese (Liu et al., 2025), SEED-Image (Li et al., 2023a), LLaVA-bench-in-the-wild (Liu et al., 2024b), hallucination benchmarks (POPE (Li et al., 2023d), SHR (Zhao et al., 2023)) and general VQA benchmarks (ScienceQA-Image (Lu et al., 2022), GQA (Hudson and Manning, 2019)).

Implementation Details The hyperparameter settings for both LLaVA-v1.5-7B and LLaVA-v1.5-13B are identical. Specifically, both models are trained for 1 epoch using a batch size of 10. The learning rate and weight decay are set as 2e-6 and 0, respectively with the learning rate adjusted by a cosine scheduler following prior research (Zhu et al., 2024). We fine-tune all linear layer parameters using LoRA (Hu et al., 2021), setting its rank to 1024 and α to 2048. The scale parameter β in ASPO is fixed as 0.1. For InstructBLIP-13B, we set LoRA's rank to 64 and α to 16 following prior research (Zhao et al., 2023). q_{proj}, k_{proj}, v_{proj} in the language model are fine-tuned. We employ a learning rate of 4e-6 with a cosine learning rate scheduler and a batch size of 1, while β is set as 0.1. The metric-weighting factor α in Equation (10) is set to 0.5. All of the experiments are conducted on 8 NVIDIA-L20Z-80GB GPUs for 1 epoch, every training takes less than 1 hour. ASPO and standard DPO share the same configuration above.

4.2 Experimental Results

As shown in Table 1, ASPO improves the performance of the base models across the board, demonstrating the effectiveness of our approach. From the experimental results, we can see that from InstructBLIP-13B, LLaVA-1.5-7B to LLaVA-1.5-13B, the performance of these models continues to improve and their improvement shows the same trend after training with our method, which is in line with our expectations. This further verifies that the credibility of the basic models is positively correlated with their performance. Therefore, as the performance of the basic models increases,

Method	MMVet	MMB^{D}	MMB^{T}	MMB^{C}	SEED^{I}	LLaVA^{W}	SQA^{I}	GQA	POPE	SHR (↓)	Avg.
LLaVA-1.5-7B	30.5	64.3	66.4	58.3	66.1	63.4	66.8	**62.0**	85.9	37.9	62.59
+DPO	33.3	64.7	66.4	58.3	66.1	65.7	66.4	61.0	86.2	40.1	63.12
+ASPO-*S*	**35.6**	65.5	67.3	59.5	66.1	74.6	67.4	61.7	86.5	34.6	64.91
+ASPO-*P*	34.9	**65.6**	**67.9**	**59.8**	**66.4**	75.4	**67.9**	61.9	**86.6**	35.2	**65.16**
+ASPO-*S*+*P*	35.3	**65.6**	67.7	59.5	66.3	**75.7**	67.7	**62.0**	**86.6**	**33.9**	**65.16**

Table 5: Ablation study of reward weight. *S* represents the Image-Text Similarity and *P* represents Textual Perplexity. The last column shows the average values of all metrics except for SHR.

Method	Level	MMVet	MMB^{D}	MMB^{T}	MMB^{C}	SEED^{I}	LLaVA^{W}	SQA^{I}	GQA	POPE	SHR (↓)	Avg.
DPO	Response	33.3	64.7	66.4	58.3	66.1	65.7	66.4	61.0	86.2	40.1	63.12
ASPO	Response	20.9	58.6	59.8	50.7	58.2	58.7	63.3	50.7	73.2	72.1	54.90
ASPO	**Sentence**	**35.3**	**65.6**	**67.7**	**59.5**	**66.3**	**75.7**	**67.7**	**62.0**	**86.6**	**33.9**	**65.16**
ASPO	Token	30.3	61.3	62.3	54.8	61.8	62.1	66.6	59.1	77.7	59.1	59.56

Table 6: Ablation study of reward level on LLaVA-1.5-7B. We calculate adaptive rewards at the response level, sentence level, and token level to explore the most appropriate solution. The last column shows the average values of all metrics except for SHR.

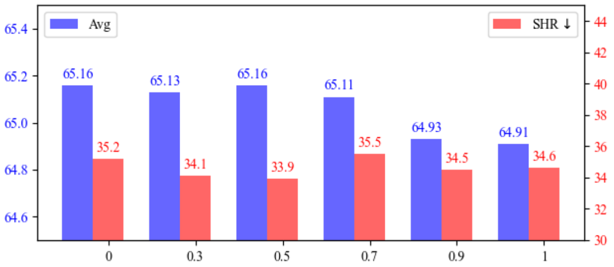


Figure 3: Ablation study of the metric-weighting α.

their credibility becomes higher and higher, and the adaptive rewards calculated based on this become much more accurate, so the improvement in model performance is much larger. We show the comparison results of our method with other preference optimization-based methods in Table 2 and ASPO achieves the highest average score of 66.18 across all evaluated datasets, which is notably higher than the baseline LLaVA-1.5-7B model's average score of 62.18. This improvement is consistent across most individual datasets as well. In addition, compared with reward mechanisms at different levels, our method shows certain advantages. These results strongly support the effectiveness of our method. The example results in Table 3 show that our method is more faithful to the image input and enhances the model's reasoning capabilities. Additionally, from the example in Table 4 on the Image Captioning task, it can be seen that after optimization with our method, the model's output is more concise and contains fewer hallucinations.

4.3 Ablation Study

In this section, we conduct comprehensive ablation experiments on the LLaVA-v1.5-7B model to verify the effectiveness of our method. As shown in Table 5, we first train the model using the standard DPO algorithm, but from the experimental results we only see a slight improvement in model performance, and the performance of the model even decreases on some benchmarks. This is mainly due to the fact that a considerable number of preference data pairs have overlapping segments, which introduces a lot of noise into our training process. However, our method can deal with this problem well. We also performed ablation on the level of adaptive rewards to explore the most appropriate reward level. The results are shown in Table 6. We first tested at the response level. Experimental results show that adaptive rewards at the response level significantly reduce model performance. This is because the training data itself contains a lot of noise, and the final implicit reward is easily affected by extremely high or low values in the overall response, which affects the stability of model training. The token-level reward is also inappropriate. This is related to the part of the speech of tokens in the response. For conjunctions, prepositions, etc., the model often shows higher confidence, while the tokens that really contain key information are not fully optimized. We will further study this issue in future work. In contrast, sentence-level adaptive rewards achieve the best performance. Figure 3 shows the ablation results of the metric-weighting factor α. We present the changes in both SHR and the average values of all other benchmarks excluding SHR. Overall, our method achieves the best results when α is set to 0.5.

5 Conclusion

In this paper, we introduce ASPO, a method designed to overcome the limitations of traditional DPO approaches, which treat entire responses as indivisible units during optimization. By employing a fine-grained optimization strategy that treats sentences as the fundamental units, ASPO enables more precise and effective alignment of MLLMs with human preferences. By moving beyond response-level optimization to embrace a sentence-level approach, our method opens new avenues for improving the comprehensive capabilities of MLLMs, paving the way for more accurate and reliable alignment with human preferences.

6 Limitations

Despite these advances, there are still areas for further exploration. One limitation of our method is the reliance on pre-defined metrics, such as image-text similarity and textual perplexity, which may not fully capture the complexity of human preferences in certain scenarios. Future work could explore incorporating more sophisticated metrics or leveraging reinforcement learning to refine the reward mechanism further. Additionally, the potential impact of ASPO on other aspects, such as robustness and generalization, warrants further investigation. Finally, due to computational resource limitations, we did not extend the experiments to the base models with much more parameters to provide a clear picture of the scalability of ASPO.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62372380 and U22B2036, Key Research and Development Program of Shaanxi Province under Grants 2024GX-ZDCYL-01-05, the Natural Science Basic Research Program of Shaanxi under Grants 2024JC-YBMS-513, Key Research and Development Program of Zhejiang Province under Grants 2024C01025, and Key Research and Development Program of Hangzhou under Grants 2024SZD1A23.

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*.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 1.

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

Jiuhai Chen, Lichang Chen, Chen Zhu, and Tianyi Zhou. 2023a. How many demonstrations do you need for in-context learning? *arXiv preprint arXiv:2303.08119*.

Jiuhai Chen and Jonas Mueller. 2024. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Annual Meeting of the Association for Computational Linguistics*, pages 5186–5200.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023b. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

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

Chenhang Cui, An Zhang, Yiyang Zhou, Zhaorun Chen, Gelei Deng, Huaxiu Yao, and Tat-Seng Chua. 2024. Fine-grained verifiers: Preference modeling as next-token prediction in vision-language alignment. *arXiv preprint arXiv:2410.14148*.

Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

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

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

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

Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. 2024. Regiongpt: Towards region understanding vision language model. In *Conference on Computer Vision and Pattern Recognition*, pages 13796–13806.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.

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*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. 2023. Language is not all you need: Aligning perception with language models. *Neural Information Processing Systems*, 36:72096–72109.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.

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 *Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

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

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Neural Information Processing Systems*, 36.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. 2024. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *International Conference on Machine Learning*.

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

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llavamed: Training a large language-and-vision assistant for biomedicine in one day. *Neural Information Processing Systems*, 36.

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

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Weibin Liao, Xu Chu, and Yasha Wang. 2024. Tpo: Aligning large language models with multi-branch & multi-step preference trees. *arXiv preprint arXiv:2410.12854*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Neural Information Processing Systems*, 36.

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

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Neural Information Processing Systems*, 35:2507–2521.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. 2025. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *European Conference on Computer Vision*, pages 395–413. Springer.

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

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

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

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*.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Daniel Cox, Yiming Yang, and Chuang Gan. 2024. Salmon: Self-alignment with instructable reward models. In *International Conference on Learning Representations*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. 2024b. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Furong Huang, et al. 2024c. Enhancing visual-language modality alignment in large vision language models via self-improvement. *arXiv preprint arXiv:2405.15973*.

Yuanhao Wang, Qinghua Liu, and Chi Jin. 2023. Is rlhf more difficult than standard rl? a theoretical perspective. *Neural Information Processing Systems*, 36:76006–76032.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multimodal large language model with modality collaboration. In *Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.

Eunseop Yoon, Hee Suk Yoon, SooHwan Eom, Gunsoo Han, Daniel Wontae Nam, Daejin Jo, Kyoung-Woon On, Mark A Hasegawa-Johnson, Sungwoong Kim, and Chang D Yoo. 2024. Tlcr: Token-level continuous reward for fine-grained reinforcement learning from human feedback. *arXiv preprint arXiv:2407.16574*.

Tianyu Yu, Jinyi Hu, Yuan Yao, Haoye Zhang, Yue Zhao, Chongyi Wang, Shan Wang, Yinxv Pan, Jiao Xue, Dahai Li, et al. 2023a. Reformulating vision-language foundation models and datasets towards universal multimodal assistants. *arXiv preprint arXiv:2310.00653*.

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

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

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

Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. 2024b. Osprey: Pixel understanding with visual instruction tuning. In *Conference on Computer Vision and Pattern Recognition*, pages 28202–28211.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

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

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

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

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

Ke Zhu, Liang Zhao, Zheng Ge, and Xiangyu Zhang. 2024. Self-supervised visual preference alignment. In *International conference on Multimedia*, pages 291–300.