DA-DPO: Cost-efficient Difficulty-aware Preference Optimization for Reducing MLLM Hallucinations

Anonymous authors
Paper under double-blind review

Abstract

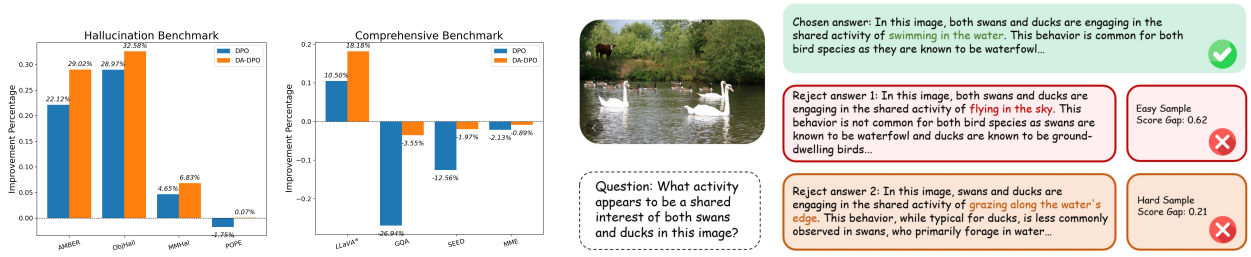
Direct Preference Optimization (DPO) has shown significant promise in reducing hallucinations in Multimodal Large Language Models (MLLMs). However, existing multimodal DPO methods suffer from overfitting due to difficulty-level imbalance in preference data. Our analysis reveals that MLLMs tend to overfit on easily distinguishable pairs, which limits their ability to remove hallucinations in a fine-grained manner and impairs the model's comprehensive ability. To address this challenge, we introduce Difficulty-Aware Direct Preference Optimization (DA-DPO), a cost-effective framework comprising two key components: (1)*Difficulty Estimation*, where we leverage pre-trained vision-language models with complementary generative and contrastive objectives, integrating their outputs through a distribution-aware voting strategy to obtain robust difficulty scores without additional training; and (2) *Difficulty-Aware Training*, where we reweight preference data according to the estimated difficulty, down-weighting easy samples while emphasizing harder ones to mitigate overfitting. This paradigm enhances preference optimization by efficiently exploiting challenging examples without requiring new data or additional fine-tuning stages. Extensive experiments demonstrate that DA-DPO significantly improves multimodal preference optimization, achieving stronger robustness against hallucinations and better generalization on standard benchmarks, all in a cost-efficient manner.

1 Introduction

Recent advancements in Multimodal Large Language Models (MLLMs) (Liu et al., 2023a; OpenAI, 2023; Li et al., 2024b) have significantly improved vision-language tasks, such as image captioning (Lin et al., 2014), visual question answering (Agrawal et al., 2015; Mathew et al., 2021; Marino et al., 2019). By combining powerful large language models with state-of-the-art vision models, MLLMs have enhanced multimodal understanding and reasoning. However, a persistent challenge for MLLMs is their tendency to produce responses that are not reliably grounded in visual inputs, often resulting in "hallucinations" where descriptions include non-existent or inaccurate visual details. This limitation affects the reliability of MLLMs, posing a significant barrier for applications that require high factual accuracy.

Recent efforts have turned to Direct Preference Optimization (DPO)(Rafailov et al., 2024) as a promising approach to mitigate hallucinations in MLLMs. DPO encourages models to align their outputs with preference data that favor faithful responses and reduce hallucinations. Crucially, the effectiveness of DPO hinges on the quality of pairwise preference data. To address this, early approaches (Sun et al., 2023c; Yu et al., 2024a) rely on manual annotation, but such data collection is both labor-intensive and difficult to scale. More recently, several works (Pi et al., 2024; Li et al., 2023d; Zhou et al., 2024c; Yang et al., 2025) have proposed automated strategies for constructing multimodal preference data. These methods exploit trained models to produce pairwise preference data at scale, significantly increasing data coverage across diverse scenarios and thereby improving the model's ability to reduce hallucinations.

Despite their effectiveness in reducing hallucinations, vanilla DPO methods trained on existing pairwise preference data often lead to noticeable degradation in general multimodal capabilities, as shown in Figure 1a. We attribute this limitation to an imbalance between easy and hard samples in the training data, as illustrated



(a) (b)

Figure 1: **(1a) Performance Comparison of DPO and DA-DPO.**. We provide the performance improvements of DPO and DA-DPO compared to the LLaVA v1.5 7B without preference optimization. The *Hallucination* indicates the performance on 4 hallucination benchmarks, and *Comprehensive* indicates the performance of 4 comprehensive MLLM benchmarks. The details are described in the experiments section. **(1b) Easy and hard pairwise samples:** "Easy Samples" have a large score gap due to clear differences between preferred and dispreferred responses, while "Hard Samples" show minor differences, making them more valuable for learning.

in Figure 1b. Easy samples typically involve clearly distinguishable faithful and hallucinated responses, whereas hard samples require more nuanced reasoning to differentiate. This imbalance leads to a training bias where models overfit to easy cases while failing to learn from more challenging examples. We provide a detailed empirical analysis of this phenomenon in Section 3, showing that while models quickly adapt to easy samples, they struggle to generalize to hard ones, ultimately limiting the effectiveness of preference-based alignment.

To address this limitation, we propose a difficulty-aware training framework that dynamically balances the contribution of easy and hard samples during preference optimization. A key challenge in implementing this strategy is the *lack of explicit supervision for estimating sample difficulty.* We tackle this by introducing a lightweight, training-free strategy: by aggregating signals from multiple pre-trained vision–language models (VLMs) trained under diverse paradigms, we obtain robust difficulty scores that estimate the difficulty of pairwise preference data without explicit training a specific model. These difficulty scores are then used to reweight preference data, enabling effective difficulty-aware training that emphasizes harder samples while preventing overfitting to easier ones.

Specifically, we propose **D**ifficulty **A**ware **D**irect **P**reference **O**ptimization (DA-DPO), a framework that consists of two steps: *difficulty estimation* and *difficulty-aware training*. The first step assesses the difficulty of each pairwise preference sample using multiple VLMs. In particular, we leverage both contrastive VLMs (e.g., CLIP (Radford et al., 2021b)) and generative VLMs (e.g., LLaVA (Liu et al., 2023b)) to estimate difficulty from complementary perspectives. Their outputs are aggregated through a distribution-aware voting strategy, in which the weight of each VLM is adaptively derived from its observed classification reliability over the training data. Building on these scores, the second step performs difficulty-aware training by dynamically adjusting the optimization strength of each sample in DPO. Specifically, the difficulty scores adjust the degree of divergence permitted between the learned policy and the initial policy. This mechanism strengthens learning from challenging samples while limiting unnecessary drift on trivial ones.

We conduct experiments on three popular MLLMs with different scales and abilities. To provide a comprehensive comparison, we report the performance comparison and analysis on two sets of benchmarks, hallucination benchmarks (Wang et al., 2023a; Rohrbach et al., 2018; Sun et al., 2023c; Li et al., 2023e) and general MLLM benchmarks (Hudson & Manning, 2019; Liu et al., 2023b; Fu et al., 2023b; Li et al., 2023a), which demonstrate the effectiveness of our approach.

Our main contributions are summarized as follows:

- We conduct analysis on the multimodal preference optimization training and empirically demonstrate the existence of an overfitting issue, which can lead to suboptimal performance.

- We propose a cost-effective framework that leverages vision-language models (VLMs) to estimate the sample difficulty without additional training and utilize the estimation to improve preference modeling via difficulty-aware training.
- We evaluate our method on hallucination and comprehensive benchmarks, and experimental results show that it significantly enhances the performance of various MLLMs in a cost-efficient manner.

2 Preliminaries

In this section, we provide a brief overview of the Reinforcement Learning from Human Feedback (RLHF) to Direct Preference Optimization (DPO) pipelines.

RLHF Reinforcement Learning from Human Feedback (RLHF) is a widely used framework for aligning LLMs with human values and intentions. The standard approach (Bai et al., 2022a; Ouyang et al., 2022) first trains a reward model and then optimizes a KL-regularized reward objective to balance preference alignment with output diversity. The optimization can be written as:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim D,\, y\sim\pi_\theta(x)}[r_\phi(x,y)] - \beta\,\mathrm{KL}[\pi_\theta(y|x)\,\|\,\pi_{\mathrm{ref}}(y|x)], \tag{1}$$

where π_{ref} is a reference policy (typically the SFT model) and β controls the trade-off between reward maximization and staying close to π_{ref}. The objective is usually optimized with PPO (Ouyang et al., 2022).

Pair-wise Preference Optimization Despite the success of the above RLHF, PPO is challenging to optimize. To enhance the efficiency of PPO, DPO (Rafailov et al., 2024) reparameterizes the reward function with the optimal policy:

$$r(x,y) = \beta \log \left(\frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)} \right) + \beta \log Z(x), \tag{2}$$

where $Z(x)$ denotes the partition function ensuring proper normalization. The hyperparameter β, analogous to the KL weight in Eq. (1), controls the trade-off: a larger value encourages π_θ to remain closer to the reference policy, preserving generalization and robustness, while a smaller value places greater emphasis on preference alignment but risks overfitting.

Building on this reward formulation, we can directly integrate it into the Bradley-Terry model, which treats pairwise preferences probabilistically. By doing so, we can optimize the preference objective without learning a separate reward model. The optimization objective is described as:

$$\min_{\pi_\theta} -\mathbb{E}_{x,y_c,y_r}[\log \sigma(r(x,y_c) - r(x,y_r))], \tag{3}$$

where $r(x,y)$ can be any reward function parametrize by π_θ, such as those defined in Eq. (2) and y_c and y_r denote the chosen and rejected responses in pairwise preference data, respectively.

3 Multimodal Preference Optimization Analysis

In this section, we present a systematic investigation of the prevalent overfitting challenge in multimodal preference optimization. Through empirical analysis, we demonstrate that models exhibit a tendency to overfit to simpler training samples, while progressively reducing their effective learning from harder instances. This phenomenon is particularly pronounced in pairwise training paradigms like DPO (Rafailov et al., 2024). This overfitting behavior ultimately compromises model performance when applied to diverse real-world scenarios. We substantiate these findings with quantitative evidence drawn from training dynamics and reward trend analyses.

Definition of Easy and Hard Samples To analyze overfitting in preference optimization, we first define what constitutes an *easy* or *hard* sample. Under large-scale pairwise preference datasets, acquiring reliable human annotations to assess sample difficulty is often infeasible due to the cost. To address this limitation, we

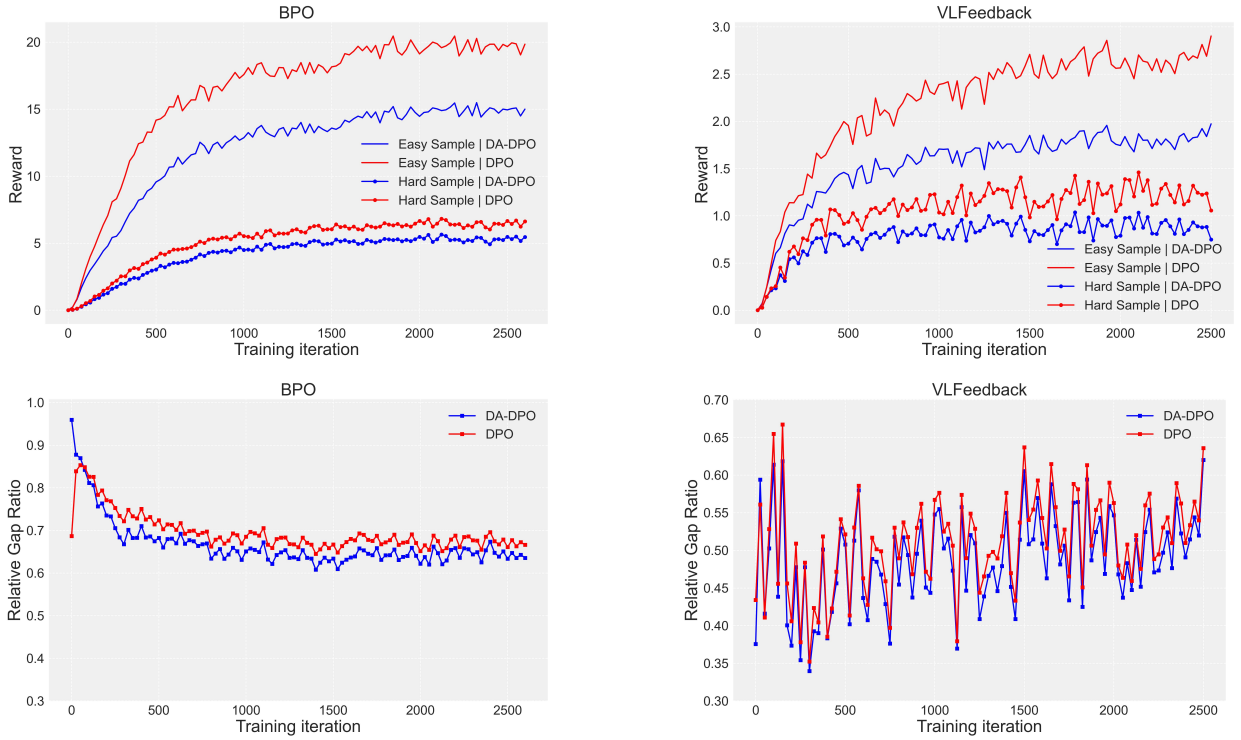


Figure 2: **Reward Trends and Relative Reward Gaps Between Easy and Hard Samples.** We present training dynamics of LLaVA-v1.5-7B for both DPO and DA-DPO on two datasets: BPO and VLFeedback. The first row shows the reward trajectories of easy and hard samples over training iterations, highlighting how the reward evolves for different difficulty levels. The second row illustrates the relative reward gap between easy and hard samples, providing a quantitative measure of the reward difference throughout training.

adopt an automatic CLIP-based difficulty estimation approach and only preserve high-confidence predictions serve as a practical alternative. The CLIP-based difficulty estimation is detailed described in the *Contrastive Estimation* in Section 4.1. We rank all pairwise preference samples by their CLIP-based difficulty scores and focus on high-confidence predictions. Specifically, samples with scores above a high threshold are treated as *easy samples*, whereas those below a low threshold are regarded as *hard samples*. Samples falling in between these thresholds are discarded, since their ambiguous difficulty would otherwise confound the analysis of easy versus hard samples. This partitioning strategy enables us to systematically analyze the behavior of easy and hard samples during preference optimization, even in the absence of ground-truth difficulty annotations.

Reward Dynamics We analyze the reward dynamics from two complementary perspectives. The first perspective examines how the rewards of easy and hard samples evolve throughout the training process. As shown in the first row of Figure 2, we observe that for both DPO and our proposed DA-DPO, the reward of hard samples remains consistently lower than that of easy samples. This trend indicates a limited capacity to fit hard samples, which typically require fine-grained understanding capabilities. Moreover, *We observe that in DA-DPO, the reward of easy samples increases more slowly compared to that in naive DPO*. This slower growth is a result of DA-DPO's adaptive weighting mechanism, which adjusts the importance of each training instance based on its estimated difficulty. By doing so, DA-DPO effectively reduces the over-optimization on easy samples, thereby alleviating the overfitting tendency frequently observed in standard DPO training.

To further quantify this phenomenon, we introduce a second perspective: the *relative reward gap* between easy and hard samples. This metric is computed by taking the difference between the reward of easy samples and that of hard samples, normalized by the reward of easy samples. It reflects the proportion of the reward difference relative to the easy sample reward. A larger relative gap indicates a stronger optimization

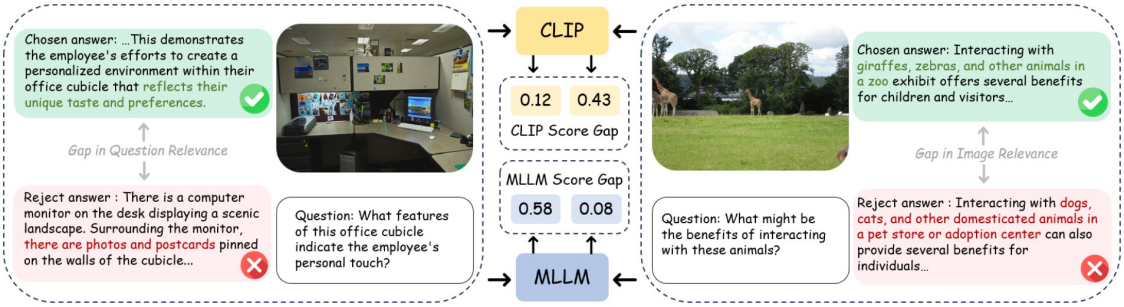


Figure 3: **Illustration of the two contrastive and generative VLMs.** As shown on the right side, CLIP captures the misalignment between the image and the answer such that the CLIP score gap is large when chosen and rejected answers differ in image relevance. However, as shown on the left side, the MLLM is better at capturing the logical connection between the question and answers, such that the MLLM score gap is large when chosen and rejected answers differ in question relevance.

preference toward easy samples. As shown in the second row of Figure 2, the relative gap in DA-DPO remains consistently smaller than that in naive DPO, indicating that DA-DPO achieves a more balanced optimization across samples of varying difficulty. These analyses collectively demonstrate the presence of overfitting in multimodal preference optimization and underscore the potential of DA-DPO in addressing this issue. Further comparative analyses on additional MLLMs are provided in the Appendix.

4 Methods

In this section, we introduce the DA-DPO framework, which addresses the overfitting issue in standard multimodal DPO training in a cost-efficient manner. In Section 4.1, we first discuss the estimation of preference data, where CLIP and MLLMs to evaluate the difficulty of preference data. In Section 4.2, we explain how the data difficulty estimation from pretrained VLMs informs and guides the difficulty-aware DPO training.

4.1 Data Difficulty Estimation

The key challenge in evaluating the difficulty of preference data is the lack of explicit supervision. To address this, we propose a lightweight, training-free strategy that leverages pre-trained contrastive and generative VLMs to estimate sample difficulty from complementary perspectives, as illustrated in Figure 3. To combine the estimates from multiple models, we adopt a distribution-aware voting strategy, where each model's contribution is proportional to its preference classification accuracy on the training set. This results in a robust difficulty score for each sample without requiring additional model training. The details are described in the following sections.

Contrastive Estimation CLIP is trained on web-scale image captions via contrastive training objectives and is proven to contain generalized knowledge regarding image and text relevance. We utilize CLIP to evaluate the difficulty of pairwise DPO data. Specifically, we first compute the CLIP text embeddings for the chosen response y_c and rejected response y_r (denoted as f_c and f_r, respectively), and the CLIP image embedding for the image in DPO data m (denoted as v_m). The CLIP scores c_c and c_r represent the image-text relevance of both responses is computed as follows:

$$c_c = \text{CosSim}(f_c, v_m), \quad c_r = \text{CosSim}(f_r, v_m). \tag{4}$$

We then introduce CLIP score gap c_g, which reflects the difficulty of a DPO sample; a larger gap indicates that the chosen and rejected responses are easily distinguishable in terms of image-text relevance. Formally, the CLIP score gap c_g is defined as follows:

$$c_g = c_c - c_r. \tag{5}$$

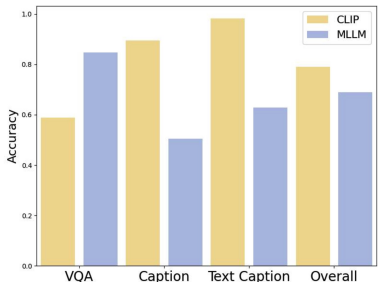


Figure 4: **Preference classification comparison.** Preference classification evaluates whether the pre-trained VLMs output a higher reward score for the chosen answer compared with the rejected answer. We report the classification accuracy on three subcategories and the overall performance on the BPO training dataset.

To integrate this metric with other VLM-based estimates, we normalize c_g to a common scale via dataset-level Gaussian projection. Given a dataset of N samples, the normalized score is computed as follows:

$$\hat{c}_g = \Phi(\frac{c_g - \mu_{c_g}}{\sigma_{c_g}}), \tag{6}$$

where μ_{c_g} and σ_{c_g} are the mean and variance of all CLIP score gaps c_g in the DPO dataset. The normalized CLIP score gap $\hat{c}_g \in [0, 1]$, and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard Gaussian distribution.

Generative Estimation Recent Multi-modal large language models (MLLMs) are built on large language models (LLMs) and learn to interact with the visual world by connecting a visual encoder to an LLM with language modeling objectives. For simplification, we denote this as a generative VLM. Although such models are prone to hallucination, they provide an informative evaluation of DPO data difficulty from another perspective complementary to CLIP. We extract the MLLM score for the chosen responses y_c and rejected responses y_r, denoted as $m_c \in \mathbb{R}$ and $m_r \in \mathbb{R}$, which is defined as follows:

$$m_c = \sum_{t=1}^{T} \log\left(\mathrm{P}(y_c^t | m, y_c^1, \ldots, y_c^{t-1}; \pi_{\text{ref}})\right), \tag{7}$$

$$m_r = \sum_{t=1}^{T} \log\left(\mathrm{P}(y_r^t | m, y_r^1, \ldots, y_r^{t-1}; \pi_{\text{ref}})\right), \tag{8}$$

where y_c^t and y_r^t are the t^{th} tokens in the chosen and rejected responses, and T is the sequence length. To evaluate the difficulty of pairwise DPO data, the MLLM score gap m_g is computed as follows:

$$m_g = m_c - m_r. \tag{9}$$

For a similar reason as the normalization of CLIP score, we utilize a Gaussian normalization to acquire a normalized MLLM score gap $\hat{m}_g \in [0, 1]$. The process is defined as follows:

$$\hat{m}_g = \Phi(\frac{m_g - \mu_{m_g}}{\sigma_{m_g}}), \tag{10}$$

where μ_{m_g}, σ_{m_g} is the mean and variance of all the MLLM score gaps m_g in the DPO dataset.

Distribution-aware Voting Fusion To this end, we evaluate the pairwise DPO data from two perspectives, resulting in two difficulty scores. As shown in Figure 4, the two VLMs perform differently in preference classification: CLIP excels on caption-related preference data, while MLLM performs better on VQA-related data. We propose a data-driven voting strategy to adaptively combine the difficulty scores based on the preference classification results. Specifically, we use the classification accuracies of CLIP and MLLM, denoted

as cls_c and cls_m, to determine the weight of *beta* for each DPO sample, as described below:

$$\hat{\beta} = [(\frac{cls_c}{cls_c + cls_m}\hat{c}_g + \frac{cls_m}{cls_c + cls_m}\hat{m}_g) + 1], \tag{11}$$

where we add a constant term of 1 to improve the stability of the training process. Without this term, when both $\hat{c}_g$ and $\hat{m}_g$ are close to zero, the resulting $\hat{\beta}$ would also approach zero, potentially leading to training collapse.

4.2 Difficulty-aware Training

After estimating the difficulty of the preference pairs, we obtain a robust score that reflects the difficulty of the pairwise DPO data. Building on previous work (Wu et al., 2024a), we perform difficulty-aware DPO training by adaptively calibrating the β in Eq. (1). This approach allows us to adjust the weight of each training sample, reducing overfitting on easy samples compared to standard DPO training. Formally, the proposed difficulty-aware DPO objective is defined as:

$$\mathcal{L}_{\text{DA-DPO}} = -\mathbb{E}_{(x,y_c,y_r,\hat{\beta})\sim D}\left[\log\sigma\left(\hat{\beta}\, r(x,y_c) - \hat{\beta}\, r(x,y_r)\right)\right], \tag{12}$$

where $\hat{\beta}$ is the difficulty-aware scaling factor, $r(x,y)$ denotes the reward function, y_c and y_r are the chosen and rejected responses, and $\sigma(\cdot)$ is the sigmoid function.

5 Experiments

5.1 Implementation Details

Training Setup To validate the effectiveness of the proposed methods, we use the pair preference datasets from BPO (Pi et al., 2024). This dataset contains 180k pairwise preference data, where the negative responses are generated by the Image-Weakened prompting and LLM Error Injection. For training parameters, we train the model for 1 epoch and set the β_0 to 0.2. For LLaVA V1.5, we follow previous work (Pi et al., 2024) to adopt the LORA (Hu et al., 2021) training with rank 32 and LORA alpha 256. The learning rate is set to 2e-6. For LLaVA-OneVision, we use the recommended official training script to perform full fine-tuning where the learning rate is 5e-7. The training takes about 7 hours for LLaVA V1.5 7B models and 22 hours for LLaVA-OneVision 7B. For the analysis section, we preserve top and bottom 20% scores as high confidence CLIP-based predictions.

Choice of VLMs For the choice of VLMs, we employ the CLIP (Radford et al., 2021a) ViTL/14@336. For the preferred and negative text responses in each data sample, we encode the longest possible text from the start and truncate the rest of the text. For MLLM, we use the LLaVA v1.5 7B (Liu et al., 2023a) to compute the probability of responses given the image and question. We also provide an ablation study on the choice of these models in Section 5.5.

5.2 Evaluation Benchmarks

To comprehensively evaluate the impact of preference optimization on MLLMs, we select two types of benchmarks. Hallucination benchmarks measure the model's ability to reduce factual errors, which is the primary goal of multimodal preference alignment. Comprehensive benchmarks assess general multimodal capabilities, ensuring that improvements in hallucination do not come at the cost of overall performance.

Hallucination Evaluation. Following previous works (Pi et al., 2024; Wang et al., 2024a; Ouali et al., 2024), we comprehensively evaluate the DA-DPO on various hallucination benchmarks such as AMBER (Wang et al., 2023a), MMHalBench (Sun et al., 2023c), Object HalBench (Rohrbach et al., 2018), and POPE (Li et al., 2023e). 1) AMBER provides a multidimensional framework suitable for assessing both generative and discriminative tasks. 2) MMHalBench is a question-answering benchmark with eight question types and 12 object topics. We follow the official evaluation scripts with GPT-4. 3) Object HalBench is a

Table 1: **Results on hallucination and comprehensive benchmarks.** For each column, the ↑ symbol indicates that higher values are better, while the ↓ symbol indicates that lower values are better. The *AMB. Gen.* and *AMB. Dis.* refer to the generative and discriminative components of the AMBER benchmark. The ∗ symbol denotes the evaluation results of the official model weights from BPO (Pi et al., 2024). We clarify that ‡ focuses on utilizing CLIP knowledge to construct preference data, whereas we leverage CLIP to address the overfitting issue in standard DPO training. The mDPO† provides an alternative approach to improving multimodal preference optimization by addressing the over-prioritization of language preference, which is orthogonal to the DA-DPO approach.

	AMB.^{G}				AMB.^{D}	ObjHal		MMHal		POPE	LLaVA^W	GQA	Seed^I	MME^P	MME^C
	C_s ↓	Cov. ↑	Hal. ↓	Cog. ↓	F1 ↑	C_s ↓	C_i ↓	Score ↑	HalRate ↓	F1 ↑	Score ↑	Acc↑	Acc↑	Score↑	Score↑
Reference Only (Not directly comparable)															
CogVLM	5.6	57.2	23.6	1.3	-	-	-	-	-	-	-	-	-	-	-
mPLUG-Owl2	10.6	52.0	39.9	4.5	-	-	-	-	-	-	56.1	-	-	-	-
InstructBLIP	8.8	52.2	38.2	4.4	-	-	-	-	-	-	58.2	-	58.8	1212.8	-
Qwen-VL	5.5	49.4	23.6	1.9	-	36.0	21.3	2.89	0.43	-	-	59.3	-	1487.6	360.7
GPT-4V	4.6	67.1	30.7	2.6	-	13.6	7.3	3.49	0.28	-	-	-	-	-	-
LLaVA-v1.5-7B	7.8	51.0	36.4	4.2	74.7	54.7	15.9	2.19	0.57	85.8	63.8	62.0	66.1	1510.7	307.5
+ HA-DPO	7.2	33.6	19.7	2.6	-	39.9	19.9	1.97	0.60	-	-	-	-	-	-
+ CLIP-DPO‡	3.7	47.8	16.6	1.3	-	-	-	-	-	85.8	-	-	-	1468.7	-
+ mDPO†	4.4	52.4	24.5	2.4	-	35.7	9.8	2.39	0.54	-	-	-	-	-	-
Main Results															
LLaVA-v1.5-7B	7.8	51.0	36.4	4.2	74.7	54.7	15.9	2.19	0.57	85.8	63.8	**62.0**	**66.1**	**1510.7**	307.5
+ DPO*	5.5	**58.4**	35.7	**2.0**	83.9	43.3	10.0	**2.24**	0.53	84.3	70.5	45.3	57.8	1409.4	315.0
+ DA-DPO	**4.3**	57.4	**28.0**	2.1	**85.6**	**39.7**	**9.9**	2.22	**0.50**	**85.9**	**75.4**	59.8	64.8	1406.6	**323.2**
LLaVA-v1.5-13B	7.1	52.4	33.9	3.8	**88.6**	56.3	15.8	2.22	0.57	**86.9**	71.2	**63.0**	**67.5**	**1496.7**	**308.9**
+ DPO	6.6	**61.9**	45.7	2.5	85.5	39.0	8.8	2.05	0.55	86.0	74.1	60.9	53.5	1450.6	276.4
+ DA-DPO	**5.1**	59.0	**32.2**	**1.9**	87.3	**37.0**	**8.3**	**2.39**	**0.48**	85.6	**74.5**	61.5	61.4	1474.2	276.4
LLaVA-OV-7B	8.4	**74.8**	70.4	9.8	90.6	41.3	8.1	2.76	0.37	**87.2**	**80.0**	**59.6**	**75.8**	**1580.4**	**406.4**
+ DPO	7.7	66.5	59.0	4.3	**91.9**	34.3	8.6	2.61	0.38	86.1	72.9	55.0	73.7	1545.4	364.6
+ DA-DPO	**7.0**	64.1	**51.8**	**3.7**	91.7	**28.0**	**8.0**	**2.78**	**0.30**	86.0	73.7	59.2	75.6	1551.4	381.8

standard benchmark for assessing object hallucination, and we follow the settings in (Wang et al., 2024a). 4) POPE utilizes a polling-based query to evaluate the model's hallucination. We report the average F1 score of three kinds of questions in POPE.

Comprehensive Evaluation. For evaluating MLLM helpfulness, we use: 1) LLaVA-Bench (Liu et al., 2023b), a real-world benchmark with 60 tasks assessing LLaVA's visual instruction-following and question-answering abilities. We use official scripts to compute scores with GPT-4. 2) Seedbench (Li et al., 2023a), which consists of 14k multiple-choice VQA samples to evaluate the comprehensive ability of MLLMs. 3) MME (Fu et al., 2023a) which measures both perception and cognition abilities using yes/no questions. 4) GQA Hudson & Manning (2019) evaluates real-world visual reasoning and compositional question-answering abilities in an open-ended answer generation format.

5.3 Baselines

The proposed DA-DPO framework is designed to improve pairwise preference optimization by introducing the pretrained VLMs in a cost-efficient manner. We mainly compare DA-DPO with standard DPO (Rafailov et al., 2024) under three MLLMs, LLaVA V1.5 7B/13B (Liu et al., 2023a) and LLaVA-OneVision 7B (Li et al., 2024b). Additionally, we provide results from other multimodal LLMs, such as GPT4-V (OpenAI, 2023), CogVLM (Wang et al., 2023b), mPLUG-Owl2 (Ye et al., 2023), InstructBLIP (Dai et al., 2023),

Table 2: **The impact of estimation from different VLMs.** We report the performance of the model trained with different strategies on hallucination benchmarks such as the AMBER benchmark and the comprehensive benchmark SeedBench. The row with the blue color indicates the adopted strategy.

	VLMs		AMB.^{G}				AMB.^{D}	Seed^I
	CLIP	MLLM	$C_s \downarrow$	Cov. $\uparrow$	Hal. $\downarrow$	Cog. $\downarrow$	F1 $\uparrow$	Score $\uparrow$
DPO	×	×	5.5	58.4	35.7	2.0	83.9	57.8
Ours	✓	×	4.7	**58.0**	31.9	2.0	85.2	63.8
	×	✓	4.6	57.6	29.9	2.3	85.3	64.3
	✓	✓	**4.3**	57.4	**28.0**	**2.1**	**85.6**	**64.8**

Table 3: **Results of LLaVA v1.5 7B trained with VLFeedback and LLaVA-RLHF dataset.** The VLFeedback is an automatically constructed dataset by collecting responses from multiple VLMs and filtering the results with GPT4-V. The LLaVA-RLHF is a human-annotated dataset.

	AMB.^{G}				AMB.^{D}	ObjHal		MMHal		POPE	LLaVA^W	GQA	Seed^I	MME^P	MME^C
	$C_s \downarrow$	Cov. $\uparrow$	Hal. $\downarrow$	Cog. $\downarrow$	F1 $\uparrow$	$C_s \downarrow$	$C_i \downarrow$	Score $\uparrow$	HalRate $\downarrow$	F1 $\uparrow$	Score $\uparrow$	Acc$\uparrow$	Acc$\uparrow$	Score$\uparrow$	Score$\uparrow$
LLaVA-v1.5-7B	7.8	51.0	36.4	4.2	74.7	54.7	15.9	2.19	0.57	85.8	63.8	62.0	66.1	1510.7	307.5
VLFeedback															
+ DPO	6.5	**55.1**	34.5	**2.3**	**84.6**	49.0	13.0	2.19	0.65	84.6	72.1	59.8	63.5	1368.5	294.2
+ DA-DPO	**5.6**	53.0	**29.7**	2.7	**85.8**	**48.3**	**12.8**	**2.23**	**0.53**	84.2	**73.0**	**60.3**	**65.0**	**1422.7**	**297.1**
LLaVA-RLHF															
+ DPO	7.3	**50.8**	33.7	3.9	**86.6**	58.3	16.9	2.10	**0.57**	84.4	68.7	60.7	64.5	1415.9	**343.9**
+ DA-DPO	**5.8**	50.7	**27.6**	**3.0**	**86.6**	**48.0**	**13.8**	**2.02**	0.59	**84.5**	**71.3**	**60.9**	**64.8**	**1464.1**	301.4

Qwen-VL (Bai et al., 2023), HA-DPO (Zhao et al., 2023), CLIP-DPO (Ouali et al., 2024), mDPO (Wang et al., 2024a) for reference.

5.4 Results Analysis

Hallucination Benchmarks To demonstrate the effectiveness of the proposed methods in reducing hallucinations, we present evaluation results on various hallucination benchmarks in Table 1. We observe that the DA-DPO improves model performance on most benchmarks compared to DPO, such as the *HalRate* in AMBER generative decrease from 35.7 to 28.0 when training with LLaVA v1.5 7B. Moreover, the performance of three MLLMs is improved significantly on the Object Hallucination benchmark, which demonstrates that the DA-DPO greatly reduces the hallucination in the caption.

Comprehensive Benchmarks We evaluate the proposed methods on five comprehensive benchmarks. The results, shown in Table 1, indicate that while preference optimization reduces hallucination, performance on general abilities suffers. However, DA-DPO mitigates this degradation on most benchmarks when compared to standard DPO. Furthermore, DA-DPO significantly enhances conversational ability, with performance on the LLaVA-Bench reaching 75.4, compared to 70.5 for DPO using LLaVA v1.5 7B.

5.5 Ablation Study

Impact of the Difficult-aware Training To better understand the proposed methods, we conduct an ablation study to assess the effectiveness of each design. As shown in Table 2, we begin with the standard Direct Preference Optimization (DPO) (Rafailov et al., 2024). We then add the reward score from the CLIP to control sample weight during training, followed by results for DPO with only the MLLM's reward score. Finally, we combine both reward scores using the adaptive fusion strategy. The results show that using

Table 4: **Choices of VLMs.** The *ViTL* indicate the CLIP ViTL/14@336 and *EVA 8B* is the EVA-CLIP 8B. The *LLaVA 7B* is the LLaVA v1.5 7B and the *OV 7B* is the LLaVA-Onevision 7B.

Model Choice		AMB.^{G}				AMB.^{D}	Seed^I
CLIP	MLLM	$C_s \downarrow$	Cov. ↑	Hal. ↓	Cog. ↓	F1 ↑	Score ↑
ViTL	LLaVA 7B	**4.3**	57.4	28.0	2.1	85.6	64.8
EVA 8B	LLaVA 7B	4.9	57.1	31.8	2.6	83.4	56.2
ViTL	OV 7B	4.6	**57.7**	31.4	2.0	85.7	65.2
EVA 8B	OV 7B	**4.3**	55.6	**25.5**	2.0	**86.0**	**65.5**

Table 5: **Comparison of DA-DPO with direct filtering baselines (10%, 25%, 50% easy samples removed).** Metrics with ↑ indicate higher is better; with ↓ indicate lower is better.

Model	AMBER^G				AMBER^D	POPE	MMHal		ObjHall		LLaVA^W	GQA	Seed	MME	
	$C_s \downarrow$	$Cov. \uparrow$	$Hal. \downarrow$	$Cog. \downarrow$	F1↑	F1↑	Score↑	HalRate↓	$C_s \downarrow$	$C_i \downarrow$	Score↑	Acc↑	Acc↑	P↑	C↑
DPO	5.5	58.4	35.7	**2.0**	83.9	84.3	2.23	0.53	43.3	10.0	70.48	45.3	57.7	**1409.4**	315.0
Filter 10%	5.6	**63.2**	39.6	2.6	85.9	84.9	2.22	0.51	39.0	9.3	72.49	53.4	50.3	1365.1	300.0
Filter 25%	5.3	61.3	37.8	2.3	85.8	84.3	**2.29**	**0.50**	38.3	**9.0**	67.04	53.3	51.1	1318.6	**330.3**
Filter 50%	5.3	62.2	38.3	2.0	**86.0**	85.5	2.03	0.56	42.7	10.2	71.21	**56.3**	52.2	1359.7	306.7
DA-DPO	**4.3**	57.4	**28.0**	2.1	85.6	**85.9**	2.22	**0.50**	**37.7**	9.8	**75.38**	**59.8**	**64.8**	1406.5	323.2

either CLIP or MLLM's reward score improves both *HalRate* and $CHAIR_s$, with the combination of both achieving the best performance, highlighting the necessity of each framework design.

Influence of Preference Datasets To demonstrate that the proposed methods mitigate data bias in different types of multimodal pairwise preference data, we use VLFeedBack (Li et al., 2023d), which consists of 80k responses generated by MLLMs with varying levels of ability, and rated by GPT4-V (OpenAI, 2023). This dataset is automatically generated through a different approach compared to BPO (Pi et al., 2024), and it covers the mainstream data generation pipeline for multimodal preference data. We follow the settings of mDPO (Wang et al., 2024a) to select 10k preference data samples. Moreover, we trained on LLaVA-RLHF (Sun et al., 2023c), the most widely used human-annotated multimodal preference dataset with 10k samples. As shown in Table 3, DA-DPO outperforms DPO on most benchmarks, demonstrating that overfitting is a common issue in multimodal preference data, and DA-DPO effectively alleviates this problem in a cost-efficient manner.

Choices of the VLMs We present an ablation study on the selection of VLMs. These models estimate the difficulty of pairwise preference data and guide difficulty-aware preference optimization during training in the proposed framework. We experiment with two CLIP models of varying parameter scales: CLIP ViT-L/14@336 (Radford et al., 2021a) and EVA-CLIP 8B (Sun et al., 2023a). For MLLMs, we use LLaVA v1.5 7B (Liu et al., 2023a) and LLaVA-OneVision 7B (Li et al., 2024b). As shown in Table 4, we observe that performance remains similar across VLMs with different capabilities, which is attributed to the Gaussian normalization in Eq. (10) and Eq. (6). This normalization ensures that the VLMs only provide a ranking of preference data, making our proposed framework robust to variations in the choice of VLMs.

Comparison with Direct Filtering Baseline To further validate the effectiveness of our sample re-weighting strategy, we compare DA-DPO against a baseline that directly filters out *easy samples* from the training data. Specifically, we remove 10%, 25%, and 50% of the easy samples from the BPO dataset based on our difficulty metric and train a model using the DPO algorithm. Table 5 summarizes the performance of DA-DPO and the direct filtering baselines across a wide range of benchmarks. We observe that DA-DPO consistently outperforms the filtering-based approaches on most metrics, especially on hallucination-related benchmarks such as POPE, MMHal, and Object Hallucination. Notably, the performance of the direct filtering baseline fluctuates across different filtering ratios, with no consistent improvement as more easy samples are removed.

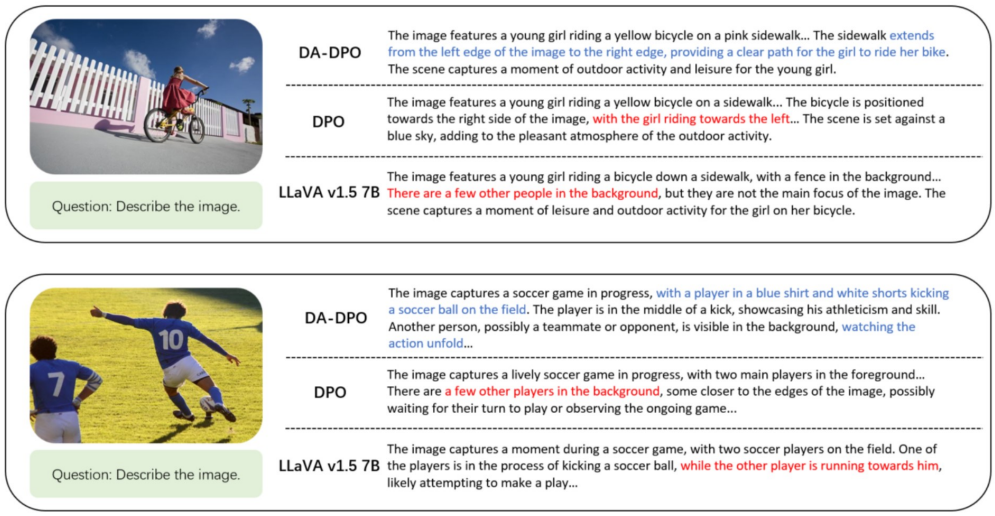


Figure 5: **Visualization of Predictions**. We present the outputs from the proposed method alongside baseline models to highlight the characteristics of our approach. The results are derived from the generative component of AMBER, which is designed to detect hallucinations in image captions.

We attribute the limited effectiveness of direct filtering to its inherent trade-off: while it removes truly uninformative samples, it also discards a portion of valuable training data, thereby reducing the diversity and coverage of preference information. In contrast, DA-DPO addresses this issue by softly down-weighting easy samples instead of hard filtering, preserving data diversity while emphasizing informative examples. This allows DA-DPO to maintain robustness across benchmarks with varying difficulty and annotation styles.

Visualization To provide a better understanding of our method's performance, we present visualizations of the model's outputs, comparing them with the results obtained from both the DPO and reference models in Figure 5. These visualizations highlight the ability of our approach to more effectively reduce hallucinations. By examining the outputs, it is evident that our method aligns better with the expected responses, demonstrating superior accuracy in scene understanding and coherence in generated content.

6 Related Works

6.1 Vision-Language Models

Vision-Language Models (VLMs) (Radford et al., 2021b; Jia et al., 2021; Li et al., 2022) have substantially advanced multimodal understanding, achieving strong performance across a wide range of downstream tasks (Guo et al., 2022; Ning et al., 2023). Contrastive VLMs, such as CLIP (Radford et al., 2021a), align images and text via large-scale contrastive objectives and demonstrate impressive zero-shot transfer on recognition tasks (Nukrai et al., 2022; Qiu et al., 2024; Liao et al., 2022; Ning et al., 2023; Yao et al., 2022; Peebles & Xie, 2022). With the rise of large language models (LLMs), subsequent works integrate LLMs with visual encoders to enhance image-conditioned text generation (Li et al., 2023b; Dai et al., 2023; Liu et al., 2023b; Zhu et al., 2023; Liu et al., 2023a; Chen et al., 2023), enabling instruction following and open-ended reasoning. More recent efforts (Li et al., 2024a;b; Gao et al., 2024) further improve visual capabilities by leveraging higher-resolution inputs, multi-image contexts, and even video sequences. Despite these advances, VLMs remain prone to hallucinations when aligning visual evidence with textual responses, motivating preference-based optimization approaches to improve faithfulness.

6.2 Preference Alignment in Large Language Models

The alignment problem (Leike et al., 2018) aims to ensure that agent behaviors are consistent with human intentions. Early approaches leveraged Reinforcement Learning from Human Feedback (RLHF) (Bai et al., 2022a;b; Glaese et al., 2022; Nakano et al., 2021; Ouyang et al., 2022; Scheurer et al., 2023; Stiennon et al., 2020; Wu et al., 2021; Ziegler et al., 2019), where policy optimization methods such as PPO (Schulman et al., 2017) were used to maximize human-labeled rewards. More recently, Direct Preference Optimization (DPO) (Rafailov et al., 2024) reformulates alignment as a direct optimization problem over offline preference data, avoiding the need for reinforcement learning. Extensions such as Gibbs-DPO (Xiong et al., 2023) enable online preference optimization, while β-DPO (Wu et al., 2024b) addresses sensitivity to the temperature parameter β by introducing dynamic calibration with a reward model. However, such approaches are less effective in multimodal domains, where reward models are vulnerable to reward hacking (Sun et al., 2023b). Other works explore alternative strategies, such as self-rewarding mechanisms or data reweighting (Wang et al., 2024b; Zhou et al., 2024a), to mitigate issues like overconfident labeling and distributional bias in preference data.

6.3 Preference Alignment in Multi-modal Models

Recent efforts have extended preference alignment from language-only models to the multi-modal domain. A major line of work focuses on constructing multimodal preference datasets. (Sun et al., 2023b; Yu et al., 2024a) collect human annotations, while others rely on powerful multimodal models such as GPT-4V (Li et al., 2023c; Yu et al., 2024b; Zhou et al., 2024d; Yang et al., 2025) to generate preference signals. However, both human annotation and large model inference incur prohibitive costs, limiting scalability. To address this, alternative approaches (Deng et al., 2024; Pi et al., 2025) explore automatic or self-training methods for preference data generation. These works synthesize dis-preferred responses from corrupted images or misleading prompts, enabling the model to learn preferences without external supervision. Another direction (Ouali et al., 2024) leverages CLIP to score diverse candidate responses, ranking preferred versus dis-preferred outputs using image–text similarity. From the perspective of training objectives, most multimodal alignment methods adopt the standard DPO objective (Li et al., 2023c; Zhao et al., 2023; Zhou et al., 2024b) to optimize preferences on paired data. Other approaches employ reinforcement learning (Sun et al., 2023b) or contrastive learning (Sarkar et al., 2024; Jiang et al., 2024) to improve alignment. To further reduce overfitting to language-only signals, (Ouali et al., 2024) extends DPO by jointly optimizing both textual and visual preferences. Despite these advances, multimodal DPO remains vulnerable to overfitting, especially when trained on imbalanced preference data. In this work, we introduce a general difficulty-aware training framework that explicitly accounts for sample difficulty, thereby mitigating overfitting and improving robustness in multimodal preference optimization.

7 Conclusion

In this work, we present an empirical analysis of the overfitting issue in multimodal preference optimization, which often stems from imbalanced data distributions. To address this, we introduce DA-DPO, a cost-efficient framework consisting of difficulty estimation and difficulty-aware training. Our method leverages pretrained contrastive and generative VLMs to estimate sample difficulty in a training-free manner, and uses these estimates to adaptively reweight data—emphasizing harder samples while preventing overfitting to easier ones. Experiments across hallucination and general-purpose benchmarks demonstrate that this paradigm effectively improves multimodal preference optimization.

Limitations Despite these promising results, our framework relies on the assumption that pretrained VLMs provide reliable evaluations of preference data. Although the adaptive voting strategy shows robustness on existing datasets, its generalizability to domains that differ substantially from the pretraining objectives of these VLMs remains uncertain. Future work may explore integrating domain-adaptive or self-improving mechanisms to further enhance robustness.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4 – 31, 2015.

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Jun Chen, Deyao Zhu1 Xiaoqian Shen1 Xiang Li, Zechun Liu2 Pengchuan Zhang, Raghuraman Krishnamoorthi2 Vikas Chandra2 Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023a.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023b.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, Wenqi Shao, Chao Xu, Conghui He, Junjun He, Hao Shao, Pan Lu, Hongsheng Li, and Yu Qiao. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *ArXiv*, abs/2402.05935, 2024.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. *arXiv preprint arXiv:2209.14169*, 2022.

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

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *ArXiv*, abs/2102.05918, 2021.

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*, pp. 27036–27046, 2024.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

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

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

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023b.

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

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *ArXiv*, abs/2312.10665, 2023d. URL `https://api.semanticscholar.org/CorpusID:266348439`.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022.

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

Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20091–20100, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *ArXiv*, abs/2310.03744, 2023a.

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

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3190–3199, 2019.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Sha Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23507–23517, 2023.

David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. *ArXiv*, abs/2211.00575, 2022.

OpenAI. Vision - openai api. `https://platform.openai.com/docs/guides/vision`, 2023.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. *arXiv preprint arXiv:2408.10433*, 2024.

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.

William S. Peebles and Saining Xie. Scalable diffusion models with transformers. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2022.

Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strengthening multimodal large language model with bootstrapped preference optimization. *ArXiv*, abs/2403.08730, 2024. URL `https://api.semanticscholar.org/CorpusID:268379605`.

Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strengthening multimodal large language model with bootstrapped preference optimization. In *European Conference on Computer Vision*, pp. 382–398. Springer, 2025.

Longtian Qiu, Shan Ning, and Xuming He. Mining fine-grained image-text alignment for zero-shot captioning via text-only training. *ArXiv*, abs/2401.02347, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021a. URL `https://api.semanticscholar.org/CorpusID:231591445`.

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

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.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Conference on Empirical Methods in Natural Language Processing*, 2018. URL `https://api.semanticscholar.org/CorpusID:52176506`.

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.

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training language models with language feedback at scale. *arXiv preprint arXiv:2303.16755*, 2023.

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

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Quan Sun, Yuxin Fang, Ledell Yu Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *ArXiv*, abs/2303.15389, 2023a.

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*, 2023b.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf. *ArXiv*, abs/2309.14525, 2023c. URL `https://api.semanticscholar.org/CorpusID:262824780`.

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*, abs/2406.11839, 2024a. URL `https://api.semanticscholar.org/CorpusID:270560448`.

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*, 2023a.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. *ArXiv*, abs/2311.03079, 2023b. URL `https://api.semanticscholar.org/CorpusID:265034288`.

Zhaoyang Wang, Weilei He, Zhiyuan Liang, Xuchao Zhang, Chetan Bansal, Ying Wei, Weitong Zhang, and Huaxiu Yao. Cream: Consistency regularized self-rewarding language models. *ArXiv*, abs/2410.12735, 2024b. URL `https://api.semanticscholar.org/CorpusID:273375180`.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*, 2021.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. -dpo: Direct preference optimization with dynamic . *arXiv preprint arXiv:2407.08639*, 2024a.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. -dpo: Direct preference optimization with dynamic . *arXiv preprint arXiv:2407.08639*, 2024b.

Wei Xiong, Hanze Dong, Chenlu Ye, Han Zhong, Nan Jiang, and Tong Zhang. Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. *arXiv preprint arXiv:2312.11456*, 2023.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10610–10620, 2025.

Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, W. Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *ArXiv*, abs/2209.09407, 2022.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Mingshi Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owi2: Revolutionizing multi-modal large language model with modality collaboration. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13040–13051, 2023.

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*, pp. 13807–13816, 2024a.

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*, 2024b.

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.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. Wpo: Enhancing rlhf with weighted preference optimization. In *Conference on Empirical Methods in Natural Language Processing*, 2024a. URL `https://api.semanticscholar.org/CorpusID:270559089`.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. Wpo: Enhancing rlhf with weighted preference optimization. *arXiv preprint arXiv:2406.11827*, 2024b.

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

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*, 2024d.

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

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

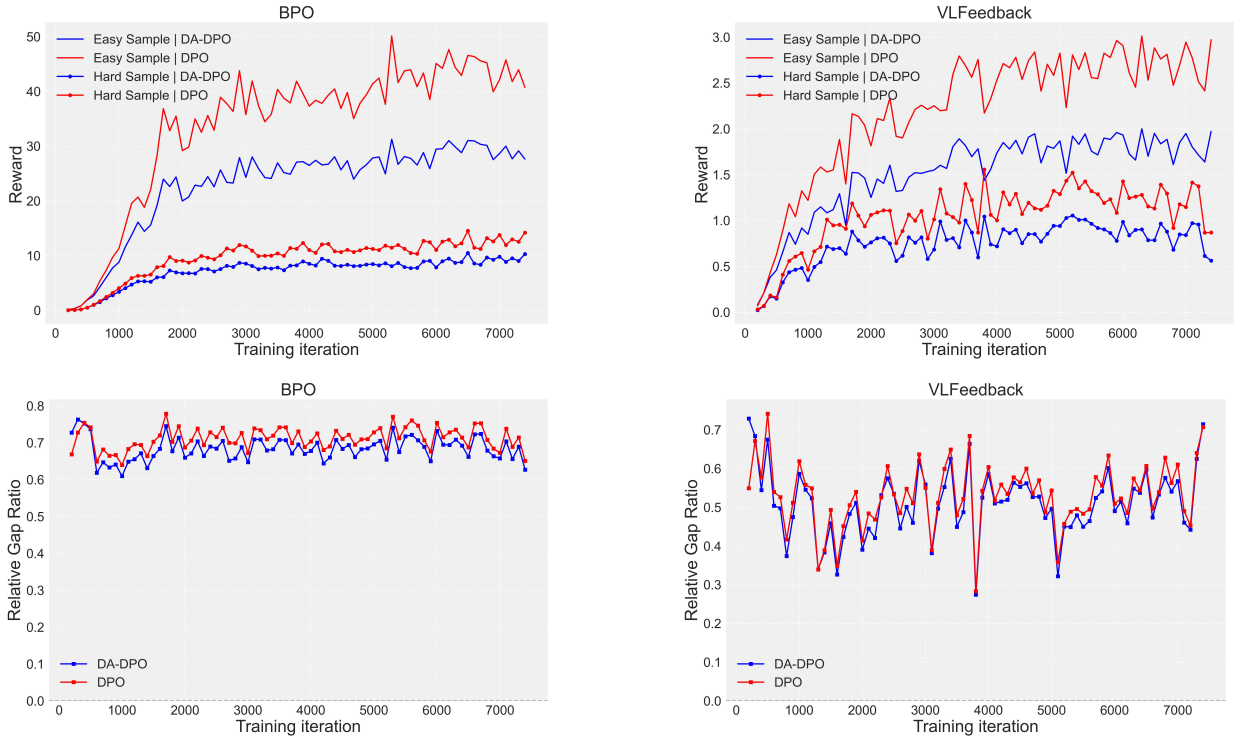


Figure 6: **Reward Trends and Relative Reward Gaps Between Easy and Hard Samples.** We present training dynamics of LLaVA-OV-7B for both DPO and DA-DPO on two datasets: BPO and VLFeedback. The first row shows the reward trajectories of easy and hard samples over training iterations, highlighting how the reward evolves for different difficulty levels. The second row illustrates the relative reward gap between easy and hard samples, providing a quantitative measure of the reward difference throughout training.

A Additional Multimodal Preference Optimization Analysis

To further validate the generality of the overfitting phenomenon discussed in the main paper, we conduct additional experiments on more recent MLLM, LLaVA-OneVision-7B, using the same DPO and DA-DPO training protocols as illustrated in Figure 6. We observe similar trends: the reward of easy samples increases steadily while the reward of hard samples plateaus or improves slowly under naive DPO, and DA-DPO helps narrow this gap. These results reinforce the universality of the overfitting issue and the effectiveness of our proposed method across different model variants.

B Influence of Adaptive Score Fusion

The proposed framework integrates two types of pretrained VLMs to assess the difficulty of pairwise preference data. We design an adaptive score fusion mechanism to determine the weight of each reward model without requiring hyperparameter selection from the data perspective. As shown in Table 6, we present the results for multiple fixed fusion scores. We observe that the adaptive score fusion achieves the best performance.

C Sensitivity of β

We propose a difficult-aware preference optimization strategy that aims to alleviate the overfitting easy sample issue during alignment training. To achieve this, we dynamically calibrate the β, which is described

Table 6: **Choices of pretrained VLMs.** The *Fusion Ratio* indicate the weight for the $\hat{c}_g$ and $\hat{m}_g$ in Equation 11. The row with blue color is the weight of adaptive score fusion weight.

Fusion Ratio		AMB.^{G}				AMB.^{D}	Seed^I
CLIP	MLLM	$C_s \downarrow$	Cov. $\uparrow$	Hal. $\downarrow$	Cog. $\downarrow$	F1 $\uparrow$	Score $\uparrow$
20%	80%	4.9	57.7	30.9	**2.0**	85.4	63.2
40%	60%	**4.3**	57.5	28.3	**2.0**	85.6	64.0
53%	47%	**4.3**	57.4	**28.0**	2.1	**85.6**	**64.8**
60%	40%	4.5	57.6	28.4	2.1	85.3	63.8
80%	20%	4.6	**58.3**	28.6	**2.0**	85.5	62.8

Table 7: **Performance with different β for DA-DPO and DPO.** We provide the results of DPO and our difficulty-aware DPO trained in the BPO dataset. We mark the reported hyperparameter with blue. For DA-DPO, we chose the best performance β, which is 0.2. For DPO, we follow previous work reports that β equals to 0.1.

β	AMBER^G				AMBER^D
	$C_s \downarrow$	Cover. $\uparrow$	H $\downarrow$	Cog. $\downarrow$	F1 $\uparrow$
DA-DPO					
0.05	5.8	62.3	41.1	2.7	79.0
0.1	4.6	59.0	32.5	2.0	85.4
0.2	4.3	57.4	28.0	2.1	85.6
0.3	4.7	53.6	25.8	2.4	86.3
0.4	4.9	54.2	28.4	2.6	86.0
DPO					
0.05	2.6	8.3	3.5	0.1	77.9
0.1	5.5	58.4	35.7	2.0	83.9
0.2	4.8	58.5	30.0	2.0	84.2
0.3	4.9	56.6	29.4	2.0	84.4
0.4	4.9	56.5	30.8	2.0	84.5

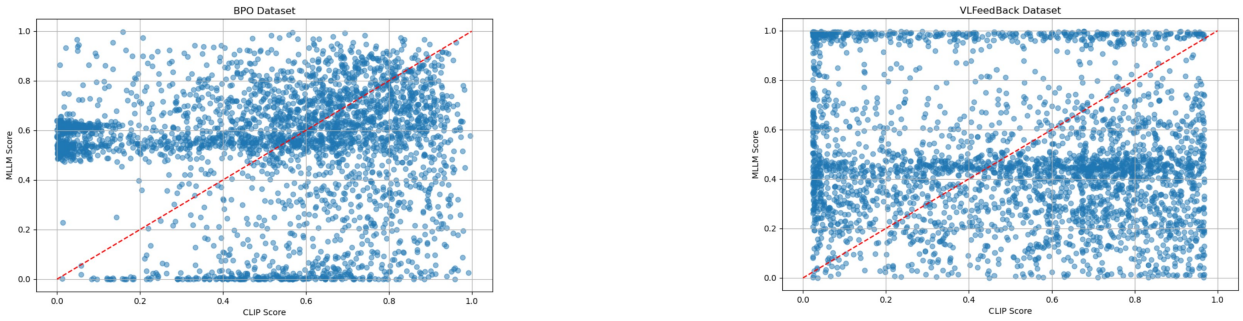
in the main paper. However, our method affects the scale of the β in the DPO objective. As shown in Table 7, we provide the ablation of the β between the DPO and DA-DPO to demonstrate the improvements of DA-DPO come from the difficulty-aware training strategy.

D Pretrained VLMs Correlation

In our proposed DA-DPO framework, we utilize the CLIP and the MLLM to evaluate the difficulty of preference data from different perspectives. To validate this claim, we provide the score of difficulty correlation between the two VLMs. As shown in Figure 7, the scores from the two models exhibit no significant positive correlation, as evidenced by their weak correlation coefficient. This indicates that the two models capture different aspects of the response evaluation, and their scoring patterns do not align consistently across the datasets.

E Fusion Score Classification Accuracy

To provide a comprehensive understanding of the score fusion process, we supplement the main text with detailed numerical results and evaluation metrics. Score fusion is conducted at the dataset level, where distinct fusion scores are employed for the BPO and VLFeedback datasets.



(a) BPO (b) VLFeedBack

Figure 7: **The visualization correlation between pretrained VLMs.** We present the normalized MLLM and CLIP scores for the same preference data sample in three datasets, BPO and VLFeedBack. The x-axis is the CLIP score and the y-axis is the MLLM score. The red line indicates the MLLM and CLIP's predictions are the same.

Table 8: Classification accuracy of fusion scores on the BPO and VLFeedback datasets.

Dataset	RM Type	VQA	Caption	Text VQA	Overall
BPO	CLIP	0.5877	0.8947	0.9827	**0.7732**
BPO	MLLM	0.8478	0.5053	0.6282	**0.6415**
VLFeedback	CLIP	0.5827	0.6352	–	**0.5897**
VLFeedback	MLLM	0.8104	0.5193	–	**0.7715**

Evaluation Metric. We compute classification accuracy on pairwise preference data using the following criterion: for each sample, the VLM assigns a score to both the chosen and rejected answers. A sample is considered correctly classified if the chosen answer receives a higher score than the rejected one.

Results. Table 8 reports classification accuracy for each category (VQA, Caption, and Text VQA, where applicable) and the overall accuracy on both datasets. We evaluate two types of VLMs: a vision encoder (CLIP) and a multimodal large language model (MLLM). Specifically, we use EVA-CLIP 8B as the CLIP model and LLaVA-1.5 7B as the MLLM.

Based on the overall accuracy, we designate the CLIP-based classification score (`cls_c`) and MLLM-based classification score (`cls_m`) as follows: For the BPO dataset, `cls_c` = 0.7732 and `cls_m` = 0.6415; for the VLFeedback dataset, `cls_c` = 0.5897 and `cls_m` = 0.7715. These results demonstrate the complementary strengths of CLIP and MLLM-based reward estimations across different categories.