DAMA: Data- and Model-aware Alignment of Multi-modal LLMs

Jinda Lu [1] **Junkang Wu** [1] **Jinghan Li** [1] **Xiaojun Jia** [2] **Shuo Wang** [1] **Yifan Zhang** [3] **Junfeng Fang** [4]
Xiang Wang [1 *] **Xiangnan He** [1 *]

Abstract

Direct Preference Optimization (DPO) has shown effectiveness in aligning multi-modal large language models (MLLM) with human preferences. However, existing methods exhibit an imbalanced responsiveness to the data of varying hardness, tending to overfit on the *easy-to-distinguish* data while underfitting on the *hard-to-distinguish* data. In this paper, we propose **Da**ta- and **M**odel-**a**ware DPO (DAMA) to dynamically adjust the optimization process from two key aspects: (1) a data-aware strategy that incorporates data hardness, and (2) a model-aware strategy that integrates real-time model responses. By combining the two strategies, DAMA enables the model to effectively adapt to data with varying levels of hardness. Extensive experiments on five benchmarks demonstrate that DAMA not only significantly enhances the trustworthiness, but also improves the effectiveness over general tasks. For instance, on the Object HalBench, our DAMA-7B reduces response-level and mentioned-level hallucination by 90.0% and 95.3%, respectively, surpassing the performance of GPT-4V. Code is available at: `https://github.com/injadlu/DAMA`.

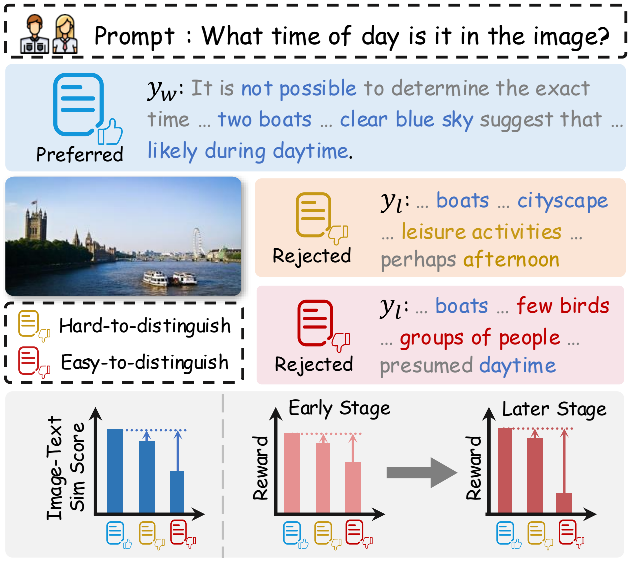

Prompt : What time of day is it in the image?
Preferred
y_w: It is not possible to determine the exact time … two boats … clear blue sky suggest that … likely during daytime.
Rejected
y_l: … boats … cityscape … leisure activities … perhaps afternoon
Rejected
y_l: … boats … few birds … groups of people … presumed daytime
Hard-to-distinguish
Easy-to-distinguish
Image-Text Sim Score
Early Stage
Reward
Later Stage
Reward

Figure 1. (1) Preference data (Prompt, Image, Preferred response y_w, Rejected response y_l) with different hardness: "easy-to-distinguish" data denotes a large Image-Text sim score gap between y_l and y_w; "hard-to-distinguish" data indicates a low score gap between y_l and y_w. (2) Implicit reward across the optimization stage: the reward gap for "easy-to-distinguish" data enhances significantly during optimization, while for "hard-to-distinguish" data, the gap remains low.

1. Introduction

Recent advances in Multimodal Large Language Models (MLLMs) have demonstrated remarkable visual understanding capabilities on the basis of large language models (Liu et al., 2023b; 2024a). However, despite their effectiveness, the hallucination issue — generating outputs inconsistent with the image content and human preference — limits their broader applicability (Liu et al., 2023a; Yu et al., 2024b). To address this, direct preference optimization (DPO) (Rafailov et al., 2024; Wu et al., 2024b;a) has been adapted into MLLM alignment (Yu et al., 2024c; Wang et al., 2024; Lu et al., 2025), achieving encouraging performance with moderate computational costs.

DPO methods (Lu et al., 2025; Yu et al., 2024c) collect preference data consisting of an image, a prompt, and two responses (y_w, y_l). The preferred response y_w is better aligned with the visual content (large Image-Text similarity score), while the rejected response y_l contains more hallucinated content (small Image-Text similarity score). They prioritize preferred responses y_w over the rejected ones y_l using DPO with a hyperparameter β, which balances retaining the reference model π_{ref} and incorporating new preferences into the

---

*Corresponding Authors. [1]MoE Key Lab of BIPC, University of Science and Technology of China [2]Nanyang Technological University [3]Institute of Automation, University of Chinese Academy of Sciences [4]National University of Singapore. Correspondence to: Xiang Wang <xiangwang1223@gmail.com>, Xiangnan He <hexn@ustc.edu.cn>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

updated model $\pi_{\boldsymbol{\theta}}$ (Gu et al., 2024; Xie et al., 2024).

However, our analysis reveals that current methods exhibit imbalanced responsiveness in handling the data with varying hardness during the optimization process, resulting in sub-optimal performance. As illustrated in Figure 1, for "easy-to-distinguish" data (large Image-Text similarity score gap between preferred y_w and rejected y_l), the reward gap amplifies during training, indicating stronger alignment. Conversely, for "hard-to-distinguish" data (small Image-Text similarity score gap), the reward gap stagnates, suggesting limited capability to distinguish y_w from y_l. This implies that current methods, which employ optimization strategies using a static β across data with varying hardness (Sun et al., 2024; Yu et al., 2024c), could fail to capture the learning dynamics inherent in multimodal preference data.

To address this imbalanced responsiveness issue, we propose **Da**ta- and **M**odel-**a**ware direct preference optimization (**DAMA**), which dynamically adapts β to both data hardness and model's responsiveness, enabling adaptively adjust model's learning behavior based on the inherent preference dynamics and model's real-time responses. Specifically, we propose two novel mechanisms:

Data-aware Preference Optimization (Section 3.1): We quantify data hardness via CLIP-based image-text similarity scores (Radford et al., 2021), decomposing responses into sub-sentences for granularity-aware estimation. Then we normalize and transform the scores into probabilities to enable effective hardness estimation. By dynamically scaling β inversely with hardness, we enforce stronger regularization on "easy-to-distinguish" samples (large β) while relaxing constraints for "hard-to-distinguish" ones (small β), preventing overfitting and underfitting, respectively.

Model-aware Preference Optimization (Section 3.2): We estimate the model's responsiveness through reward gaps between preferred (y_w) and rejected responses y_l, incorporating this into the optimization process by dynamically adjusting β. Larger β values are assigned to larger reward gaps between y_w and y_l, indicating the model's proficiency in those cases, while smaller β values correspond to smaller reward gaps, suggesting that the model needs to improve its responsiveness on those cases. This adaptive strategy allows the model to focus more on cases with less confidence, while maintaining its performance on well-learned cases.

By combining these strategies via element-wise multiplication, DAMA enables real-time adaptation to both data hardness and model responsiveness, demonstrating strong alignment performance across various evaluation benchmarks. Our contributions are summarized as follows:

- We pioneer the exploration of imbalanced responsiveness in Multi-modal LLM (MLLM) alignment, where MLLMs tend to overfit on "easy-to-distinguish" data while struggling to capture the "hard-to-distinguish" data.
- We introduce DAMA, a data- and model-aware strategy that dynamically adapts the optimization process based on both data hardness and model responsiveness, facilitating more nuanced and adaptive optimization.
- We demonstrate DAMA's effectiveness through comprehensive evaluations across various benchmarks and model sizes. For instance, on the LLaVA-7B model, it achieves more than 16% performance gains in Non-Response levels compared to vanilla DPO.

2. Preliminary

In this section, we briefly review the MLLM preference learning procedure, which starts by sampling pairwise preference data with a supervised fine-tuned (SFT) model and then optimizing on such data. Specifically, we categorize this process into the following aspects:

• **Preference construction.** Preference learning of an MLLM $\boldsymbol{\pi}$ begins by sampling pair-wise responses (y_w, y_l) from an SFT model $\boldsymbol{\pi}_{\text{SFT}}$, given an image $\mathcal{I}$ and a prompt question x. This can be formalized as $(y_w, y_l) \sim \boldsymbol{\pi}_{\text{SFT}}(y|\mathcal{I}, x)$. These pairs are then labeled, either by humans or other models, to indicate preference, such that y_w is preferred over y_l given $(\mathcal{I}, x)$, denoted as $(y_w \succ y_l|\mathcal{I}, x)$.

• **Preference optimization.** Given pair-wise preference data $(y_w, y_l) \sim \boldsymbol{\pi}_{\text{SFT}}(y|\mathcal{I}, x)$, the preference optimization process traditionally involves explicit reward modeling followed by policy optimization.

In this work, we employ Direct Preference Optimization (DPO) (Rafailov et al., 2024), which simplifies this process by directly optimizing a policy against the preference data without an explicit reward model. Instead, DPO defines an implicit reward function, formalized as:

$$\boldsymbol{r}(y|\mathcal{I}, x) = \beta \log \frac{\pi_{\boldsymbol{\theta}}(y|\mathcal{I}, x)}{\pi_{\text{ref}}(y|\mathcal{I}, x)}. \tag{1}$$

Therefore, the DPO loss can be defined as:

$$\begin{aligned} \mathcal{L}_{\text{dpo}} = -\mathbf{E}_{(\mathcal{I},x,y_w,y_l)}[\log \sigma(\beta \log \frac{\pi_{\boldsymbol{\theta}}(y_w|\mathcal{I}, x)}{\pi_{\text{ref}}(y_w|\mathcal{I}, x)} \\ -\beta \log \frac{\pi_{\boldsymbol{\theta}}(y_l|\mathcal{I}, x)}{\pi_{\text{ref}}(y_l|\mathcal{I}, x)})]. \end{aligned} \tag{2}$$

3. Approach

In this section, we describe our DAMA in detail. Specifically, we first illustrate our data-aware preference optimization, then we describe our model-aware preference optimization, and finally, we show our combination strategies for robust preference optimization. Our approach algorithm is listed in Algorithm. 1.

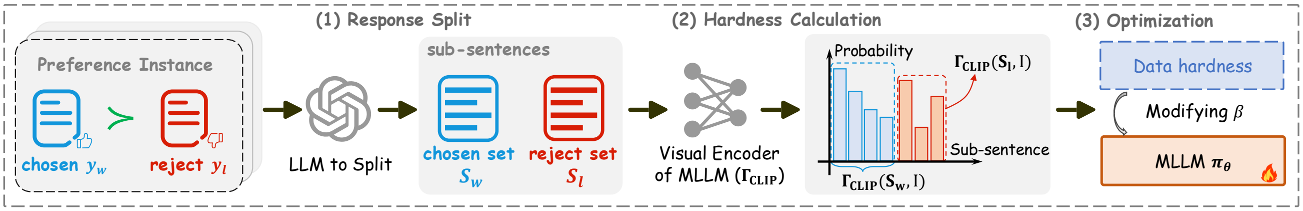


Figure 2. Overview of our data-aware preference optimization. For each preference instance: (1) We first break the preferred and rejected response into sub-sentences by prompting a large language model (LLM); (2) Next, we estimate the similarity scores between each sub-sentence and the given image using the CLIP classifier, and then calculate the differences between the preferred and rejected response as the hardness of the data; (3) Finally, we incorporate the estimated hardness into the preference optimization process by modifying β in Equ (2), allowing the model to adjust based on the data hardness.

3.1. Data-aware Preference Optimization

An overview of our data-aware preference optimization is shown in Figure 2. Given a preference instance from the dataset $\mathcal{D}$ as $\{(\mathcal{I}, x, y_w, y_l)\} \sim \mathcal{D}$, where $\mathcal{I}$, x, y_w, y_l denotes the image, question, preferred response, and rejected response, respectively, it firstly splits the responses into simple and self-contained sub-sentences. Next, it calculates the image-text similarity scores between the sub-sentences and the image by the CLIP classifier. Then, it combines the scores of each response and compares the difference between the preferred and rejected responses as the data hardness. This hardness is embedded into the preference optimization process by modifying the β in Equ (2). The following are detailed descriptions.

We employ the CLIP classifier $\mathbf{\Gamma}_{\mathrm{CLIP}}$ (Radford et al., 2021), to calculate similarity scores. For each preference instance, we aim to effectively capture the similarity between the responses (y_w, y_l) with the given image $\mathcal{I}$, while alleviating the 77 token length constraints in CLIP. To achieve this, we decompose the complex responses, which contain various objects and relations, into simple and self-contained sub-sentences. Concretely, we prompt the open-source large language model, such as LLaMA-3 (Dubey et al., 2024), to split (y_w, y_l) into sub-sentences $\mathbf{S}_w = \{\mathbf{S}_{w,j} | j = 1, 2, \dots p\}$ and $\mathbf{S}_l = \{\mathbf{S}_{l,k} | k = 1, 2, \dots q\}$, where p and q denotes the number of sub-sentences for $\mathbf{S}_w$ and $\mathbf{S}_l$.

Subsequently, we employ the CLIP classifier $\mathbf{\Gamma}_{\mathrm{CLIP}}$ to calculate the similarity score between the given image $\mathcal{I}$ and the sub-sentences $\mathbf{S}_w, \mathbf{S}_l$ as:

$$\begin{aligned} \mathbf{C}_w &= [\mathbf{\Gamma}_{\mathrm{CLIP}}(\mathcal{I}, \mathbf{S}_{w,j})]_{j=1}^{p}\,, \\ \mathbf{C}_l &= [\mathbf{\Gamma}_{\mathrm{CLIP}}(\mathcal{I}, \mathbf{S}_{l,k})]_{k=1}^{q}\,, \end{aligned} \tag{3}$$

where $\mathbf{C}_w \in \mathbb{R}^p$ and $\mathbf{C}_l \in \mathbb{R}^q$ represents the corresponding similarity scores of $\mathbf{S}_w$ and $\mathbf{S}_l$, respectively. To effectively quantify the difference between preferred response $\mathbf{C}_w$ and rejected response $\mathbf{C}_l$ for each instance, we normalize the corresponding score by the softmax probabilities as :

$$\begin{bmatrix} \mathbf{P}_w \\ \mathbf{P}_l \end{bmatrix} = \mathbf{Softmax}\left(\begin{bmatrix} \mathbf{C}_w \\ \mathbf{C}_l \end{bmatrix}\right), \tag{4}$$

$\mathbf{P}_w \in \mathbb{R}^p$ and $\mathbf{P}_l \in \mathbb{R}^q$ represents the probabilites. The difference between the preferred and rejected probabilities demonstrates the data hardness. A large difference implies that the preference data is "easy-to-distinguish", where the rejected response includes more elements that are not present in the image, conversely, a small difference suggests that the preference data is "hard-to-distinguish", and the rejected response exhibits minimal hallucination. Then we define the hardness based on the probabilities difference as:

$$\delta = \sum_{j=1}^{p} \mathbf{P}_{w,j} - \sum_{k=1}^{q} \mathbf{P}_{l,k}, \tag{5}$$

$$\alpha_{\mathrm{D}} = \sigma(\delta)/\sigma(\bar{\delta}), \tag{6}$$

where α_{D} denotes the data hardness, δ measures the difference between $\mathbf{P}_w$ and $\mathbf{P}_l$, and $\bar{\delta}$ denotes the mean difference across the dataset. The Sigmoid function $\sigma(\cdot)$ is employed to transform the response divergence $\mathbf{P}_w$ and the mean $\bar{\delta}$ into the range $(0, 1)$ for convenient comparison.

Finally, we adapt β of Equ (2) to incorporate the hardness into the optimization procedure, and each preference instance corresponding to a specific β as:

$$\beta_{\mathrm{D}} = \beta \cdot \alpha_{\mathrm{D}}. \tag{7}$$

This adjustment allows the model to optimize based on the data hardness, further enhancing its adaptability to the data.

3.2. Model-aware Preference Optimization

An overview of our model-aware preference optimization is shown in Figure 3. Given a batch of preference instances from dataset $\mathcal{D}$, as $\mathcal{B} = \{(\mathcal{I}_i, x_i, y_{w,i}, y_{l,i}) | i = 1, 2, \dots, N\} \sim \mathcal{D}$, it firstly calculates the reward gaps between the preferred $y_{w,i}$ and rejected $y_{l,i}$, and then filters

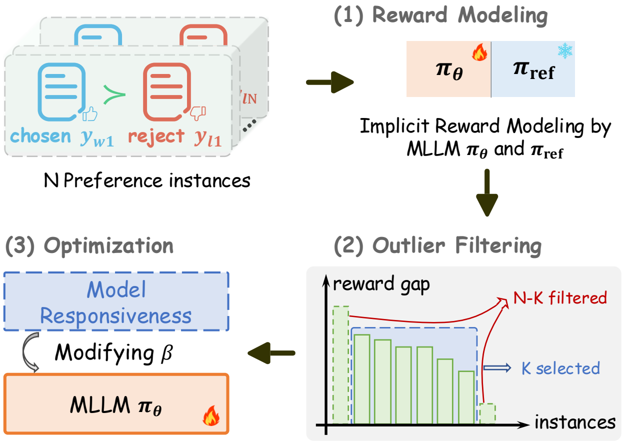


Figure 3. Overview of our model-aware preference optimization. Given N preference instances: (1) we first calculate the reward gap of each instance using the implicit reward model; (2) To ensure stable modeling, we filter out the outliers (*i.e.* the instance with excessively high or low gaps) and then estimate the average gap; (3) To enable the model to be aware of its current responsiveness, we integrate such estimation into the preference optimization process by modifying β in Equ (2).

out the outliers (*i.e.* the instance with excessively high or low gaps) for stable estimation. Such estimations are embedded into the preference optimization process by integrating into β in Equ (2), enabling the model to be aware of its current responsiveness. Details of model-aware preference optimization are as follows.

We employ current implicit reward gaps between the preferred and rejected responses of the given $\mathcal{B}$ instances to measure the current model's responsiveness. Specifically, the reward gap $\mathcal{R}$ for the i-th instance in $\mathcal{B}$ is formalized as:

$$\mathcal{R}_i = \left[\beta \log \frac{\pi_\theta(y_{w,i}|\mathcal{I}_i,x_i)}{\pi_{\text{ref}}(y_{w,i}|\mathcal{I}_i,x_i)} - \beta \log \frac{\pi_\theta(y_{l,i}|\mathcal{I}_i,x_i)}{\pi_{\text{ref}}(y_{l,i}|\mathcal{I}_i,x_i)}\right], \tag{8}$$

where π_θ and π_{ref} represent the optimizing model and reference model, respectively. We then normalize the reward gaps using the estimated mean as follows:

$$\bar{\mathcal{R}}_i = \mathcal{R}_i / \bar{\mathcal{R}}, \tag{9}$$

where $\bar{\mathcal{R}}$ represents the estimated average reward gap, and $\bar{\mathcal{R}}_i$ is the normalized one for the i-th instance.

However, the estimation remains sensitive to outliers despite normalization, especially in the full fine-tuning scenario, where the batch size is relatively small. To mitigate this issue, we filter out instances with exceptionally high or low gaps using a mask vector $\mathcal{M} \in \mathbb{R}^N$, defined as:

$$\boldsymbol{M}_i = \begin{cases} 1, & (\bar{\mathcal{R}}_i - \bar{\mathcal{R}})^2 \leq \tau, \\ 0, & (\bar{\mathcal{R}}_i - \bar{\mathcal{R}})^2 > \tau, \end{cases} \tag{10}$$

where $(\bar{\mathcal{R}}_i - \bar{\mathcal{R}})^2$ implies the squared distances from the mean, and τ represents the sorted K-th distance. With the filtering, current responsiveness of π_θ can be formalized as:

$$\bar{\mathcal{R}}_{\mathcal{B}} = \frac{1}{N-K} \sum_{i=1}^{N} \mathcal{M}_i \times \bar{\mathcal{R}}_i, \tag{11}$$

$$\alpha_{\text{M}} = \sigma(\bar{\mathcal{R}}_{\mathcal{B}}) / \sigma(\bar{\mathcal{R}}). \tag{12}$$

$\sigma(\cdot)$ is the Sigmoid function, which transforms both the filtered gaps $\bar{\mathcal{R}}_{\mathcal{B}}$ and the estimated mean $\bar{\mathcal{R}}$ into the range $(0,1)$ for convenient comparison, and α_M refers to the estimated model responsiveness.

We then integrate α_M into the optimization procedure by modifying β of Equ (2) as:

$$\beta_{\text{M}} = \beta \cdot \alpha_{\text{M}}. \tag{13}$$

By utilizing β_M to optimize π_θ with the filtered batch $\mathcal{B} \cdot \mathbf{M}$, the model can effectively adapt to its current responsiveness to the preference data.

Finally, we update the estimated mean $\bar{\mathcal{R}}$ using a moving average after optimization over batch B as follows:

$$\bar{\mathcal{R}} \leftarrow \gamma \cdot \bar{\mathcal{R}} + (1-\gamma) \cdot \bar{\mathcal{R}}_{\mathcal{B}}, \tag{14}$$

the momentum γ is set to 0.9, and $\bar{\mathcal{R}}$ is initialized to 0.

3.3. Combining for Preference Optimization

In this section, we present the combination of both data- and model-aware strategies for robust preference optimization. Specifically, given a batch of preference instances $\mathcal{B} = \{(\mathcal{I}_i, x_i, y_{w,i}, y_{l,i}) | i = 1, 2, \ldots, N\} \sim \mathcal{D}$, where $\mathcal{I}_i$, $x_i, y_{w,i}, y_{l,i}$ denotes the image, question, preferred response, and rejected response, respectively, we first compute data hardness offline based on our data-aware preference optimization, and then obtain the corresponding instance-wise hardness by Equ (6) as $\alpha_{\text{D}}^{\mathcal{B}} \in \mathbb{R}^N$. Next, we calculate the reward gaps using Equ 12 as α_{M}. To facilitate the optimization process to be both data- and model-aware, we propose an element-wise combination strategy. Specifically, we combine $\alpha_{\text{D}}^{\mathcal{B}}$ and α_{M} as:

$$\alpha = \alpha_{\text{D}}^{\mathcal{B}} \cdot \alpha_{\text{M}}, \tag{15}$$

where $\alpha \in \mathbb{R}^N$ represents the combined factor. Subsequently, we adjust β of Equ (2) to incorporate both components into optimization procedure as:

$$\beta_{\text{C}} = \beta \cdot \alpha, \tag{16}$$

where $\beta_{\text{C}} \in \mathbb{R}^N$, and each preference instance in $\mathcal{B}$ corresponds to a specific β. Finally, our combination strategy can be achieved by employing the β_{C} to optimize π_θ with the filtered batch $\mathcal{B} \cdot \boldsymbol{M}$, where $\boldsymbol{M}$ is obtained by Equ (10). Thus, the optimization process can become more adaptive, allowing the model to refine its preferences based on both pre-computed data hardness and real-time model responsiveness, further enhancing the robustness.

4. Experiment

In this section, we elaborate on the effectiveness of our **Da**ta- and **Mo**del-aware Direct Preference Optimization (DAMA). Specifically, we first introduce the details of our experimental settings. Next, we illustrate the ablation studies, and finally, we compare the results with the state-of-the-art methods over various benchmarks.

4.1. Experimental Settings

In this section, we describe the experimental settings.
Backbone: We employ the LLaVA-1.5 7B and 13B for performance comparison (Liu et al., 2023b).
Dataset: Our focus is not on the preference data construction, thus we directly utilize the released dataset by (Yu et al., 2024c), which contains 22k preference data totally.

Baselines. In this work, we compare against state-of-the-art baselines across various categories:
(1) Hallucination-specific baselines. In this category, we mainly compare with VCD (Leng et al., 2024), Less-is-more (Yue et al., 2024), OPERA (Huang et al., 2024), and CCA-LLaVA (Xing et al., 2024).
(2) Preference Optimization-based baselines. In this category, we mainly compare with HA-DPO (Zhao et al., 2023), POVID (Zhou et al., 2024a), LLaVA-RLHF (Sun et al., 2024), RLHF-V (Yu et al., 2024b), RLAIF-V (Yu et al., 2024c), AMP-MEG (Zhang et al., 2024), CSR (Zhou et al., 2024b), V-DPO (Xie et al., 2024), and TPO (Gu et al., 2024).
(3) Commercial baseline. We include GPT-4V as a strong reference to evaluate the performance gap between the open-source and commercial models.

Benchmarks. We conduct experiments on five benchmarks, including three hallucination benchmarks reflecting trustworthiness, and two general benchmarks:
(1) Object HalBench is employed to evaluate object hallucination by detailed descriptions of the image content. We report the response-level and mentioned-level **non**-hallucination rates to evaluate its capability to reduce hallucination (Rohrbach et al., 2018).
(2) AMBER is a multi-dimensional hallucination benchmark, which contains more than 15k samples. We report the Accuracy and F1 metric by its discriminative component (Wang et al., 2023).
(3) MMHal-Bench assesses response-level hallucination rate and informativeness by GPT-4 compare model outputs with human responses and object labels (Sun et al., 2024).
(4) LLaVA Bench consists of 24 images and 60 questions including conversation, detailed description, and complex reasoning ability (Liu et al., 2023b).
(5) MM-Vet is designed to evaluate six integrated competencies, including OCR, recognition, knowledge, language generation, spatial awareness, and math (Yu et al., 2024d).

Implementation Details. For both LLaVA-1.5 7B and 13B models, we employ full parameter-tuning over the preference dataset with four epochs. Specifically, for reproducibility, we adopt the same hyperparameters as provided in the official LLaVA GitHub repository [1]. The batch size N is set to 16, the selected size K is set to 12, and the penalty hyperparameter β is set to 0.1 by following (Rafailov et al., 2024; Yu et al., 2024c). All experiments are conducted with four A100 80GB GPUs, and four epochs of fine-tuning cost seven hours for both backbones.

4.2. Ablation Studies

In this section, we evaluate the effects of different components of our DAMA. To this end, we utilize the LLaVA-1.5 7B model as the backbone. For clear illustration, we report both the response and the mentioned-level non-hallucination rate on the Object Hallucination benchmark, where the non-hallucination rate is defined as $100\%-$ hallucination rate. The following are detailed illustrations.

- **Influences of different components.** The performance of various components of DAMA is reported in Table 1, where "DPO" refers to Direct Preference Optimization (Rafailov et al., 2024), "MDPO" represents our Model-aware Preference Optimization, "D^{2}PO" denotes Data-aware Preference Optimization, and "DAMA" corresponds to our combined strategy. The experimental results demonstrate: (1) All strategies significantly outperform the baseline method (DPO), especially our final "DAMA", achieving more than 10% response level performance gains, highlighting the effectiveness of our method; (2) Compared with "MDPO", the performance gain of "D^{2}PO" is relatively modest, suggesting that the quality of the preference data is already high, which further validates the efficacy of the preference data construction strategy (Yu et al., 2024c).

- **Influences of probability transformation in data-aware preference optimization.** Table 2 summarizes the performance of different response inconsistency estimation strategies, where "CLIP Scores" denotes that we directly estimate δ based on $\mathbf{C}_w$ and $\mathbf{C}_l$, with $\delta = \sum_{j=1}^{p} \mathbf{C}_{w,j} / \sum_{k=1}^{q} \mathbf{C}_{l,k}$, and "CLIP Probs" represents our strategy, which transforms $\mathbf{C}_w$ and $\mathbf{C}_l$ into probabilities. The results indicate that: firstly, both strategies improve the performance, underscoring the effectiveness of integrating data inconsistency into the optimization process, which allows the model to better

---

[1] https://github.com/haotian-liu/LLaVA

Table 1. Experimental results of different components of DAMA.

Method	Object HalBench	
	Response (↑)	Mention (↑)
LLaVA-1.5-7B	47.75	73.08
+DPO	78.29	89.48
+D^{2}PO	82.54+4.25	90.64+1.16
+MDPO	88.00+9.71	93.74+4.26
+DAMA	**90.87**+12.58	**95.33**+5.85

Table 2. Experimental results of different preference inconsistency construction strategies.

Method	Object HalBench	
	Response (↑)	Mention (↑)
DPO	78.29	89.48
+CLIP Scores	80.65+2.36	89.62+0.14
+CLIP Probs	**82.54**+4.25	**90.64**+1.16

Table 3. Experimental results of different filtering strategies.

Method	Object HalBench	
	Response (↑)	Mention (↑)
No Filter	86.66	92.62
Bottom	76.36-10.30	87.20-5.42
Top	86.34-0.32	92.48-0.14
Bottom & Top	**88.00**+1.34	**93.74**+1.12

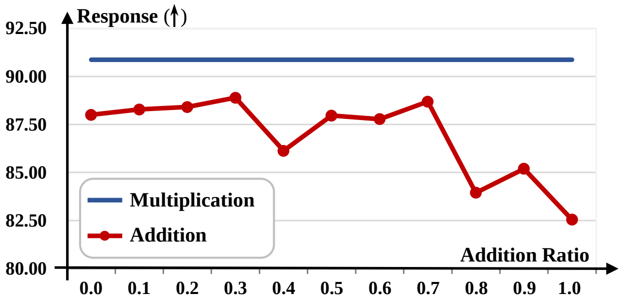


Figure 4. Experimental results of the combination strategies with the response-level non-hallucination rates.

handle varying levels of data hardness, thereby improving overall robustness; furthermore, transforming $\mathbf{C}w$ and $\mathbf{C}_l$ into probabilities yields a larger performance gain, as probabilities smooth the estimation of δ, mitigating noise from large gaps between $\mathbf{C}w$ and $\mathbf{C}_l$, and preventing the influences by the outliers.

• **Effects of the data filtering in model-aware preference optimization.** Table 3 presents the performance of different data filtering strategies, where “No Filter” denotes that we directly utilize the mean gaps to estimate the model state without filtering, “Bottom” shows that we remove the $N-K$ samples with the largest distances in the batch, “Top” is filtering the $N-K$ samples with the smallest distances, “Bottom & Top” refers to our filtering strategy, which filters both extremes based on the squared distances. Specifically, we can observe that: firstly, filtering solely from the bottom or top leads to performance degradation, indicating that such data introduces bias in estimating the model state. Moreover, filtering only the bottom samples results in significant performance drops due to overfitting on the top-ranked data, which misguides the estimation of model responsiveness to focus excessively on potentially less representative instances. Furthermore, filtering both bottom and top samples yields performance improvements, demonstrating the effectiveness of our proposed strategy, as it balances the influence of extreme data points.

• **Effects of the combination strategy.** Figure 4 and Figure 5 illustrate the performance of different combination strategies from the Response and Mention, respectively. “Multiplication” refers to combining both effects using a weighted sum strategy, where $\alpha = (1-\rho)\cdot\alpha_M + \rho\cdot\alpha_{\mathrm{D}}^{\mathcal{B}}$, with ρ ranging from 0.0 to 1.0, where $\rho = 0.0$ is the MDPO, and $\rho = 1.0$ refers to the “D^{2}PO”. We can observe that multiplication outperforms the weighted sum by a significant margin. This is because the weighted sum balances the two components without effectively merging their influences, whereas multiplication amplifies their contributions, allowing for more robust and superior performance gains.

• **Sensitivity to hyper-parameter.** Table 5 shows the effects of varying K, where we select K = 8, 10, 12, and 14 responses from a batch of 16, where no filter denotes that K is set to 16. We observe that the performance improves as K increases from 8 to 12, peaking at K = 12, and then slightly declines as K approaches 16. Notably, K = 8 performs worse than the baseline ($K = 16$), which can be attributed to the introduced bias by too few samples (50% samples). These suggest that moderate filtering—removing a small number of low-quality or outlier samples—can effectively improve performance. Based on this observation, we adopt K = 12 in our experiments.

4.3. Comparison with state-of-the-art methods

In this subsection, we compare our method with state-of-the-art methods under three trustworthy benchmarks: Object HalBench, AMBER, and MMHal-Bench, and two general benchmarks: LLaVA-Bench and MM-Vet. We compare our method against baselines from various aspects, including hallucination-specific baselines, preference optimization-based baselines, and the commercial baseline GPT-4V. Our DAMA is adapted to both the LLaVA-1.5 7B and 13B models. The experimental results, along with results of vanilla DPO and the improvements achieved by DAMA over DPO,

Table 4. Performance comparisons with state-of-the-art methods on different benchmarks. We report non-hallucination rates in different levels including response level (Non-Rsp.) and mentioned-level (Non-Men.) for Object HalBench (Rohrbach et al., 2018). Hall. refers to the Hallucination Rate for MMHal Bench (Sun et al., 2024). The best results of all methods are indicated in **bold**, and the second best results are underlined. The compared results are sourced from (Yu et al., 2024c; Gu et al., 2024), and the reported results of LLaVA-1.5, DPO, and DAMA are evaluated using GPT-4-turbo-2024-04-09.

Method	Size	Object HalBench		AMBER		MMHal-Bench		LLaVA Bench(↑)	MM-Vet(↑)
		Non-Rsp.(↑)	Non-Men.(↑)	Acc(↑)	F1(↑)	Scores(↑)	Hall.(↓)		
Method (Hallucination-specific)									
VCD (CVPR'24)	7B	51.2	75.7	71.8	74.9	2.12	54.2	61.6	-
Less-is-more (ACL'24)	7B	59.7	82.2	72.4	75.8	2.33	50.0	-	-
OPERA (CVPR'24)	7B	54.9	77.7	75.2	78.3	2.15	54.2	61.3	-
CCA-LLaVA (NIPS'24)	7B	53.3	76.2	77.7	81.9	1.92	61.5	64.3	-
Method (Preference optimization)									
HA-DPO (arXiv'23)	7B	60.1	80.1	75.2	79.9	1.98	60.4	-	-
POVID (arXiv'24)	7B	51.9	75.6	82.9	87.4	2.08	56.2	68.2	31.7
RLHF-V (CVPR'24)	7B	-	-	74.8	78.5	2.02	60.4	68.0	32.3
RLAIF-V (arXiv'24)	7B	89.5	94.8	76.8	84.5	2.95	**32.3**	-	-
CSR (NIPS'24)	7B	-	-	73.2	76.1	2.05	60.4	68.9	31.0
V-DPO (EMNLP'24)	7B	-	-	-	81.6	2.16	56.0	-	-
TPO (arXiv'24)	7B	-	-	79.3	85.0	2.47	51.0	70.2	33.0
TPO (arXiv'24)	13B	-	-	83.9	88.0	2.72	45.8	72.8	36.2
LLaVA-RLHF (ACL'24)	13B	61.9	81.1	79.7	83.9	2.02	62.5	**95.6**	-
RLHF-V (CVPR'24)	13B	87.8	92.5	72.6	75.0	2.45	51.0	76.7	**38.5**
AMP-MEG (NIPS'24)	13B	68.3	79.4	79.5	84.6	**3.08**	36.5	-	-
LLaVA-1.5	7B	47.8	71.2	73.9	77.7	1.95	63.5	62.3	31.6
+ DPO	7B	78.3	89.5	75.5	79.2	2.15	57.0	64.6	30.4
+ DAMA	7B	**90.9**	**95.3**	83.3	87.0	2.76	41.0	68.0	32.8
Improvements (%)		**+16.1%**	**+6.5%**	**+10.3%**	**+9.8%**	**+28.4%**	**+28.1%**	**+5.3%**	**+7.9%**
LLaVA-1.5	13B	50.0	76.4	71.2	73.0	2.36	56.0	66.1	36.1
+ DPO	13B	84.5	92.4	78.5	83.5	2.59	48.0	74.0	35.4
+ DAMA	13B	89.1	94.4	**84.3**	**88.1**	2.89	43.0	75.1	36.4
Improvements (%)		**+5.1 %**	**+2.1 %**	**+ 5.8%**	**+ 4.6%**	**+15.4%**	**+10.4%**	**+1.4 %**	**+2.8 %**
GPT-4V	-	86.4	92.7	83.4	87.4	3.49	28.1	98.0	67.7

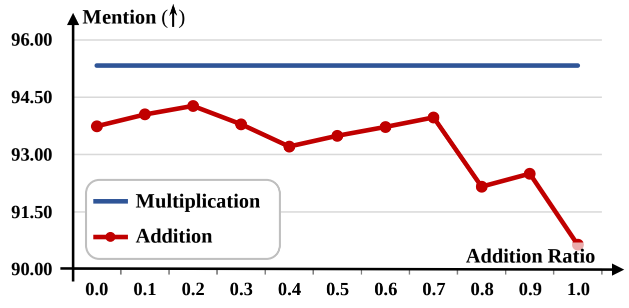


Figure 5. Experimental results of the combination strategies with the mentioned-level non-hallucination rates.

Table 5. Experimental results of different selected number(K), with the batch size N=16.

Method	Object HalBench	
	Response (↑)	Mention (↑)
No Filter	86.66	92.62
K = 8	85.34-1.32	92.25-0.37
K = 10	87.50+0.84	93.19+0.57
K = 14	87.96+1.30	93.45+0.83
K = 12	**88.00**+1.34	**93.74**+1.12

are listed in Table 4.

From the experimental results, we can observe that: (1) Compared to DPO, our DAMA achieves substantial performance improvements over all compared benchmarks. It's noteworthy that on MMHal-Bench, it achieves more than 10% improvements for the 13B model, and more than 28% gain for 7B models, which is significant; (2) DAMA achieves new state-of-the-art over various benchmarks. Our DAMA-7B reduces the response-level and mentioned-level hallucination on the generative Object HalBench by 90.9% and 95.3%. Moreover, DAMA-13B achieves 84.3% Accuracy and 88.1 F1 score on the discriminative AMBER benchmark. These attained results surpass those of GPT-4V, fully demonstrating the effectiveness of DAMA.

5. Related Work

In this section, we first briefly summarize the hallucination mitigation methods for multi-modal large language models, and then, we illustrate related methods for aligning with human preferences. Finally, we enumerate the differences between ours and related methods.

5.1. Hallucination mitigation in MLLMs

Hallucination, as a key indicator of trustworthiness, refers to that the MLLM outputs are not aligned with the image content (Liu et al., 2024b). Current mitigation strategies generally fall into the following four key areas:

- **Data cleaning.** Works such as LRV-Instruction (Liu et al., 2023a) and Hallucidoctor (Yu et al., 2024a) identify noise within instruction tuning data as a cause of hallucination and propose fine-grained data cleaning and curation strategies.
- **Visual representation enhancement.** Methods like (Jain et al., 2024; Tong et al., 2024) suggest that insufficient visual cues as a cause of hallucination, and they incorporate more intricate visual features to enrich the visual representations.
- **Inference-time enhancement.** Works like VCD (Leng et al., 2024) and MARINE (Zhao et al., 2024) employ visual contrastive decoding mechanisms to enhance the model's focus on visual details, often by contrastively sampling from the original and visually distorted distribution.
- **Preference Optimization.** Inspired by the great success of preference optimization in language models (Rafailov et al., 2024; Shao et al., 2024), recent works utilizing preference optimization have demonstrated significant performance in hallucination mitigation (Lu et al., 2025; Yu et al., 2024b;c). They achieve this by designing strategies to prioritize the preferred responses (better aligned with visual content) over the rejected responses (containing more hallucinated content).

5.2. Preference Optimization

To align model outputs with human preferences, preference optimization methods have garnered significant attention (Ouyang et al., 2022; Rafailov et al., 2024; Wu et al., 2024b). As a pioneering work, RLHF (Ouyang et al., 2022) firstly trains a reward model on preference data and then optimizes the policy with PPO (Schulman et al., 2017). However, obtaining an effective reward model remains challenging. To address this, DPO (Rafailov et al., 2024) simplifies the process with an implicit reward function, allowing for direct optimization on preference data. β-DPO (Wu et al., 2024b) further enhances DPO by introducing dynamic β to more effectively leverage preference data of varying quality.

In MLLMs, current efforts typically focus on two aspects:

- **Collecting high-quality preference data.** Existing works employ diverse strategies for preference construction. These include human-based methods (Yu et al., 2024b; Sun et al., 2024), and large language model-based strategies (Yu et al., 2024c; Li et al., 2024). Among the latter, RLAIF-V (Yu et al., 2024c) utilizes open-source models (e.g., LLaVA-NeXT (Liu et al., 2024a)), while works like (Li et al., 2024; Zhang et al., 2025) employ closed-source models such as GPT-4V (Achiam et al., 2023) for annotation.

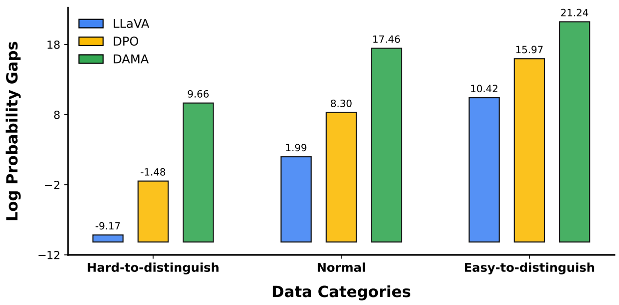


Figure 6. Log probability gaps of different models (LLaVA-1.5-7B, DPO, and DAMA) between "easy-to-distinguish" (top 25% image-text similarity), "hard-to-distinguish" (bottom 25% image-text similarity), and "normal" data (middle 50%).

- **Emphasizing visual detail.** Recent approaches that construct vision-based rejected samples via image manipulation include (Xie et al., 2024; Wang et al., 2024; Lu et al., 2025) (e.g., object removal or replacement (Lu et al., 2025)). Additionally, works like (Gu et al., 2024; Cui et al., 2025) identify and emphasize key language tokens from preference responses to better attend to visual details.

Differences. Compared with existing MLLM preference optimization strategies, which focus on data curation and introduce fine-grained regularizations, our work aims to design an improved optimization strategy that can effectively leverage model responses and preference data, enabling the model to adaptively respond to the data, and thereby enhancing the alignment performance.

6. Discussion and Future Work

6.1. Discussion

Although DPO has demonstrated its effectiveness in mitigating hallucination issues in multi-modal LLMs (Yu et al., 2024b;c), it inherently treats all training samples equally. As illustrated in Figure 6, this tends to bias the model toward "easy-to-distinguish" data—those with larger log probability gaps—while neglecting the more challenging "hard-to-distinguish" data, which often exhibit negative log probability gaps. This imbalance constrains the model's expressive capacity and generalization ability.

To address this limitation, we introduce DAMA, a data- and model-aware strategy that enhances the model's sensitivity to data hardness. As shown in Figure 6, by explicitly accounting for data hardness and model responses, DAMA achieves consistently higher log probability gaps, with notable gains on "hard-to-distinguish". This leads to a more balanced optimization process and substantially enhances the model's overall reliability.

6.2. Future Work

While DAMA marks a significant advancement towards adaptive and robust MLLM alignment, several promising avenues remain for future exploration:
(1) Adaptive strategies in online scenarios. DAMA currently operates on offline data; future work could investigate its application in online settings, leveraging approaches like GRPO (Huang et al., 2025) to enable continuous model refinement with evolving preference data.
(2) Advanced hardness and responsiveness estimations. Our current estimation of data hardness (CLIP-based similarity (Radford et al., 2021)) and model responses (reward gaps) is simple. Future research could explore more sophisticated metrics, *e.g.* semantic entropy (Farquhar et al., 2024), fine-grained visual representations (Tong et al., 2024).
(3) Scalability to broader architectures and modalities. A crucial next step involves assessing DAMA's efficacy and scalability on commercial MLLMs (Google, 2025; OpenAI, 2025), diverse multi-modal architectures (*e.g.*, Omni model (Xu et al., 2025), Unified model (Deng et al., 2025)), and reasoning models (Kimi et al., 2025), which promises to broaden its practical impact significantly.

7. Conclusion

This paper introduced **Da**ta- and **M**odel-**a**ware Direct Preference Optimization (DAMA), a simple yet effective strategy to address the imbalanced responsiveness to varying data hardness in Multimodal Large Language Models (MLLMs). DAMA adaptively adjusts the model's learning behavior through two key strategies: (1) a data-aware strategy that dynamically incorporates data hardness, and (2) a model-aware strategy that integrates the model's real-time responses. Experiments across five benchmarks and various model sizes consistently demonstrated DAMA's effectiveness.

Impact Statement

This paper aims to enhance the robustness of multi-modal large language model alignment, thereby advancing the field of visual understanding. By tackling challenges like hallucination, we strive to improve the reliability of AI systems in alignment with human preferences. We hope that our research will contribute to facilitating the development of AI systems that are both effective and reliable, ultimately delivering meaningful societal benefits.

Acknowledgment

This paper is supported by the National Science and Technology Major Project (2023ZD0121102), and the National Natural Science Foundation of China (U24B20180, 62121002).

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Cui, C., Zhang, A., Zhou, Y., Chen, Z., Deng, G., Yao, H., and Chua, T.-S. Fine-grained verifiers: Preference modeling as next-token prediction in vision-language alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=cJQ1K2fjpD`.

Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., and Fan, H. Emerging properties in unified multimodal pretraining, 2025. URL `https://arxiv.org/abs/2505.14683`.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Google. Gemini 2.5: Our most intelligent ai model. Large language model, 2025. URL `https://deepmind.google/models/gemini/`.

Gu, J., Wang, Y., Cao, M., Bu, P., Song, J., He, Y., Li, S., and Zheng, B. Token preference optimization with self-calibrated visual-anchored rewards for hallucination mitigation. *arXiv preprint arXiv:2412.14487*, 2024.

Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., and Yu, N. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., and Lin, S. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025. URL `https://arxiv.org/abs/2503.06749`.

Jain, J., Yang, J., and Shi, H. Vcoder: Versatile vision encoders for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27992–28002, 2024.

Kimi et al. Kimi k1.5: Scaling reinforcement learning with llms, 2025. URL `https://arxiv.org/abs/2501.12599`.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., and Bing, L. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Li, L., Xie, Z., Li, M., Chen, S., Wang, P., Chen, L., Yang, Y., Wang, B., Kong, L., and Liu, Q. Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6227–6246, 2024.

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 2023b. URL `https://openreview.net/forum?id=w0H2xGHlkw`.

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llava-next: Improved reasoning, ocr, and world knowledge, 2024a.

Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., and Peng, W. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024b.

Lu, J., Li, J., Gao, Y., Wu, J., Wu, J., Wang, X., and He, X. Adavip: Aligning multi-modal llms via adaptive vision-enhanced preference optimization. *arXiv preprint arXiv:2504.15619*, 2025.

OpenAI. Introducing openai o3 and o4-mini. Large language model, 2025. URL `https://openai.com/index/introducing-o3-and-o4-mini/`.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL `https://arxiv.org/abs/2402.03300`.

Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.-Y., Wang, Y.-X., Yang, Y., et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics*, 2024.

Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., and Xie, S. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Wang, F., Zhou, W., Huang, J. Y., Xu, N., Zhang, S., Poon, H., and Chen, M. mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088, 2024.

Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Yan, M., Zhang, J., and Sang, J. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Wu, J., Wang, X., Yang, Z., Wu, J., Gao, J., Ding, B., Wang, X., and He, X. α-dpo: Adaptive reward margin is what direct preference optimization needs. *arXiv preprint arXiv:2410.10148*, 2024a.

Wu, J., Xie, Y., Yang, Z., Wu, J., Gao, J., Ding, B., Wang, X., and He, X. β-dpo: Direct preference optimization with dynamic β. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL `https://openreview.net/forum?id=ZfBuhzE556`.

Xie, Y., Li, G., Xu, X., and Kan, M.-Y. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 13258–13273, 2024.

Xing, Y., Li, Y., Laptev, I., and Lu, S. Mitigating object hallucination via concentric causal attention. *Advances in neural information processing systems*, 2024.

Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., Zhang, B., Wang, X., Chu, Y., and Lin, J. Qwen2.5-omni technical report, 2025. URL `https://arxiv.org/abs/2503.20215`.

Yu, Q., Li, J., Wei, L., Pang, L., Ye, W., Qin, B., Tang, S., Tian, Q., and Zhuang, Y. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12944–12953, 2024a.

Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Sun, M., 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, 2024b.

Yu, T., Zhang, H., Yao, Y., Dang, Y., Chen, D., Lu, X., Cui, G., He, T., Liu, Z., Chua, T.-S., et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024c.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International Conference on Machine Learning*, pp. 57730–57754. PMLR, 2024d.

Yue, Z., Zhang, L., and Jin, Q. Less is more: Mitigating multimodal hallucination from an eos decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 11766–11781, 2024.

Zhang, M., Wu, W., Lu, Y., Song, Y., Rong, K., Yao, H., Zhao, J., Liu, F., Sun, Y., Feng, H., et al. Automated multi-level preference for mllms. *Advances in Neural Information Processing Systems*, 2024.

Zhang, R., Wei, X., Jiang, D., Guo, Z., Li, S., Zhang, Y., Tong, C., Liu, J., Zhou, A., Wei, B., et al. Mavis: Mathematical visual instruction tuning with an automatic data engine. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=MnJzJ2gvuf`.

Zhao, L., Deng, Y., Zhang, W., and Gu, Q. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv preprint arXiv:2402.08680*, 2024.

Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., and He, C. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

Zhou, Y., Cui, C., Rafailov, R., Finn, C., and Yao, H. Aligning modalities in vision large language models via preference fine-tuning. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024a. URL `https://openreview.net/forum?id=GRGvC0rpA8`.

Zhou, Y., Fan, Z., Cheng, D., Yang, S., Chen, Z., Cui, C., Wang, X., Li, Y., Zhang, L., and Yao, H. Calibrated self-rewarding vision language models. In *Advances in Neural Information Processing Systems*, 2024b. URL `https://openreview.net/forum?id=nXYedmTf1T`.

A. Appendix

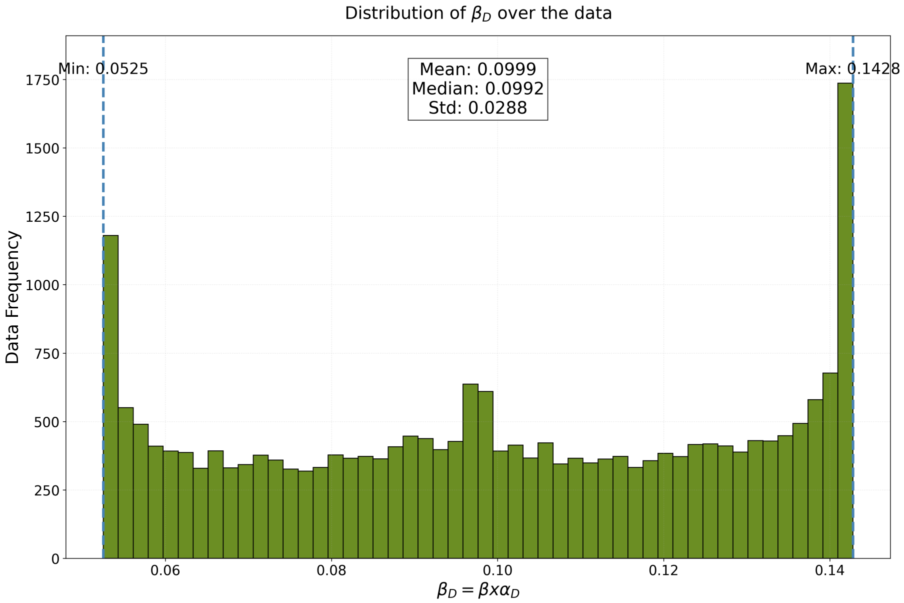


Figure 7. Dynamic data-aware β_D over each data point, we also show the mean, std, min, and max value.

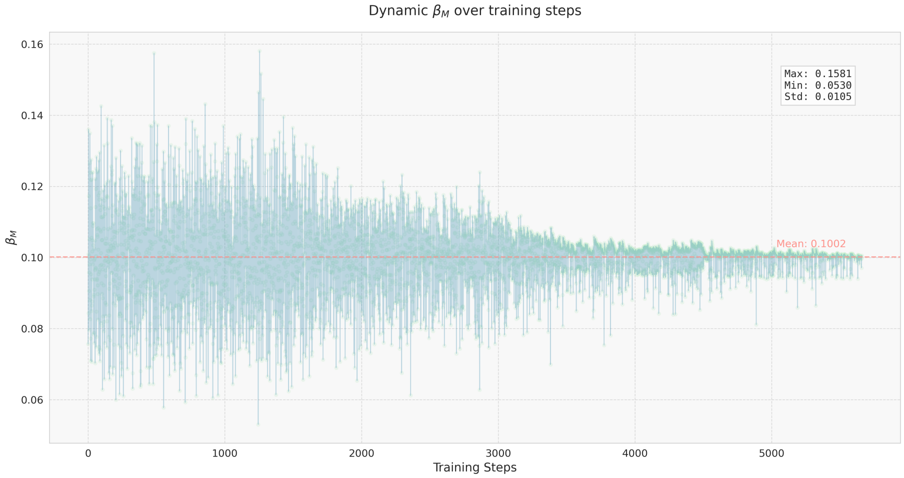


Figure 8. Dynamic model-aware β_M over each training step, we also show the mean, std, min, and max value.

A.1. Prompt templates to split the sentences

In this section, we introduce the prompt template to split the sentences into sub-sentences with LLaMA3 (Dubey et al., 2024). The prompt is: *You are an expert in extracting facts from the given question-answer pair about an image. Your task is to: Analyze the provided question-answer pair based on the image, extract all factual statements from the answer, and rewrite them into self-contained sentences. Requirements for each sentence are: 1. complete, each sentence must be self-contained; 2. factual (omit opinions, subjective statements); 3. concise (no more than 77 tokens). Format your result strictly as:### Facts:- Fact 1 (e.g., "A red shoe sits on a wooden floor.")- Fact 2 (e.g., "The shoe has laces and a white sole.")- ...### Question-answer pair: Question: "{Question}" Answer: "{Answer}"*

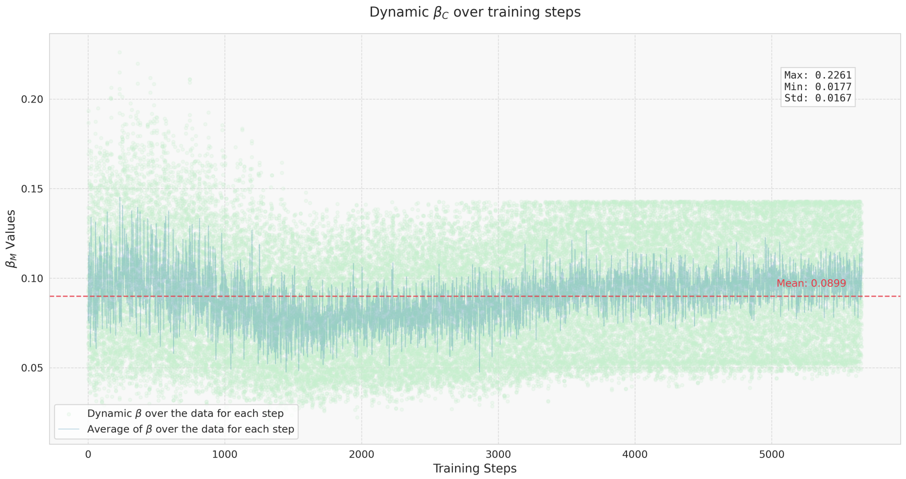


Figure 9. Dynamic β_{C} by combining the data- and model-aware strategies, we also show the mean, std, min, and max value.

A.2. Dynamic β with different strategies.

In this section, we analyze the dynamic β with our data- and model-aware strategies.

- Data-aware strategy (Figure 7,): The range of data-aware β_{D} falls within (0.0524, 0.1428) with the original β initialized as 0.1. Moreover, the mean value is 0.0999 with a standard deviation of 0.0288. These observations demonstrate that maintains proximity to the original, while adaptively adjusting based on the data characteristics, enabling a more effective capture of the data.
- Model-aware strategy (Figure 8): as the training progresses, the model-aware β_{M} gradually converges to the original β. Meanwhile, we find that the β_{M} fluctuates within a moderate range of (0.0530, 0.1580), demonstrating controlled adaptivity. These observations suggest that as the model training stabilizes, its responsiveness becomes more consistent and eventually approaches a steady state.
- Combined strategy (Figure 9): Combining both the data- and model-aware strategies yields a more dynamic range of combined β_{C}, spanning from 0.0177 to 0.2261, which is wider than either. Moreover, while the value eventually stabilizes around 0.1, we notice that the mean value during the training stage is a bit lower than β_{M}. This suggests that it relaxes the constraints during training based on the data hardness, enabling the model to better capture fine-grained data patterns and thereby adaptively enhancing its responsiveness to data characteristics.

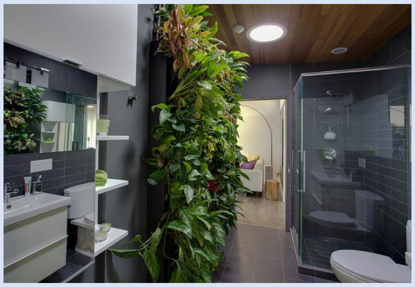
A.3. Qualitative analysis

In this section, we provide qualitative analysis between our Data- and Model-aware preference optimization (DAMA) and the DPO method (Rafailov et al., 2024). The case studies are shown in Figure 10 and Figure 11, and we also include the evaluations of GPT-4. From the case studies, we can observe two key points: (1) while both methods generate correct outputs, DPO tends to be more general, whereas DAMA offers richer, more detailed descriptions. This can be attributed to the integration of data hardness, quantified by the CLIP score between the Image and responses, into the optimization process, enabling DAMA to capture and describe more nuanced visual details; and (2) as evidenced by the GPT-4 evaluation, DAMA generates more vivid and immersive visual descriptions, this can be attributed to the enhancement of the responsiveness to the visual content, resulting in a more expressive and contextually aligned representation of the scene.

Algorithm 1 Algorithm of DAMA.

```
Input: Preference dataset $\mathcal{D}$, hyper-parameter $\beta$, SFT model $\pi_{\text{SFT}}$, CLIP classifier $\Gamma_{\text{CLIP}}$.
Output: The optimized model $\pi_{\boldsymbol{\theta}}$.
Initialize model $\pi_{\boldsymbol{\theta}}$ and reference model $\pi_{\text{ref}}$ as $\pi_{\text{SFT}}$.
for $\{(\mathcal{I}, x, y_w, y_l)\}$ in $\mathcal{D}$ do
    $\mathbf{S}_w \leftarrow \text{LLM}\{y_w\}$, $\mathbf{S}_l \leftarrow \text{LLM}\{y_l\}$;
    obtains $\delta$ with $\{\mathcal{I}, \mathbf{S}_w\}$, $\{\mathcal{I}, \mathbf{S}_l\}$;        ▷ Equ (3) → (5);
    $\alpha_{\text{D}} \leftarrow \sigma(\delta)/\sigma(\bar{\delta})$;        ▷ Equ (6);
end for
repeat
    for $\mathcal{B} = \{(\mathcal{I}_i, x_i, y_{w,i}, y_{l,i})\}_{i=1}^{N} \sim \mathcal{D}$ do
        obtain $\mathcal{R}_i$ with $y_{w,i}$ and $y_{l,i}$;        ▷ Equ (8);
        obtain $\bar{\mathcal{R}}_{\mathcal{B}}$ with $\mathcal{R}_i$;        ▷ Equ (9) → (11);
        $\alpha_M \leftarrow \sigma(\bar{\mathcal{R}}_{\mathcal{B}})/\sigma(\bar{\mathcal{R}})$;        ▷ Equ (12);
        $\alpha \leftarrow \alpha_{\text{D}}^{\mathcal{B}} \cdot \alpha_M$, where $\alpha_{\text{D}}^{\mathcal{B}} = \{\alpha_{\text{D,i}}\}_{i=1}^{N}$;        ▷ Equ (15);
        $\beta_{\text{C}} \leftarrow \beta \cdot \alpha$;        ▷ Equ (16);
        Compute loss w.r.t. $\beta_{\text{C}}, \pi_{\boldsymbol{\theta}}$;        ▷ Equ (2);
        Compute the gradient and update the model $\pi_{\boldsymbol{\theta}}$.
        $\bar{\mathcal{R}} \leftarrow \gamma \cdot \bar{\mathcal{R}} + (1 - \gamma) \cdot \bar{\mathcal{R}}_{\mathcal{B}}$;        ▷ Equ (14);
    end for
until The optimization is converged.
```

Q: Provide a thorough description of the given image.



DPO: The image features a bathroom with a unique and eye-catching element - **a large green plant climbing up the wall**. The plant adds a touch of nature to the space, making it feel more alive and vibrant.

In the bathroom, there is a **toilet** located towards the right side of the image. A **shelf** can also be seen in the room, possibly for storing bathroom essentials. The presence of the plant and the shelf contribute to the overall cozy and inviting atmosphere of the bathroom.

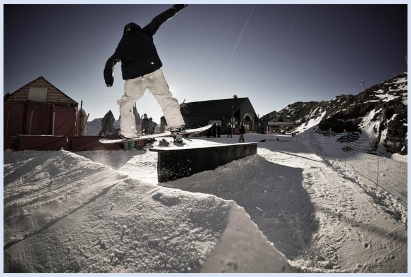
DAMA: The image features a bathroom with a unique and eye-catching element - **a large green plant climbing up the wall**. The plant **covers a significant portion of the wall**, adding a touch of nature to the space. In the bathroom, there is **a toilet** located towards the right side of the image. **A potted plant** can also be seen in the room, further enhancing the greenery in the space. The bathroom appears to **have a modern design, giving it a contemporary feel**.

Prompt: Given two different descriptions of a given image, with the question "Provide a thorough description of the given image", tell me which one you like, and show me reasons.

GPT-4: **DAMA is more effective due to the richness in its description.** It not only describes the plant's effect but gives a clearer picture of its scale and placement, **creating a more vivid visual.** The added mention of another potted plant and the modern design **gives a stronger sense of the room's style and enhances the overall atmosphere.** These details would likely help someone visualize the space better.

Figure 10. Qualitative Comparisons between DAMA and vanilla DPO. Data source: Object HalBench.

Q: What is this photo about? Please answer in great detail.



DPO: In the image, **a person is performing a trick on a snowboard** in the snow. They are in the middle of the action, showcasing skills. The snowboarder is the main focus of the scene, and their snowboard is prominently visible beneath them.
The background features **a building**, adding to the wintery atmosphere of the scene. The snowboarder's impressive trick and the presence of the building create a dynamic and exciting image.

DAMA: The image features **a snowboarder performing a trick in the snow**, with the snowboard in mid-air near a **ramp**. The snowboarder is the main focus of the scene, showcasing skills.
There are **several other people and a building** present in the background, **possibly watching the snowboarder's performance or engaging in their own snowboarding activities**. The overall atmosphere of the image conveys an exciting and dynamic winter sports scene.

Prompt: Given two different descriptions of an given image, with the question "What is this photo about? Please answer in great detail", tell me which one you like, and show me reasons.

GPT-4: **DAMA provides a more comprehensive, detailed, and immersive portrayal of the image.** It not only highlights the snowboarder's skill and trick but also **fleshes out the background, giving a sense of the environment and the people around**. The **extra information** about the ramp, the other people, and the overall excitement makes the scene feel more dynamic and real, which is why I prefer it over DPO.

Figure 11. Qualitative Comparisons between DAMA and vanilla DPO. Data source: Object HalBench.