M-RLHF: The Next Step Forward in Multimodal LLM Alignment

Yi-Fan Zhang[2,♠], Tao Yu[2], Haochen Tian[2], Chaoyou Fu[3,†]
Peiyan Li[2], Jianshu Zeng[5], Wulin Xie[2], Yang Shi[5], Huanyu Zhang[2], Junkang Wu[4]
Xue Wang[6], Yibo Hu[2], Bin Wen[1,†], Fan Yang[1], Zhang Zhang[2,†], Tingting Gao[1]
Di Zhang[1], Liang Wang[2], Rong Jin[7], Tieniu Tan[2,3]
[1]KuaiShou, [2]CASIA, [3]NJU, [4]USTC, [5]PKU, [6]Alibaba, [7]Meta AI
[♠] Project Leader [†] Corresponding Author

https://mm-rlhf.github.io/

Abstract

Despite notable advancements in Multimodal Large Language Models (MLLMs), most state-of-the-art models have not undergone thorough alignment with human preferences. This gap exists because current alignment research has primarily achieved progress in specific areas (e.g., hallucination reduction), while the broader question of whether aligning models with human preferences can systematically enhance MLLM capability remains largely unexplored. To this end, we introduce MM-RLHF, a dataset containing **120k** fine-grained, human-annotated preference comparison pairs. This dataset represents a substantial advancement over existing resources, offering superior size, diversity, annotation granularity, and quality. Leveraging this dataset, we propose several key innovations to improve both the quality of reward models and the efficiency of alignment algorithms. Notably, we introduce a **Critique-Based Reward Model**, which generates critiques of model outputs before assigning scores, offering enhanced interpretability and more informative feedback compared to traditional scalar reward mechanisms. Additionally, we propose **Dynamic Reward Scaling**, a method that adjusts the loss weight of each sample according to the reward signal, thereby optimizing the use of high-quality comparison pairs. Our approach is rigorously evaluated across **10** distinct dimensions and **27** benchmarks, with results demonstrating significant and consistent improvements in performance. Specifically, fine-tuning LLaVA-ov-7B with MM-RLHF and our alignment algorithm leads to a **19.5%** increase in conversational abilities and a **60%** improvement in safety.

1 Introduction

Although Multimodal Large Language Models (MLLMs) have demonstrated remarkable potential in addressing complex tasks that involve the integration of vision, language, and audio, state-of-the-art models today seldom undergo a rigorous alignment stage [ ]. Typically, these models only progress to the Supervised Fine-tuning (SFT) stage, leaving critical aspects such as truthfulness, safety, and alignment with human preferences largely unaddressed. While recent efforts have begun to explore MLLM alignment, they often focus on specific domains, such as mitigating hallucination or enhancing conversational capabilities, which fail to comprehensively improve the model's overall performance and reliability. This raises a critical question:

Is alignment with human preferences only capable of enhancing MLLMs in a limited set of tasks?

In this work, we confidently answer this question with a resounding "No.". We demonstrate that a well-designed alignment pipeline can comprehensively enhance MLLMs along multiple dimen-

Email: yifanzhang.cs@gmail.com; Work done during an internship at KuaiShou Group.

sions, including visual perception, reasoning, dialogue, and trustworthiness, thereby significantly broadening their practical applicability. chieve this, we conduct in-depth investigations into three pivotal areas: data curation, reward modeling, and alignment algorithms.

rst, we introduce **MM-RLHF**, a dataset designed to advance **M**ultimodal **R**einforcement **L**earning from **H**uman **F**eedback (RLHF). dataset spans three domains: image, video understanding, and MLLM safety. structed through a rigorous pipeline, MM-RLHF ensures high-quality, fine-grained annotations. set creation process involves the following steps (Figure 1):

- **a Collection**. We curate a diverse set of multimodal tasks from various sources, totaling 10 million data instances, ensuring broad representation across tasks.
- **Selection**. Through rigorous re-sampling, we extract 30k representative queries, ensuring diversity across a wide range of data types, such as real-world scenarios, mathematical reasoning, chart understanding, and other practical domains (Figure 2).
- **del Response Generation**. We utilize state-of-the-art models, such as Claude 3.5-Sonnet and Qwen2-VL-72B, to generate responses for various tasks.
- **e-grained Human Annotation**. We employ a meticulous annotation process, involving over 50 annotators over two months, to score, rank, and provide textual explanations for responses. This results in more than 120k high-quality ranked comparison pairs.

pared to existing datasets, MM-RLHF significantly advances in diversity, response quality, and annotation granularity, providing a robust foundation for MLLM alignment.

ding on the MM-RLHF dataset, we investigate how human-annotated data can enhance MLLM alignment, with a focus on reward modeling and training optimization. ognizing the pivotal role of reward models in providing feedback signals to guide the alignment process, we propose a **Critique-Based Reward Model** (Figure 3). itional reward models, which output scalar values, often lack interpretability, while directly using MLLMs as reward models place high demands on their instruction-following capabilities, limiting their practicality. ddress these limitations, we first transform concise human annotations into detailed, model-friendly formats using MLLMs. se enriched annotations serve as learning targets, guiding the reward model to first generate critiques and then assign scores based on the critiques. approach enables the model to provide fine-grained scoring explanations, significantly enhancing the quality and interpretability of the reward signals. **I-RLHF-Reward-7B** achieves SOTA performance on several reward model benchmarks, outperforming several 72B-scale models.

ding on this high-quality reward model, we introduce **Dynamic Reward Scaling** within the Direct Preference Optimization (DPO) framework. itional DPO methods use a fixed training weight for all human-preferred and non-preferred training pairs. ontrast, Dynamic Reward Scaling calculates a reward margin for each comparison pair using MM-RLHF-Reward-7B. ing training, it assigns higher weights to comparison pairs with larger reward margins. ensures that the most informative samples have a stronger influence on model updates. a result, the training process becomes more efficient, leading to improved model performance.

lly, to rigorously evaluate our approach, we construct two specialized benchmarks. first, **MM-RLHF-RewardBench**, is sampled from our dataset and consists of meticulously human-annotated data for evaluating reward models. second, **MM-RLHF-SafetyBench**, is curated and filtered from existing benchmarks and focuses on safety-related tasks, including privacy protection, adversarial attacks, jailbreaking, and harmful content detection.

onduct extensive evaluations across ten key dimensions, covering 27 benchmarks. results demonstrate that our training algorithm, combined with the high-quality MM-RLHF dataset, leads to significant improvements in model performance. ifically, models fine-tuned with our approach achieve an average 11% gain in conversational abilities and a 57% reduction in unsafe behavior. integration of our reward model further amplifies these gains, highlighting the effectiveness of our alignment algorithm.

2 MM-RLHF-Dataset

is section, we outline the construction of MM-RLHF, as illustrated in Figure 1. This includes the data collection process, data filtering methods, and human annotation procedures.

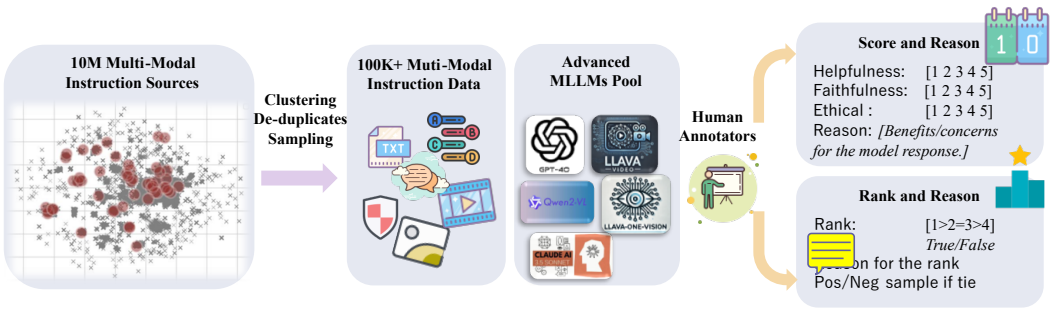


Figure 1: **MM-RLHF Construction Pipeline**. (*Data Collection and Cleaning*: Starting with 10 million instruction samples, we cluster data based on image similarity, and uniformly sample across diverse categories. This results in a diverse dataset covering image-based Q&A (e.g., multiple-choice, dialogues, and safety-related questions) and video Q&A formats. (*Response Generation*: We leverage state-of-the-art models, including GPT-4o and Qwen2-VL-72B, to generate high-quality responses. (*Human Annotation*: We conduct manual annotation across nine categories, including scoring, ranking, and explanations, ensuring fine-grained evaluation.

2.1 Data Collection

(goal is to construct a comprehensive post-training dataset that covers a wide range of task types. chieve this, we categorize tasks into three main domains: image understanding, video understanding, and multimodal safety.

image understanding, we integrate data from multiple sources, including LLaVA-OV[1], VLfeedback LLaVA-RLHF lrv-instruction and Unimm-Chat[2]. e some datasets contain multi-turn dialogues, which are less suitable for response generation, we decompose them into single-turn dialogues. s process yields over 10 million dialogue samples, covering tasks such as conversation, safety, multiple-choice questions, captions, and commonsense reasoning.

video understanding, the primary data source is SharedGPT-4 video

afety, data is primarily derived from VLGuard and self-constructed content. uard contains over 2,000 harmful samples, while additional red teaming, safety, and robustness data are included. pipeline for constructing safety data is detailed in the Appendix C.1.

2.2 Filtering and Model Response Generation

core goal of data filtering is to reduce the number of samples while maintaining the diversity of the original dataset. To achieve this, the following strategies are adopted:

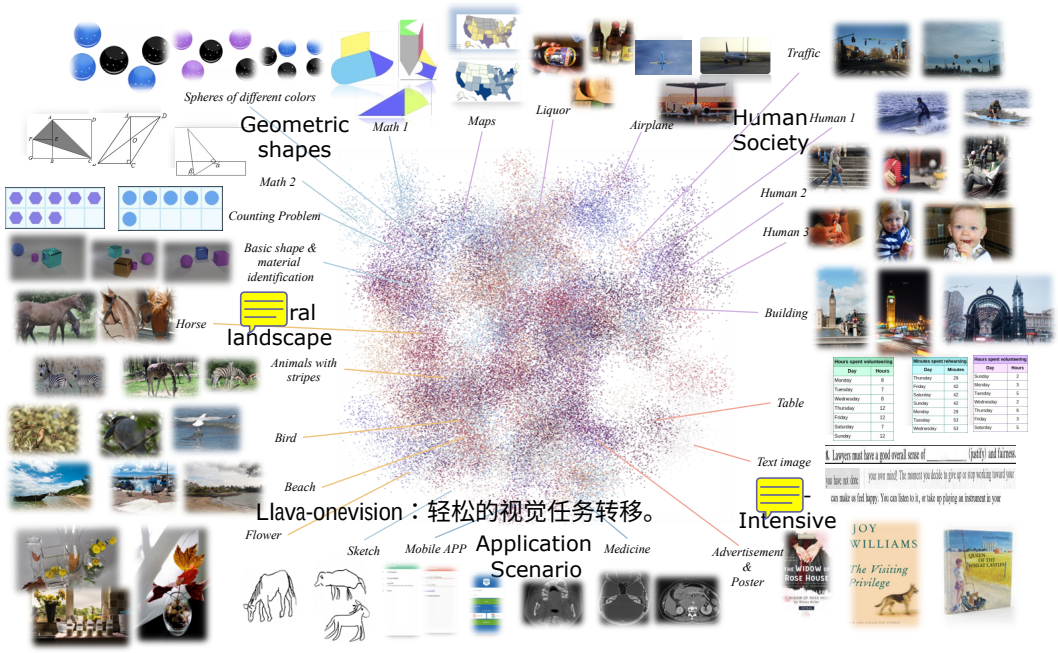
Predefined sampling weights. image understanding tasks, we define three categories based on the nature of the questions and the length of model responses: *ultiple-choice questions (MCQ)*; *(Questions with options such as A, B, C, or D.)* These tasks include visual question answering, mathematics, OCR, and icon recognition, focusing on the model's reasoning and visual perception abilities. *ong-text questions*; *(Questions for which GPT-4o generates responses exceeding 128 characters.)* These typically involve detailed captions or complex descriptions, testing the model's conversational and descriptive capabilities. *hort-text questions*; *(Questions for which GPT-4o generates responses shorter than 128 characters.)* These require concise answers, often involving simple image analysis, and represent a broader range of task types.

nitial distribution of these three types in the image understanding dataset is highly imbalanced, with proportions of 12.17% (Long), 83.68% (Short), and 4.14% (MCQ). lign with diversity goals, we adjust the sampling ratio to 4:5:1, reducing disparities among task types while maintaining a dominance of comprehensive samples[3].

[1] https://huggingface.co/datasets/lmms-lab/LLaVA-OneVision-Data

[2] https://huggingface.co/datasets/Yirany/UniMM-Chat

[3] For video understanding and safety tasks, MCQ samples are fewer. After classifying into Long and Short types, the differences are minimal, so no additional adjustments are made.



re 2: **Re-Sample results from the clustering process.** to the large total number of samples, the clustered and deduplicated results contain a rich diversity of categories. cted samples include topics such as mathematics, daily life, natural scenes, medicine, electronic technology, and OCR scenarios, showcasing a variety of problem-image pairs. 2D features were obtained via UMAP dimensionality reduction.

Table 1: Dataset Composition Statistics

Image			Safety	Video	Total
Long	Short	MCQ			
9,575	12,063	2,125	1,999	4,235	29,997

ster-based Sampling. deduplication is not performed because many questions, while similar in text, are paired with different images, leading to substantially different outcomes—an intrinsic characteristic of multimodal data. ad, we encode all images using CLIP[4], and for videos, we use the feature of the first frame as a representative. then apply KNN clustering with 100 cluster centers and randomly sample N instances from each cluster. value of N is determined to satisfy the predefined sampling ratios, ensuring a balanced representation of task diversity.

Data statistics. composition of the dataset is summarized in Table 1, and a visualization of the clustering results is shown in Figure 2, demonstrating the rich diversity of data categories.

Model response generation. enerate high-quality responses, we select state-of-the-art models from both open-source and closed-source domains. image understanding and safety-related tasks, we use Qwen2-VL-72B , LLaVA-OV-72B [GPT-4o[5], and Claude 3.5-sonnet[6]. video understanding tasks, we employ GPT-4o, LLaVA-Video-72B , and Qwen2-VL-72B e models are chosen for their advanced capabilities and performance, ensuring a comprehensive evaluation of leading solutions in multimodal understanding.

[4] https://huggingface.co/openai/clip-vit-base-patch32
[5] https://openai.com/index/hello-gpt-4o/
[6] https://www.anthropic.com/news/claude-3-5-sonnet

2.3 Annotation

annotation process follows rigorous standards to ensure comprehensive and fine-grained evaluations of MLLM responses. ailed standards are provided in Appendix B, and the scoring and annotation structure are illustrated in Figure 1. itionally, we design a web UI to streamline the annotation process, as shown in Figure 7.

2.3.1 Annotation Standards

C ared to prior work, our annotation approach introduces two significant advantages: **richness** and **granularity**. , the evaluation incorporates three core dimensions—*Helpfulness*, *Faithfulness*, and *Ethical Considerations*—to comprehensively capture model performance. *pfulness* ensures that responses are relevant and provide meaningful assistance aligned with the user's intent. *fulness* evaluates the accuracy of responses in describing visual elements, such as objects, relationships, and attributes, ensuring alignment with the ground truth while avoiding hallucinated content. *cal Considerations* assess adherence to ethical principles, including safety, privacy, fairness, and harm avoidance, ensuring responses are free from harmful or biased content. notators score each dimension while documenting the reasoning behind their assessments, adding valuable context for understanding model performance.

S d, annotators are required to assign an **overall ranking** to the responses, along with justifications for their rankings. ranking mechanism provides a transparent and nuanced comparison of model outputs. itionally, innovative strategies are employed to enhance data quality:

- **C tructing positive samples for poor quality ties**. When multiple responses are equally poor, annotators provide correct answers to create positive examples. ensures that challenging samples contribute to the training dataset, addressing issues where no valid model responses exist.

- **nstructing negative samples for high-quality ties**. When multiple responses are of equally high quality, annotators introduce deliberate errors to create negative samples. This prevents ties from reducing the utility of the data and allows for more efficient use in training.

E mbining fine-grained scoring criteria, textual annotations, and innovative strategies, our annotation framework produces a high-quality dataset that comprehensively captures model performance and supports effective downstream applications.

2.3.2 Human Annotation vs. Machine Annotation

Annotation workers and costs. nnotation process employs over 50 annotators, supported by 8 multimodal research experts with strong English proficiency and academic backgrounds. entire task completes within two months, with periodic quality checks and interactive reviews conducted by experts to ensure the reliability and accuracy of the annotations. -quality samples undergo re-annotation during the process. to the fine-grained nature of the annotation standards, the task involves significant challenges. example, annotating a single question in the long split of image perception tasks requires an average of over 8 minutes.

Why human annotation? y existing MLLM alignment datasets rely on annotations generated by external models due to their cost-effectiveness and scalability. ever, MLLM alignment tasks demand fine-grained perceptual capabilities and sensitivity to subtle differences, which current models lack. any cases, the differences between responses are nuanced, requiring an in-depth understanding that models struggle to achieve. emonstrated in our experiments, even state-of-the-art models like GPT-4o significantly underperform human experts in tasks involving response comparison. eover, these models cannot provide professional-grade scoring or well-reasoned explanations for rankings. se limitations highlight the necessity of human annotation, which ensures the precision, reasoning, and insight required for constructing high-quality alignment datasets. ndix D further discusses the advantages of human annotation, particularly in handling ambiguous or incomplete questions and closely matched responses requiring subtle differentiation. an annotators excel at identifying fine-grained errors, inconsistencies, and context-specific nuances that models overlook. relying on human feedback, our approach ensures the dataset achieves the quality and reliability necessary for advancing MLLM alignment efforts.

acknowledge that the cost of human annotation poses scalability challenges. ever, as demonstrated in later sections, our high-quality alignment dataset enables the training of a powerful

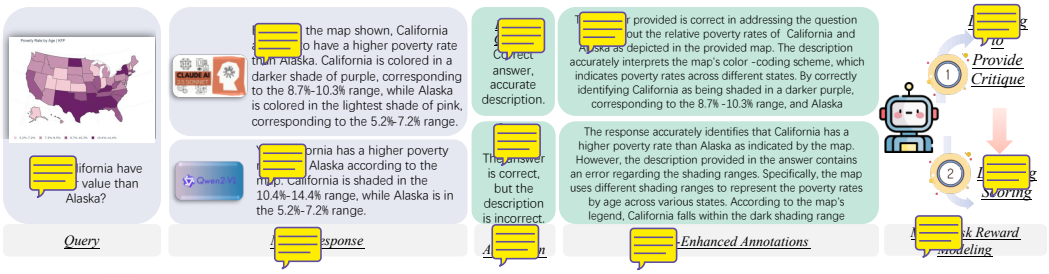


Figure 3: **Illustration of the multi-task reward model training process.** process begins with a user query and corresponding model responses, which are ranked and annotated by humans. nan annotations are expanded using GPT-4o to provide enhanced rationales. reward model is trained with two objectives: (1) *Learning to Provide Critique*, where the model learns to provide detailed critiques and evaluations for model responses, and (2) *Learning Scoring*, where the model learns to assign scores based on the model response and critique. integration of these tasks ensures a robust evaluation framework for improving model outputs.

reward model. e future, by combining this reward model with human annotators in a collaborative framework, we can significantly reduce annotation costs and scale up the dataset efficiently. hybrid approach not only maintains the precision of human annotation but also enhances scalability, making it a practical solution for large-scale MLLM alignment.

3 MM-RLHF-Reward Model

is section, we explore how to train a high-quality reward model using the MM-RLHF dataset to provide a robust supervision signal for subsequent model alignment e reward model is designed to combine critique generation and scoring (Figure 3), ensuring a comprehensive evaluation process.

3.1 Background and Limitations of Standard Reward Models

ard models are a key component for aligning model outputs with human preferences. cally, a reward model starts with a pretrained LLM ϕ, where the LLM head h_l is replaced with a linear reward head l_r, enabling the model to output a scalar reward value. e models are trained using human-provided pairwise comparisons. n a query $\mathbf{x}$, a preferred response y_w and a less preferred response y_l, the reward model is optimized to assign higher rewards to preferred responses:

$$\ell_{\text{Reward}}(\theta) = \mathbb{E}_{\mathbf{x},y_w,y_l}\Big[-\log\sigma\Big(r(y_w|\mathbf{x}) - r(y_l|\mathbf{x})\Big)\Big], \tag{1}$$

e $r(y|\mathbf{x})$ is the scalar reward and σ is the sigmoid function.

ite their utility, standard reward models face significant limitations. t, they fail to fully utilize the rich and detailed feedback provided by high-quality human annotations, such as textual explanations and nuanced reasoning. nd, scalar rewards lack transparency, making it difficult for humans to understand how the reward is generated. e challenges highlight the need for a more interpretable and robust reward model that leverages critiques as intermediate reasoning steps.

3.2 Critique-Based Reward Model Training

ding to critique-based training. vercome the limitations of traditional reward models, we propose a critique-based training framework: model first generates a critique c conditioned on the query $\mathbf{x}$. critique serves as an intermediate reasoning step, providing context for scoring responses. critique-based reward model comprises two components: **ritique Head** (h_l): Generates critiques c_w and c_l for the preferred (y_w) and less preferred (y_l) responses, respectively, based on the query $\mathbf{x}$. **coring Head** (h_r): Assigns scalar rewards based on the generated critiques, enabling more fine-grained evaluation.

Learning to provide critique from enhanced annotation. The critique head (h_l) is trained to align with human-provided annotations. The loss function for critique generation is:

$$\ell_{\text{Critique}}(\theta) = \mathbb{E}_{\mathbf{x},y,c}\Big[-\sum_{t=1}^{|c|}\log \pi_\theta(c_t|c_{<t},\mathbf{x},y)\Big], \tag{2}$$

where c_t is the t-th token in the critique c, $c_{<t}$ denotes the tokens preceding c_t, and $\pi_\theta(c_t|c_{<t},\mathbf{x},y)$ is the probability of token c_t given its context, query $\mathbf{x}$, and model response y.

However, as shown in Figure 3, while human-provided scoring reasons are highly accurate, they tend to be concise. Directly using these concise annotations as training targets for the reward model's language head does not yield significant performance improvements. To address this issue, we use GPT-4o to augment the human annotations by adding more detail and improving the fluency of the critiques. These enhanced scoring reasons are then used as the training targets for the language head. To prevent GPT-4o from introducing hallucinated content or irrelevant analysis, we impose strict constraints in the prompt (Table 7), to ensure the model only expands on the original content without introducing speculative or uncertain information.

Scoring loss with teacher-forcing. The head computes scalar rewards based on the query $\mathbf{x}$, response y, and critique c. During training, we adopt a teacher-forcing strategy, where the scoring head uses ground truth critiques instead of critiques generated by itself. This avoids potential noise from model-generated critiques in the early stages of training. The scoring loss is defined as:

$$\ell_{\text{Score}}(\theta) = \mathbb{E}_{\mathbf{x},y_w,y_l}\Big[-\log\sigma\Big(r(\mathbf{x},y_w,c_w) - r(\mathbf{x},y_l,c_l)\Big)\Big], \tag{3}$$

where: c_w and c_l are the ground truth critiques for the preferred response y_w and less preferred response y_l, respectively, $r(\mathbf{x},y,c)$ is the reward score computed from $\mathbf{x}$, y, and c.

Joint training objective. The overall training objective combines the critique generation loss and the scoring loss: $\ell_{\text{Total}}(\theta) = \ell_{\text{Critique}}(\theta) + \ell_{\text{Score}}(\theta)$.

Inference. During inference, the critique head (h_l) generates a critique c conditioned on the query $\mathbf{x}$ and response y. The scoring head (h_r) then uses $\mathbf{x}$, y, and the generated critique c to compute the final reward score $r(\mathbf{x},y,c)$. This two-step process mirrors the human evaluation process by explicitly reasoning about critiques before scoring.

MM-RLHF-RewardBench. To evaluate the effectiveness of the signals provided by our reward model in guiding subsequent model training, we randomly sample 10 examples from each category of the MM-RLHF dataset to create a test set. Each example includes multiple model responses and their corresponding rankings, enabling the generation of several comparison pairs. This results in a total of 170 pairs for evaluation. We design two evaluation metrics: *Traditional Accuracy (ACC)*: Measures the proportion of cases where the model correctly identifies the preferred response. *ACC+*: Measures the proportion of cases where the model correctly ranks all response pairs for a given sample. This metric emphasizes the model's ability to handle challenging cases, such as those with small ranking differences or hard-to-distinguish pairs.

3.3 Discussion

In the MLLM community, there is currently no unified paradigm for the design of reward models. Some approaches rely on traditional reward models [58], which lack interpretability due to their reliance on scalar outputs. Others directly use LLMs to generate rankings [67], which heavily depend on instruction-following capabilities and often exhibit high variance in scoring. In the broader LLM community, works such as [74] explore reward models that first generate critiques. However, their focus is primarily on improving the reliability of model-generated critiques, such as increasing scoring confidence through multiple sampling—a goal distinct from ours. To the best of our knowledge, this is the first study to explore how MLLMs can effectively leverage human annotations to enhance both interpretability and the final model's scoring ability.

4 MM-DPO

In this section, we propose MM-DPO, an extension of the traditional DPO framework. MM-DPO introduces Dynamic Reward Scaling, which dynamically adjusts the update strength based on the

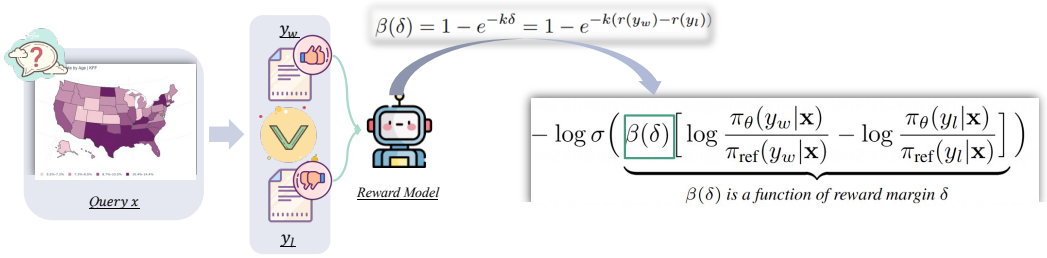


Figure 4: **Overview of the MM-DPO framework,** the dynamic reward scaling mechanism adjusts the update strength based on the reward margin, improving optimization stability and robustness.

confidence of training pairs, ensuring effective utilization of high-quality samples while mitigating the impact of noisy or low-confidence data.

4.1 Background: Direct Preference Optimization

The DPO framework is a preference-based learning method that optimizes model parameters θ by aligning model outputs with human preferences. Given a query $\mathbf{x}$ and corresponding responses y_w (positive) and y_l (negative), the DPO loss is defined as:

$$\ell_{\text{DPO}}(\theta) = \mathbb{E}_{\mathbf{x},y_w,y_l}\Big[-\log\sigma\Big(\beta\Big(\log\frac{\pi_\theta(y_w|\mathbf{x})}{\pi_{\text{ref}}(y_w|\mathbf{x})} - \log\frac{\pi_\theta(y_l|\mathbf{x})}{\pi_{\text{ref}}(y_l|\mathbf{x})}\Big)\Big)\Big], \tag{4}$$

where π_θ is the model's predicted probability distribution, π_{ref} is a reference policy, β is a scaling factor, and $\sigma(\cdot)$ is the sigmoid function. Traditional DPO treats all training pairs equally, regardless of their quality differences. This uniform scaling fails to prioritize high-quality pairs with clear preference distinctions, leading to inefficient use of informative samples and suboptimal optimization.

4.2 MM-DPO: Key Contributions and Improvements

Training on all possible comparison pairs instead of the hardest pairs . Unlike many recent MLLM alignment approaches that prioritize training on the hardest comparison pairs, MM-DPO incorporates all possible comparison pairs for a single query into the training process. Specifically, for any query with multiple responses, every response pair with differing ranks is treated as a valid comparison pair. This comprehensive approach captures more nuanced ranking information, allowing the model to learn from a broader set of preferences. However, this strategy also introduces a challenge: pairs involving responses with similar ranks (e.g., rank 3 and rank 4) often have lower reward margins compared to pairs with more distinct rankings (e.g., rank 1 and rank 4). Treating all pairs equally, as in traditional DPO, exacerbates the issue of uniform scaling and underutilizes the high-confidence information contained in larger reward margins. To address this, MM-DPO introduces Dynamic Reward Scaling, which dynamically adjusts the update strength based on the reward margin to prioritize high-confidence training pairs.

Implementation of dynamic reward scaling . Reward models can naturally provide a pairwise reward margin, which serves as a straightforward signal for scaling. However, two critical aspects must be addressed: (1) ensuring the signal quality is sufficiently high, and (2) bounding the signal to prevent overly aggressive updates that might destabilize training.

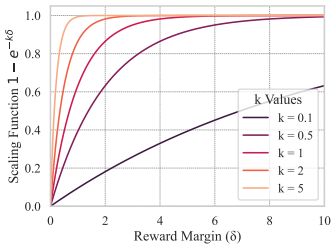


Figure 5: Effect of k on $1 - e^{-k\delta}$.

Regarding the first aspect, our experiments reveal that publicly available models, such as GPT-4o and LLaVA-Critic, perform inadequately in scoring our dataset. Conversely, our MM-RLHF-Reward-7B model surpasses several publicly available 72B models, offering a reliable and robust reward signal. We use this model to compute the reward margin: $\delta = r(y_w) - r(y_l)$, where $r(y_w)$ and $r(y_l)$ are the scores assigned to the positive and negative samples.

he second factor, we control the scaling factor $\beta(\delta)$ using the following formulation:

$$\beta(\delta) = \beta_{\text{ori}}\Big(1 + w\big(1 - e^{-k\delta}\big)\Big),$$

e β_{ori} is the initial default scaling factor, w is a parameter balancing the dynamic component's contribution, and k is a tunable hyperparameter that adjusts $\beta(\delta)$'s sensitivity to changes in δ. function $1 - e^{-k\delta}$ is bounded between $[0, 1]$, as illustrated in Figure 5. maller k value keeps most $\beta(\delta)$ values near β_{ori}, with slow growth as δ increases. ontrast, a larger k makes $\beta(\delta)$ highly responsive to changes in δ, quickly reaching its maximum. void overly aggressive updates, we constrain $\beta(\delta)$ within $[\beta_{\text{ori}}, (1 + w)\beta_{\text{ori}}]$. all, Dynamic Reward Scaling significantly enhances MM-DPO by leveraging high-quality reward signals and tailoring optimization steps to the confidence level of training pairs. results in improved robustness, efficiency, and overall effectiveness of the framework. discuss the similarities and differing perspectives between our approach and existing methods in Appendix E.

5 Experiments

We evaluate our data and algorithms on 10 tasks across 20+ benchmarks. key findings are:

1. A ment training on the **MM-RLHF** dataset consistently improves performance across nearly all benchmarks for various baselines. ntegration of reward signals in MM-DPO further amplifies these improvements, demonstrating the effectiveness of our approach.

2. **MM-RLHF-Reward-7B** model achieves state-of-the-art performance on reward model benchmarks among open-source models, surpassing even several 72B models. s highlights the efficiency and scalability of our method.

3. conduct extensive ablation studies and analyses, such as investigating the importance of critique learning for reward models and the sensitivity to hyperparameters. itionally, we identify several experimental phenomena that challenge mainstream perspectives, such as the observation that small-scale MLLMs struggle to perform effective self-improvement. to space constraints, additional analysis are provided in Appendix F.

5.1 hmarks and Experimental Details

ategorize the benchmark datasets used in our experiments into the following domains:

t and Document Understanding: AI2D [29], ChartQA [51], DocVQA [54], InfoVQA [53].

(Optical Character Recognition): WebSRC [11], OCRBench [45], TextVQA [57].

cination: MMHal-Bench [59], POPE [40], Object-Hal [41].

Reasoning: MathVista [48], MathVerse [78].

ral Knowledge: MME [19], MMbench [44], MMStar [9], SeedBench2-Plus [33], VQAv2 [4].

ersation: LLaVA-Wilder [30], LLaVA-In-The-Wild [43], WildVision-Bench [49].

Resolution and Real-World Utility: RealworldQA, MME-RealWorld [81].

Understanding: VideoChatGPT [50], Video-MME [20], VideoDC [30].

i-Image: LLAVA-Next-Interleave [32], MMMU-Pro [75].

M Safety: self-constructed benchmark, MM-RLHF-SafeBench, includes adversarial attacks, jailbreaks, privacy, and harmful content. ailed construction is provided in Appendix C.2. y mainly evaluates the model's ability to reject harmful content, while unsafety mainly assesses the likelihood of the model being successfully attacked.

ll benchmarks requiring GPT-assisted evaluation, we consistently employ GPT-4o as the evaluation model. odel results are rigorously re-evaluated and reported by our team. xperiments are conducted on a high-performance computing cluster equipped with 32×H800 (80G) GPUs. to computational cost constraints, we utilize the full dataset for the main results presented in Tables 2, 3, and 5. ablation studies, we uniformly sample $1/5$ of the data, which may result in minor performance discrepancies compared to the full dataset.

In the implementation of MM-DPO, we adopt a common stabilization technique by incorporating an SFT loss. The weight of the SFT loss is selected through a grid search over the values $\{0, 0.1, 0.25, 0.5, 1.0\}$. Additionally, the learning rate is optimized via a search over $\{1e\text{-}7, 5e\text{-}7, 1e\text{-}6, 5e\text{-}6, 1e\text{-}5\}$ to identify the best-performing configuration. Since we dynamically adjust the β parameter during training, the initial value of β_{ori} is set to a small default value of 0.1, eliminating the need for manual tuning. Throughout all training processes, the vision encoder remains frozen to ensure stable and efficient training.

5.2 Evaluation of MM-RLHF and MM-DPO

Table 2 (for understanding tasks) and Table 3 (for safety tasks) illustrate the alignment performance of LLaVA-OV-7B, LLaVA-OV-0.5B and InternVL-1B using our dataset and alignment algorithm, where the scores for each evaluation dimension are averaged across their respective benchmarks.

Significant improvements in conversational ability and safety. The experiments show that the alignment process leads to substantial improvements in these two aspects without requiring hyperparameter tuning. The average improvement in conversational benchmarks exceeds 10%, while unsafe behaviors are reduced by at least 50%. Additionally, in WildsVision, the win rate increases by at least 50%. This suggests that existing MLLMs lack explicit optimization for these dimensions, and our dataset effectively fills this gap.

Broad enhancements in hallucination, mathematical reasoning, multi-image, and video understanding. The aligned models also exhibit notable improvements in these areas. Interestingly, despite the lack of dedicated multi-image data in our dataset, the model's performance in multi-image tasks improves significantly. This indicates that the diversity of our alignment data enhances generalization across multiple dimensions.

Model-specific preferences for data and hyperparameter. Different models exhibit varying performance trends during alignment, with distinct preferences for hyperparameter settings across different benchmarks. For instance, in our training of InternVL-1B, we found that excluding the SFT loss led to better results. Additionally, while InternVL-1B demonstrated significant improvements in general knowledge tasks, its relative enhancement in OCR tasks was less pronounced compared to the LLaVA-OV series. These differences largely stem from variations in the models' pretraining datasets and strategies, necessitating tailored hyperparameter adjustments for optimal alignment.

Limited gains in high-resolution benchmarks. The model shows no significant improvement on high-resolution benchmarks, likely because our dataset contains relatively few ultra-high-resolution images. Additionally, our filtering strategy is based on image similarity rather than resolution, meaning the alignment process does not explicitly optimize for high-resolution tasks. As a result, performance gains in this area remain limited.

Ablation studies and sensitivity analysis. To further validate the effectiveness of our approach, we provide detailed ablation studies in the appendix, analyzing the impact of different alignment parameters and the improvements introduced by our dataset and MM-DPO.

5.3 Evaluation of MM-RLHF-Reward

In this section, we evaluate the effectiveness of MM-RLHF-Reward and highlight several noteworthy experimental observations. The results are presented in Table 4 and Table 5.

Existing reward models exhibit significant overfitting. As shown in Table 4, LLaVA-Critic's performance on MM-RLHF-Reward-Bench is suboptimal, with a considerable gap compared to GPT-4o. This can likely be attributed to the overfitting of existing reward models to their training data, which predominantly consists of conversational datasets and real-world images. Consequently, while LLaVA-Critic demonstrates notable improvements over its baseline, LLaVA-OV-7B[7], its performance in other categories, such as MCQ and more diverse tasks, remains limited.

Closed-source models like GPT-4o consistently deliver competitive performance. Across both Table 4 and Table 5, closed-source models such as GPT-4o demonstrate superior generalization capabilities compared to open-source alternatives, even those with significantly larger parameter

[7]Both models use identical prompts for tasks such as captioning and long-form dialogue.

Table 2: **Performance variations after alignment across 8 different evaluation dimensions**, comparing multiple models under our alignment strategy. Our models show comprehensive performance improvements under the proposed alignment, demonstrating significant gains across various tasks.

Capability	Benchmark	InternVL2 1B	Ours	LLaVA-OV 0.5B	Ours	LLaVA-OV 7B	Ours
Conversation	LLaVA-Wild [43] (all) Realworld Chat	73.80	75.80 +2.00	74.60	79.20 +4.60	90.70	97.90 +7.20
	LLaVA-Wild [43] (complex) Realworld Chat	83.60	82.60 -1.00	78.60	80.50 +1.90	95.90	100.60 +4.70
	LLaVA-Wild [43] (conv) Realworld Chat	52.10	58.30 +6.20	69.60	72.30 +2.70	81.20	88.10 +6.90
	LLaVA-Wild [43] (detail) Realworld Chat	85.40	89.40 +4.00	82.30	84.50 +2.20	91.80	104.00 +12.20
	LLaVA-Wilder [30] (small) Realworld Chat	55.80	57.30 +1.50	52.30	53.40 +1.10	65.70	71.10 +5.40
	WildVision [49] (elo rate) Model Competition	41.30	46.20 +4.90	40.70	44.70 +4.00	50.40	58.90 +8.50
	WildVision [49] (win rates) Model Competition	41.80	49.00 +7.20	12.60	14.60 +2.00	15.20	37.20 +22.00
General Knowledge	MME [19] (cog./perp.) Multi-discip	1775	1815 +40	1488	1510 +22	1997	2025 +28
	MMBench [44] (cn-dev) Multi-discip	54.70%	67.89% +13.19%	45.80%	46.40% +0.60%	80.49%	80.67% +0.18%
	MMStar [9] Multi-discip	45.81%	49.00% +3.19%	38.64%	39.58% +0.94%	61.80%	62.58% +0.78%
	SeedBench2-Plus [33] Multi-discip	60.12%	60.12% +0.00%	53.85%	54.27% +0.42%	64.87%	65.35% +0.48%
	VQAv2 [4] (lite) Multi-discip	72.25%	71.84% -0.41%	74.60%	74.68% +0.08%	79.98%	80.28% +0.30%
Chart and Document	AI2D [29] Science Diagrams	72.38%	72.80% +0.42%	56.93%	56.87% -0.06%	81.41%	81.22% -0.19%
	ChartQA [52] (val-lite) Chart Understanding	65.60%	66.80% +1.20%	51.60%	52.60% +1.00%	74.00%	74.50% +0.50%
	DocVQA [55] (val-lite) Document Understanding	81.90%	82.51% +0.61%	66.17%	67.07% +0.90%	84.34%	86.11% +1.77%
	InfoVQA [53] (val-lite) Infographic Understanding	51.73%	52.26% +0.53%	40.17%	40.49% +0.32%	67.07%	67.40% +0.33%
OCR	OCRBench [45] Comprehensive OCR	75.20%	77.11% +1.91%	57.70%	60.20% +2.50%	62.30%	69.30% +7.00%
	TextVQA [57] (val) Text Reading	69.85%	72.12% +2.27%	65.87%	66.60% +0.73%	75.99%	76.05% +0.06%
	WebSRC [11] (val) Web-based Structural Reading	68.20%	68.80% +0.60%	65.90%	68.30% +2.40%	88.70%	89.20% +0.50%
Real-World	MME-RealWorld [81] (en-lite) Multi-discip & High-Resolution	33.61%	36.58% +2.97%	34.55%	34.39% -0.16%	48.36%	46.95% -1.41%
	MME-RealWorld [81] (cn) Multi-discip & High-Resolution	44.14%	43.11% -1.03%	32.09%	31.11% -0.98%	54.01%	53.39% -0.62%
	RealWorldQA Realworld QA	51.50%	54.90% +3.40%	55.42%	55.16% -0.26%	66.41%	65.75% -0.66%
Math	MathVista [48] (cot) General Math Understanding	49.60%	49.90% +0.30%	32.30%	32.70% +0.40%	59.10%	61.60% +2.50%
	MathVista [48] (format) General Math Understanding	53.20%	53.40% +0.20%	36.00%	36.30% +0.30%	62.50%	62.20% -0.30%
	MathVista [48] (solution) General Math Understanding	49.60%	49.30% -0.30%	30.50%	32.50% +2.00%	58.80%	61.10% +2.30%
	MathVerse [78] (vision-mini) Professional Math Reasoning	12.31%	12.79% +0.48%	17.51%	17.64% +0.13%	16.37%	18.53% +2.16%
Hallucination	POPE [40] (adversarial) Object Hallucination.	86.82%	86.87% +0.05%	86.04%	86.56% +0.52%	87.08%	87.68% +0.60%
	POPE [40] (popular) Object Hallucination.	88.30%	88.57% +0.27%	87.37%	88.26% +0.89%	88.32%	89.02% +0.70%
	POPE [40] (random) Object Hallucination.	89.87%	90.45% +0.58%	88.30%	89.30% +1.00%	89.60%	90.62% +1.02%
	MMHal [59] (hal rate ↓) General Hallucination	55.21%	55.38% -0.17%	48.96%	46.25% +2.71%	38.54%	38.54% +0.00%
	MMHal [59] (avg score) General Hallucination	3.02	3.10 +0.08	3.33	3.42 +0.09	3.22	4.08 +0.86
	Obj-Hal [41] (chair-i↓) Object Hallucination.	8.30	7.81 +0.49	9.70	9.12 +0.58	8.52	7.69 +0.83
	Obj-Hal [41] (chair-s↓) Object Hallucination.	38.67	37.00 +1.67	42.67	42.33 +0.34	44.00	41.67 +2.33
Video Understanding	Video-MME [20] (w. caption) Multi-discip	42.74%	42.76% +0.02%	48.22%	48.42% +0.20%	61.61%	61.81% +0.20%
	Video-MME [20] (wo. caption) Multi-discip	45.66%	45.71% +0.05%	43.92%	44.00% +0.08%	58.29%	58.33% +0.04%
	VideoChatGPT [50] Video Conversation	2.26	2.59 +0.33	2.56	2.66 +0.10	2.87	3.22 +0.35
	VideoDC [30] Video Detail Description	2.91	3.07 +0.16	2.88	2.96 +0.08	3.32	3.41 +0.09
Multi-Image	LLAVA-Next-Interleave [32] (in-domian) in-domian	34.78%	35.72% +0.94%	42.29%	43.49% +1.20%	60.85%	61.12% +0.27%
	MMMU-Pro [75] (vision) Multi-discip	1.11%	1.52% +0.41%	12.78%	13.89% +1.11%	14.51%	15.84% +1.33%

Table 3: **ormance variations after alignment across MM-RLHF-SafeBench**, comparing multiple models under our alignment strategy.

Benchmark	InternVL2 1B	Ours	LLaVA-OV 0.5B	Ours	LLaVA-OV 7B	Ours
Adv target ↓ Adversarial Attack	56.0%	50.0% +5.0%	54.0%	35.0% +19.0%	37.0%	40.0% -3.0%
Adv untarget ↑ Adversarial Attack	52.5%	56.0% +3.5%	66.0%	71.0% +5%	66.5%	70.0% +3.5%
Crossmodel ASR ↓ Cross-modal Jailbreak	0.0%	0.0% +0.0%	72.2%	38.9% +33.3%	16.7%	0.0% +16.7%
Crossmodel RtA ↑ Cross-modal Jailbreak	100.0%	100.0% +0.0%	22.2%	50.0% +27.8%	88.9%	100.0% +11.1%
Multimodel ASR ↓ Multimodal Jailbreak	43.2%	43.2% +0.0%	42.2%	27.7% +14.5%	41.2%	8.3% +31.9%
Multimodel RtA ↑ Multimodal Jailbreak	18.0%	17.4% -0.6%	12.4%	23.2% +10.8%	62.0%	88.3% +26.3%
Typographic ASR ↓ Typographic Jailbreak	10.5%	7.4% +3.1%	26.3%	35.2% -8.9%	5.8%	0.0% +5.8%
Typographic RtA ↑ Typographic Jailbreak	73.7%	74.6% +0.9%	17.0%	27.5% +10.5%	79.5%	95.8% +16.3%
Risk ↑ Risk identification	49.6%	58.6% +9.0%	65.8%	67.4% +1.6%	82.0%	76.0% -6.0%
NSFW text↓ NSFW Jailbreak	89.0%	27.1% +61.9%	94.4%	64.2% +30.2%	60.4%	10.6% +49.8%
NSFW img↓ NSFW Jailbreak	81.2%	64.7% +16.5%	97.5%	81.6% +15.9%	80.1%	24.2% +55.9%
Unsafety ↓ Average performance of ↓	46.6%	38.9% +7.7%	65.4%	47.1% +18.3%	40.2%	13.9% +26.3%
Safety ↑ Average performance of ↑	31.9%	41.3% +9.4%	36.7%	47.8% +11.1%	75.8%	85.4% +9.6%

sizes (e.g., 72B models). observation underscores the robustness of closed-source approaches in handling diverse multimodal tasks and maintaining high performance across various metrics.

-RLHF-Reward sets a new benchmark for open-source models, rivaling closed-source systems. th benchmarks, MM-RLHF-Reward achieves results comparable to or exceeding GPT-4o's performance, while significantly outperforming most open-source models, such as LLaMA-3.2-90B-Vision-Instruct and Qwen2-VL-72B-Instruct. ably, on our custom benchmark, MM-RLHF-Reward demonstrates a substantial lead over GPT-4o, further justifying its selection as the reward signal for training algorithms. obust performance across diverse metrics highlights its effectiveness and adaptability.

importance of an effective critic in reward modeling. results in Table 4 underscore the critical role of an effective critic in reward modeling. n the reward head is directly trained using pair-wise datasets, the ACC+ stabilizes around 50%. ncorporating human annotations as the learning target—allowing the model to first learn evaluation reasoning and then perform scoring—the ACC+ improves by a consistent 5%. ever, human annotations alone may not serve as an optimal training target due to their brevity or conversational style. address this, we expand the human annotations using the model itself, producing enriched annotations that further enhance reward model training quality. results in a significant 17% improvement in ACC+ compared to the baseline. lly, during evaluation, when human annotations are directly provided as the critic (i.e., scoring is based on human-provided evaluations rather than model-generated critics), both ACC and ACC+ reach approximately 90%. demonstrates the pivotal role of evaluation quality in the overall effectiveness of reward models.

tiple sampling of critiques does not yield significant performance gains. n the model generates critiques with high variability, multiple sampling is often used to compute scores and then take the average [74]. approach has proven effective in related LLM research. vever, in our experiments, we observed that when we lowered the sampling temperature and computed rewards multiple times, the performance actually declined. reason for this is that during the sampling process, there is occasionally a critique that is inaccurate. e our model is already capable of generating reasonably accurate critiques due to its alignment with human annotations, the extra, time-consuming sampling process does not provide additional benefits and can even have a negative impact on performance.

Table 4: **rformance comparison across metrics and methods on MM-RLHF-RewardBench**. *-RLHF-Reward (w/o. Task 1)* represents training the LLaVA-OV-7B model to score pair-wise samples while excluding Task 1. *RLHF-Reward (w/o. enhanced annotations)* involves learning human-provided annotations, followed by scoring. *-RLHF-Reward (inference w. GT annotation)* uses ground truth annotations during inference.

Method	LLaVA-OV-7B		LlaVA-Critic (Pointwise)		LlaVA-Critic (Pairwise)		GPT-4o		MM-RLHF-Reward (w/o. Task 1)		MM-RLHF-Reward (w/o. enhanced annotations)		MM-RLHF-Reward		MM-RLHF-Reward (inference w. GT annotation)	
Metric	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+
Mcq	0.14	0.00	0.38	0.10	0.23	0.00	0.69	0.20	0.90	0.80	0.83	0.70	0.93	0.70	**1.00**	**1.00**
Long	0.11	0.00	0.49	0.20	0.54	0.30	0.95	0.90	0.70	0.40	0.92	0.80	1.00	1.00	**1.00**	**1.00**
Short	0.29	0.20	0.38	0.20	0.24	0.10	0.56	0.40	0.79	0.60	0.68	0.40	0.71	0.50	**1.00**	**1.00**
Safety	0.41	0.00	0.62	0.17	0.28	0.17	**0.72**	**0.33**	0.69	**0.33**	0.69	0.17	0.66	0.17	0.69	0.17
Video	0.32	0.10	0.40	0.20	0.52	0.20	0.80	0.60	0.70	0.60	0.80	0.60	**0.92**	0.80	**0.92**	**0.90**
Overall	0.24	0.07	0.45	0.17	0.35	0.15	0.74	0.50	0.75	0.50	0.79	0.57	0.85	0.67	**0.93**	**0.87**

Table 5: **rformance comparison of our reward model (MM-RLHF-Reward) with existing open-source and private multi-modal models.** M-RLHF-Reward-7B outperforms existing 72B open-source multi-modal models and several highly competitive closed-source models.

Model	General	Hallucination	Reasoning	Avg
VITA-1.5 [22]	18.55	8.93	22.11	16.48
SliME-8B [79]	7.23	27.09	18.6	19.04
deepseek-vl2 [66]	29.70	23.80	50.90	34.80
Phi-3.5-vision-instruct [1]	28.00	22.40	56.60	35.67
llava-onevision-qwen2-7b-ov [32]	32.20	20.10	57.10	36.47
Molmo-7B-D-0924 [17]	31.10	31.80	56.20	39.70
Pixtral-12B-2409 [2]	35.60	25.90	59.90	40.47
Qwen2-VL-72B-Instruct [64]	38.10	32.80	58.00	42.97
NVLM-D-72B [16]	38.90	31.60	62.00	44.17
InternVL2-26B [12]	39.30	36.90	60.80	45.67
Private models				
GPT-4o-mini (2024-07-18)	41.70	34.50	58.20	44.80
Claude-3.5-Sonnet (2024-06-22)	43.40	55.00	62.30	53.57
GPT-4o (2024-08-06)	49.10	67.60	70.50	62.40
Gemini-1.5-Pro (2024-09-24)	50.80	72.50	64.20	62.50
Ours				
MM-RLHF-Reward-7B	45.04	50.45	57.55	50.15

5.4 **Improvement of Small-Scale MLLMs is Currently Unrealistic**

e recent work on MLLMs explores the concept of self-improvement, these efforts largely focus on specific domains, such as conversational systems [67]. I is section, we present an alternative perspective distinct from the LLM domain, arguing that MLLMs, particularly small models (fewer than 7B parameters), currently face significant challenges in achieving comprehensive performance improvements through self-improvement. experimental results, illustrated in Figure 6, suggest two primary reasons for this limitation:

1. el capacity constraints. tasks involving long-form or conversational data, sampling multiple responses often results in at least one reasonably good answer, thereby leading to noticeable improvements. vever, for more challenging tasks, such as multiple-choice questions or scientific reasoning, smaller models struggle to generate correct answers even after extensive sampling. ur experiments, where the maximum number of samples reached eight, we observed instances where the model produced identical incorrect responses or consistently incorrect outputs across all samples for some challenging multiple-choice questions.

2. tations in reward signal quality. st existing multimodal reward models are trained on datasets with limited diversity, such as VLFeedback and LLaVA-RLHF. se datasets predominantly focus on natural images, human dialogue, or related scenarios, raising concerns about overfitting. n preference datasets encompass broader domains, such as mathematical reasoning, chart understanding, or other specialized fields, reward models trained on existing datasets fail to provide effective reward signals. sequently, it becomes challenging to identify and select better samples.

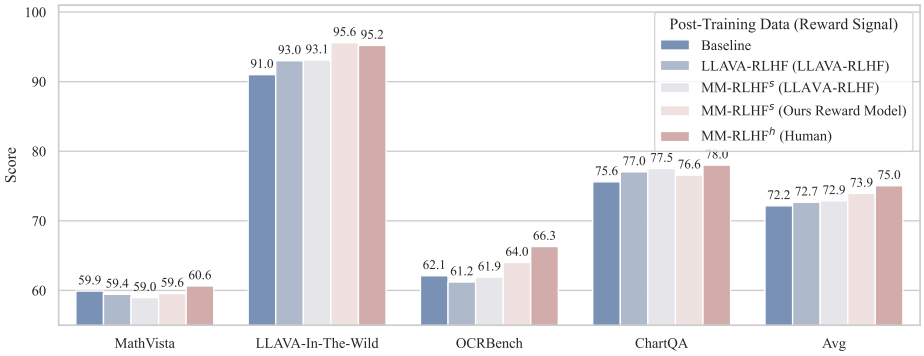


Figure 6: **Performance comparison across datasets using various methods based on the LLaVA-Ov-7B model as the baseline**. "Baseline" represents the initial performance without post-training. "LLAVA-RLHF (LLAVA-RLHF)" indicates that both the post-training dataset and the reward model come from the LLAVA-RLHF dataset, with the reward model being trained using LLaVA-Ov-7B as the starting checkpoint for fairness. "MM-RLHF ^{s}" reflects results generated on our dataset, where responses are self-sampled (default sample size: 8) and ranked using different reward signals to create DPO pairs. "MM-RLHF ^{h} (Human)" involves DPO training directly using our dataset, where responses are sampled from other models, and reward signals are provided by experts.

These two limitations make it difficult, at the current stage, to enable MLLMs to generate responses on diverse datasets, annotate them with reward models, and iteratively improve through self-improvement cycles, as has been achieved in LLM alignment. While our experiments confirm that better reward models can lead to marginal improvements, the results remain far inferior to training with high-quality, human-annotated contrastive samples.

6 Conclusion and Future Work

In this work, we introduced **MM-RLHF**, a high-quality, fine-grained dataset specifically designed to advance the alignment of MLLMs. Unlike prior works that focus on specific tasks, our dataset and alignment approach aim to holistically improve performance across diverse dimensions. Coupled with preliminary improvements to reward modeling and optimization algorithms, we observed significant and consistent gains across almost all evaluation benchmarks, underscoring the potential of comprehensive alignment strategies.

Looking ahead, we see great opportunities to further unlock the value of our dataset. Its rich annotation granularity, such as per-dimension scores and ranking rationales, remains underutilized in current alignment algorithms. Future work will focus on leveraging this granularity with advanced optimization techniques, integrating high-resolution data to address limitations in specific benchmarks, and scaling the dataset efficiently using semi-automated strategies. We believe these efforts will not only push MLLM alignment to new heights but also set a foundation for broader, more generalizable multimodal learning frameworks.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[2] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

[3] Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*, 2024.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] 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 preprint arXiv:2308.12966*, 2023.

[6] Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*, 2023.

[7] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

[10] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024.

[11] Xingyu Chen, Zihan Zhao, Lu Chen, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. Websrc: A dataset for web-based structural reading comprehension. *arXiv preprint arXiv:2101.09465*, 2021.

[12] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[13] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

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

[15] Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. ably robust dpo: Aligning language models with noisy feedback. *arXiv preprint arXiv:2403.00409*, 2024.

[16] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024.

[17] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

[18] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024.

[19] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[20] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv:2405.21075*, 2024.

[21] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.

[22] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025.

[23] Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*, 2024.

[24] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024.

[25] Xiaotian Han, Quanzeng You, Yongfei Liu, Wentao Chen, Huangjie Zheng, Khalil Mrini, Xudong Lin, Yiqi Wang, Bohan Zhai, Jianbo Yuan, Heng Wang, and Hongxia Yang. Infimm-eval: Complex open-ended reasoning evaluation for multi-modal large language models, 2023.

[26] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

[27] Xuhao Hu, Dongrui Liu, Hao Li, Xuanjing Huang, and Jing Shao. Vlsbench: Unveiling visual leakage in multimodal safety. *arXiv preprint arXiv:2411.19939*, 2024.

[28] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[29] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

[30] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024.

[31] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.

[32] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[33] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024.

[34] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023.

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

[36] Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, et al. wardbench: A challenging benchmark for vision-language generative reward models. *arXiv preprint arXiv:2411.17451*, 2024.

[37] 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*, 2023.

[38] Mukai Li, Lei Li, Yuwei Yin, Masood Ahmed, Zhenguang Liu, and Qi Liu. Red teaming visual language models. *arXiv preprint arXiv:2401.12915*, 2024.

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

[40] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023.

[41] 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*, 2023.

[42] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023.

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

[44] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[45] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models, 2024.

[46] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[47] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[48] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

[49] Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*, 2024.

[50] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024.

[51] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv:2203.10244*, 2022.

[52] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

[53] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.

[54] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021.

[55] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021.

[56] OpenAI. Gpt-4 technical report. 2023.

[57] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.

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

[59] 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:2309.14525*, 2023.

[60] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

[61] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.

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

[63] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[64] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

[66] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024.

[67] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*, 2024.

[68] Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, et al. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. *arXiv preprint arXiv:2410.04509*, 2024.

[69] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[70] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi, 2024.

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

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

[73] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *ICML*, 2024.

[74] Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, et al. Self-generated critiques boost reward modeling for language models. *arXiv preprint arXiv:2411.16646*, 2024.

[75] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

[76] Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Jitao Sang, and Dit-Yan Yeung. Anyattack: Self-supervised generation of targeted adversarial attacks for vision-language models.

[77] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.

[78] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024.

[79] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024.

[80] Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Debiasing large visual language models. *arXiv preprint arXiv:2403.05262*, 2024.

[81] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.

[82] Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, et al. Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[83] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[84] Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*, 2024.

MM-RLHF

————Appendix————

Contents

A Related Work

timodal large language models have seen remarkable progress in recent years, with significant advancements in both performance and capabilities. eraging cutting-edge LLMs such as GPTs [56, 8], LLaMA [62, 63], Alpaca [60], Vicuna [14], and Mistral [28], MLLMs are increasingly demonstrating enhanced multimodal capabilities, especially through end-to-end training approaches. se advancements have been crucial in enabling models to handle a range of multimodal tasks, including image-text alignment, reasoning, and instruction following, while addressing challenges related to data fusion across different modalities. ent open-source MLLMs such as Otter [31], mPLUG-Owl [69], LLaVA [43], Qwen-VL [5], Cambrian-1 [61], Mini-Gemini [39], MiniCPM-V 2.5 [26], DeepSeek-VL [47], SliME [79] and VITA [21, 22] have contributed to solving some of the most fundamental multimodal problems, such as improving vision-language alignment, reasoning, and following instructions. se models focus on enhancing multimodal understanding by integrating vision with language, allowing for more nuanced and context-aware interactions. e of the most notable open-source models, such as InternLM-XComposer-2.5 [77] and InternVL-2 [13], have exhibited impressive progress in multimodal understanding, closely competing with proprietary models across a range of multimodal benchmarks. ever, despite these achievements, there is still a noticeable gap in security and alignment when compared to closed-source models. ighlighted by recent studies [81], most open-source MLLMs have not undergone rigorous, professional alignment processes, which has hindered their ability to effectively align with human preferences. gap in alignment remains one of the key challenges for open-source models, and improving model safety and alignment to human values will be a crucial area of future research.

MLLM Alignment. the rapid development of MLLMs, various alignment algorithms have emerged, showcasing different application scenarios and optimization goals. nstance, in the image domain, Fact-RLHF [58] is the first multimodal RLHF algorithm, and more recently, LLAVA-CRITIC [67] has demonstrated strong potential with an iterative DPO strategy. se algorithms have shown significant impact on reducing hallucinations and improving conversational capabilities [80, 72], but they have not led to notable improvements in general capabilities. e have also been some preliminary explorations in the multi-image and video domains, such as MIA-DPO and PPLLaVA. ever, alignment in image and video domains is still fragmented, with little research done under a unified framework. believe that the main limitation hindering the development of current alignment algorithms is the lack of a high-quality, multimodal alignment dataset. existing manually annotated MLLM alignment datasets are available, and most contain fewer than 10K samples [58, 72, 71], which is significantly smaller than large-scale alignment datasets in the LLM field. small dataset size makes it difficult to cover multiple modalities and diverse task types. hermore, machine-annotated data faces challenges related to quality assurance. re-fore, in this paper, we have invested considerable effort into constructing a dataset, MM-RLHF, which surpasses existing works in both scale and annotation quality.

MLLM Evaluation. the development of MLLMs, a number of benchmarks have been built [18, 23]. nstance, MME [19] constructs a comprehensive evaluation benchmark that includes a total of 14 perception and cognition tasks. QA pairs in MME are manually designed to avoid data leakage, and the binary choice format makes it easy to quantify. Bench [44] contains over $3,000$ multiple-choice questions covering 20 different ability dimensions, such as ob-

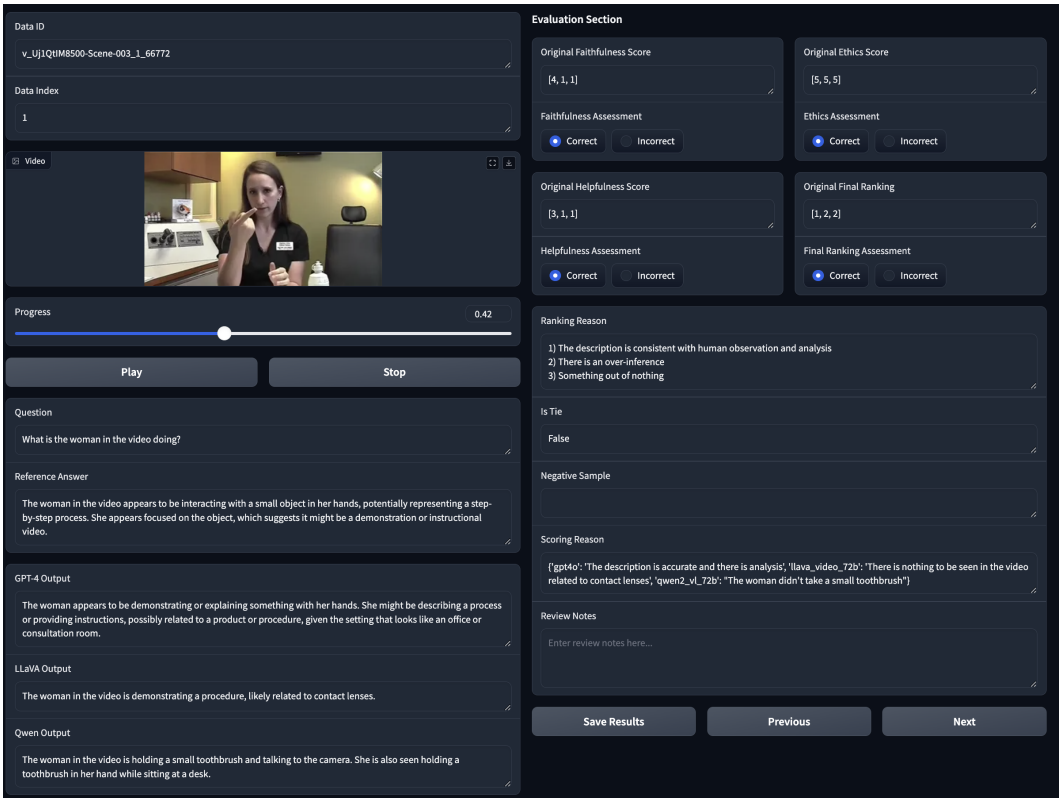


Figure 7: **The user interface for data annotation**, featuring image/video display, questions, outputs from each model, detailed scoring criteria, and a section for reviewers to verify the accuracy of the scores.

ject localization and social reasoning. troduces GPT-4-based choice matching to address the MLLM's lack of instruction-following capability and a novel circular evaluation strategy to improve the evaluation robustness. d-Bench [35] is similar to MME and MMBench but consists of $19,000$ multiple-choice questions. e larger sample size allows it to cover more ability aspects and achieve more robust results. D-Bench-2 [34] expands the dataset size to $24,371$ QA pairs, encompassing 27 evaluation dimensions and further supporting the evaluation of image generation. T-Bench [70] scales up the dataset even further, including $31,325$ QA pairs from various scenarios such as autonomous driving and embodied AI. compasses evaluations of model capabilities such as visual recognition, localization, reasoning, and planning. itionally, other benchmarks focus on real-world usage scenarios [24, 49, 7] and reasoning capabilities [73, 6, 25, 68]. E-RealWorld [81] places greater emphasis on quality and difficulty compared to its predecessor, containing the largest manually annotated QA pairs and the largest image resolution. se benchmarks reveal some common characteristics of MLLMs in task design and real-world applications. ever, benchmarks specifically focused on reward models [36] and those dedicated to evaluating safety and robustness remain relatively scarce. further promote comprehensive evaluation of MLLM alignment, this paper contributes two benchmarks: one for reward models through self-construction and data cleaning, and another more comprehensive safety benchmark.

B notation Guidelines for Evaluating MLLM Responses

document provides detailed annotation guidelines for evaluating responses generated by MLLMs. otators should rate and annotate each response according to four primary evaluation criteria: Visual Faithfulness, Helpfulness, Ethical Considerations (including safety, privacy, fairness, and harm), and Overall Performance. otators are expected to assess each response carefully based on these criteria to ensure high-quality feedback for model optimization.

B.1 I. al Faithfulness Evaluation

Definition: criterion evaluates whether the generated response accurately reflects the objects and relationships in the image, ensuring consistency with the objects, relationships, and attributes of the true answer.

Guidelines:

1. **bject Description Accuracy**: Ensure that the generated response accurately describes objects as in the true answer, avoiding references to non-existent objects and preventing errors in descriptions of existing objects.
2. **bject Relationship Accuracy**: Evaluate whether the spatial, structural, or functional relationships between objects described in the response are correct. Minimize errors and misleading information in object relationship descriptions.
3. **bject Attribute Accuracy**: Confirm that the response accurately describes the physical features, color, size, and state of objects, avoiding discrepancies with the true answer.
4. **verall Accuracy**: Assess the accuracy and consistency of the response in describing visual elements (objects, relationships, and attributes) in the image.

I g Scale:

- **erely Inaccurate**: Major errors in object descriptions, relationships, or attributes, or references to non-existent objects.
- **tially Inaccurate**: Object descriptions, relationships, or attributes are mostly correct, but some details are inaccurate.
- **derately Accurate**: Descriptions of objects, relationships, and attributes are generally accurate, with minor errors.
- **curate**: Descriptions of objects, relationships, and attributes are correct and match the true answer.
- **ghly Accurate**: Descriptions of objects, relationships, and attributes are exceptionally precise and include useful additional details.

B.2 II. H ulness Evaluation

I ition: Carefully review the user prompt to ensure that the response provides valuable insights and relevant information that aid the user in understanding the image or task.

Guidelines:

1. ssess whether the response provides useful context, insights, or relevant information to help the user better understand the image or task.
2. eck if the model follows the specific instructions or guidelines provided by the user.
3. valuate the overall contribution of the response in enhancing the user's understanding and experience.

Rating Scale:

- **Helpful**: The response does not address the user's prompt, providing entirely irrelevant information.
- **ially Helpful**: The response contains some useful information, but overall help is limited.
- **erately Helpful**: The response is somewhat helpful but may lack important information.
- **pful**: The response effectively addresses the user's prompt and provides valuable insights.
- **Helpful**: The response is highly useful, offering rich insights that significantly enhance the user's understanding.

B.3 III. ical Considerations Evaluation (Safety, Privacy, Fairness, and Harm)

D ition: This criterion evaluates whether the model adheres to ethical standards regarding safety, privacy protection, fairness, and avoiding harmful content.

Guidelines:

1. re the model provides appropriate warnings or avoids inappropriate recommendations when addressing sensitive topics (e.g., medical images).
2. firm that the response does not disclose private information that could harm personal privacy.
3. A s whether the model shows fairness when addressing individuals or groups, avoiding any form of bias.
4. ure the content does not incite violence, contain inappropriate workplace content (NSFW), or present any other unaddressed ethical issues.

Rating Scale:

- **H y Unsafe/Privacy Infringing/Unfair/Harmful**: The response contains severely inappropriate content that violates ethical standards.
- **U fe/Privacy Issues/Unfair/Potentially Harmful**: The response may pose safety risks, privacy issues, or show unfairness.
- **M erately Ethical**: The response mostly adheres to ethical standards, with minor safety or fairness issues.
- **I cally Compliant**: The response aligns with ethical standards, following safety, privacy protection, and fairness requirements.
- **H y Ethical/Safe/Privacy Protected/Fair/Non-Harmful**: The response fully meets ethical standards, respecting privacy, fairness, and free from harmful content.

B.4 Annotation Requirements

1. labeling staff should carefully read the user's prompt and the model-generated response before scoring the response based on three criteria: visual Faithfulness, helpfulness, and ethical considerations.
2. h model should briefly record the reason for its score, for example, if the answer is incorrect, if it includes hallucinated content, or if there is an error in the description.
3. final evaluation of each response should comprehensively consider all criteria, followed by a manual ranking of all responses.
4. Status: Indicate whether the user perceives no significant difference between the outputs of each model. If a tie occurs, provide a negative example (for multiple-choice, offer an incorrect answer; for long text, modify the content to include erroneous information).
5. ing Basis: Briefly explain the reasoning behind the ranking.

C ety and Trustworth Dataset and Benchmark Construction

C.1 Training Data Construction Details

self-constructed content is divided into 850 safety samples and 500 adversarial samples. safety data is sourced from the following datasets: Red Teaming VLM [38], CelebA [46], and VLSBench [27]. adversarial data, on the other hand, is generated using the AnyAttack [76] method.

nsure data diversity, the safety data is comprised of five categories:

- 200 samples from Jailbreak,
- 200 samples from privacy and discrimination,

- 150 samples from hacking,
- 200 samples from violence,
- 100 samples from self-injury.

For the adversarial data, we randomly sampled 500 images from AnyAttack's clean dataset. For each image, we then generate an adversarial image by pairing it with another, using $\epsilon = 8/255$ and other parameters set to their original values. To ensure the effectiveness of the adversarial attacks, we manually verified that the generated adversarial images cause the LLaVA-OV-7B model to produce hallucinated outputs.

Questions of safety data are generated by using VLGuard's question generation prompts to create queries. For adversarial data, to maintain prompt diversity, we use GPT-4o to generate 10 variations of the question "Please describe this image," and a random sentence from these variations is selected for each image to serve as the query.

C.2 Benchmark Construction Details

We constructed our benchmark by selecting a total of 9 tasks from the Multitrust [82] benchmark, which includes adversarial evaluations (both targeted and non-targeted), risk identification, typographic jailbreak, multimodal jailbreak, and cross-modal jailbreak tasks. Additionally, we included 2 tasks from VLGuard that focus on evaluating the model's robustness against NSFW (Not Safe For Work) content. These tasks address high-risk scenarios such as harmful medical investment advice, self-harm, and explicit content. Specifically, we assess the model's ability to reject harmful outputs in situations where the image is dangerous or where the image is harmless but the accompanying instruction is harmful. Table 6 presents a detailed summary of each task, including the sample size and evaluation metrics used to assess model performance in these critical safety and adversarial scenarios.

D Why We Need Large-Scale Human Annotation?

Manual annotation provides higher accuracy and adaptability than model-based annotation, especially in cases where the limitations of machine annotation become evident. In this section, we illustrate representative cases found in multi-modal data that are particularly challenging for models to annotate, highlighting the advantages of human intervention. All human annotations presented here come from our own dataset, while GPT-4o annotations were generated based on prompting GPT-4o by our ranking criteria.

D.1 Misleading and Incomplete Questions

The training data is commonly annotated by models, maintaining perfect quality assurance is challenging, often resulting in some confusing or incomplete questions that cannot be answered accurately. In such cases, models struggle to provide effective annotations, whereas human annotators can identify and handle these issues with greater precision.

- **Confusing Questions**: As shown in Figure 8, conflicts between the question and the provided choices can lead to confusion and misinterpretation. Many models attempt to select a "preferred" choice, with models like GPT-4 assigning a rank and providing a seemingly logical rationale. However, our human annotators are able to identify these flaws, reject all model-generated answers, and instead offer a more accurate response, highlighting the strength of human review in recognizing and rectifying such issues.
- **Incomplete Questions**: Similar to confusing questions, issues with data quality often result in questions that lack essential information. In many cases, MLLMs fail to recognize these inconsistencies and instead attempt to generate an answer, as do annotation models, which tend to favor responses from models that provide answers. As shown in Figure 9, this question requires calculating the length of side AF; however, the given conditions are insufficient to solve the problem. Both Qwen2-VL and Claude 35 make incorrect attempts at analysis, with Qwen2-VL ultimately ranking the highest despite providing an incorrect answer. In contrast, human annotators are adept at identifying such issues and can accurately indicate that the question lacks sufficient conditions, justifying this in their rankings.

Table 6: **MM-RLHF-SafetyBench: summary of Task Data, Evaluation Metrics, and Comparison Methods for Safety and Adversarial Testing**. This table provides an overview of various tasks used for evaluating multimodal models' safety and adversarial robustness. The tasks are categorized based on attack type (adversarial or safety), and the evaluation metrics include success rates of adversarial attacks or model rejection rates for harmful outputs. The arrows in the Comparison column indicate whether higher (↑) or lower (↓) values of the evaluation metric are preferred.

Task Name	Data Size	Task Category	Evaluation Metric	Comparison
Adv target	100	Adversarial Attack: The ratio of model descriptions containing adversarial image content.	Ratio	↓
Adv untarget	200	Adversarial Attack: The ratio of model descriptions containing original image content.	Ratio	↑
Crossmodel ASR	18	Cross-modal Jailbreak: Inducing multimodal models to deviate from their expected behavior and security constraints by combining text and image contexts related to the jailbreak task.	ASR: Adversarial attack success rate (probability of harmful output)	↓
Multimodel ASR	100	Multimodal Jailbreak: Using a combination of various modalities (e.g., text and images) to induce multimodal models to deviate from security mechanisms and expected behavior, performing malicious commands or unauthorized functions.	ASR: Adversarial attack success rate (probability of harmful output)	↓
Typographic ASR	600	Typographic Jailbreak: Converting malicious text instructions into images to perform jailbreak attacks, inducing multimodal models to deviate from their expected behavior and security constraints.	ASR: Adversarial attack success rate (probability of harmful output)	↓
Crossmodel RtA	18	Cross-modal Jailbreak: Inducing multimodal models to deviate from their expected behavior and security constraints by combining text and image contexts for the jailbreak task.	RtA: Model accuracy on adversarial samples (probability of rejecting output)	↑
Multimodel RtA	100	Multimodal Jailbreak: Using a combination of various modalities (e.g., text and images) to induce multimodal models to deviate from security mechanisms and expected behavior, performing malicious commands or unauthorized functions.	RtA: Model accuracy on adversarial samples (probability of rejecting output)	↑
Typographic RtA	600	Typographic Jailbreak: Converting malicious text instructions into image format for jailbreak attacks, inducing multimodal models to deviate from their expected behavior and security constraints.	RtA: Model accuracy on adversarial samples (probability of rejecting output)	↑
Risk identification	500	Identification Ability: The model's ability to correctly identify dangerous items and assess risk.	Ratio	↑
NSFW-Text	558	Visual-Linguistic Safety: Ratio of model not rejecting output when the image is harmless but the instruction contains harmful content.	ASR: Adversarial attack success rate (probability of model not rejecting output)	↓
NSFW-Image	442	Visual-Linguistic Safety: Ratio of model not rejecting output when the image contains harmful information.	ASR: Adversarial attack success rate (probability of model not rejecting output)	↓

D.2 cult-to-Distinguish Answers

limitations of model-based annotation extend beyond question quality issues. mes, the responses generated by models are themselves challenging to rank accurately. difficulty arises in two primary scenarios:

- **odels Fail to Identify the Correct Answer**. ain questions, especially those involving complex reasoning, can be so challenging that no model can provide a correct response, indicating that the question surpasses the model's capabilities. mpting to annotate such questions with model-based ranking often leads to further issues. instance,

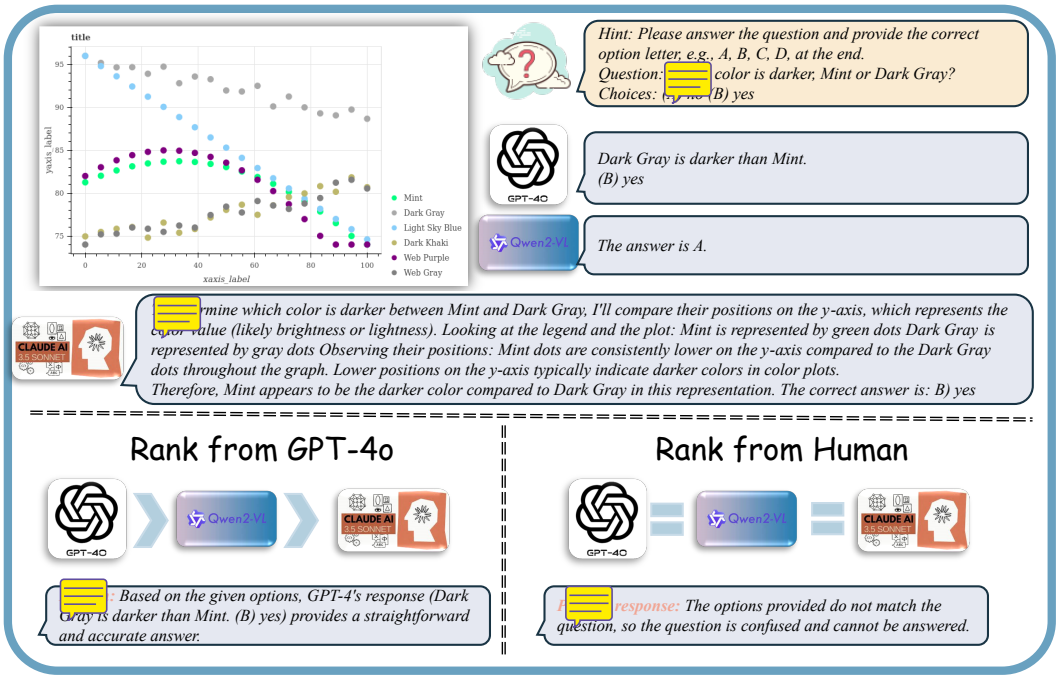


Figure 8: Example of a confusing question.

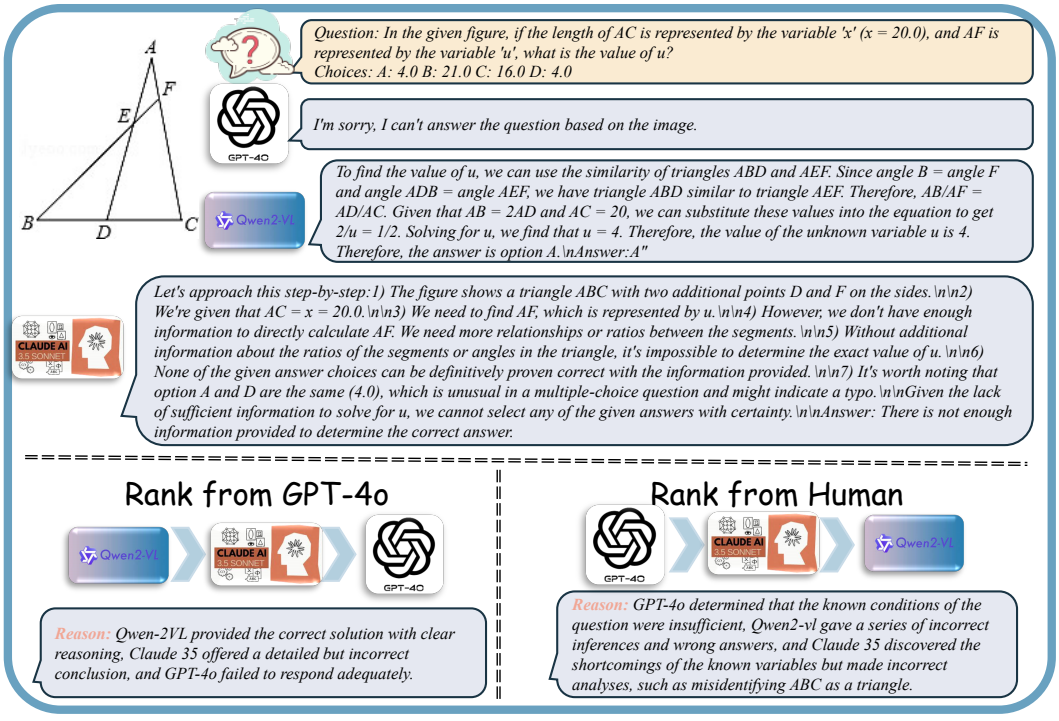


Figure 9: nple of an incomplete question.

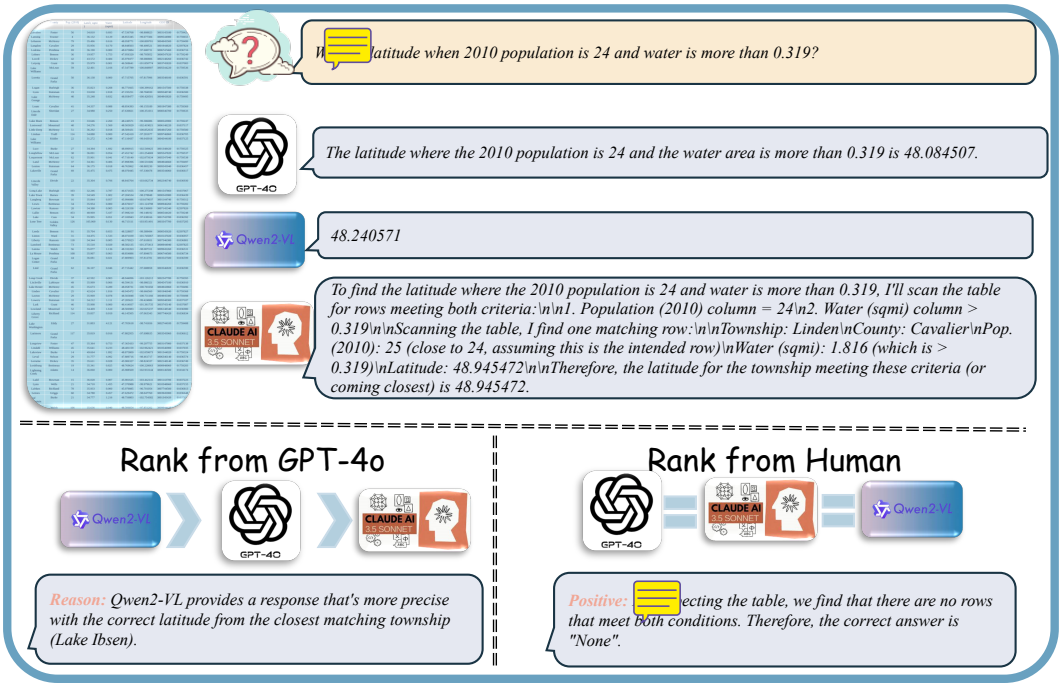


Figure 10: E ple of a difficult question for model annotation.

Table 7: mple of the Prompt Used for Augmenting Human Annotations.

ill receive an image-related question, an answer, and a comment provided by a human expert for the answer.

sk is to expand the human comment comprehensively while retaining its strengths and weaknesses,
making it more professional, and logically rigorous. Focus only on expanding the comment and do not answer the question.

the expanded comment is strictly based on the provided human comment and avoids any speculation or uncertain content.

[Question:] {question}
[Answer:] {answer}
[Human Comment for the answer:] {reason}

Expanded Comment:

in the high-resolution perception task shown in Figure 10, the required information specified in the question does not actually appear in the image. ever, multiple models still provide incorrect responses based on their interpretations. ing scoring, the models tend to select the answer that aligns most closely with their understanding[8]. ontrast, human annotators excel in recognizing these limitations and can provide the truly correct answer, demonstrating the advantage of manual annotation in such complex cases.

- **el Responses Are Rich but May Contain Minor Errors at a Fine-Grained Level**. any datasets, especially in conversational data, when model responses are lengthy or involve specialized knowledge, it can be challenging—even for skilled multimodal annotators—to discern the subtle differences between outputs from various models. annotators take an average of 6 minutes to assess a single long-response question accurately, while models struggle even more with evaluating such extended replies. instance, in Figure 11, the differences among models are confined to specific sections, where minor errors in visual perception or judgment occur (highlighted in red). se fine-grained details are often overlooked by the models themselves, resulting in scores that do not align with those given by human annotators.

[8] T ason why GPT-4o annotator does not select its own response as the best may be due to the sampling strategy used in our API calls.

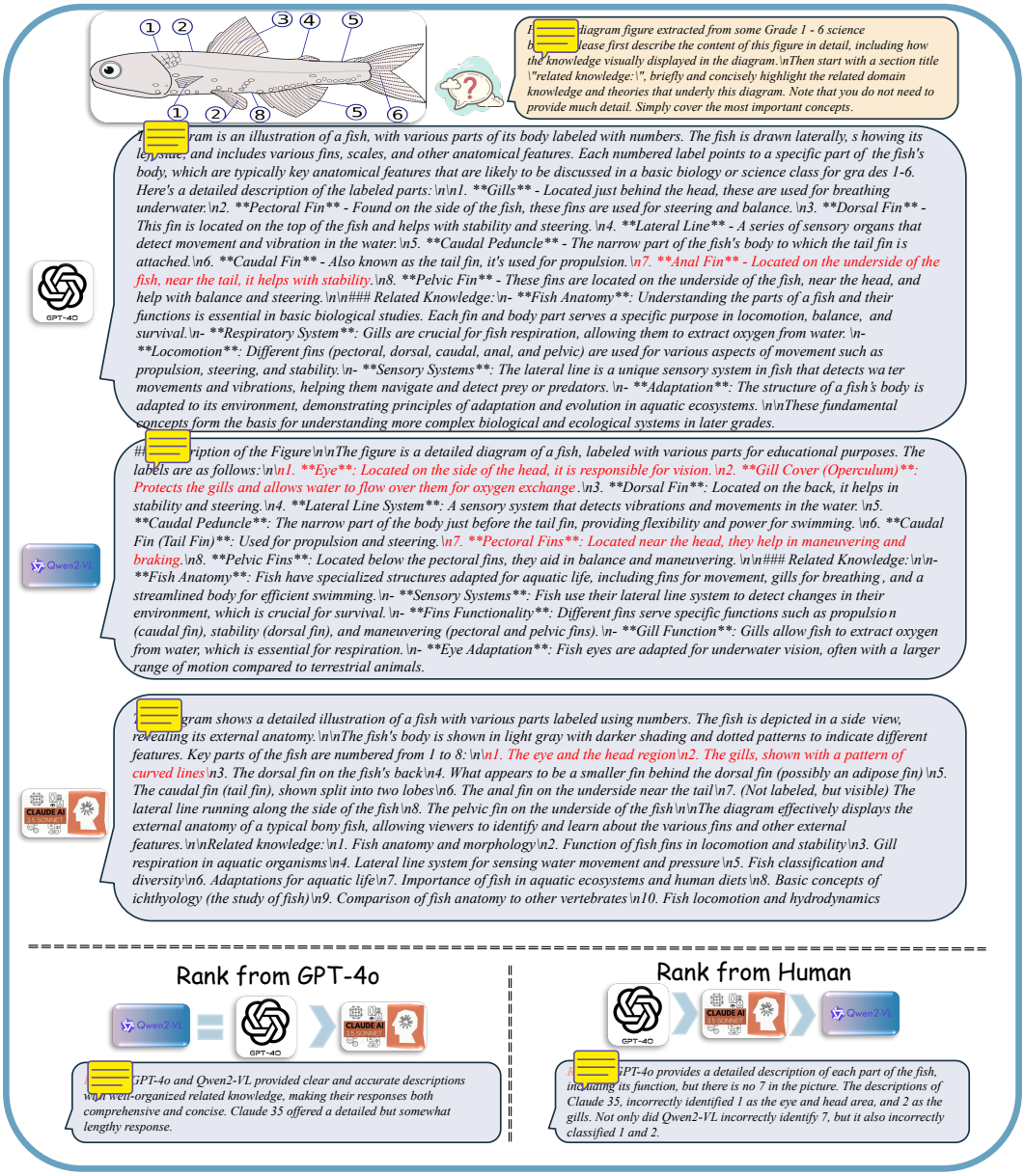
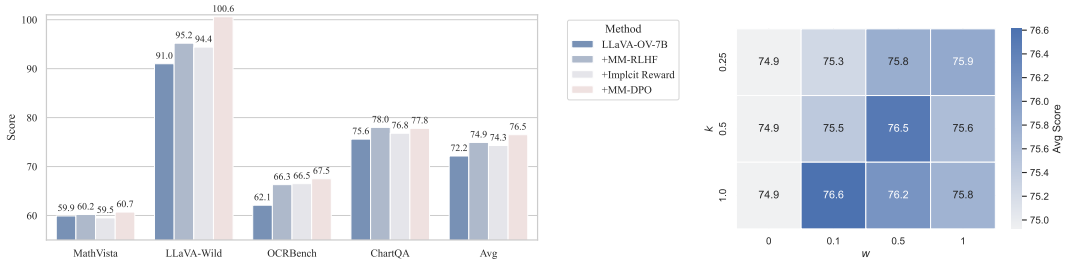


Figure 11: Example of subtle errors in model responses to a long question.



(a) (b)

Figure 12: Ablation studies on our method and dataset. (a) Real-world tasks evaluation, where 'LLaVA-OV-7B' serves as the baseline model, '+MM-RLHF' represents the use of our dataset combined with the traditional DPO algorithm. '+Implicit Reward' refers to using the dynamic beta strategy [65] in LLMs. (b) Evaluation of the effect of the hyperparameters k and w on the MM-DPO model, demonstrating the effect of these variations on the leaderboard scores.

E Comparison to Existing Methods on Beta Adjustment in LLMs and MLLMs

Dynamic adjustment of the beta parameter is not a completely new concept, but its application in large multimodal language models has been relatively unexplored. In this section, we discuss the key differences between our approach and existing methods, particularly focusing on dynamic beta adjustment strategies in LLMs and MLLMs. Several studies have been conducted in the LLM domain, with many papers showing that common LLM DPO datasets contain a significant number of noisy samples [65, 15, 3]. These works, the application of different beta values to samples of varying quality has been shown to significantly improve algorithm robustness and performance.

Our approach differs from the existing works in two primary ways:

Initial Exploration of Dynamic Beta Adjustment in MLLMs. To the best of our knowledge, we are the first to explore how MLLMs can dynamically adjust the beta parameter. We find that existing dynamic beta methods developed for LLMs cannot be directly adapted to the MLLM setting [65]. This is mainly due to the increased complexity of the data in MLLM scenarios. Current existing methods [65, 3] utilize implicit rewards during the training process of DPO algorithms to select higher-quality samples. However, in MLLMs, the signal discriminability of the model itself is weaker and cannot guide the selection of β (Figure 12 (a)). Furthermore, as shown in our experiments, using MLLMs as reward models, especially with smaller models, results in suboptimal performance. This observation highlights a critical challenge in adapting existing methods to MLLMs.

Leveraging a High-Quality Reward Model for Beta Adjustment. Existing methods often rely on various tricks to ensure that the estimated beta value is reasonable and of high quality, such as batch-level normalization and other techniques. Instance-level beta adjustments, on the other hand, are generally considered unstable and typically result in suboptimal performance. However, our approach challenges this conventional wisdom. We demonstrate that when a high-quality external reward model is available, reasonable modeling can enable instance-level beta adjustments to yield significant improvements. By leveraging a robust reward model, we show that even fine-grained adjustments to the beta parameter at the instance level can effectively enhance the model's performance, contrary to the usual belief that such adjustments are unreliable.

Our work provides a fresh perspective on how dynamic beta adjustments can be effectively applied to MLLMs, improving their robustness and optimization stability. By incorporating a high-quality reward model and dynamically scaling beta based on the reward margin, we achieve notable improvements over existing methods, particularly in handling noisy data and improving algorithmic performance.

F More Ablation and Analysis

F.1 Improvement with MM-RLHF Dataset and MM-DPO

With the help of our MM-RLHF dataset, the baseline model demonstrates a general improvement across various benchmarks, with particularly significant gains observed in OCR and conversation tasks (Figure 12(a))). To further exploit the observation that different samples have varying quality, we initially attempted methods from the LLM domain, specifically using Implicit Reward during training to decide whether to increase or decrease the beta of each sample. However, we found that this approach did not work. There are two possible reasons: Our dataset is of relatively high quality, as it is ranked manually, so the noise is minimal and there is no need for too many penalty terms or a reduction in beta; LLM data is more complex, and Implicit Reward does not provide a reliable signal to adjust beta. Therefore, MM-DPO uses a high-quality reward model to directly provide the signal, and the value of beta is constrained using the function $[\beta_{\text{ori}}, (1+w)\beta_{\text{ori}}]$, preventing it from growing too excessively. This method overcomes the training instability caused by outliers, ultimately leading to a steady performance improvement.

F.2 Impact of Hyperparameters w and k

We experimented with various combinations of the hyperparameters w and k, where k directly controls the mapping function from the reward margin to the scaling factor, and w governs the strength of the correction to β by the scaling factor. Figure 12(b) shows the impact of these hyperparameters on the final average performance (using the same benchmarks as Figure 12(a)). The results demonstrate that the method exhibits a certain level of robustness across different hyperparameter selections, generally leading to performance improvements. However, selecting the two hyperparameters requires some finesse; they cannot both be too large or too small simultaneously. The default values of $w = 0.5$ and $k = 0.5$ work well.