Multi-modal Preference Alignment Remedies Degradation of Visual Instruction Tuning on Language Models

Anonymous ACL submission

Abstract

In production, multi-modal large language models (MLLMs) are expected to support multi-turn queries of interchanging image and text modalitie. However, the current MLLMs trained with visual-question-answering (VQA) datasets could suffer from degradation, as VQA datasets lack the diversity and complexity of the original text instruction datasets which the underlying language model had been trained with. To address this challenging degradation, we first collect a lightweight (6k entries) VQA preference dataset where answers were annotated by Gemini for 5 quality metrics in a granular fashion, and investigate standard Supervised Fine-tuning, rejection sampling, Direct Preference Optimization (DPO), and SteerLM. Our findings indicate that the with DPO we are able to surpass instruction-following capabilities of the language model, achieving a 6.73 score on MT-Bench, compared to Vicuna's 6.57 and LLaVA's 5.99 despite small data scale. This enhancement in textual instruction proficiency correlates with boosted visual instruction performance (+4.9% on MM-Vet, +6% on LLaVA-Bench), with minimal alignment tax on visual knowledge benchmarks compared to previous RLHF approach. In conclusion, we propose a distillation-based multi-modal alignment model with fine-grained annotations on a small dataset that reconciles the textual and visual performance of MLLMs, restoring and boosting language capability after visual instruction tuning.

1 Introduction

Recent advancements in artificial intelligence have led to the rise of multi-modal large language models (MLLMs), which combine textual and visual interpretation capabilities in a single model (Shen et al., 2023). However, effectively blending multimodality in one system has proven non-trivial. The integration of diverse data forms often creates internal representation conflicts, giving rise to the issue known as "catastrophic forgetting" (Kirkpatrick et al., 2017). The diversity constraint in visual question answering (VQA) datasets could be attributed as a source of the issue. VQA tasks typically focus on descriptive queries about image contents, whereas textual datasets encompass a broader range of complex cognitive tasks, including reasoning, writing, summarization, and coding. This discrepancy in dataset complexity is a key factor contributing to the observed performance degradation in MLLMs. Our evaluation of models such as BLIP-2, InstructBLIP, and LLaVA against language instruction-following benchmarks like MT-Bench (Zheng et al., 2023) and AlpacaEval (Li et al., 2023b) revealed diminished language capabilities in comparison to their linguistic backbones. For instance, LLaVA, built on the Vicuna-13b LLM, demonstrated a decline in MT-Bench performance from 6.57 to 5.92, even underperforming the Vicuna-7B model.

Driven by the limitations observed in distillation-based instruction tuning, particularly its constrained generalizability and the narrow performance improvements on tasks outside the training distribution, this study investigates the efficacy of distillation-based preference alignment in addressing modality conflict in MLLMs. The decision to explore this avenue is predicated on the hypothesis that integrating AI-generated preference data can provide a more granular and nuanced alignment with human expectations, potentially mitigating the adverse effects of modality conflict.

This study rigorously evaluates three baseline methodologies—Direct Preference Optimization (DPO), SteerLM, and Rejection Sampling—as potential solutions to utilize the distilled preference data and enhance the instruction-following capabilities and address the modality conflict inherent in MLLMs. Each of these methods offers a unique approach to model alignment, from the direct optimization of preferences in DPO to the conditional supervision in SteerLM and the selective acceptance in Rejection Sampling. Our empirical analysis reveals that DPO, in particular, demonstrates a pronounced efficacy in reconciling the performance disparities observed between textual and visual modalities. By leveraging a refined preference dataset, fine-tuned with the DPO objective and supplemented with comprehensive annotations from advanced AI models, DPO not only addresses the modality conflict but also significantly enhances the model's performance across a spectrum of benchmarks. The results indicate that, through the application of DPO, MLLMs can achieve a more robust alignment with human-like preferences, thereby mitigating the adverse effects of catastrophic forgetting and modality conflict, and elevating the models' capabilities to a level that surpasses traditional instruction tuning methods.

Our main contributions are:

1. **Exploration of Modality Degradation:** This work is at the forefront of identifying and addressing modality degradation in MLLMs, a phenomenon where visual instruction tuning detrimentally impacts language instruction capabilities. Our systematic investigation into this issue contributes novel insights to the field, laying the groundwork for further research in mitigating such degradation.

2. **Innovative Preference Alignment Methodology:** We propose a groundbreaking preference alignment framework that not only mitigates the negative effects of visual instruction tuning on text-based tasks but also enhances the MLLM's performance beyond its original language model backbone. This method also demonstrates significant improvements in visual instruction benchmarks, underscoring the efficacy of preference alignment in augmenting MLLM functionalities.

3. **Efficient Data Annotation Scheme:** Our data collection strategy employs a granular quality metric annotation format, leveraging cost-effective commercial APIs. This scalable approach enables the efficient production of high-quality datasets, addressing a critical challenge in MLLM development and facilitating extensive model training and refinement.

2 Related Work

2.1 MLLMs and Visual Instruction Tuning

Incorporating another modality into large language models represents a natural evolution for these systems. Modality expansion can be achieved through system-level enhancements at inference time, with approaches such as Mm-react (Yang et al., 2023), Visual ChatGPT (Wu et al., 2023), and HuggingGPT (Shen et al., 2023) enabling the LLM to invoke off-the-shelf vision models and APIs. An alternative strand of research involves the training of end-to-end MLLMs. To avoid the prohibitive costs associated with pre-training from scratch, these models often integrate pre-trained vision models with large language models, applying various degrees of modality adaptation. Mini-GPT4 (Gong et al., 2023) focuses solely on training a linear projection matrix to connect CLIP-based (Radford et al., 2021) vision representations with the LLaMA model (Touvron et al., 2023a); BLIP-2 introduces a cross-attention module to extract vision tokens relevant to the query. Both LLaVA (Liu et al., 2023c) and mPlug-OWL (Ye et al., 2023a) feature cross-modality connectors between the vision and language domains, but they also fine-tune the LLM and vision encoder, respectively. Flamingo (Alayrac et al., 2022), in contrast, incorporates new cross-attention layers directly into the LLM.

In the language domain, Wei et al. (2022) discovered that fine-tuning a base LLM with instructions described in natural language enhances the model's ability to follow those instructions. In a similar vein, MLLMs are typically fine-tuned with instructions; Mini-GPT4 (Gong et al., 2023) utilized template instructions based on image-text pairs, while InstructBLIP (Dai et al., 2023), Otter (Li et al., 2023a), and LLaVA (Liu et al., 2023c) employed human-written visual question-answers or synthetically generated question-answer pairs by prompting GPT-4 with COCO captions and bounding boxes (Liu et al., 2023c). However, considering that both LLaVA and Instruct-BLIP utilize Vicuna (Chiang et al., 2023)—an instruction-tuned LLaMA—it remains a topic of debate whether their steps of visual instruction tuning genuinely add to the model's instruction-following capabilities or merely conform to the instruction-following format used in Vicuna's training.

2.2 Mitigating Modality Conflict in MLLMs

To preserve the ability to follow language instructions, mPLUG-OWL (Ye et al., 2023b) and LLaVA 1.5 (Liu et al., 2023c) incorporate language-only instruction data back into their mixed visual-language instruction datasets, specifically ShareGPT. It is noteworthy that the LLM backbone of LLaVA 1.5, Vicuna, had been previously trained on this identical ShareGPT dataset. Further investigation reveals that, despite the integrated dataset, LLaVA 1.5 exhibits degradation in language instruction-following capabilities; the MT-Bench score for LLaVA-1.5-13b is notably lower than that for Vicuna-V1.5-7b.

While mPLUG-OWL-2 (Ye et al., 2023b) presents promising solutions to the challenges of modality conflict, and has shown superior performance on text-based benchmarks, it also introduces increased parameter count and more complex implementation, which poses practical challenges. Specifically, mPLUG-OWL-2 implements modality-adaptive modules that include distinct layer normalization, as well as separate key and value projection matrices for text and visual tokens, whilst maintaining a shared structure for query projection matrices. The model is fine-tuned from the non-instruction-tuned LLaMA-2-7B base model, incorporating 548K textual instruction data samples from a total of 1.2M, derived from both SlimOrca (Lian et al., 2023) and ShareGPT. This approach has enabled mPLUG-OWL-2 to excel in language and visual-language tasks, outperforming RLHF-augmented LLaMA-2 Chat on text-centric evaluations such as MMLU and BBH (Ye et al., 2023b). However, the introduction of modality-specific modules has led to an increase in the model's parameter count from 7.2 billion to 8.2 billion. Additionally, the separate processing paths for visual and language tokens have resulted in a more intricate compute graph, complicating the utilization of fused GPU kernels to achieve efficient inference. The methodologies discussed so far are reliant on standard fine-tuning practices, necessitating a significant augmentation of computational resources to integrate text instruction data effectively, with the aim of resolving the challenges posed by modality conflict.

2.3 Distillation-based Instruction Tuning

Leveraging the output of large proprietary models, smaller open-source models such as Vicuna (Chiang et al., 2023), Alpaca, and more recently ShareGPT4V (Chen et al., 2023), have been fine-tuned, although this approach has limitations in terms of generalization capabilities. Gudibande et al. (2023) observed that models fine-tuned through instruction tuning by imitation barely bridge the performance gap in tasks beyond the scope of the training data. They contend that imitation as a strategy is a *false promise*, asserting that only a significant volume of imitation data or a larger base model can close the disparity between open and closed-source models (Gudibande et al., 2023). While recognizing the utility of the expansive GPT4V dataset like ShareGPT4V (Chen et al., 2023), it is posited that such scaling of distillation-based instruction tuning primarily extends the model's competency within the distribution it was trained on rather than its out-of-distribution generalizability. Further research indicated that distillation instruction tuning on a smaller scale tends to skew the model's performance towards a niche subset and significantly impair its broader applicability. This was evidenced by a baseline experiment in which fine-tuning LLaVA with a 6k VQA dataset, sourced from Gemini Pro-generated answers, resulted in pronounced performance declines across both textual and visual benchmarks.

2.4 Preference Alignment

The Instruct-GPT series (Ouyang et al., 2022) has shown that merely employing supervised fine-tuning (SFT) on Large Language Models (LLMs) is insufficient for aligning them with human preferences. The technique of Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) addresses this by constructing a reward model that encapsulates human preferences and then applying reinforcement learning to maximize this reward. The Direct Preference Optimization (DPO) approach posits that directly tuning the preference dataset can serve as an effective substitute for reward modeling, offering the added benefit of reduced computational complexity. Another novel method, known as rejection sampling SteerLM, has recently been identified to achieve performance akin to RLHF by incorporating human-annotated quality metrics before generation, serving as a conditional SFT-based strategy for alignment (Dong et al., 2023). Our experiments with DPO, SteerLM, and rejection sampling reference the prior work on LLaVA-RLHF (Sun et al., 2023), using it as a benchmark for RLHF performance.

2.5 Distilling AI Feedback for Preference Alignment

In the realm of alignment methods, reliance on human-annotated preference annotations is common. While effective on a large scale, this approach incurs substantial costs and operational complexities (Touvron et al., 2023b). The effectiveness of reward models based on pairwise ranking is constrained by the inherent subjectivity of human preferences, with LLaMA's reward model achieving an accuracy range of 64.3-70.6%, and the LLaVA-RLHF model reaching 67%. In response to these limitations, Zephyr (Tunstall et al., 2023) and Ultra-Feedback (Cui et al., 2023) have utilized preference annotations distilled from GPT-4 to train models with 7B parameters, achieving performance levels comparable to those of 70B parameter models. Motivated to adapt this distillation-preference alignment approach for MLLMs, our work introduces a nuanced chain-of-thought prompting technique, coupled with a detailed annotation guide, spanning five assessment metrics.

3 Method

The focus of this study is to investigate improving the text instruction following of MLLMs while retaining or potentially enhancing their multi-modal reasoning. For this purpose, we propose to harness alignment methods that utilize self-sampled responses and preference annotations. Therefore, Pure Supervised Fine-Tuning (SFT) is deemed unsuitable and thus excluded from our methodology.

3.1 Data Collection

Sources of Multi-modal Prompts. We have curated 3,000 samples from SciGraphQA (Li and Tajbakhsh, 2023) and an equal number from LRV-Instruct (Liu et al., 2023a) to assemble the image-text prompts for our multi-modal preference dataset. The LRV-Instruct dataset is a visual instruction dataset aimed at mitigating hallucinations by incorporating both positive prompts (inquiring about objects present in the image) and negative prompts (requesting information about absent objects) (Liu et al., 2023a). The inclusion of negative examples encourages the multi-modal LLM to critically evaluate the prompts and identify instances where the requested objects are not present. The LRV-Instruct dataset enhanced its performance in reducing hallucinations, as evidenced by its application in the Pope framework (Liu et al., 2023a).

Granular Annotation: We generate 4 completions for the dataset above using LLaVA-1.5-13B with a temperature of 0.7 and then prompt Gemini-Vision (Team et al., 2023) with the labeling guide of HelpSteer gave to Scale AI workers (Wang et al., 2023), images, questions, and the 4 completions, obtaining granular annotations for various quality metrics, including helpfulness, correctness, and coherence, providing multifaceted insights into the model's outputs. We leveraged Zero-Shot Chain-of Thought prompting (Kojima et al., 2022) such that the Gemini gives the reasoning for rating each response, an inner calibration monologue, and ratings as shown in 1. Appendix provides two annotation examples on the LLaVA-RLHF dataset (Sun et al., 2023) such that we can visualize how Gemini reasons and rate each metric compared to a binary crow-sourced worker-provided preference.

We selected Gemini Pro (dated 01/01/2024) for its performance, cost efficiency, and bias mitigation (Team et al., 2023). Leading the OpenCompass multi-modal leaderboard with an average rank of 1.89, outperforming GPT-4V's 2.89 (ope, 2023), its generous free tier supports extensive data collection. Crucially, using GPT-4(V) for both benchmark judging and data labeling could introduce bias, potentially skewing our models towards GPT-4's preferences. This consideration led us to opt for Gemini to ensure a more objective evaluation of our fine-tuning efforts.

3.2 Alignment Methods

Self-sampled SteerLM: SteerLM, a conditional Supervised Fine-Tuning (SFT) technique, aligns LLMs by augmenting prompts with a description of the desired response quality, as introduced by (Dong et al., 2023). This method conditions SFT on granular annotation generated by Gemini and surpasses traditional SFT and alignment strategies. We construct a conditional prompt template for this technique by incorporating a conditional prompt following HelpSteer guidelines. For instance,

Rejection Sampling: Following Constitutional AI (Bai et al., 2022) and LLaMA-2 (Touvron et al., 2023b), we adopt a simplified rejection sampling approach. Specifically, we select the top-scoring response from the four completions described above, based on Gemini's aggregated scores for helpfulness, correctness, and coherence. And we apply standard SFT based on the selected responses with-

Metric	Description
Helpfulness	Measures if the response fully addresses the prompt's request.
Correctness	Assesses the accuracy and relevance of the information, ensuring it's free from errors or misinformation.
Coherence	Evaluates the clarity and logical consistency of the response.
Complexity	Looks at the level of sophistication in the language used, from simple to advanced.
Verbosity	Considers the brevity or lengthiness of the response in relation to the prompt's needs.

Table 1: Granular annotation format and labelling guide proposed in HelpSteer (Wang et al., 2023) used in VQA annotation collection. LLaVA-1.5-13b generates 4 candidates. Using the HelpSteer labeling guide, images, and questions, Gemini rates each completion with a score of 0-4 in each metric.

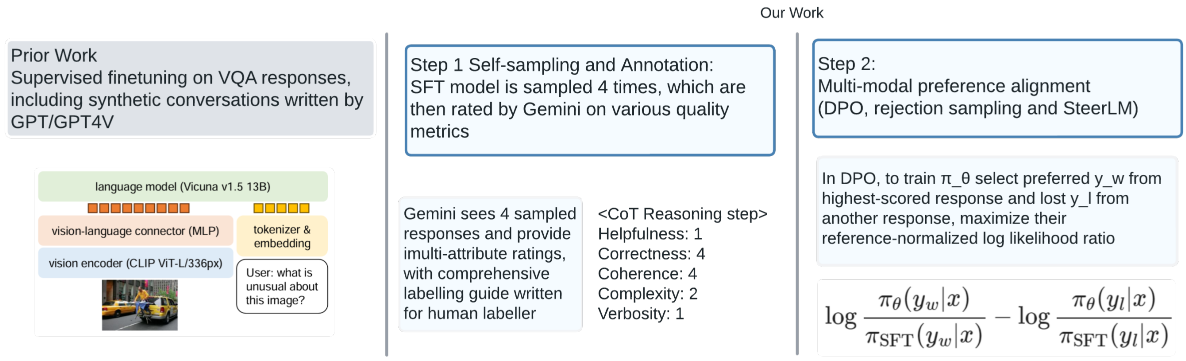

Figure 1: Starting from an SFT-ed checkpoint, we generate 4 completions for a given image-question prompt. These answers are then presented to Gemini to obtain granular annotations given a labeling guide. We construct a preference dataset of (image-text prompt, preferred completion) and (image-text prompt, rejected completion). We benchmarked DPO, Rejection sampling, and SteerLM alignment methods, in addition to a pure SFT baseline using Gemini provided answer directly

out any additional prompt conditions, following a straightforward format: (image, prompt, best response).

Direct Preference Optimization (DPO): We convert our dataset of granular annotation into a preference set by selecting the highest score and the worst response. Specifically, we generate an aggregated score per response by summing Helpfulness, and Correctness. When prompting Gemini for annotations, Gemini reviews four responses and provides absolute quality metrics, which we converted into a preference dataset. We form (preferred, rejected) pairs by selecting the best response and randomly selecting another from the four responses, with a filter specifying that the preferred response is at least 2 points lower in summed scores across helpfulness, correctness, and coherence than the rejected response.

SFT from Gemini Responses This is an important baseline not using self-sampling, but using answers from Gemini directly for pure SFT. Gemini was prompted with the questions and images from our dataset. We gather the answer directly from Gemini answering the question and use the same hyperparameters in SteerLM and Rejection Sampling.

4 Experiments

4.1 Experimental Setup

For training, we utilize Azure Cloud (NC-A100 series) with 4 A100-80G GPUs. In all experiments, LoRA is employed for parameter-efficient tuning. For the SFT experiments, including standard SFT with Gemini responses, rejection sampling, and self-sampled SteerLM, we adhered to the same hyperparameters used in LLaVA-v1.5's instruction tuning. In our DPO experiments, we performed a hyperparameter search based on 1000 samples on LLaVA Bench, exploring various values for beta (0.1, 0.2, and 0.3, averaged and non-averaged log probabilities, and learning rates of 5e-5 and 5e-6. We pre-computed the log probabilities of the reference model (LLaVA-v1.5-13b) for our preference dataset before training. Complete list is provided in the Appendix.

Table 2: Performance comparison among alignment strategies. The results demonstrate DPO-13B's superior performance, particularly in reconciling language capabilities while enhancing visual task performance, validating the DPO methodology's efficacy in multi-modal alignment.

Model Name	Visual Instruction Benchmark		Visual Multi-Choice Benchmark		Text Instruction Benchmark	
	MM-Vet	LLaVA-bench	PoPe	MM-Bench	MT-bench	AlpacaEval
Vicuna-1.5-13b (Chiang et al., 2023)	-	-	-	-	6.57	81.4
LLaVA-1.5-13b (Liu et al., 2023c)	36.3	73.1	0.859	67.4	5.99	79.3
LLaVA-RLHF-13b (Sun et al., 2023)	37.2	76.8	0.869	60.1	6.18	81.0
Alignment method we benchmarked, finetuning LLaVA-1.5-13b						
Standard SFT	36.5	63.7	0.850	65.4	5.01	50.2
SteerLM	35.2	67.0	0.878	65.1	5.70	68.8
Rejection-sampling	38.0	70.6	**0.883**	**67.6**	6.22	74.9
DPO	**41.2**	**79.1**	0.870	66.8	**6.73**	**86.4**

4.2 Benchmarks

We have the following three sets of benchmarks.

Visual Instruction Benchmarks

- **MM-Vet**: MM-Vet measures six core visual-language capabilities across 128 tasks, providing a comprehensive evaluation of multi-modal understanding. It blends math, reasoning, and visual knowledge. (Yu et al., 2023).

- **LLaVA-Bench**: LLaVA-Bench (in the wild) is a dev benchmark for LLaVA comprising 60 tasks, designed to test visual instruction-following and question-answering capabilities in natural settings (Liu et al., 2023c).

Visual Multi-Choice Benchmarks

- **PoPE**: An object hallucination benchmark with 8,440 samples, aimed at evaluating the model's ability to discern and describe visual content accurately. The multi-modal LLM is prompted to answer yes or no to objects that could appear in the images (Li et al., 2023c).

- **MM-Bench**: Serving as a comprehensive multi-modal benchmark, MM-Bench is a multi-choice visual knowledge and reasoning benchmark with 4.7K samples (Liu et al., 2023d).

Language Instruction-Following Benchmarks

- **MT-Bench**: Utilizing LLM to approximate human preferences with over 80% agreement, MT-Bench focuses on measuring the helpfulness of responses across 160 samples in single-turn and multi-turn settings. (Zheng et al., 2023).

- **Noisy-image-context MT-Bench:** Inspired by (Zhou et al., 2023b) where the language commonsense capability of MLLMs was evaluated by sending non-informative images including the blank and random images to the Multi-modal LLM to assess their language commonsense capability, we propose **MT-Bench with non-informative image context** to assess the language instruction-following capability of MLLM. We evaluate LLaVA, Blip-2, InstructBLIP. Within the theme of this paper, we emphasize multi-modal LLM in production.

- **AlpacaEval**: AlpacaEval leveraged GPT-4 to assess the percentage of cases where the candidate LLM outperforms GPT-3 API (text-davinci-003) across 160 evaluations (Li et al., 2023b).

We run benchmark code once on target model, with greedy decoding unless a benchmark use a different temperature setting.

4.3 Results

Table 2 illustrates a comparative analysis of various alignment methods, including RLHF, Standard SFT, SteerLM, Rejection Sampling, and DPO, aimed at enhancing the language capabilities of the LLaVA model that were compromised post-visual instruction tuning. Vicuna, the base language model of LLaVA is tested on language tasks. After visual instruction tuning, LLaVA experiences a decline in language benchmark scores from Vicuna (from 6.57 to 5.99 on MT-Bench, and from 81.4 to 79.3 on AlpacaEval). However, DPO, among the alignment strategies on the 5k multi-modal dataset, not only mitigates the degradation problem but also

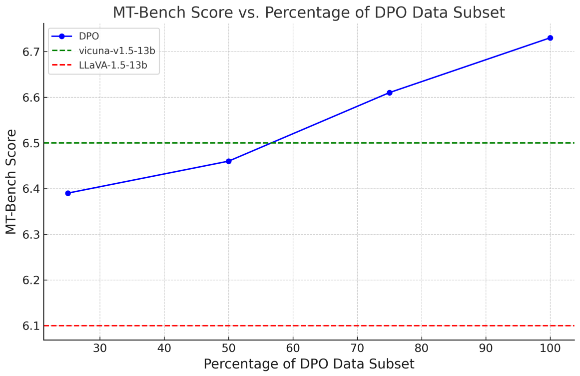

Figure 2: Advances in MT-Bench scores via DPO data scaling

surpasses Vicuna's performance on both benchmarks.

In open-ended visual instruction tasks, DPO significantly outperforms the baseline LLaVA and LLaVA-RLHF models on both the MM-Vet and LLaVA-bench. These open-ended benchmarks, structured similarly to MT-Bench where GPT-4 assesses the responses against a gold standard, revealed a relatively high 0.73 Pearson Correlation between MM-Vet and MT-Bench (n=7, p=0.1).

In visual multi-choice benchmarks, PoPe, which evaluates for hallucination, and MM-Bench, which assesses world knowledge and reasoning, rejection sampling emerged as the most effective method, whereas DPO showed lesser efficacy. While DPO improved open-ended benchmarks, it slightly decreased LLaVA's MM-Bench score from 67.4 to 66.8, contrasting with LLaVA-RLHF's notable drop from 67.4 to 60.1, indicating a less significant alignment tax. Figure 2 illustrates the effect of scaling the DPO dataset on the MT-Bench scores, signifying the efficacy of DPO in advancing the performance of the LLaVA-v1.5-13b model after visual instruction tuning. The baseline at 0% DPO data, marked at a score of 5.99, represents the initial performance of LLaVA-v1.5-13b. As the DPO dataset's size increases, a notable rise in MT-Bench scores is observed, peaking at 6.73 with 100% data utilization. Remarkably, the performance surpasses that of the Vicuna-v1.5-13b benchmark using less than 75% or 4.2K of the DPO data, underlining DPO's data efficiency. This data scaling trend emphasizes DPO's potential as an effective alignment strategy for MLLMs, addressing the challenge of performance degradation due to visual instruction tuning.

5 Discussion

The scarcity, high cost, and inconsistency of existing multi-modal preference datasets present significant challenges to model alignment, as evidenced by our subjective assessments and objective evaluations using Gemini. Our manual labeling of a selected subset from the LLaVA-RLHF human preference dataset underscored the difficulty in achieving consensus or identifying clear preferences, shedding light on the inherent subjectivity of these datasets. To quantify these observations, we embarked on a targeted experiment involving 500 samples, wherein Gemini annotated two distinct responses from the dataset employed in the LLaVa-RLHF reward model's data collection (Sun et al., 2023). This facilitated a direct comparison between Gemini's annotations and the human preferences. The correlation heatmap depicted in Figure 3 reveals a notably weak correlation between human preferences from LLaVA-RLHF and Gemini scores, with correlation coefficients falling below 0.1. This observation was corroborated by further manual inspections, which frequently resulted in an inability to definitively determine clear preferences. This finding underscores the high subjectivity and individual bias within the LLaVA-RLHF preference data. Contrarily, the detailed annotation framework employed by Gemini represents a scalable and more objective method for collecting preference data, offering a viable solution to the limitations observed in current datasets.

The superficial alignment hypothesis states that a model's knowledge capabilities are largely acquired during its initial pre-training stage (Zhou et al., 2023a). A corollary of this hypothesis is that alignment tuning refines the model output generation with a preferred response format rather than knowledge acquisition. As a result, models can be effectively realigned post-visual instruction using a relatively small set of examples (Kirstain et al., 2022). This principle applies to MLLMs as well, which acquire multi-modal knowledge representation via visual instruction tuning (Liu et al., 2023c). However, existing work mixed large-scale text instruction data (518K out of 1.23 million in case of mPlug-OWL 2, and 40K in case of LLaVA-1.5). We hypothesize that the data-inefficiency above is attributed to the underlying alignment strategy and demonstrate that one would need only a small alignment dataset so long as a proper alignment strategy such as DPO is utilized.

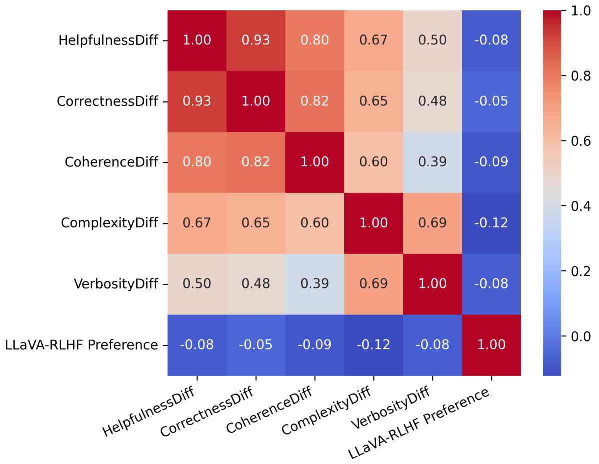

Figure 3: Pearson Correlation Heatmap among the difference in Gemini-Annotated data attributes and LLaVA-RLHF human annotated preference (n=500).

Table 3: This benchmarks adds irreverent image in context when benchmarking MT-Bench, testing for robustness in a real-world condition. Vicuna-7B and Vicuna-13B are as baseline reference.

Model	**LLM**	**Noisy-image MT-Bench**
Vicuna 13B v1.5	Vicuna$_{13B}$	6.57
Vicuna 7B v1.5	Vicuna$_{7B}$	6.17
BLIP-2	FlanT5	1.93
InstructBLIP	Vicuna$_{7B}$	4.73
LLaVA-v1.5-13b	Vicuna$_{13B}$	5.92
DPO (ours)	Vicuna$_{13B}$	6.63

As suggested by Table 2, Direct Preference Optimization (DPO) emerges as a computationally efficient solution for enhancing model performance in the mixed-modal alignment space. Unlike the mixing text instruction as described above or LLaVA-RLHF, which used a large 82K dataset and complex training pipeline involving reward modeling and PPO, DPO achieves significant improvements in language capabilities with a smaller dataset and one-stop training setup. A notable advantage of DPO is its minimal alignment tax, which curtails the degradation of existing knowledge, as evidenced by its performance on benchmarks like MM-Bench, where DPO shows minimal impact. This method not only enables effective alignment of multi-modal models post-visual instruction tuning but also ensures the preservation of model performance. We also note that as 3 showed, our DPO model is most more robust than other baselines in a real-world condition where user may have irrelevant image in the context. Our methodology exhibits notable proficiency in value alignment and data efficiency, yet it is imperative to acknowledge certain limitations and potential risks. One key consideration is the scalability of our approach. While our data scaling analysis suggests significant improvements up to a 6K preference dataset, the full extent of scalability beyond this threshold remains unexplored. As the foundational open-sourced models like LLaVA evolve in complexity and size, the effectiveness of our distillation-based approach might encounter diminishing returns.

Addressing true human preference accurately is another pivotal concern. The preference dataset distilled from artificial annotations may not fully encapsulate the nuanced spectrum of human values, raising ethical considerations regarding model alignment with societal norms and individual expectations. Moreover, the inherent risks associated with safety and bias propagation are magnified when models are fine-tuned on artificially-labeled data, potentially reinforcing existing prejudices.

6 Conclusion

In this paper, we addressed the performance decline of the widely-used multi-modal LLM, LLaVA-1.5, on language instruction-following benchmarks. We investigated various alignment strategies to enhance the model's language instruction capabilities, including standard SFT with rejection sampling, using gold responses from Gemini, SteerLM, and DPO. Our comprehensive benchmark evaluations revealed that DPO stands out as the most effective method, not only restoring but also surpassing the language capabilities of LLaVA's foundational Vicuna model on both MT-Bench and AlpacaEval. Specifically, LLaVA's performance on MT-Bench had dipped from Vicuna's 6.57 to 5.99, whereas our DPO model advanced to 6.73. Beyond language proficiency, DPO also bolstered performance on multi-modal benchmarks, enhancing accuracy by 4.9% on MM-Vet and 6% on LLaVA-Bench. Remarkably, DPO achieved this with just 5k preference examples, proving more data-efficient than other strategies like RLHF, which typically demand over 80k training examples. Our findings highlight DPO's minimal 'alignment tax,' effectively boosting one performance area without substantially detracting from another. This aspect is particularly valuable for practitioners aiming to maintain the pre-training knowledge throughout multi-modal alignment.

References

2023. OpenCompass. [Online; accessed 24. Jan. 2024].

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback.

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

Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *ArXiv preprint*, abs/2305.04790.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Yuval Kirstain, Patrick Lewis, Sebastian Riedel, and Omer Levy. 2022. A few more examples may be worth billions of parameters. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1017–1029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *ArXiv preprint*, abs/2305.03726.

Shengzhi Li and Nima Tajbakhsh. 2023. SciGraphQA: A Large-Scale Synthetic Multi-Turn Question-Answering Dataset for Scientific Graphs. *arXiv*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models.

Wing Lian, Guan Wang, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Slimorca: An open dataset of gpt-4 augmented flan reasoning traces, with verification.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Aligning large multi-modal model with robust instruction tuning. *ArXiv preprint*, abs/2306.14565.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. *arXiv*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua

Lin. 2023d. Mmbench: Is your multi-modal model an all-around player?

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented rlhf.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *ArXiv preprint*, abs/2312.11805.

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

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

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. 2023. HelpSteer: Multi-attribute Helpfulness Dataset for SteerLM. *arXiv*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *ArXiv preprint*, abs/2303.04671.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *ArXiv preprint*, abs/2303.11381.

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

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023b. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. LIMA: Less Is More for Alignment. *arXiv*.

Kankan Zhou, Eason Lai, Wei Bin Au Yeong, Kyriakos Mouratidis, and Jing Jiang. 2023b. Rome: Evaluating pre-trained vision-language models on reasoning beyond visual common sense. *ArXiv preprint*, abs/2310.19301.

A Appendix

A.1 data and model details

Table 4: DPO preference data mixture

Data Type	Data Name	Size
VQA	LRV-Insturct (Liu et al., 2023b)	2562
	SciGraphQA (Li and Tajbakhsh, 2023)	2522
Total		5084

Model	Data Name	Size
Model architecture	Pretrained model checkpoint	llava-v1.5-13b
	Language Backbone Architecture	LLaMA-13B (finetuned to Vicuna and LLaVa)
	Vision Tower	openai/clip-vit-large-patch14-336
	Multi-modal Adapter	2-layer Multi-layer perception
Lora Setting	Lora Rank	128
	Lora Alphra	256
DPO Setting	Use Average Log Probablity	False
	Use Fine-grained DPO	False
	Beta	0.1
	Learning Rate	5e-5
	Gradient Acculumulation Steps	4
	Per device train batch size	8
	Model Max length	300
SFT Setting	Learning Rate	4e-4
	Per device train batch size	16
	Model Max Length	2048
Common settings	Harwdware	4X A100 80G
	Distributed Learning	Zero-3
	Use BF-16	True
	Learning Rate Scheduler	Cosine
	Learning Rate Warm up	0.003
	Weight Decay	False

Table 5: Model parameters

Gemini-Annotation example		
		
Question	What aspects of the room might have artistic or creative significance?	What do you see happening in this image?
Response 1	In this image, the room has some artistic or creative significance due to the presence of several sculptures placed on a long wooden table. The sculptures appear to be made from various materials, such as stone and wood. Additionally, the room features a chair and a light fixture that resembles a boot. These elements, combined with the assortment of sculptures, give the room an eclectic and artistic atmosphere, suggesting that the space might be used for displaying art, inspiration, or as a workspace for an artist or designer. The presence of the boot-shaped light also adds a unique and imaginative touch to the room's overall aesthetic.	The image features a person running along a beach near the water's edge. Behind the person, multiple birds are flying in the sky, creating a dynamic and lively scene. The man is wearing a swimsuit and appears to be enjoying his time at the beach. The birds' positions vary across the sky, with some flying closer to the person and others further away, spreading out across the background.
Gemini-Rating for Reponse 1	Helpfulness:4, Correctness 4, Coherence: 4, Complexity: 3, Verbosity: 3.	Helpfulness:3, Correctness 4, Coherence: 4, Complexity: 2, Verbose: 2
Response 2	In the room, a collection of wooden shoes on a long table next to a chair can be considered as having artistic or creative significance. The display of shoes may represent a unique form of artistic expression or a design concept. Additionally, the white-walled setting with a dim light allows for a minimalist and visually striking atmosphere that enhances the shoes' display and creates a sense of depth and artistic interest in the space.	The image captures a beach scene where a man is running along the shore near the water, with waves crashing behind him. The man is in motion, as he is probably enjoying an invigorating run on the beach. In the sky above, several birds fly in a line, adding a sense of movement and liveliness to the scene. The birds' positions vary, with some being closer to the water and others higher up in the sky. The overall atmosphere of the image suggests a pleasant, active day at the beach.
Gemini-Rating for Reponse 2	Helpfulness:3, Correctness 3, Coherence: 3, Complexity: 2, Verbosity: 3.	Helpfulness:4, Correctness 4, Coherence: 4, Complexity: 3, Verbosity: 3.

Table 6: Examples of Annotation

Gemini-Annotation example		
Gemini-Reasoning	Both responses are relevant to the question and provide good descriptions of the room. However, the first response is more detailed and provides more analysis of the potential artistic or creative significance of the room. I will give the first response a slightly higher rating for Helpfulness, Correctness, and Complexity.	I think that Response 2 is the best response because it is the most detailed, engaging, and accurate. It provides the most information about the image and uses the most vivid language. Response 1 is also good, but it is not as detailed or engaging as Response 2.
LLaVA-RLHF rated preference	Reponse 2	Response 2
Our preference	Response 1	Response 2

Table 6: Examples of Annotation (continued)