Strengthening Multimodal Large Language Model with Bootstrapped Preference Optimization

Renjie Pi[1⋆], Tianyang Han[3*], Wei Xiong[2]

Jipeng Zhang[1], Runtao Liu[1], Rui Pan[1], and Tong Zhang[2]

[1] The Hong Kong University of Science and Technology
[2] University of Illinois at Urbana-Champaign
[3] The Hong Kong Polytechnic University

Abstract. Multimodal Large Language Models (MLLMs) excel in generating responses based on visual inputs. However, they often suffer from a bias towards generating responses similar to their pretraining corpus, overshadowing the importance of visual information. We treat this bias as a "preference" for pretraining statistics, which hinders the model's grounding in visual input. To mitigate this issue, we propose Bootstrapped Preference Optimization (BPO), which conducts preference learning with datasets containing negative responses bootstrapped from the model itself. Specifically, we propose the following two strategies: 1) using distorted image inputs to the MLLM for eliciting responses that contain signified pretraining bias; 2) leveraging text-based LLM to explicitly inject erroneous but common elements into the original response. Those undesirable responses are paired with original annotated responses from the datasets to construct the preference dataset, which is subsequently utilized to perform preference learning. Our approach effectively suppresses pretrained LLM bias, enabling enhanced grounding in visual inputs. Extensive experimentation demonstrates significant performance improvements across multiple benchmarks, advancing the state-of-the-art in multimodal conversational systems.

Keywords: Multimodal Learning · Preference Learning

1 Introduction

The emergence of Large Language Models (LLMs) has marked a significant milestone in the field of AI, revolutionizing natural language processing and understanding [8, 10, 21, 41, 49, 56, 57]. These models, trained on vast text corpus datasets, possess rich world knowledge, making them excel in generating helpful and contextually relevant text. With the advancement of LLMs, Multimodal

⋆ Equal Contribution.

Large Language Models (MLLMs) have seen rapid improvements [1,12,31,36,41, 54,69], which typically process the images using a pretrained visual encoder (e.g., vision transformer) and feed them to the LLM as token embeddings along with the text token embeddings. These models extend the capabilities of LLMs to engage in interesting conversations with image inputs, which enables various potential applications such as autonomous driving [14] and medical assistants [30].

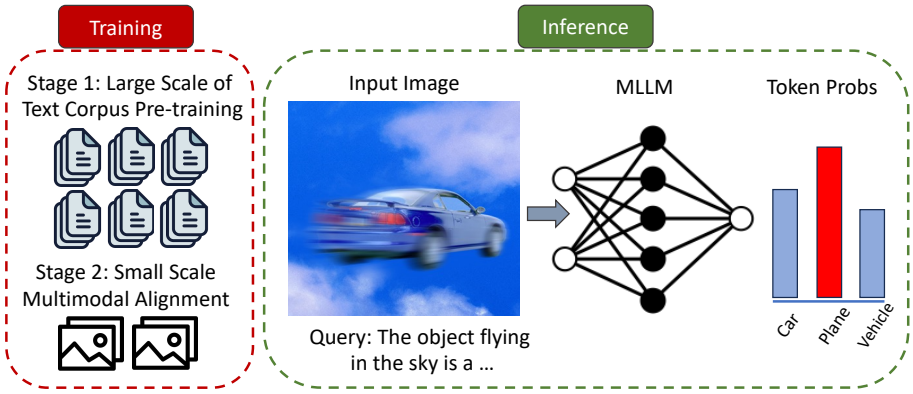

Fig. 1: Illustration of pretraining bias during MLLM's inference. Due to the difference in data scales between text-based pretraining and multimodal alignment, the MLLM is prone to generating contents that are frequently seen during its pretraining stage.

Despite the fascinating capabilities of state-of-the-art Multimodal Large Language Models (MLLMs), they exhibit a susceptibility to producing erroneous or hallucinatory responses that do not correspond to the input image. For instance, MLLMs often generate non-existent objects, incorrectly identify attributes such as shape or color, or provide inaccurate object counts. This issue renders MLLMs unreliable and impractical for real-world applications, particularly those with high stakes, such as autonomous driving systems [15] or medical assistants [30].

We hypothesize one of the major causes of this phenomenon is the bias inherited from LLM's pre-training stage. Inspired by recent research in jailbreaking of LLMs [17], we point out that MLLMs can be treated as mixture models, consisting of both distributions learned from the pretraining text corpus, as well as multi-modal alignment tuning. Specifically, the LLM undergoes an extensive pretraining stage with the large scale text corpus. Comparatively, the multimodal alignment stage in current SOTA MLLMs utilizes much fewer training samples and shorter training period. The gap between the training scales of the two phases inevitably makes the pretraining distribution dominate the generation of MLLM under certain scenarios, especially when the image is of lower quality or is not sufficiently trained during multi-modal alignment.

Motivated by the reasons above, we introduce a novel stand point to tackle the aforementioned problem. Our study draws an analogy between a blind person

who, even after a cornea transplant, still instinctively prefers walking on tactile paving. We argue that the distribution bias of MLLM stemming from pretraining can be viewed as an inherited "preference" derived from past prevalent behavior. Conversely, generating responses based on image inputs represents a new "preference" that the model must adapt to. To effectively address the current challenges faced by MLLMs, we propose to use the preference learning techniques from reinforcement learning (RL) [11, 70], which is the leading technique to adapt the model generation toward the goals of being preferred. The effectiveness of the preference learning has been showcased with its tremendous success in ChatGPT [41], Claude [2], and Gemini [23], and is known to be far more efficient than the SFT [47]. The primary goal of this paper is to extend these techniques to align the different modality of MLLMs. Specifically, the most standard and popular preference learning [2, 42, 57] consists of three steps:

- construct a preference dataset, which consists of a pair of samples and the preference signal indicating which one is more preferred;
- model a reward function based on the preference dataset;
- optimize the reward function using proximal policy optimization (PPO) [51].

While there are a diverse set of preference datasets in the LLMs, the preference learning in MLLMs is largely under-explored. To this end, our first contribution is an innovative strategy to obtain comparison pairs based on existing datasets with ground truth annotations. Specifically, we regard the existing datasets with ground truth annotations as positive responses, and generate negative responses by 1) Image-weakened prompting: we utilize distorted images as "weakened visual prompts" to elicit responses from the MLLM, revealing the inherent bias from pretraining. These responses contain a higher degree of erroneous patterns and align more closely with the pretraining distribution, while still being relevant to the image input. 2) LLM bias injection, we leverage the LLM component of the MLLM to directly modify the original responses using carefully designed prompts and few-shot examples, resulting in negative responses that exhibit similarities but differ in specific details from the original annotations. This collection of negative responses reveals a more pronounced bias towards the pretraining distribution, thereby exposing potential weaknesses and unreliability of the MLLM.

In terms of algorithmic design, it is known that the PPO algorithm is unstable and sample-inefficient in aligning LLMs [9] and imposes a heavy burden on the GPU resources as it requires loading multiple (typically four) models at the same time [16, 66]. In contrast, the recently proposed direct preference optimization (DPO) combines the reward modeling with the policy optimization into one step, and directly learns from the preference dataset (hence the name). The DPO algorithm has emerged as a promising alternative to RLHF due to its stability and competitive performance. Motivated by this, we propose a variant of DPO, referred to as the **Bootstrapped Preference Optimization (BPO)**, to extend the techniques to the MLLMs, which can significantly boosts the model performance as evaluated by multiple popular visual understanding benchmarks, while reducing object hallucinations by a large margin. To summarize, we make the following contributions in this paper:

– Firstly, make take a novel view and formulate the multimodal alignment into preference learning task, where the pretraining bias and visual grounding are treated as the old and new preferences, respectively.
– Secondly, we introduce a novel approach to construct preference datasets automatically at scale. The collected negative samples effectively expose the pretraining bias of MLLM.
– Lastly, we demonstrate through empirical evidence that our approach effectively enhances the grounding of MLLM on image inputs and results in performance boost in multiple benchmarks.

Fig. 2: We demonstrate a few examples of responses generated before and after BPO. The responses generated by the MLLM after BPO improves the grounding on visual inputs, which improves visual faithfulness and results in less erroneous outputs.

2 Related Work

Multi-Modal Large Language Model. In recent years, transformative advancements have been witnessed in the development of large language models (LLMs) [2, 5, 10, 26, 42, 49, 52, 57]. These advancements have greatly elevated the capabilities

of language understanding and generation, showcasing near-human proficiency across diverse tasks. Concurrently, the success of LLMs has inspired explorations into the incorporation of visual modality into LLM, leading to the emergence of multi-modal large language models (MLLMs) [1, 12, 12, 19, 20, 31, 36, 41, 43, 45, 54, 69]. These models have demonstrated remarkable abilities in engaging in dialogue based on visual inputs.

Alignment of Large Language Model. Alignment in agent behavior, originally introduced by [27] ensures that actions align with human intentions. Reinforcement Learning from Human Feedback (RLHF) approaches, such as those presented in [2, 3, 22, 40, 42, 50, 53, 59, 70], utilize methods like PPO [51] to maximize the rewards of model outputs. RRHF [66] and RAFT [13, 16, 58] leverage the capabilities of large language models (LLMs) to bootstrap responses, followed by fine-tuning the model on the subset of collected samples with high rewards. [46] propose direct preference optimization (DPO), which directly learns from the offline dataset with a clever reparameterization technique. The DPO is later extended to the online setting by Xiong. et. al [60]. Recently, several works have investigated the vulnerability of MLLM against malicious image inputs [37, 44] More recently, Silkie [32] suggests curating preference data to fine-tune multimodal large language models (MLLMs) using responses generated by a pool of different MLLMs, which are scored by GPT4-V.

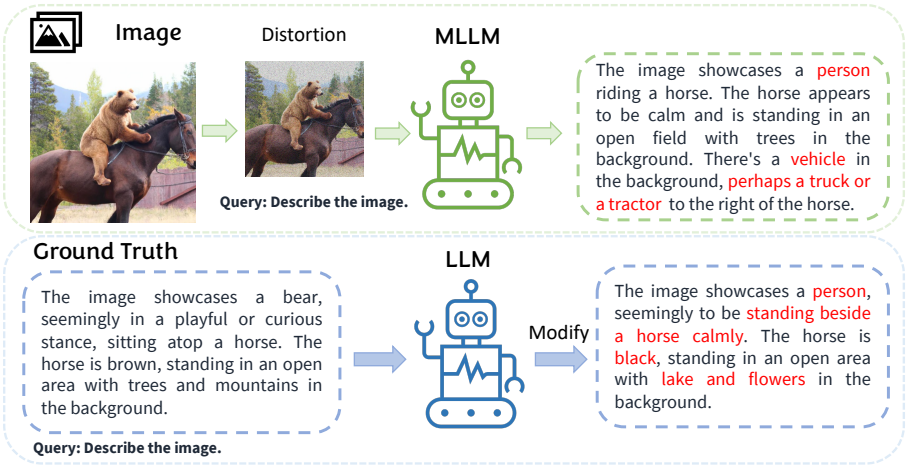

Fig. 3: The generation pipeline for negative response. Top: Image weakened prompting, which elicits responses containing pretraining bias by injecting noises into the image features; Bottom: LLM-bias injection, which explicitly modifies the details of the ground truth responses by injecting erroneous but common elements.

Hallucination in Multimodal Large Language Models. Recently, many efforts have been dedicated to alleviate the hallucination of MLLMs, which may manifest across various aspects, including object existence, object count, attribute, and relation between objects [24, 33, 35]. Woodpecker [62] leverages external tools such as object detectors and LLM to correct the hallucination in MLLM's responses. Despite the effectiveness, the external tools make this method inflexible and does not improve the MLLM's true performance. LRV-Instruction [35] proposes to conduct supervised fine-tuning (SFT) on the MLLM with positive and negative instructions that focus on the semantics of objects. However, SFT on these instructions hinder the capability of MLLMs for generating detailed responses. Visual contrastive decoding [28] proposes to correct the MLLM's output bias by subtracting the output from the MLLM using distorted image as input, which requires inference two models for each token. LLaVA-RLHF [55] introduce RLHF into MLLM training pipeline by training a reward model, which reduces hallucinatin via PPO. However, this approach requires manually labelled data for reward model training.

Generating Training Data from LLM. Owing to the advent of powerful large language models (LLMs), a new line of research that aim to automatically bootstrapping training data from LLMs has drawn lots of attention. For instance, several works propose generating training data from a more powerful LLM to tune another student language model [38, 39, 61]. Other works propose to synthesize data with better quality using more advanced techniques [18, 39], such as bi-level optimization. Recently, LLMs are used to synthesize data to improve the model's reasoning ability [19, 63]. In our work, we propose leveraging the model to bootstrap negative responses for preference learning.

3 Scalable Preference Dataset Generation

A preference dataset $\mathcal{D}$ consists of numerous tuples, such as (I, q, r^1, r^2, p), where I is the image, q is the query, r^1 and r^2 are the two responses, and p is the preference signal where $p = 1$ indicates that $r^1 \succ r^2$ given (I, q), while $p = 0$ stands for $r^1 \prec r^2$.

Annotating preference datasets manually can be a laborious and time consuming process. For instance, previous work [55,64] hire crowd workers to identify potential hallucinations in the model's responses, where the responses associated with less hallucination are assigned with higher scores. The labelled responses are subsequently leveraged to construct the preference dataset for training the reward model. This costly human-labelling process prohibits the scalability of such approaches.

On the other hand, there are abundant existing datasets targeted for supervised fine-tuning, which are annotated with high-quality image-question-answer triplets. For example, LLaVA [36] and MiniGPT4 [69] utilize the fine-grained annotations (e.g., captions, bounding boxes) to generate high-quality captions and QA pairs that are associated with images. ShareGPTV [6] leverages the powerful

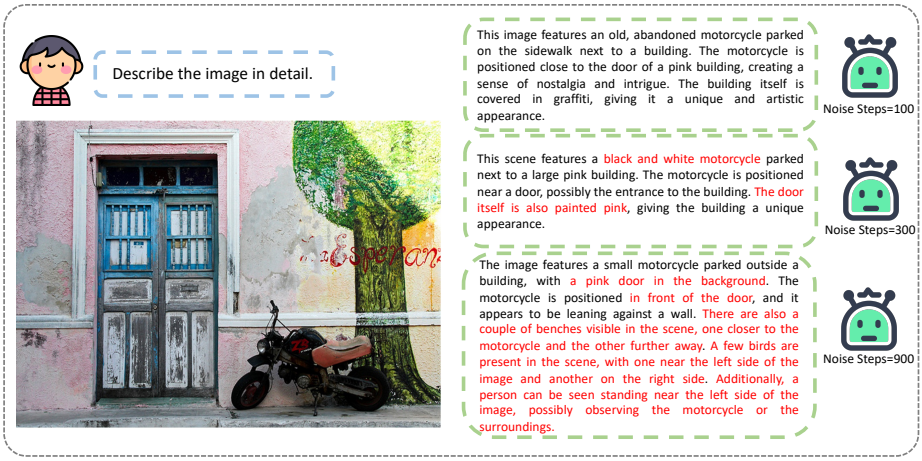

Fig. 4: The MLLM-generated responses with continuously growing steps of added noise. We can see that higher level of noise leads to more decline in visual faithfulness and the generation of hallucinated objects. These responses expose the over-reliance on knowledge learning from pretraining corpus, which is leveraged to suppress the unwantede pretraining bias via our BPO.

GPT4-V to produce high-quality captions for images. Given that the annotations in these high-quality datasets are well grounded to the image contents, they can readily serve as the positive responses in preference pairs.

3.1 Negative Response Collection

To automatically collect negative responses at scale without excessive cost, we propose the following strategies.

Image-Weakened prompting: To expose the pretraining bias and potential weaknesses of MLLMs, we apply distortions to the image features before providing them to the MLLMs for inference. Specifically, inspired by [28], we apply gaussian mask on the image embeddings from CLIP model, which is analogous to the forward process of diffusion models [25].

In the context of MLLMs, image input can be treated as a part of the prompt, which brings the MLLM's output distribution to the visual domain. After being applied with distortion, the strength of the image becomes weaker, which makes the model more likely to be overwhelmed by the pretraining distribution, further leading to inaccurate responses. As shown in figure 5, the MLLM becomes more likely to generate tokens commonly seen in pretraining corpus when the image is weakened by noise. Therefore, those responses can well expose the pre-training bias of the MLLM. We also visualize the responses generated using distorted image inputs in figure 4.

Error Injection: We take a more direct approach by leveraging the LLM component of the MLLM to explicitly modify the ground truth responses. Specif-

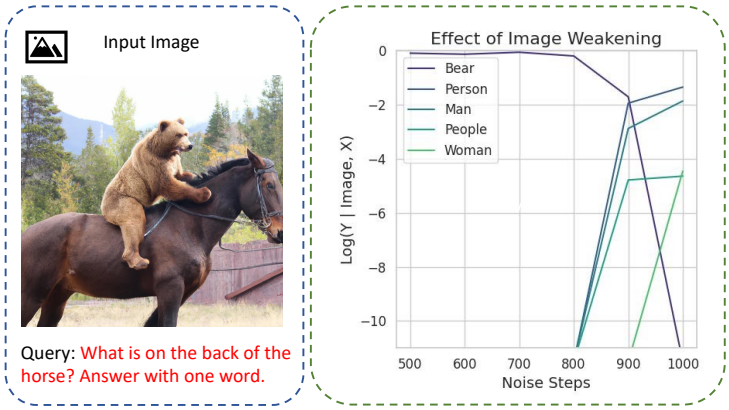

Fig. 5: The effect of image-weakened prompting. We observe the change in logits from the MLLM's output by continuously injecting higher level of noise into the image features. The log likelihood of the bear starts to decrease when the noise gets higher, and the likelihood of words such as "Person", "Man", "People" and "Woman" starts to increase, and finally take over "bear". This demonstrates the pretraining bias starts to overwhelm image information when the image is weakened.

ically, we prompt the LLM to tweak the details in the responses, such that the modified responses are similar to the original ones, but different from some aspects (e.g., object existence, object attributes, object counts). We prompt the LLM to ensure the modified response is logical and common in reality, which is likely to be close to the pretraining distribution (prompts shown in the figure 6).

3.2 Data Sources

We collect ground truth annotations from the following datasets as shown in table 1: 1) ShareGPT-V [6]: a captioning dataset constructed by prompting responses from the powerful GPT4-V, which contains detailed responses with rich visual concepts; 2) LLaVAR [68]: a VQA dataset consisting of images with rich texts; 3) LLaVA-Instruct [36]: the original instruction tuning dataset from LLaVA, which comprises of image-based conversations.

Table 1: Data sources of our preference dataset. We uniformly sample data from the popular visual instruction tuning datasets.

Source	Content	Samples Num
ShareGPT-V [6]	High-quality image captioning dataset annotated by GPT4-V	57906
LLaVAR [68]	VQA dataset consisting of images with rich texts	55445
LLaVA-Instruct [36]	Instruction dataset comprising of image-based conversations	54359

Prompt for Error Injection via LLM Component

I will give you a question and a response about an image. Pretend that you can see the image. You must modify the response by changing the original details, such as adding more objects or change the attributes of objects. Note that the modified responses should still be common in reality. I will give you some examples, then you need to paraphrase the new response.
Examples:
Question: Describe the image.
Response: the image shows an old man walking across the street. There are many people on the sidewalks, and cars are running on the street.
Modified response: the image shows a woman running towards a bus, there are no cars on the street.

Question: What is the color of the apple?
Response: the color of the apple is black.
Modified response: the color of the apple is yellow.

...

Question: New Question
Response: New Response
Modified Response:

Fig. 6: Prompting and few-shot examples used during error injection process.

4 Direct Preference Optimization

To facilitate the preference learning, one common assumption is that Bradley-Terry (BT) model [4, 42, 46, 60], which states that there exists a reward function $\phi^*(I, q, r) \to [0, 1]$ so that the preference satisfies

$$\begin{aligned}\mathbb{P}(r^1 \succ r^2 | x, r^1, r^2) &= \frac{\exp(\phi^*(I, q, r^1))}{\exp(\phi^*(I, q, r^1)) + \exp(\phi^*(I, q, r^2))} \\ &= \sigma\big(\phi^*(I, q, r^1) - \phi^*(I, q, r^2)\big),\end{aligned} \tag{1}$$

where $\sigma(z) = 1/(1+\exp(-z))$ is the sigmoid function. Essentially, the BT model implies that the preference probability is a non-decreasing and non-linear transformation (the sigmoid function) of the reward difference. This also partially explains why we choose to generate negative samples by error injection or image-weakened prompting. Otherwise, if the two samples are of similar quality, even the preference signal from the human can also be noisy, which may hurt the subsequent preference learning.

Under the BT model, the learning objective of preference learning [2,42,46,60] is

$$J(\pi) = \mathbb{E}_{I,q\sim d_0}\Big[\underbrace{\mathbb{E}_{r\sim\pi(\cdot|I,q)}[\phi^*(I, q, r)]}_{\text{Optimize the reward}} - \underbrace{\eta D_{\mathrm{KL}}(\pi_\theta(\cdot|I, q)\|\pi_0(\cdot|I, q))}_{\text{Stay close to the initial model}}\Big], \tag{2}$$

where $\eta > 0$ is the KL penalty coefficient, π_0 is the initial model, π_θ is the model to optimize. After obtaining the preference dataset $\mathcal{D} = \{(I, q, r^1, r^2, p)\}$

, in the classic framework [2, 42, 57], two stages of training can be performed: 1) the reward model π_θ can be trained under the Bradley-Terry (BT) model; 2) train the model π_θ with online RL algorithm, such as proximal policy optimization (PPO) [51]. However, previous research has found that training the reward model with multi-modal inputs leads to more severe reward hacking [55], since the continuous nature of image inputs makes the modelling of preference more challenging. Meanwhile, the instability of DRL-based PPO requires extensive efforts to tune the model to its best performance.

Recently, a more easy-to-tune approach has been proposed for aligning the preference, which is termed Direct preference optimization (DPO). This method prevents the need for training an external reward model by directly fine-tuning it on an offline preference dataset. The key insight is that the maximization of Equation (2) admits a computationally intractable solution [67]:

$$\pi^*(r|I,q) = \frac{1}{Z(I,q)}\pi_0(r|I,q)\exp(\frac{1}{\eta}\phi^*(I,q,r)),$$

where $Z(I,q) = \sum_{r'}\pi_0(r'|I,q)\exp(\frac{1}{\eta}\phi^*(I,q,r'))$ is the normalization constant that cannot be computed in practice. Then, we may solve the reward as

$$\phi^*(I,q,r) = \eta\log\frac{\pi^*(r|I,q)}{\pi_0(r|I,q)} + \eta\log Z(I,q). \tag{3}$$

Plugging Equation (3) back into the Bradley-Terry model in Equation (1), we can now conduct maximal likelihood estimation (MLE) in the policy space, i.e., the space of generative model, directly by minimizing the following loss function (the negative log-likelihood):

$$\sum_{(I,q,r_w,r_l)\in\mathcal{D}} -\Big[\log\sigma\Big(\eta\log\frac{\pi_\theta(r_w|I,q)}{\pi_0(r_w|I,q)} - \eta\log\frac{\pi_\theta(r_l|I,q)}{\pi_0(r_l|I,q)}\Big)\Big], \tag{4}$$

where r_w, r_l is the positive (winning) and negative (losing) response, respectively.
Interpretation As shown in Equation (4), the first term boosts the reference-normalized log-likelihood of the positive response, while the second term penalizes that of the negative response. Optimizing Equation (4) increases the margin between the positive sample and negative sample, thus improving the preference for grounding on visual inputs. One notable feature is the presence of the KL divergence, which is critical for preventing the model from overfitting and distribution collapse. Without the KL-penalty, the optimal policy of Equation (2) is greedy and deterministic in terms of the reward function, which deviates from the principle of generative models and can lead to an inferior performance without additional regularization [16, 34].

The DPO formulation is related to contrastive learning [7] to some degree, which also leverages pairs of positive and negative samples, i.e., a sample's representation should be closer to its positive references, and further from negative references. The difference mainly lies in the following: contrastive learning is an

unsupervised learning framework, which utilizes the samples' representation distances between their positive and negative references to optimize the decision boundary. On the other hand, DPO utilizes preference datasets with labelled preference dataset and directly optimize the model's output probability.

Table 2: Results on MM-Vet and LLaVA-Wild benchmarks. We observe consistent performance boosts over baseline models across all tasks on the two benchmarks. Notably, our tuned LLaVA1.5-7B-BPO even surpasses the larger LLaVA1.5-13B .

Model	MM-Vet							Object-HalBench		LLaVA-Wild
	Rec	OCR	Know	Gen	Spat	Math	Total	Resp$^{\downarrow}$	Obj$^{\downarrow}$	All
MiniGPT-4-8B [69]	27.4	15.0	12.8	13.9	20.3	7.70	22.1	-	-	-
BLIP-2-12B [31]	27.5	11.1	11.8	7.00	16.2	5.80	22.4	-	-	38.1
LLaVA-7B [36]	28.0	17.1	16.3	18.9	21.2	11.5	23.8	-	-	-
MiniGPT-4-14B [69]	29.9	16.1	20.4	22.1	22.2	3.80	24.4	-	-	-
Otter-9B [29]	27.3	17.8	14.2	13.8	24.4	3.80	24.7	-	-	-
InstructBLIP-14B [12]	30.8	16.0	9.80	9.00	21.1	10.5	25.6	-	-	58.2
LLaVA-13B [36]	30.9	20.1	23.5	26.4	24.3	7.70	26.4	63.0	29.5	67.3
LLaVA1.5-7B [36]	37.0	22.9	16.8	20.2	25.7	7.70	31.7	45.9	23.7	63.8
LLaVA1.5-13B [36]	41.1	29.1	23.0	24.2	35.6	7.70	36.8	45.2	21.8	71.2
LLaVA1.5-7B-BPO [36]	**41.3**	**29.5**	**24.8**	**27.8**	**34.8**	**11.5**	**36.8**	**31.9**	**15.1**	**71.6**
LLaVA1.5-13B-BPO [36]	**46.9**	**31.6**	**34.6**	**37.2**	**36.1**	**11.5**	**41.4**	**27.3**	**12.9**	**74.4**

5 Experiments

5.1 Implementation Details

We finetune the MLLM from checkpoints of LLaVA1.5 [36]. We adopt parameter efficient training technique to save computational cost and alleviate catastrophic forgetting. Specifically, we use LoRA with rank set to 64. We use learning rate of $2e^{-6}$ and train the model for 2 epochs. The model is trained on 8 A40 GPUs with 48G memory each, the batch size per GPU is set to 4. The training takes around 17 hours to complete for 7B model, ad 28 hours for 13B model.

5.2 Evaluation Benchmarks and Metrics

Helpfulness Evaluation. We use the following benchmarks for evaluation of MLLM's helpfulness: 1) *LLaVA-Bench* [36] is a real-world benchmark consisting of 60 tasks for testing LLaVA's visual instruction-following and question-answering abilities in natural environments; 2) *MM-Vet* [65] evaluates multimodal understanding by measuring six core visual-language capabilities across 128 tasks. It offers a comprehensive assessment that combines math, reasoning, and visual knowledge;

Visual Truthfulness Evaluation. For evaluation of visual truthfulness, we leverage *Object HalBench* [48], which aims to assess the MLLM's hallucination in their generated image descriptions, we follow [64] to apply 8 diverse prompts

for providing detailed descriptions of images. We evaluate hallucinations at both the response level (the percentage of responses that contain hallucinations) and the object level (the percentage of hallucinated object mentions compared to all object mentions). For all benchmarks, we leverage the powerful GPT4 as judge.

Qualitative Results We showcase a few examples of MLLM-generated responses before and after BPO tuning in Table 2. We observe that after BPO tuning, the MLLM is able to produce responses that are more grounded with the visual inputs and contain less erroneous elements.

5.3 Results on Visual Helpfulness and Truthfulness

We evaluate the effectiveness of our proposed BPO on the popular MM-Vet and LLaVA-Wild benchmarks for helpfulness, and Object-Hallucination bench for visual truthfulness in table 2. Compared with the baseline models, we observe consistent performance boosts across all tasks on the three benchmarks. Surprisingly, our tuned LLaVA1.5-7B-BPO even surpasses the larger LLaVA1.5-13B baseline on majority of the tasks. Therefore, after strengthening the preference of MLLM over visual inputs, both the helpfulness and truthfulness of the MLLM can be greatly boosted.

Table 3: Comparison with SFT baselines. The SFT datasets are constructed by extracting the positive responses from the preference dataset.

Model	Rec	OCR	Know	Gen	Spat	Math	Total
LLaVA1.5-7B	37.0	22.9	16.8	20.2	25.7	7.70	31.7
LLaVA1.5-13B	41.1	29.1	23.0	24.2	35.6	7.70	36.8
LLaVA1.5-7B-SFT	35.9	28.4	21.0	25.0	33.1	7.70	33.3
LLaVA1.5-13B-SFT	44.3	27.0	28.1	29.9	32.1	7.70	38.3
LLaVA1.5-7B-BPO	**41.3**	**29.5**	**24.8**	**27.8**	**34.8**	**11.5**	**36.8**
LLaVA1.5-13B-BPO	**46.9**	**31.6**	**34.6**	**37.2**	**36.1**	**11.5**	**41.4**

5.4 Comparison with SFT

A straightforward baseline approach would involve supervised fine-tuning, aiming to address the questions: "How effective is the preference learning algorithm?" and "How significant are the negative responses?" To validate this, we extract solely the positive responses from our preference dataset and proceed with SFT. As demonstrated by the results in table 3, we notice only a marginal improvement in performance compared to the baseline methods. This demonstrates the indispensability of negative responses and preference learning.

5.5 Comparison with Self-generated Response

Another straightforward question is whether signifying the pre-training bias in MLLMs is necessary. We compare the results achieved on LLaVA-7B by directly using the responses bootstrapped from MLLMs as negative samples with responses generated using image-weakened prompting in table 4, which verifies that purposefully exposing pretraining bias helps achieve better performance.

Table 4: Comparison with self-generated responses without image weakening.

Model	Rec	OCR	Know	Gen	Spat	Math	Total
Baseline	37.0	22.9	16.8	20.2	25.7	7.70	31.7
Self-generated	38.6	21.7	21.5	20.8	28.3	1.90	32.4
Image-Weakened	**38.9**	**23.5**	**24.8**	**21.2**	**28.5**	**7.70**	**34.3**

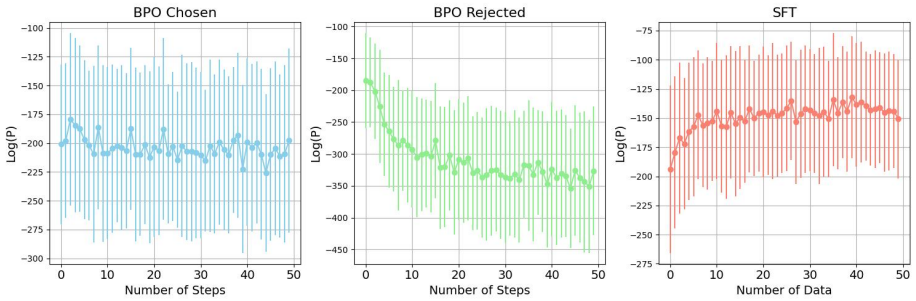

Fig. 7: The value of log probabilities for responses throughout the training process. Left: positive responses for BPO; Middle: negative responses for BPO; Right: ground truth responses for supervised fine-tuning (SFT).

5.6 Comparison of Loss Curve

Figure 7 visualizes the log probabilities of responses throughout the training process. The left graph displays the log probabilities corresponding to positive responses during the training of our model. In the middle graph, the log probabilities for negative responses during the same training phase are presented. On the right side, the log probabilities for ground truth responses during supervised fine-tuning (SFT) are shown. We plot the mean value of every interval spanning 100 steps, and add visualize the standard deviation with the error bars.

It is noteworthy that the log probability for negative responses consistently decreases during the process of BPO, while the log probability for positive responses remains relatively stable. Conversely, the log probability for ground truth

responses steadily increases for SFT. This observation suggests that preference learning aims to establish a clear margin between the likelihoods of positive and negative responses, effectively suppressing the biases originating from pre-training. On the other hand, the supervised fine-tuning (SFT) approach aims to memorize the ground truth annotations, which continuously increases the likelihood of ground truth responses, potentially leading to overfitting and the occurrence of catastrophic forgetting [34].

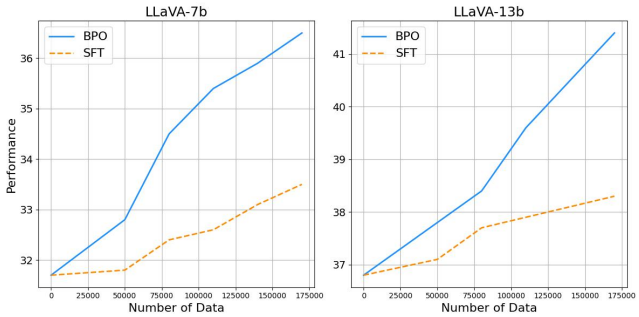

Fig. 8: We show the performance gain introduced by BPO and SFT on LLaVA-7B and LLaVA-13B, respectively. BPO consistently outperforms SFT across all dataset sizes.

5.7 Effect of Dataset Sizes

We examine the improvement brought by BPO with various sizes for preference dataset and compared to that of supervised fine-tuning in table 8. We find that larger scale of preference data indeed leads to better performance. In addition, the effectiveness of BPO consistently dominate that of SFT, which verifies that preference learning has better sample efficiency than the SFT counterpart.

6 Conclusion

In conclusion, our paper introduces Bootstrapped Preference Optimization (BPO) as a solution to mitigate bias in Multimodal Large Language Models (LLMs) when generating responses based on visual inputs. By curating paired preference datasets through bootstrapping negative responses from the model itself, we encourage the model's grounding in visual information by suppressing the preference for pretraining bias. Our approach leads to significant performance improvements across multiple benchmarks and advancing the state-of-the-art in multimodal conversational systems. We hope that our method will encourage more future research towards stronger multimodal alignment.

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023) 2, 5
2. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al.: Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862 (2022) 3, 4, 5, 9, 10
3. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al.: Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073 (2022) 5
4. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952) 9
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 4
6. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions (2023) 6, 8
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations (2020) 10
8. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/ 1
9. Choshen, L., Fox, L., Aizenbud, Z., Abend, O.: On the weaknesses of reinforcement learning for neural machine translation. arXiv preprint arXiv:1907.01752 (2019) 3
10. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022) 1, 4
11. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. Advances in neural information processing systems **30** (2017) 3
12. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023) 2, 5, 11
13. Diao, S., Pan, R., Dong, H., Shum, K.S., Zhang, J., Xiong, W., Zhang, T.: Lmflow: An extensible toolkit for finetuning and inference of large foundation models. arXiv preprint arXiv:2306.12420 (2023) 5
14. Ding, X., Han, J., Xu, H., Zhang, W., Li, X.: Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. arXiv preprint arXiv:2309.05186 (2023) 2
15. Ding, X., Han, J., Xu, H., Zhang, W., Li, X.: Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving (2023) 2
16. Dong, H., Xiong, W., Goyal, D., Zhang, Y., Chow, W., Pan, R., Diao, S., Zhang, J., Shum, K., Zhang, T.: Raft: Reward ranked finetuning for generative foundation model alignment (2023) 3, 5, 10

17. Douglas, R., Draguns, A., Gavenčiak, T.: Mitigating the problem of strong priors in lms with context extrapolation (2024) 2
18. Gao, J., Pi, R., Lin, Y., Xu, H., Ye, J., Wu, Z., Zhang, W., Liang, X., Li, Z., Kong, L.: Self-guided noise-free data generation for efficient zero-shot learning (2023) 6
19. Gao, J., Pi, R., Zhang, J., Ye, J., Zhong, W., Wang, Y., Hong, L., Han, J., Xu, H., Li, Z., Kong, L.: G-llava: Solving geometric problem with multi-modal large language model (2023) 5, 6
20. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., Qiao, Y.: Llama-adapter v2: Parameter-efficient visual instruction model (2023) 5
21. Geng, X., Liu, H.: Openllama: An open reproduction of llama (May 2023), https://github.com/openlm-research/open_llama 1
22. Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., et al.: Improving alignment of dialogue agents via targeted human judgements. arXiv preprint arXiv:2209.14375 (2022) 5
23. Google: Gemini: A family of highly capable multimodal models (2023), https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf 3
24. Han, T., Lian, Q., Pan, R., Pi, R., Zhang, J., Diao, S., Lin, Y., Zhang, T.: The instinctive bias: Spurious images lead to hallucination in mllms (2024) 6
25. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020) 7
26. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022) 4
27. Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., Legg, S.: Scalable agent alignment via reward modeling: a research direction. arXiv preprint arXiv:1811.07871 (2018) 5
28. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding (2023) 6, 7
29. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning (2023) 11
30. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day (2023) 2
31. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023) 2, 5, 11
32. Li, L., Xie, Z., Li, M., Chen, S., Wang, P., Chen, L., Yang, Y., Wang, B., Kong, L.: Silkie: Preference distillation for large visual language models (2023) 5
33. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models (2023) 6
34. Lin, Y., Tan, L., Lin, H., Zheng, Z., Pi, R., Zhang, J., Diao, S., Wang, H., Zhao, H., Yao, Y., et al.: Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. arXiv preprint arXiv:2309.06256 (2023) 10, 14
35. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning (2023) 6
36. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023) 2, 5, 6, 8, 11
37. Liu, X., Zhu, Y., Lan, Y., Yang, C., Qiao, Y.: Query-relevant images jailbreak large multi-modal models (2023) 5

38. Meng, Y., Huang, J., Zhang, Y., Han, J.: Generating training data with language models: Towards zero-shot language understanding. arXiv preprint arXiv:2202.04538 (2022) 6
39. Meng, Y., Michalski, M., Huang, J., Zhang, Y., Abdelzaher, T., Han, J.: Tuning language models as training data generators for augmentation-enhanced few-shot learning (2023) 6
40. Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al.: Webgpt: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332 (2021) 5
41. OpenAI: Gpt-4 technical report (2023) 1, 2, 3, 5
42. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022) 3, 4, 5, 9, 10
43. Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., Yao, L., Han, J., Xu, H., Kong, L., Zhang, T.: Detgpt: Detect what you need via reasoning (2023) 5
44. Pi, R., Han, T., Xie, Y., Pan, R., Lian, Q., Dong, H., Zhang, J., Zhang, T.: Mllm-protector: Ensuring mllm's safety without hurting performance (2024) 5
45. Pi, R., Yao, L., Gao, J., Zhang, J., Zhang, T.: Perceptiongpt: Effectively fusing visual perception into llm (2023) 5
46. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct preference optimization: Your language model is secretly a reward model (2023) 5, 9
47. Ramamurthy, R., Ammanabrolu, P., Brantley, K., Hessel, J., Sifa, R., Bauckhage, C., Hajishirzi, H., Choi, Y.: Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. arXiv preprint arXiv:2210.01241 (2022) 3
48. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning (2019) 11
49. Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilic, S., Hesslow, D., Castagne, R., Luccioni, A.S., Yvon, F., Galle, M., et al.: Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100 (2022) 1, 4
50. Scheurer, J., Campos, J.A., Korbak, T., Chan, J.S., Chen, A., Cho, K., Perez, E.: Training language models with language feedback at scale. arXiv preprint arXiv:2303.16755 (2023) 5
51. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms (2017) 3, 5, 10
52. Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., et al.: Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. arXiv preprint arXiv:2201.11990 (2022) 4
53. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F.: Learning to summarize with human feedback. Advances in Neural Information Processing Systems **33**, 3008–3021 (2020) 5
54. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: Pandagpt: One model to instruction-follow them all (2023) 2, 5
55. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.Y., Wang, Y.X., Yang, Y., Keutzer, K., Darrell, T.: Aligning large multimodal models with factually augmented rlhf (2023) 6, 10

56. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca (2023) 1
57. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 1, 3, 4, 10
58. Wang, H., Lin, Y., Xiong, W., Yang, R., Diao, S., Qiu, S., Zhao, H., Zhang, T.: Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. arXiv preprint arXiv:2402.18571 (2024) 5
59. Wu, J., Ouyang, L., Ziegler, D.M., Stiennon, N., Lowe, R., Leike, J., Christiano, P.: Recursively summarizing books with human feedback. arXiv preprint arXiv:2109.10862 (2021) 5
60. Xiong, W., Dong, H., Ye, C., Zhong, H., Jiang, N., Zhang, T.: Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. arXiv preprint arXiv:2312.11456 (2023) 5, 9
61. Ye, J., Gao, J., Li, Q., Xu, H., Feng, J., Wu, Z., Yu, T., Kong, L.: Zerogen: Efficient zero-shot learning via dataset generation. In: Empirical Methods in Natural Language Processing (2022) 6
62. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: Hallucination correction for multimodal large language models (2023) 6
63. Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., Kwok, J.T., Li, Z., Weller, A., Liu, W.: Metamath: Bootstrap your own mathematical questions for large language models (2023) 6
64. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.T., Sun, M., Chua, T.S.: Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback (2023) 6, 11
65. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities (2023) 11
66. Yuan, Z., Yuan, H., Tan, C., Wang, W., Huang, S., Huang, F.: Rrhf: Rank responses to align language models with human feedback without tears. arXiv preprint arXiv:2304.05302 (2023) 3, 5
67. Zhang, T.: Mathematical Analysis of Machine Learning Algorithms. Cambridge University Press (2023). https://doi.org/10.1017/9781009093057 10
68. Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., Sun, T.: Llavar: Enhanced visual instruction tuning for text-rich image understanding (2024) 8
69. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models (2023) 2, 5, 6, 11
70. Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019) 3, 5