SBSPO : Self-Bootstrapping Mitigates Object Hallucination via Preference Optimization

Qixin Xu
Dept. of CS
Tsinghua University
xqx23@mails.tsinghua.edu.cn

Siyuan Chen
Dept. of CS
Tsinghua University
siyuan-c23@mails.tsinghua.edu.cn

Haoyu Dong
Dept. of EE
Tsinghua University
donghy23@mails.tsinghua.edu.cn

Abstract

*Traditional feedback-based approaches for hallucination reduction often rely on labor-intensive manual annotation or expensive proprietary models, leaving a significant gap in the foundational understanding of how to construct high-quality, yet easily accessible, feedback mechanisms for open-source Multimodal Large Language Models (MLLMs). In this work, we propose a **S**elf-**B**oot**S**trapping **P**reference **O**ptimization algorithm. SBSPO leverages the full potential of open-source MLLMs from two key perspectives: (1) as a policy model within a reinforcement learning framework, and (2) as an implicit reward model that facilitates the generation of high-quality feedback data for preference learning. Extensive experiments across two benchmark datasets demonstrate that SBI significantly improves the trustworthiness and reliability of multimodal models during inference.*

1. Introduction

Recent advancements in multimodal large language models (MLLMs) represent a major breakthrough in AI research [2, 4, 12, 13, 14, 26]. These models, which are trained on vast multimodal datasets, possess extensive world knowledge and demonstrate exceptional abilities in addressing a wide range of multimodal tasks [9, 15, 18]. Despite these achievements, MLLMs are often observed to confidently produce incorrect content that diverges from human preferences [24, 32, 27, 6]. To better align these models with human values, reinforcement learning from human feedback (RLHF) has been widely adopted, yielding significant improvements [24, 27]. However, RLHF is heavily reliant on labor-intensive human annotations, making it difficult to address the broad misalignment between model behavior and human preferences. In response, recent work has explored reinforcement learning from AI feedback (RLAIF), which leverages preferences from labeler models as a proxy for human feedback, showing promising potential as an alternative to RLHF [8].

Current RLAIF techniques rely on highly advanced proprietary models to generate feedback from [10, 30, 33, 31]. These approaches encounter significant scalability challenges due to their dependence on expensive API access. A critical limitation of these methods is their inherent requirement for at least two distinct Multimodal Large Language Models (MLLMs) in the pipeline: one serving as the policy model and the other as the judge model. This dual-model architecture not only increases computational and financial costs but also introduces practical challenges in resource-constrained environments, further exacerbating scalability issues. Moreover, even when high-quality feedback is available, existing training methods often reach performance plateaus and fail to fully utilize the data [24, 1], primarily due to the reward distribution shift problem. This issue arises because the reward model or feedback model remains static during training, while the model's output distribution evolves continuously, leading to a misalignment between the feedback and the model's current state.

To tackle these challenges, we introduce the Self-Bootstrapping Preference Optimization (SBSPO) framework, which aligns Multimodal Large Language Models (MLLMs) through **self-generated feedback**, achieving performance comparable to methods that rely on feedback from more advanced label models. Specifically, the SBSPO framework solves above two challenges upon one key innovation: We propose a self-bootstrapping strategy to streamline the generation of preference data pairs for Direct Preference Optimization (DPO) training. In this framework,

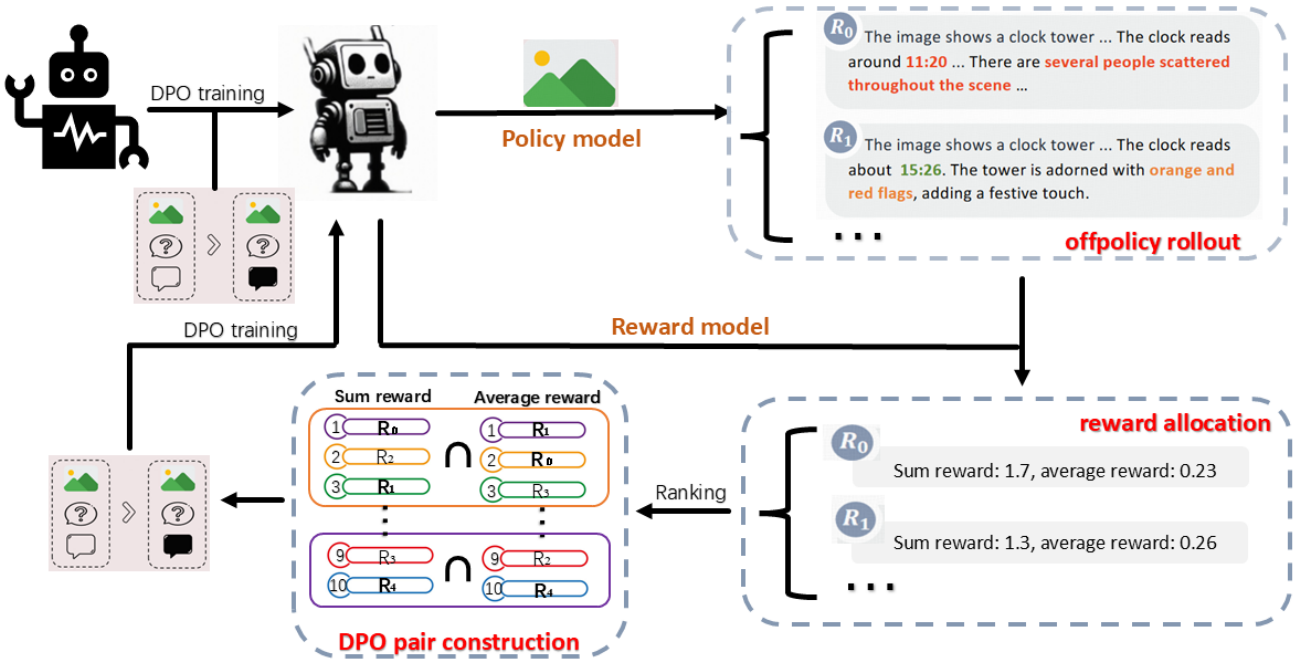


Figure 1: SBSPO pipeline. (1) Given the input image and prompt, multiple candidate responses for each image-prompt pair are generated during the offpolicy rollout period. (2) Using the initial model itself as reward model to score the candidiate responses. (3) Forming DPO pairs with the reward union approach. (4) Apply DPO training.

the policy model also functions as the reward model, autonomously evaluating and assigning rewards to its own candidate responses. Reward allocation is conducted from two complementary dimensions: cumulative reward and average reward. While each metric individually suffers from a length bias problem, their integration mitigates this issue by leveraging their combined strengths to establish a more balanced and reliable evaluation criterion. These two-dimensional rewards are then utilized to construct DPO pairs, where responses with higher combined rewards are selected as preferred choices, while those with lower rewards are designated as rejected alternatives. This self-contained, dual-reward mechanism not only simplifies the feedback generation process but also enhances its efficiency, making it more accessible and scalable compared to methods requiring external models or complex pipelines. Unlike conventional reward allocation or feedback generation mechanisms, our approach employs an on-policy reward system where the reward model is intrinsically integrated with the policy model itself. This innovative design effectively eliminates the reward distribution shift problem that typically arises between feedback mechanisms and the model's current state. Throughout the training process, the reward model dynamically evolves in tandem with the policy model's updates through preference learning, thereby maintaining an adaptive and responsive alignment process that continuously reflects the model's evolving state.

Comprehensive experiments on two benchmarks show that SBSPO can substantially enhance the trustworthiness of models with only one model involved, regardless of any human or proprietary model intervention. Using the model which we applied DPO training on LLaVA 1.5 7B [12] as both policy model and reward model, we significantly reduce the object hallucination rate of the base model on Object HalBench [27] by 81.2%. This performance not only underscores the effectiveness of our approach but also surpasses that of labeler-model-dependent algorithms by a substantial margin. Meanwhile, experimental results on the MMhal Bench demonstrate that our method achieves competitive performance compared to other approaches while incurring significantly lower training costs. This further underscores the self-alignment potential of open-source Multimodal Large Language Models (MLLMs), highlighting their capability to achieve high-quality alignment without relying on external proprietary models or extensive computational resources.

The contributions of this work can be summarized

as threefold: (1) We present SBSPO, a novel framework that aligns MLLMs with its own feedback. (2) We propose a self-bootstrapping strategy to simplify the feedback generation process. (3) We conduct comprehensive experiments to demonstrate the effectiveness of the proposed framework, achieving competitive performaces with lower training costs.

2. SBSPO

Algorithm 1 alignment of SBSPO

Input: Initial instruction model M_1 (LLaVA 1.5-7B), image-question pair set X_1
Output: Aligned model M_2
  $M_1' \leftarrow$ RLAIF-V(M_1)
  for all $x_i \in X_1$ **do**
    $y_i = [y_{i1}, y_{i2}, \cdots] \leftarrow$ Get10RandomAns(M_1', x_i)
    $z_i = [z_{i1}, z_{i2}, \cdots] \leftarrow$ ScoreByLogp(M_1', y_i)
    $a_i = [a_{i1}, a_{i2}, \cdots] \leftarrow$ SortByAvgScore(z_i)
    $s_i = [s_{i1}, s_{i2}, \cdots] \leftarrow$ SortBySumScore(z_i)
    $cho_i \leftarrow$ GetTop3UnionSet(a_i, s_i)
    $rej_i \leftarrow$ GetLast3UnionSet(a_i, s_i)
    $p_i \leftarrow [cho_i, rej_i]$
  end for
  $P \leftarrow [p_1, p_2, \cdots]$
  $M_2 \leftarrow$ TrainDPO(M_1', P)
  return M_2

In this section, we first outline the process of obtaining a DPO-aligned model, which then serves as the foundational policy model and reward model for subsequent stages. Next, we elaborate on the self-bootstrapping strategy, which streamlines feedback collection through a self-reward mechanism. This innovative approach allows the model to autonomously generate and assess its own feedback, thereby significantly simplifying the acquisition of preference data.

2.1. Direct Preference Optimization (DPO)

To ensure computational efficiency, we adopt Direct Preference Optimization (DPO) as the foundational method for training the policy model and reward model. DPO represents a lightweight alternative to traditional Reinforcement Learning with Human Feedback (RLHF), which typically relies on labor-intensive human annotations, complex multi-model pipelines, and computationally expensive reinforcement learning processes [20]. The DPO loss function is defined as:

$$L = -\mathbb{E}_{(x,y_w,y_l)}\left[\log\sigma\left(\beta\log\frac{\pi^*(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi^*(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \quad (1)$$

where x represents the input, y_w and y_l denote the preferred and less preferred responses, respectively, and $\pi_{\text{ref}}(y|x)$ serves as the fixed reference model.

In the SBSPO framework, DPO plays a dual role by serving as both a preference-learning mechanism and a means to address reward distribution shift challenges. Its simplified objective, which effectively transforms the RLHF process into a classification loss, ensures that the model can leverage self-generated feedback data efficiently and with minimal computational overhead.

2.2. RLAIF-V

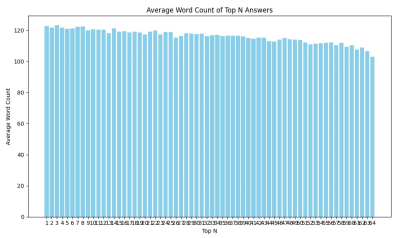
RLAIF-V [28] introduces a novel method for collecting high-quality AI feedback from open-source MLLMs and creating DPO pairs through a divide-and-conquer strategy. This approach breaks down holistic response evaluation into individual atomic claims by prompting a large language model to extract factual claims from the response. These claims are then transformed into polar questions, and their scores are calculated to form DPO pairs for later alignment.

Since RLAIF-V achieves superior trustworthiness and reduced hallucination rates compared to proprietary models, while maintaining scalability and cost-effectiveness, we utilize its open-source dataset to train the LLaVA 1.5 7B [12] with the DPO algorithm for two iterations, then adopting the aligned model as our base model for testing SBSPO algorithm.

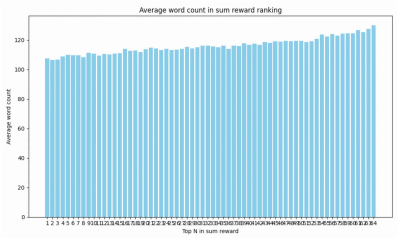
2.3. Off-policy Rollout

The feedback collected for preference learning is in the form of comparison pairs, where each pair includes a preferred response y_w and an inferior response y_l to the same input x (including the image and prompt). During training, the model learns preferences by distinguishing the differences between y_w and y_l. However, these differences can be complex and consist of many factors including not only the meaning of content but also textual styles such as the use of specific words or structure of the text, making the learning more difficult.

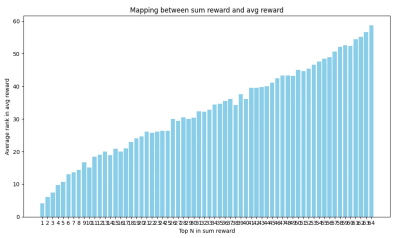
To expose the genuine differences in trustworthiness between responses, we follow [28], using a deconfounded strategy to generate candidate responses. Specifically, we ask the model to generate 10 candidate responses $\{y_1, y_2, \cdots, y_{10}\}$ through sampling decoding with different random seeds, where input x and decoding parameters are invariant. In this way, y_w and y_l are sampled from the same distribution and consequently share similar textual styles and linguistic patterns. During training, the model can effectively



(a) Average Word Count of Avg Reward Ranking



(b) Average Word Count of accumalative reward Ranking



(c) Mapping between accumalative reward and Average Reward

Figure 2: Statistical analysis of the length of favored responses regarding different reward allocation methods

concentrate on the differences in trustworthiness.

2.4. Reward allocation

After iterative learning using RLAIF-V on open-source feedback, the MLLM itself is not only a trustworthy policy model but also a reward function by the optimization objective of DPO [21] and the reward is formulated as:

$$r^{sum}(y) = \beta \log \frac{\pi_\theta(y)}{\pi_{\text{ref}}(y)} = \beta \sum_t^T \log \frac{\pi_\theta(y_t|y_{<t})}{\pi_{\text{ref}}(y_t|y_{<t})} \quad (2)$$

where β is a parameter controlling the deviation from the base reference policy π_{ref}, y is the response token sequence, T is the length of response and π_θ is the model after DPO training. We hide the prompt condition x in equations for simplicity. The reward score r, which represents the total reward allocated for the entire response (accumalative reward), can be utilized to construct training sample pairs for alignment. However, previous works have shown that DPO-aligned reward $r(y)$ can be biased towards shorter responses due to its objective formulation [19]. We aim to tackle this bias by averaging all token-level scores to get the average response score:

$$r^{avg}(y) = \frac{\beta}{T} \log \frac{\pi_\theta(y)}{\pi_{\text{ref}}(y)} \quad (3)$$

Nevertheless, our experimental findings reveal that the average reward metric exhibits a bias towards longer responses. As illustrated in Figure 2, average reward metric introduces a contrasting length bias compared to the accumalative reward metric. Specifically, under the average reward ranking, top-ranked answers tend to be longer, whereas shorter responses are favored in the accumalative reward ranking. We hypothesize that this phenomenon arises from the propensity of longer responses to incorporate a higher frequency of stop words, which typically yield relatively higher rewards (negative but closer to zero), thereby artificially inflating the overall reward for lengthier responses.

However, statistical analysis presented in Figure 2 demonstrates that responses ranking highly under the accumalative reward metric also tend to rank highly under the average reward metric, indicating a degree of consistency between the two evaluation frameworks despite their differing biases. In order to develop a more comprehensive reward allocation method for ranking responses generated from off-policy rollouts, we propose a **reward union approach**. Specifically, given a question x and its corresponding responses $\{s_1, s_2, \ldots, s_m\}$, we rank these responses based on two reward principles: **accumalative reward** and **average reward**, both in descending order. This yields two distinct permutations of the responses, denoted as:

$$\begin{aligned} \mathcal{S}_{\text{sum}} &= \{s_1^{sum}, s_2^{sum}, \ldots, s_m^{sum}\} \\ \mathcal{S}_{\text{avg}} &= \{s_1^{avg}, s_2^{avg}, \ldots, s_m^{avg}\} \end{aligned} \quad (4)$$

For each permutation, we select the top λm responses and compute their union to construct the chosen answer set $\mathcal{A}_x$for the question x, where λ is a hyperparameter within [0, 0.5) representing the trade-off between the quality and quantity of DPO pairs, as lower λ means less elements within the union set and higher threshold of the gap between chosen responses and rejected ones. Similarly, to form the rejected set $\mathcal{R}_x$ for x, we take the union of the bottom λm responses from each permutation. The process can be formulated as:

$$\begin{aligned} \mathcal{A}_x &= \mathcal{S}_{\text{sum}}^{1:\lambda m} \cup \mathcal{S}_{\text{avg}}^{1:\lambda m} \\ \mathcal{R}_x &= \mathcal{S}_{\text{sum}}^{(1-\lambda)m:m} \cup \mathcal{S}_{\text{avg}}^{(1-\lambda)m:m} \end{aligned} \quad (5)$$

We then perform a complete pairwise matching

between $\mathcal{A}_x$ and $\mathcal{R}_x$ to construct the training data. Specifically, for each response in $\mathcal{A}_x$ paired with every response in $\mathcal{R}_x$, we generate a distinct DPO (Direct Preference Optimization) sample pair. We then use these DPO pairs to train the aligned DPO model described in 2.2 which is also the policy model and reward model in the previous iteration.

3. Experiments

In this section, we empirically investigate the effectiveness of SBSPO in aligning MLLMs through self-reward feedback. In addition to evaluating model performance regarding trustworthiness and helpfulness, we also analyze the compatibility with other methods, and the generalizability of feedback data collected with SBSPO.

3.1. Experimental Setup

We introduce models, training data, evaluation benchmarks, baselines, and other implementation details. All experiments are conducted based on LLaVA 1.5 7B [12] unless otherwise specified.

Base Models. To obtain the DPO-aligned base model required for our SBSPO framework, we first introduce a setup to align MLLMs using the RLAIF-V framework. Specifically, we leverage LLaVA 1.5 7B [12] as the instruction model and LLaVA-NeXT [13] as the labeler model. The model is trained across two iterations of DPO training, forming both the base model for SBSPO and the baseline for evaluation.

Models. In our experiments, the base model trained with RLAIF-V serves as both the instruction model and the policy model at the same iteration in the SBSPO framework. This setup highlights the effectiveness of SBSPO and the lightweight nature of the pipeline. Notably, no stronger models are introduced at this stage, further underscoring the bootstrapping capability of our approach.

Training Data. The diversity of instructions can be critical for models to learn generalizable preferences. In practice, we use instructions collected from a diverse range of datasets, including MSCOCO [11], ShareGPT-4V [3], MovieNet [7], Google Landmark v2 [25], VQA v2 [5], OKVQA [16], and TextVQA [23]. In addition, we adopt image description prompts introduced in [27] to construct long-form image describing instructions.

Evaluation. We evaluate models from the perspectives of trustworthiness reflecting the hallucination degree. In detail, we perform evaluation on two benchmarks:

(1) **Object HalBench** [22] is a widely adopted benchmark for assessing common object hallucination in detailed image descriptions. We follow [27] to use 8 diverse prompts to improve the evaluation stability. We report the response-level hallucination rate (i.e., the percentage of hallucinated responses) and the mention-level hallucination rate (i.e., the percentage of hallucinated objects).

(2) **MMHal-Bench** [24] evaluates response-level hallucination rate and informativeness. It asks GPT-4 [17] to compare model outputs with human responses and object labels for evaluation.

Baselines. We compare our model with state-of-the-art baselines of different types, including general baselines with strong performance, baselines trained with feedback data and proprietary baselines.

(1) **General Baselines.** We adopt LLaVA 1.5 7B [12] as representative general baselines. These models are mostly pre-trained on large-scale multimodal data and fine-tuned on high-quality instruction data, achieving strong performance across various multimodal tasks.

(2) **Baselines tailored for feedback learning.** LLaVA-RLHF [24] trains the reward model on human-labeled preference data and performs proximal policy optimization to train the model. RLHF-V [27] collects fine-grained correctional human feedback and trains the model with dense direction preference optimization. Silkie [10] utilizes GPT-4V to collect feedback. POVID [31] and AMP-MEG [29] apply heuristic rules to pair responses generated under difference condition. More importantly, we address the model after two iteration of RLAIF-V (the base model of SBSPO) as the most important baseline.

(3) **Proprietary baseline.** We also include GPT-4V [18] as strong reference to evaluate the gap between the open-source models and proprietary models.

3.2. Main Results

The main experimental results are reported in Table 1, from which we observe that: (1) SBSPO achieves state-of-the-art performance in trustworthiness among open-source models and even surpasses proprietary models such as GPT-4V. Our framework significantly reduces the object hallucination rate of LLaVA 1.5 and RLAIF-V 2 iter by 74.3% and 14% relative points on Object HalBench, surpassing GPT-4V by a large margin. The reduction of hallucination can also be seen in MMHal-Bench, even it may not perform such well. (2) Among all the open-source multimodels, SBSPO performs a higher accuracy compared to those applied rule feedback; SBSPO reduces

Model	Size	Feedback	Object-HalBench		MMHal-Bench	
			Resp. ↓	Ment. ↓	Score	Hall.↓
HA-DPO (*arXiv'23*)	7B	Rule	39.9	19.9	1.98	60.4
POVID (*arXiv'24*)	7B	Rule	48.1	24.4	2.08	56.2
LLaVA-RLHF (*arXiv'23*)	13B	Human	38.1	18.9	2.02	62.5
Silkie (*EMNLP'24*)	10B	GPT-4V	27.1	13.4	3.19	32.3
RLHF-V (*CVPR'24*)	13B	Human	12.2	7.5	2.45	51.0
AMP-MEG (*NeurIPS'24*)	13B	Rule	31.7	20.6	3.23	34.4
LLaVA 1.5	7B	✗	53.6	25.2	2.36	51.0
+ RLAIF-V 1 iter	7B	LLaVA-NeXT	13	7.5	2.67	40.6
+ RLAIF-V 2 iter	7B	LLaVA-NeXT	11.5	5.7	2.82	37.5
+ SBSPO 1 iter (our method)	7B	**Self as Reward model**	**10.1**	**4.8**	2.96	36.5
+ RLAIF-V 3 iter	7B	LLaVA-NeXT	10.6	5.1	3.04	37.5
GPT-4V [18]	-	Unknown	13.6	7.3	**3.49**	**28.1**

Table 1: Main experimental results. We report hallucination rates in different granularities including response-level (Resp.) and mention-level (Ment.). The best results are shown in **bold** and underline.

the need for human labor force compared those that using human feedback; SBSPO is much economical compared to those that rely on GPT-4V feedback. (3) SBSPO surpasses RLAIF-V 3 iter on Object-HalBench and MMHal-Bench by 5.3% and 2.6% relative points respectively, indicating that our framework also outperforms the third iteration of LLaVA-Next feedback on the same base model.

3.3. Analysis

In this part, we further discuss (1) how the reward allocation method can possibly introduce length bias and our approach to tackle it, and (2) a possible explanation for improved performance of SBSPO.

Reward union strategy mitigates length bias.

When constructing the dataset for SBSPO, we employ a reward union strategy to select answer pairs ranked based on $\log p$. Specifically, we first sample 10 random answers and rank them according to two criteria: the sum of rewards and the average reward. From each ranking, we extract the top three and bottom three answers, then take the union of these answers to form the DPO pairs.

This strategy is designed to address the length bias observed in different sampling techniques. To demonstrate this, we randomly selected a set of answers and applied the sum reward, average reward, and reward union strategies, respectively. The results, summarized in Table 2, highlight the differences in average answer lengths between the chosen and rejected answers under each strategy.

As shown in the statistics, the average reward strategy results in chosen answers being significantly longer than rejected answers, while the opposite occurs under the sum reward strategy. In contrast, the reward union strategy effectively balances the length bias seen in these two approaches, keeping the length differences between chosen and rejected answers within a reasonable range. This balance ensures that the selected answer pairs are less influenced by length and better aligned with the intended goals of the dataset.

Strategy	Average Answer Length	
	Chosen	Rejected
Average Reward	120.77	110.22
Sum Reward	109.68	122.62
Reward Union	103.96	104.31

Table 2: Comparison of Chosen and Rejected Answer Lengths Across Strategies

Feature space similarity as a possible explanation for improved performance of SBSPO.

We hypothesize that the capability of a specific model M can be intrinsically described by a specific feature space V, characterized by its eigenvectors $\{v_1, v_2, \ldots, v_n\}$. The process of alignment aims to modify this feature space or adjust its eigenvectors to achieve improved performance.

In our work, we construct DPO pairs using a re-

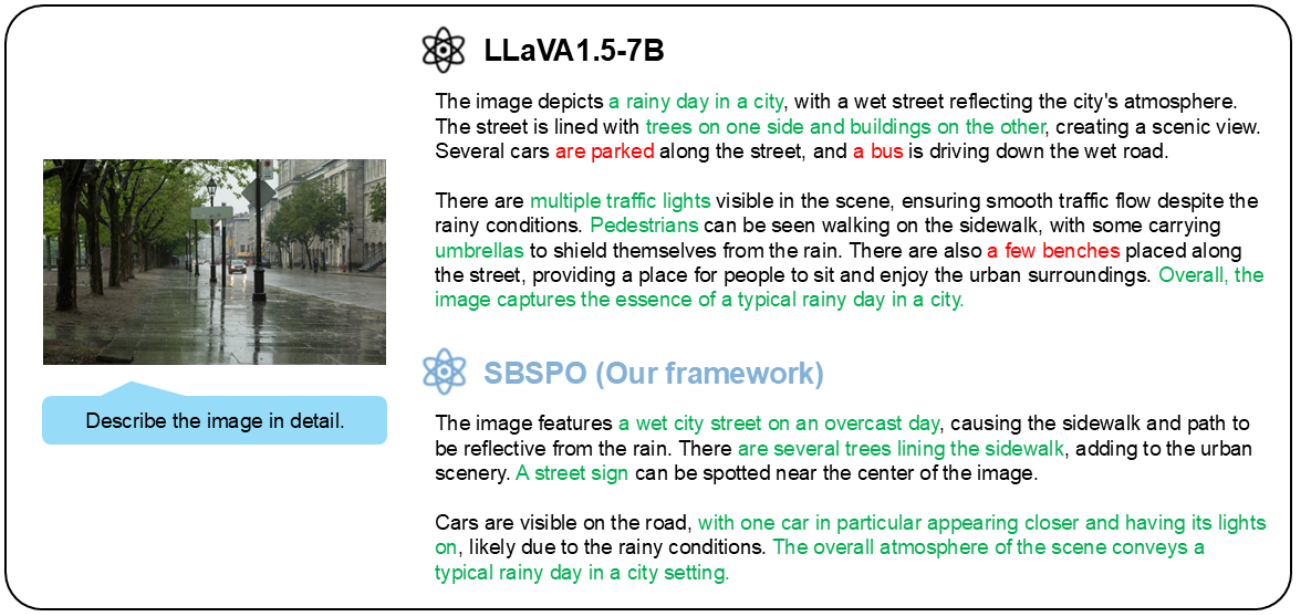


Figure 3: Case Study: Comparison between LLaVA 1.5-7B and SBSPO

ward model M_r, which is described by its eigenvectors $\{v'_1, v'_2, \ldots, v'_n\}$. Subsequently, we employ the DPO algorithm to align the model M_i, characterized by its eigenvectors $\{v_1, v_2, \ldots, v_n\}$. Our findings suggest that the higher the similarity between the eigenvectors of M_r and M_i, the greater the alignment in the learned feature space and, consequently, the more significant the improvements in the model's performance. This indicates that the degree of eigenvector similarity serves as a critical factor in determining the extent to which the model learns and enhances its capabilities.

Our proposed method, SBSPO, utilizes the model M_i itself as the reward model during dataset construction. This approach ensures that the two models share the highest degree of similarity in their eigenvectors, enabling M_i to achieve improvements across the entire feature map. This improvement can be represented as the change in the probability distribution, denoted by ΔD_i. However, if the reward model M_r differs from M_i—specifically, if their eigenvectors $\{v_1, v_2, \ldots, v_n\}$ and $\{v'_1, v'_2, \ldots, v'_n\}$ are dissimilar, or even mutually exclusive—then the distribution shift ΔD_i of M_i tends to deviate from the optimal scenario and fails to achieve the best alignment.

By leveraging the same model M_i as the reward model in SBSPO, we maximize eigenvector similarity, achieving optimal alignment and significant enhancements in the feature space.

3.4. Case Study

To provide an intuitive understanding and comparison of different frameworks, we provide a detailed result in Figure 3. In this example, both LLaVA 1.5-7B and SBSPO successfully identify the image depicts a wet street scene in a rainy day. But LLaVA mistakenly believes there are several cars parked along the street, a bus driving on the road and a few benches, while SBSPO successfully avoids these mistakes.

4. Conclusion

Aligning models with human preferences represents a pivotal objective in the development of trustworthy machine learning systems. In this study, we introduce SBSPO, a novel framework designed to enhance the reliability and trustworthiness of Multimodal Large Language Models (MLLMs) through the integration of self feedback. Extensive experimental evaluations demonstrate that our framework achieves competitive results on discriminative trustworthiness metrics with lower costs.

To address the challenges of resource-intensive feedback acquisition and feedback efficiency, we propose a self-bootstrapping strategy to collect self-reward feedback. The methodology enables the systematic refinement of feedback mechanisms, ensuring high-quality alignment while maintaining computational efficiency. Through alignment of the model with this optimized yet easily accessible feedback, we

observe significant improvements in trustworthiness.

Looking ahead, we aim to extend this framework by incorporating more sophisticated feedback mechanisms to enhance the model's capabilities in logical reasoning and complex task-solving. This future direction holds promise for advancing the robustness and applicability of MLLMs in real-world scenarios.

5. Acknowledgements

The authors wish to convey their deepest gratitude to Professor Xiaolin Hu for his exceptional guidance and unwavering support throughout the course of this research. They would also like to extend special thanks to Dr. Guo Chen for his invaluable insights and substantial contributions to the advancement of this work.

All authors contributed equally to this work. Each author participated in the conception and design of the study, data analysis and interpretation, and the drafting and revising of the manuscript. All authors have read and agreed to the published version of the manuscript.

References

[1] Leonard Adolphs, Tianyu Gao, Jing Xu, Kurt Shuster, Sainbayar Sukhbaatar, and Jason Weston. The CRINGE loss: Learning what language not to model. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of ACL*, pages 8854–8874. Association for Computational Linguistics, 2023. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966, 2023. 1

[3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *CoRR*, abs/2311.12793, 2023. 5

[4] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Proceedings of NeurIPS*, 2023. 1

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of CVPR*, pages 6325–6334. IEEE Computer Society, 2017. 5

[6] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Processing of CVPR*, 2024. 1

[7] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Proceedings of ECCV*, volume 12349 of *Lecture Notes in Computer Science*, pages 709–727. Springer, 2020. 5

[8] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. RLAIF: scaling reinforcement learning from human feedback with AI feedback. *CoRR*, abs/2309.00267, 2023. 1

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023. 1

[10] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *CoRR*, abs/2312.10665, 2023. 1, 5

[11] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Proceedings of ECCV*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. 5

[12] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744, 2023. 1, 2, 3, 5

[13] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. 1, 5

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Proceedings of NeurIPS*, 2023. 1

[15] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *Processing of ICLR*, 2024. 1

[16] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of CVPR*, pages 3195–3204. Computer Vision Foundation / IEEE, 2019. 5

[17] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 5

[18] OpenAI. GPT-4V(ision) system card, 2023. 1, 5, 6

[19] Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^*: Your language model is secretly a q-function, 2024. 4

[20] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv e-prints*, page arXiv:2305.18290, May 2023. 3

[21] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Proceedings of NeurIPS*, 2023. 4

[22] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of EMNLP*, pages 4035–4045. Association for Computational Linguistics, 2018. 5

[23] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of CVPR*, pages 8317–8326. Computer Vision Foundation / IEEE, 2019. 5

[24] 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. Aligning large multimodal models with factually augmented RLHF. *CoRR*, abs/2309.14525, 2023. 1, 5

[25] Tobias Weyand, André Araújo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 - A large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of CVPR*, pages 2572–2581. Computer Vision Foundation / IEEE, 2020. 5

[26] Tianyu Yu, Jinyi Hu, Yuan Yao, Haoye Zhang, Yue Zhao, Chongyi Wang, Shan Wang, Yinxv Pan, Jiao Xue, Dahai Li, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. Reformulating vision-language foundation models and datasets towards universal multimodal assistants. *CoRR*, abs/2310.00653, 2023. 1

[27] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. RLHF-V: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of CVPR*, 2024. 1, 2, 5

[28] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. RLAIF-V: Aligning MLLMs through Open-Source AI Feedback for Super GPT-4V Trustworthiness. *arXiv e-prints*, page arXiv:2405.17220, May 2024. 3

[29] Mengxi Zhang, Wenhao Wu, Yu Lu, Yuxin Song, Kang Rong, Huanjin Yao, Jianbo Zhao, Fanglong Liu, Yifan Sun, Haocheng Feng, and Jingdong Wang. Automated multi-level preference for mllms, 2024. 5

[30] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *CoRR*, abs/2311.16839, 2023. 1

[31] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024. 1, 5

[32] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *Proceedings of ICLR*, 2024. 1

[33] Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. IBD: alleviating hallucinations in large vision-language models via image-biased decoding. *CoRR*, abs/2402.18476, 2024. 1