Automated Multi-level Preference for MLLMs

Mengxi Zhang[1,2], **Wenhao Wu**[3], **Yu Lu**[4], **Yuxin Song**[1], **Kang Rong**[1], **Huanjin Yao**[1,5]
Jianbo Zhao[6], **Fanglong Liu**[1], **Haocheng Feng**[1], **Jingdong Wang**[1], **Yifan Sun**[1*]
[1]Baidu Inc. [2]Tianjin University [3]The University of Sydney
[4]University of Technology Sydney [5]Tsinghua University [6]Chinese Academy of Science

Abstract

Current multimodal Large Language Models (MLLMs) suffer from "hallucination", occasionally generating responses that are not grounded in the input images. To tackle this challenge, one promising path is to utilize reinforcement learning from human feedback (RLHF), which steers MLLMs towards learning superior responses while avoiding inferior ones. We rethink the common practice of using binary preferences (*i.e.*, superior, inferior), and find that adopting multi-level preferences (*e.g.*, superior, medium, inferior) is better for two benefits: 1) It narrows the gap between adjacent levels, thereby encouraging MLLMs to discern subtle differences. 2) It further integrates cross-level comparisons (beyond adjacent-level comparisons), thus providing a broader range of comparisons with hallucination examples. To verify our viewpoint, we present the Automated Multi-level Preference (**AMP**) framework for MLLMs. To facilitate this framework, we first develop an automated dataset generation pipeline that provides high-quality multi-level preference datasets without any human annotators. Furthermore, we design the Multi-level Direct Preference Optimization (MDPO) algorithm to robustly conduct complex multi-level preference learning. Additionally, we propose a new hallucination benchmark, MRHal-Bench. Extensive experiments across public hallucination and general benchmarks, as well as our MRHal-Bench, demonstrate the effectiveness of our proposed method. Code is available at `https://github.com/takomc/amp`.

1 Introduction

Multimodal Large Language Models (MLLMs) [1, 2, 3, 4, 5, 6] have achieved remarkable advancement in vision-language understanding tasks, *e.g.*, vision question answering [7], image captioning [8], and human-machine conversation. Despite MLLMs achieving significant breakthroughs, they still suffer from hallucinations [9, 10], referring to responses that are not accurately anchored to the context provided by images. This problem shrinks the performance of MLLMs and draws considerable research attention. To mitigate the hallucinations, some existing methods [11, 12, 13, 14] adopt Reinforcement Learning from Human Feedback (RLHF) methods, which collect human/AI preferences and integrate them into the MLLMs optimization process via reinforcement learning.

Existing RLHF methods have demonstrated that comparing superior and inferior responses within a binary-level preference framework can improve the performance of optimized MLLMs. However, *Is a single comparison between superior and inferior responses sufficient for preference learning in MLLMs?* Upon consideration, we find that a multi-level preference framework offers greater benefits for preference learning, primarily due to two main intuitive advantages. Firstly, reducing the gap between adjacent levels helps mitigate the challenge of distinguishing micro hallucinations in responses. As depicted in Fig. 1, in the baseline method (*i.e.*, binary preference), significant differences exist between the superior response A and the inferior response C. By introducing an additional medium response and shifting the focus to multiple comparisons between adjacent levels

*Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-0823

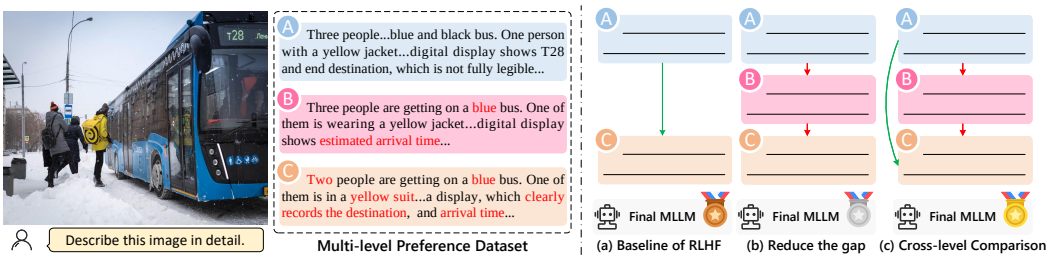

Figure 1: **Left:** Depicted are the input image, text prompt, and corresponding multi-level preference dataset. Contents highlighted in red signify hallucinations. Responses range from A to C, representing varying degrees of quality from superior to inferior. **Right:** Illustrating the strategy for leveraging inferior responses. (a) displays the conventional RLHF baseline, which adpots the binary-level preference. (b) To mitigate the gap between adjacent levels, we first split a single comparison into multiple comparisons by inserting extra medium responses. (c) Furthermore, we introduce the cross-level comparison to augment the dataset with more hallucination examples.

("A>B", "B>C"), as highlighted by the red arrows in Fig. 1(b), we mitigate this issue. Interestingly, under certain conditions, the MLLM's performance with "A>C" comparison is even worse than that with "B>C", indicating that reducing the gap between adjacent levels is sometimes more effective than enhancing the quality of superior responses. Secondly, cross-level comparisons can further enhance performance. In the comparisons between adjacent levels, the only comparison utilizing the inferior response A is "A>B", which may lead the model to focus more on suppressing hallucinations in response B. To address this, we introduce the cross-level comparison "A>C" (the green arrow in Fig. 1(c)) to provide more negative examples, thereby helping the model suppress more possible hallucinations. By integrating these strategies, we evolve the conventional binary-level preference learning into a more sophisticated multi-level preference learning framework.

However, exploring multi-level preferences for MLLMs poses significant challenges: 1) Labeling multi-level preference datasets is expensive and laborious. While some methods [11, 12] utilize human annotators to obtain preference labels, this approach is effective for binary datasets but falls short for multi-level preference datasets. Specifically, establishing a K-level preference dataset requires human annotators to make $K(K-1)/2$ comparisons. For example, with $K = 5$, this results in 10 comparisons, significantly more than is required for binary datasets. On the other hand, datasets annotated by humans or AI often contain significant noise and bias. To investigate this, we collected preferences from both humans and GPT-4V [15] on a subset of ShareGPT4V [16], using three MLLMs to generate varied responses. Setting K to 3, we compared pairs of responses (A&B, B&C, A&C) through three comparisons. However, we observed a frequent contradictory pattern (A>B, B>C, C>A), with rates of approximately 14% and 11% in human and GPT-4V annotations, respectively, resulting in a low-quality multi-level preference dataset. 2) The optimal multi-level preference learning objective remains unclear. While multi-level preference is more beneficial for optimizing MLLMs, it introduces greater complexity than binary preference. Therefore, it requires an effective algorithm to fully utilize the knowledge embedded within multi-level preference datasets.

To overcome the challenges outlined above, we introduce innovative strategies at both the data and method levels: 1) At the data level, we propose two novel methods for generating initial multi-level preference datasets without human or AI annotators. Furthermore, we implement an auto-check mechanism to further refine these datasets by evaluating the scores and accuracy of the generated responses. 2) At the method level, we introduce the Multi-level Direct Preference Optimization (MDPO) algorithm, a derivative of the traditional Direct Preference Optimization (DPO) algorithm [17]. The MDPO algorithm extends the capabilities of the DPO algorithm to facilitate multi-level preference optimization. Additionally, we incorporate a tailored penalty term into the MDPO learning objective to ensure robust multi-level preference learning. 3) Finally, we introduce a new evaluation benchmark, MRHal-Bench, which is the first designed specifically to evaluate hallucinations in multi-round dialogues. In summary, our contributions are as follows:

- Contrary to prior RLHF studies that focused solely on enhancing the quality of superior responses, our findings indicate that inferior responses can also play a crucial role in reducing hallucinations under the multi-level preference learning framework.

https://doi.org/10.52202/079017-0823

- To support effective multi-level preference learning, we develop two novel methods and an auto-check mechanism, enabling the creation of high-quality multi-level preference datasets without the need for human or AI annotators. Furthermore, we design the Multi-level Direct Preference Optimization (MDPO) algorithm with a specifically crafted learning objective, allowing MLLMs to robustly learn from the multi-level preference dataset.
- Our extensive experiments across various hallucination benchmarks confirm the effectiveness of our framework. Additionally, we have introduced MRHal-Bench, the first benchmark specifically designed to evaluate hallucinations in multi-round dialogues.

2 Related Work

2.1 Multimodal Large Language Models

Recently, the multimodal learning community has witnessed the great success of MLLMs [1, 2, 3, 4, 5, 6, 18], which employ a cross-modal alignment module to connect the visual encoder [19, 20, 21] and the language model [22, 23]. Typically, MLLMs undergo a standard training strategy involving two stages. First, to bridge the gap between visual and textual representations, the cross-modal alignment module is trained on a large-scale multimodal dataset [1, 24, 25], which endows the LLMs with visual-understanding ability. Then, MLLMs are further fine-tuned on specific visual instruction datasets [2, 16, 18, 26] to facilitate various downstream tasks [7, 8]. Despite the significant advancement, MLLMs still suffer from hallucinations, which decrease their performance on multiple tasks and attract increasing attention from researchers.

2.2 Hallucinations in MLLMs

Hallucinations in MLLMs [9, 10] denote inconsistencies between the input image and the generated response. Unlike hallucinations in LLMs [27, 28], those observed in MLLMs are more complicated, which attracts more attention from researchers. Some methods [26, 29] focus on reducing hallucinations by constructing high-quality datasets, while others employ specialized mechanisms such as decoding strategies [30, 31], retrieval augmented generation [32], and chain-of-thought [33] to mitigate hallucinations. However, due to the inherent limitations of cross-entropy loss, these methods may provide insufficient guidance for modality alignment. Recently, reinforcement learning-based methods [11, 12, 13, 14, 34], leveraging techniques like DPO [17] and PPO [35], have emerged as promising direction. Yet, these methods rely on preference datasets annotated by humans or AI, which are costly and susceptible to noise. Besides, they follow the traditional binary-level preference framework, which is insufficient for preference learning of MLLMs. To address these problems, we propose a novel AMP framework, utilizing a human-free multi-level preference dataset and the MDPO algorithm to guide MLLMs.

3 Methods

In this section, we delve into the Automated Multi-level Preference (AMP) framework. Initially, we outline two strategies for constructing an initial multi-level preference dataset, aligning with two perspectives of the scaling law [36, 37]. Subsequently, we introduce the auto-check mechanism aimed at refining the initial dataset based on relevant metrics. Lastly, we introduce the Multi-level Direct Preference Optimization (MDPO) algorithm, featuring a novel and robust learning objective.

3.1 Human-free Multi-level Preference Dataset Generation

The quality of the preference dataset significantly influences the refined model's performance. Constructing a high-quality initial preference dataset relies on two fundamental principles. Firstly, the ranking between superior and inferior responses should be correct in most cases. Secondly, the language style among different responses is expected to be consistent. Specifically, inconsistent language styles can introduce biases that mislead the MLLM, resulting in reward hacking and performance degradation [12, 36]. Considering these factors, we propose the ***Multi-size Expert Generation (MEG)*** and ***Incremental Generation (IG)*** strategies to build reliable preference datasets from the perspectives of model size and dataset size, respectively.

https://doi.org/10.52202/079017-0823

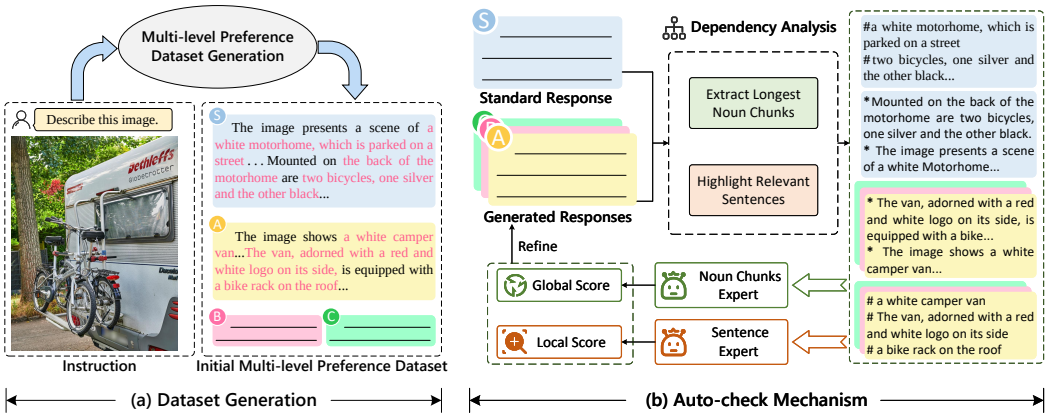

Figure 2: Pipeline for Constructing Human-free Multi-level Preference Dataset. We initiate the process with *Multi-size Expert Generation* and *Incremental Generation* to establish the initial dataset. Then, to enhance the quality of the initial preference dataset, we introduce the *Auto-check Mechanism*, which calculates both global and local metrics based on sentences and noun chunks, respectively.

3.1.1 Multi-size Expert Generation

Scaling laws suggest that the performance of the model improves as the model size increases. Thus, a logical strategy is to generate various responses using models of different sizes. For consistency in language style, it's preferable that these models stem from the same family. Specifically, we adpot LLaVA-based models, such as LLaVA-2B [38], LLaVA-7B, LLaVA-13B, and LLaVA-34B [3]. Leveraging the standard response in the instruction tuning dataset [29], we procure up to 5 responses of differing quality.

3.1.2 Incremental Generation

In *Multi-size Expert Generation*, our focus lies on employing models of various sizes, while *Incremental Generation* involves training datasets of different sizes. In practice, we partition the entire fine-tuned dataset $\mathcal{F} = \{\mathcal{I}; P; R\}$ into $K-2$ equal parts for the K-rank preference dataset, where $\mathcal{I}$, P, and R symbolize the image, text prompt, and standard response, respectively. Then, we use subsets $\mathcal{S}_i = [\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_i]$ to fine-tune the pre-trained MLLM $\mathcal{M}$, yielding $K-2$ fine-tuned MLLMs, where $i \in [1, K-2]$. Hence, the $K-2$ responses generated by fine-tuned MLLMs, along with the response generated by the pre-trained MLLM and the standard response constitute the K-rank preference dataset. The entire process is documented in Algorithm 1.

Algorithm 1 The Pseudocode of Incremental Generation for K-rank Preference Dataset.

Input: Image $\mathcal{I}$, text prompt P, and standard response R for fine-tuned dataset $\mathcal{F} = \{\mathcal{I}; P; R\}$, annotated dataset $\mathcal{A} = \{\mathcal{I}_a; P_a; R_a\}$, pre-trained MLLM $\mathcal{M}$.
Output: K-level preference dataset $\mathcal{D} = \{\mathcal{I}_a; P_a; [R_0, R_1, ..., R_{K-1}]\}$.

```
Split F into K − 2 equal parts [F_1, F_2, ..., F_{K−2}];
for (i = 1 to K − 2) do
    Train M with subset S_i = [F_1, F_2, ..., F_i] ⇒ Get fine-tuned MLLM M_i;
    R_i = M_i(I_a, P_a);        {Generate response R_i via fine-tuned MLLM M_i}
end for
R_0=R_a, R_{K−1}=M(I_a, P_a);
return D = {I_a; P_a; [R_0, R_1, ..., R_{K−1}]}.
```

3.1.3 Auto-check Mechanism

In the aforementioned process, we devised two strategies to establish the initial multi-level preference dataset. While the rankings in this dataset are generally accurate, occasional anomalies may lead to incorrect preferences. To enhance the ranking accuracy, we introduce the auto-check mechanism.

First, we identify all nouns in the various responses, including terms like "motorhome", "street", *etc*. Note that certain nouns are deprecated (further details are provided in Appendix A.1). Next, we analyze the dependency relationships within the sentence to extend each noun into the longest possible noun chunks. For example, "a white motorhome, which is parked on a street" would be represented (denoted by pink color in Fig. 2).

After extracting all noun chunks, we send them into the noun chunk expert (*i.e*., the text encoder of CLIP [19]) to obtain text features $F_S = \{f_{S_1}, f_{S_2}, ..., f_{S_\mathrm{M}}\}$ and $F_G = \{f_{G_1}, f_{G_2}, ..., f_{G_\mathrm{N}}\}$, where M and N denote the number of noun chunks in standard and generated responses, respectively. We then calculate the similarity score as outlined in Eq. 1:

$$\mathbf{S}[m,n] = \frac{f_{S_m} \cdot f_{G_n}}{\|f_{S_m}\| \times \|f_{G_n}\|}, \qquad \mathbf{s}[m] = max(\mathbf{S}[m,:]), \tag{1}$$

where $\mathbf{S} \in \mathbb{R}^{\mathrm{M}\times\mathrm{N}}$ is the similarity matrix between standard and generated responses. $\mathbf{s} \in \mathbb{R}^{\mathrm{M}}$ represents the similarity score of generated response F_G. We further introduce the accuracy metric:

$$\mathbf{p}[m] = \begin{cases} 1 & \text{if} \quad \mathbf{s}[m] > \tau \\ 0 & \text{otherwise} \end{cases}, \qquad \mathrm{Acc} = \mathrm{Sum}(\mathbf{p})/\mathrm{M}, \tag{2}$$

where τ is the threshold, set to 0.85. Accuracy (Acc) reflects the completeness of the credible components within the generated response.

While noun chunks represent the consistency at a local level, entire sentences represent global consistency, such as the relationships between multiple objects and the actions of objects, *etc*. To assess global consistency, we retrieve the sentences where each noun chunk is located as the global representation. The relevant metrics of sentences are the same as those of noun chunks.

Finally, we compute the final accuracy and scores by averaging the local and global metrics. Among the generated responses, the one with the highest accuracy is regarded as the best. In cases where multiple responses achieve equal accuracy, the one with the highest scores is considered superior.

3.2 Multi-level Direct Preference Optimization (MDPO)

Reinforcement learning algorithms [11, 12, 13, 14, 34] have demonstrated promising results in training MLLMs with human-preference datasets. Encouraged by the success of these pioneers, we delve deeper into the potential of multi-level preferences. In this section, we design the Multi-level Direct Preference Optimization (MDPO) algorithm, furnishing a novel and robust learning objective.

3.2.1 Preliminary

Prevalent methods [11, 39, 40] leverage the Proximal Policy Optimization (PPO) algorithm to align with preference data. However, the performance of this approach highly depends on the extra reward model, which is sensitive to noises within the preference dataset. Besides, the last stage of PPO fine-tunes the actor and critic model with the online strategy, resulting in high computational costs and unstable procedures. To mitigate these issues, DPO [17] excludes the reward model by analytically expressing reward functions with optimal policy π_* and initial policy π_{ref}. Denote x and y as the inputs and outputs of MLLMs, the reward function is converted into:

$$r(x,y) = \beta \log \frac{\pi_*(y|x)}{\pi_{\mathrm{ref}}(y|x)} + \beta \log Z(x), \tag{3}$$

where $Z(\cdot)$ is the partition function, β is a constant. Under the Bradley-Terry model, the policy objective becomes:

$$\begin{aligned} \mathcal{L}_{\mathrm{DPO}}(\pi_\theta; \pi_{\mathrm{ref}}) &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(r(x,y_w) - r(x,y_l)\right)\right], \\ &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)}\right)\right], \end{aligned} \tag{4}$$

where $\sigma(\cdot)$ represents the Sigmoid function, and x, y_w, and y_l denote inputs, superior and inferior responses, respectively. In practice, π_{ref} remains frozen during DPO training. Thus, only the policy model π_θ is updated in the training process, ensuring efficiency and cost-effectiveness.

https://doi.org/10.52202/079017-0823

3.3 Learning Objective of MDPO Algorithm

To facilitate the multi-level preference dataset, we revise the learning objective for K-rank with $K(K-1)/2$ comparisons:

$$\begin{aligned}\mathcal{L}_{\mathrm{MDPO}}(\pi_\theta;\pi_{\mathrm{ref}}) &= -\sum_{i=0}^{K-1}\sum_{j=i+1}^{K-1}\mathcal{L}_{\mathrm{DPO}(x,y_i,y_j)\sim\mathcal{D}},\\ &= -\sum_{i=0}^{K-1}\sum_{j=i+1}^{K-1}\mathbb{E}_{(x,y_i,y_j)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_i|x)}{\pi_{\mathrm{ref}}(y_i|x)}-\beta\log\frac{\pi_\theta(y_j|x)}{\pi_{\mathrm{ref}}(y_j|x)}\right)\right],\end{aligned} \tag{5}$$

where the quality of response y_i is superior to y_j.

During MDPO training, we observed that despite the loss decreasing normally, the optimized MLLM sometimes generates certain words or phrases repetitively. This occurs because the probability of the policy model producing both superior and inferior responses simultaneously decreases. While the probability of generating inferior responses declines more rapidly, the policy model's capability to generate superior responses also diminishes, leading to an overall deterioration in performance. To mitigate this risk, we introduce an additional penalty term, modifying Eq. 4 as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{DPO-P}}(\pi_\theta;\pi_{\mathrm{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[\log\sigma\Bigg(\beta\log\frac{\pi_\theta(y_w\mid x)}{\pi_{\mathrm{ref}}(y_w\mid x)} &-\beta\log\frac{\pi_\theta(y_l\mid x)}{\pi_{\mathrm{ref}}(y_l\mid x)}\Bigg)\\ &+\log\frac{\pi_\theta(y_w\mid x)}{\pi_{\mathrm{ref}}(y_w\mid x)}\Bigg].\end{aligned} \tag{6}$$

With this penalty term, the probability of generating superior responses is explicitly improved. To minimize the impact of medium-quality responses, we apply the penalty term exclusively to the best response. Consequently, the learning objective of MDPO is formulated as:

$$\mathcal{L}_{\mathrm{MDPO}}(\pi_\theta;\pi_{\mathrm{ref}}) = -\left[\sum_{j=1}^{K-1}\mathcal{L}_{\mathrm{DPO-P}(x,y_0,y_j)\sim\mathcal{D}}+\sum_{i=1}^{K-1}\sum_{j=i+1}^{K-1}\mathcal{L}_{\mathrm{DPO}(x,y_i,y_j)\sim\mathcal{D}}\right]. \tag{7}$$

4 Experiments and Analysis

4.1 Implementation Details

We adpot LLaVA-v1.5 [3] as our base model for all experiments, which is built upon Vicuna [22, 23] and utilizes ViT-L/14 [19] as the image encoder. Our training dataset contains 1k detailed captions from ShareGPT4V [16], 4k image-text pairs from [34], 4k human-annotated data from [12] and 2k multi-round dialogues annotated by us (the annotated process is detailed in Appendix A.2), forming a total of 11k training instances. For training MDPO, we employ the AdamW [41] optimizer for 4 epochs and apply a peak learning rate of 5×10^{-5} with the cosine decay strategy. To enhance learning efficiency, we incorporate LoRA-based [42] fine-tuning, with a low-rank r set to 64 for both attention and feed-forward modules. All experiments are conducted with a batch size of 16 on 8 Nvidia A100 GPUs with 40G VRAM. Further implementation details of the Human-free Multi-level Preference Dataset generation are provided in Appendix A.3.

4.2 Evaluation Benchmarks

To verify the effectiveness of our proposed AMP framework, we conduct comprehensive comparisons with various baselines across several benchmarks, including QA-based hallucination benchmark POPE [9], fine-grained hallucination benchmark MMHal-Bench [11], general benchmark LLaVA-Bench [2], and our newly developed multi-round dialogue hallucination benchmark MRHal-Bench. Specifically, POPE assesses the object existence hallucinations by prompting MLLMs to provide

Table 1: Comparison of conventional MLLMs and RLHF-based MLLMs across MMHal-Bench, MRHal-Bench, and LLaVA-Bench. "MEG" represents training data generated via Multi-size Expert Generation, while "IG" indicates training data produced using Incremental Generation.

Methods	MMHal-Bench		MRHal-Bench		LLaVA-Bench		
	Score↑	Hal.↓	Score (c/m)↑	Hal. (c/m)↓	Conv.↑	Detail↑	Comp.↑
LLaVA_{13B} [2]	1.11	0.84	3.01 / 3.01	0.40 / 0.37	85.4	74.3	96.3
InstructBLIP_{7B} [4]	1.80	0.62	3.00 / 3.00	0.39 / 0.38	83.2	67.6	90.6
LLaVA-v1.5_{7B} [3]	2.01	0.61	3.38 / 3.39	0.32 / 0.29	80.2	75.9	89.2
DeepSEEK-VL [44]	2.22	0.56	3.54 / 3.53	0.29 / 0.25	74.4	76.5	78.2
$\text{LLaVA-V1.6 [3]}_{7B}$	2.30	0.59	3.80 / 3.78	0.27 / 0.26	82.3	85.3	96.9
MiniCPM-V [45]	2.34	0.50	3.31 / 3.31	0.39 / 0.34	80.8	75.6	89.2
LLaVA-v1.5_{13B} [3]	2.44	0.53	3.58 / 3.59	0.29 / 0.27	81.6	75.5	95.2
Qwen-VL-Chat [6]	2.70	0.46	3.71 / 3.68	0.27 / 0.21	81.9	77.1	92.3
$\text{LLaVA-V1.6 [3]}_{13B}$	3.04	0.43	3.73 / 3.79	0.30 / 0.25	89.2	90.3	98.3
LLaVA-RLHF_{7B} [11]	2.04	0.68	3.58 / 3.56	0.34 / 0.29	85.3	74.7	105.6
LLaVA-RLHF_{13B} [11]	2.53	0.57	3.26 / 3.27	0.45 / 0.38	93.8	74.3	**111.4**
RLHF-V [12]	2.66	0.52	2.54 / 2.60	0.52 / 0.56	93.1	75.3	91.6
POVID [14]	2.69	–	3.46 / 3.47	0.28 / 0.28	75.7	75.2	89.5
SILKIE [34]	3.02	–	3.71 / 3.70	0.30 / 0.29	86.3	76.4	95.3
FGAIF [13]	3.09	0.36	3.77 / 3.79	0.30 / 0.31	**98.2**	**93.6**	110.0
AMP-MEG_{7B}	3.17	0.35	4.07 / 4.06	0.20 / 0.15	89.7	89.1	98.8
AMP-MEG_{13B}	**3.23**	**0.34**	**4.21 / 4.21**	**0.15 / 0.11**	94.4	91.2	95.6
AMP-IG_{7B}	3.12	0.41	4.02 / 4.04	0.22 / 0.13	90.2	85.9	99.8
AMP-IG_{13B}	3.18	0.36	3.96 / 4.01	0.22 / 0.20	91.3	86.8	99.4

Hal.: Hallucination rate, Conv.: Conversation, Detailed: Detailed Description, Comp.: Complex Question, c/m: cumulative/mean.

binary responses ('yes' or 'no'). MMHal-Bench is designed to quantify hallucinations with the assistance of GPT-4 [43]. Different from the simple questions in conventional benchmarks, MMHal-Bench contains more general, open-ended, and fine-grained questions. LLaVA-Bench serves as a general benchmark for systematic comprehension, encompassing three categories: conversation, detailed description, and complex questions. In addition to these benchmarks, we introduce MRHal-Bench to evaluate hallucinations in multi-round dialogues, covering six aspects: attribute, description, existence, counting, reasoning, and spatial relation. For further details, please refer to Appendix A.4.

4.3 Comparisons with Leading Methods

We compare our method with multiple MLLMs, including two types of state-of-the-art models: 1) General MLLMs. We include LLaVA [2], InstructBLIP [4], LLaVA-V1.5 [3], and Qwen-VL-Chat [6] as high-performing, open-sourced general models. These models are trained on extensive datasets and demonstrate promising results across various tasks. 2) RLHF-based MLLMs. Our comparisons also extend to RLHF models such as LLaVA-RLHF [11], RLHF-V [12], POVID [14], SILKIE [34], and FGAIF [13]. Specifically, LLaVA-RLHF employs the PPO algorithm on 10k human-preference data and 72k factually augmented data for reward and policy models, respectively. RLHF-V utilizes 1.4k human-annotated, fine-grained preference data to optimize the policy model using the DPO algorithm. Both [13] and [34] apply the DPO algorithm to align MLLMs with GPT-4V preferences. POVID [14] generates hallucination examples through two strategies and also uses the DPO algorithm.

The quantitative results are shown in Table 1 and 2. We observe that our AMP surpasses all general MLLMs across all benchmarks, highlighting the benefits of further fine-tuning with the preference dataset. Besides, our method also achieves state-of-the-art performance among RLHF-based methods, which comes from two aspects. First, our MDPO algorithm facilitates multi-level preference learning, which enables the MLLM to discern semantic granularity among different responses. Second, the accurate ranking of our human-free preference dataset ensures reliable guidance for the MLLM, leading to more promising performance. We also provide some qualitative case studies in Fig 3. For more cases, please refer to Appendix A.5.

https://doi.org/10.52202/079017-0823

Table 2: Comparisons on the POPE benchmark. * indicates evaluations using the official model.

Methods	Adversarial		Popular		Random		Overall	
	F1↑	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑	Yes
LLaVA$_{13B}$ [2]	74.4	67.2	78.2	73.6	78.8	73.7	77.1	73.7
InstructBLIP$_{7B}$ [4]	70.4	65.2	80.2	79.7	89.3	88.6	80.0	59.0
DeepSeek-VL [44]	72.2	65.4	71.3	63.8	76.4	71.5	71.7	73.3
Qwen-VL-Chat [6]	80.7	83.2	81.6	84.2	82.1	84.2	81.5	36.7
MiniCPM-V [45]	83.5	83.4	86.2	86.5	88.9	89.2	86.2	47.8
LLaVA-V1.5$_{7B}$ [3]	84.5	85.5	86.0	87.1	87.2	88.0	85.9	42.2
LLaVA-V1.5$_{13B}$ [3]	84.5	85.5	86.3	87.4	87.1	88.0	86.0	42.2
LLaVA-V1.6$_{7B}$ [3]	**85.2**	**86.4**	86.4	87.6	87.6	88.5	86.4	41.5
LLaVA-V1.6$_{13B}$ [3]	**85.2**	**86.4**	86.4	87.7	87.2	88.2	86.3	41.0
LLaVA-RLHF$_{7B}$ [11]	79.5	80.7	81.8	83.3	83.3	84.8	81.5	41.8
LLaVA-RLHF$_{13B}$ [11]	80.5	82.3	81.8	83.9	83.5	85.2	81.9	39.0
RLHF-V [12]	83.6	84.6	85.3	86.4	87.2	88.1	85.4	42.7
POVID* [14]	84.0	84.7	85.8	86.8	87.7	88.5	85.8	43.6
SILKIE [34]	80.3	83.0	81.3	84.0	81.6	83.9	81.1	36.1
FGAIF [13]	79.9	79.6	83.7	84.0	86.7	87.0	83.4	48.3
AMP-MEG$_{7B}$	83.4	83.1	87.7	**88.2**	89.6	89.9	86.9	48.0
AMP-MEG$_{13B}$	83.4	82.8	**88.0**	**88.2**	**90.3**	**90.4**	**87.2**	49.8
AMP-IG$_{7B}$	82.3	82.5	87.0	87.8	87.7	88.3	85.7	45.1
AMP-IG$_{13B}$	83.0	82.7	86.0	86.3	89.6	90.0	86.2	46.3

Figure 3: Case studies including our AMP-MEG, LLaVA-V1.5 [3], and LLaVA-RLHF [11]. Hallucinations, correct responses are highlighted in different colors. Please zoom in for the best view.

Table 3: Study on preference quantity.

Settings	MMHal-Bench		MRHal-Bench		LLaVA-Bench		
	Score↑	Hal.↓	Score (c/m)↑	Hal.↓	Conv.↑	Detail↑	Comp.↑
2-level preference	2.69	0.47	3.71 / 3.72	0.27 / 0.22	81.5	74.7	83.5
3-level preference	2.88	0.42	3.83 / 3.87	0.24 / 0.18	84.1	84.6	94.1
4-level preference	**3.17**	**0.35**	**4.07 / 4.06**	**0.20 / 0.15**	**89.7**	**89.1**	**98.8**
5-level preference	2.96	0.41	3.93 / 3.95	0.22 / 0.17	88.5	84.8	92.9

4.4 Ablation Studies

Impact of Preference Quantity. We explore the effects of varying the number of preferences from 2 to 5, with detailed implementation found in Appendix A.6. As indicated in Table 3, a 4-level preference is identified as the optimal setting. We hypothesize that the diminished performance observed with a 5-level preference dataset may be due to increased hidden noise. Unless stated otherwise, all subsequent experiments are conducted on an MLLM using the Vicuna-7B, trained on the 4-level preference dataset produced through Multi-size Expert Generation.

Table 4: Impact of the gap between adjacent levels and cross-level comparison. Preferences are ranked from most superior to most inferior in the following order: S, A, B, C.

Settings	MMHal-Bench		MRHal-Bench		LLaVA-Bench		
	Score↑	Hal.↓	Score (c/m)↑	Hal.↓	Conv.↑	Detail↑	Comp.↑
S>B	2.50	0.50	3.56 / 3.57	0.28 / 0.28	79.7	71.5	80.3
S>A	2.61	0.51	3.63 / 3.62	0.27 / 0.25	82.8	74.9	84.1
S>A & A>B	2.68	0.43	3.69 / 3.71	0.29 / 0.22	83.0	79.6	87.0
+Cross-level Comparison	2.79	0.44	3.75 / 3.73	0.27 / 0.22	87.7	78.6	90.1
S>A & A>B & B>C	2.85	0.40	3.86 / 3.90	0.24 / 0.17	**90.2**	81.3	92.4
+Cross-level Comparison	**3.17**	**0.35**	**4.07 / 4.06**	**0.20 / 0.15**	89.7	**89.1**	**98.8**
A>C	2.33	0.57	3.37 / 3.37	0.34 / 0.34	75.7	70.4	80.1
B>C	2.45	0.51	3.50 / 3.50	0.31 / 0.27	77.3	72.2	81.6

Table 5: Ablations on the human-free multi-level preference dataset using different annotations, including AI (*i.e.*, GPT-4V), Auto-check, and initial annotations from MEG and IG.

Dataset	Preference Annotation	MMHal-Bench		MRHal-Bench		LLaVA-Bench		
		Score↑	Hal.↓	Score (c/m)↑	Hal.↓	Conv.↑	Detail↑	Comp.↑
AMP-MEG	AI	2.87	0.44	3.87 / 3.88	0.25 / 0.21	92.3	79.6	89.9
AMP-MEG	Auto Check	**3.17**	**0.35**	**4.07 / 4.06**	**0.20** / 0.15	89.7	**89.1**	98.8
AMP-IG	Auto Check	3.12	0.41	4.02 / 4.04	0.22 / **0.13**	**90.2**	85.9	**99.8**
AMP-MEG	Initial	2.79	0.49	3.69 / 3.71	0.29 / 0.22	87.1	80.1	81.7
AMP-IG	Initial	2.80	0.48	3.75 / 3.77	0.26 / 0.21	89.2	78.3	81.9

Impact of Gap between Adjacent Levels. The effectiveness of multi-level preference learning is partly attributed to reducing the gaps between adjacent levels. As shown in Table 4, we reduce the gap by "S>B ⇒S>A" and "A>C ⇒B>C", both of which result in performance enhancements.

Impact of Including More Comparisons. We further introduce more inferior responses, *i.e.*, 'Response B' and 'Response C', by "S>A ⇒ S>A & S>B" and "S>A & S>B ⇒ S>A & A>B & B>C". The improvements depicted in Table 4 verify that inferior responses are also beneficial for preference learning. To provide more comparisons between the best response and hallucination examples, we devise cross-level comparisons based on settings "S>A & S>B" and "S>A & A>B & B>C". As illustrated in Table 4, this strategy brings extra performance improvement across multiple benchmarks, indicating the necessity of cross-level comparisons.

Comparisons with AI-Annotated Preference. Similar to reinforcement learning methods using AI feedback [46], we use GPT-4V [15] to directly rank the responses generated by MEG based on their visual faithfulness and helpfulness. Table 5 illustrates that training with our preference dataset yields more effective results compared to the AI-annotated preference dataset. This suggests that our human-free multi-level preference dataset contains less noise. Furthermore, the performance of the MLLM significantly decreases in the absence of our Auto-check mechanism, highlighting its crucial role in accurately refining the ranking of the multi-level preference dataset.

4.5 Comparisons with other Rank-based Preference Alignment Approaches.

We make some empirical comparisons by replacing our MDPO with the learning objectives of [47] and [48]. As reported in Table 6, our MDPO surpasses these two learning objectives on all hallucination benchmarks, $e.g.$, 3.01 -> 3.17 on MMHal-Bench. The superiority of our MDPO comes from two aspects. First, our MDPO mitigates the challenge of distinguishing micro hallucinations in responses. Taking 3-level preference as an example, the comparisons of other methods are 'A>BC, B>C', while comparisons made by our AMP are 'A>B, A>C, B>C'. More specifically, our AMP splits 'A>BC' into 'A>B, A>C', which enables MLLMs to perceive the subtle differences between different responses. Second, our penalty term explicitly increases the probability of MLLMs generating good answers, ensuring the stability of the training process.

We also conduct some experiments about perturbation-based (PB) methods. Our implementation details are as follows. We randomly change the noun, adjective, preposition, and numeral. We obtain

https://doi.org/10.52202/079017-0823

Table 6: Performance on three hallucination benchmarks across other loss functions (#1, #2), MLLMs from different families (#3, DF), perturbation-based methods (#4, PB).

Settings	MMHal-Bench		MRHal-Bench		LLaVA-Bench		
	Score↑	Hal.↓	Score↑	Hal.↓	Conv.↑	Detail↑	Comp.↑
r. [47]	2.96	0.41	3.85 / 3.82	0.26 / 0.23	84.1	81.7	88.2
r. [48]	3.01	0.38	3.95 / 3.91	0.24 / 0.19	86.2	84.3	91.9
PB	2.83	0.46	3.61 / 3.52	0.33 / 0.35	78.4	75.1	81.3
MDPO	**3.17**	**0.35**	**4.07 / 4.06**	**0.20 / 0.15**	**89.7**	**89.1**	**98.8**

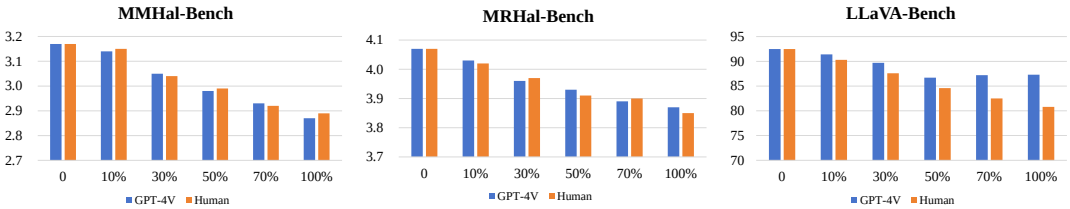

Figure 4: Performance on three hallucination benchmarks across different proportions of GPT-4V/human annotations.

answers of varying quality by controlling the proportion of perturbations (10%, 30%, 50% based on 4-level preference setting). As shown in Table 6, perturbation-based methods is still inferior to our MDPO. We infer the hallucination pattern generated by random perturbation is different from the real MLLM and is thus not informative enough for preference learning.

4.6 Evaluation of the Automated Multi-level Preference Dataset

We estimate the inconsistency rate of our AMP dataset to be 2.25% (through manual evaluation on 2000 random samples). The 2.25% inconsistency rate is significantly lower than the human (14.40% inconsistency) and GPT-4V (11.95% inconsistency) annotations.

Moreover, we conduct another experiment to validate the superiority of our AMP dataset. We mix the AMP and human/GPT-4V data for training the model. Fig. 4 shows that as the proportion of human/GPT-4V annotated data increases, the performance of MLLMs decreases accordingly.

5 Limitations

Our AMP framework offers more effective preference learning from the human-free multi-level preference dataset. However, several challenges remain: 1) The quality of standard responses limits the performance of the optimized MLLM. A portion of the standard responses in our dataset comes from superior responses generated by language models, potentially containing imperceptible hallucinations. Besides, the standard response is less helpful despite its high faithfulness, further restricting the performance. 2) Although our AMP successfully reduces hallucinations and promotes the truthfulness of MLLMs, the essence of preference learning is pushing the model to bias the preference dataset, causing a decrease in generalization ability. Therefore, finding a balance between preference learning and maintaining the capabilities of MLLMs is yet to be explored.

6 Conclusions

In this paper, we introduce the Automated Multi-level Preference (AMP) framework, achieving promising performance on several hallucination benchmarks, which benefits from the reduction of gaps in adjacent levels and the introduction of cross-level comparison. To enable the AMP framework, we propose a multi-level preference dataset generation pipeline, aiming to construct a high-quality preference dataset automatically. Furthermore, we design the Multi-level Direct Preference Optimization algorithm, which furnishes a novel learning objective to ensure robust and efficient preference learning. Lastly, we conduct the first hallucination benchmark in multi-round dialogues and devise the relevant metrics, which may stimulate future research.

References

[1] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[3] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[4] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint*, 2023.

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

[8] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[9] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[10] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

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

[12] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2023.

[13] Liqiang Jing and Xinya Du. Fgaif: Aligning large vision-language models with fine-grained ai feedback. *arXiv preprint arXiv:2404.05046*, 2024.

[14] 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.

[15] OpenAI. Gpt-4v(ision) system card. *OpenAI*, 2023.

[16] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[17] 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. *Advances in Neural Information Processing Systems*, 36, 2024.

 https://doi.org/10.52202/079017-0823

[18] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[20] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[21] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

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

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

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

[25] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[26] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

[27] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[28] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

[29] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M ^{3} it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023.

[30] 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. *arXiv preprint arXiv:2311.17911*, 2023.

[31] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023.

[32] Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. Preflmr: Scaling up fine-grained late-interaction multi-modal retrievers. *arXiv preprint arXiv:2402.08327*, 2024.

[33] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[34] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023.

[35] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[36] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[37] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[38] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024.

[39] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

[40] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[42] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[43] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

[45] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[46] 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. *arXiv preprint arXiv:2309.00267*, 2023.

[47] Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR, 2023.

[48] Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998, 2024.

[49] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

 https://doi.org/10.52202/079017-0823

[50] 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 the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

A Appendix

A.1 Omitted Nouns in Auto-check Mechanism

Similar to [5], we exclude some abstract nouns, *e.g.*, "time", "effect", *etc*. Besides, some high-frequency but unnecessary nouns, such as "image", "photo", *etc*, are also deprecated. The complete list is depicted in Fig. 5.

Unnecessary Nouns: 'photo', 'image', 'photograph', 'picture', 'figure', 'painting'.

Abstract Nouns: 'harmony', 'beauty', 'feel', 'concentration', 'life', 'matter', 'feature', 'effect', 'skill', 'detail', 'emotion', 'nature', 'time', 'background', 'placement', 'perspective', 'universality', 'attention', 'features', 'essence', 'peace', 'composition', 'element', 'foreground', 'appreciation', 'atmosphere', 'emotions', 'artwork', 'subject'.

Figure 5: Omitted nouns in the auto-check mechanism.

A.2 Annotated Process of Multi-round Dialogues in Training Dataset

We introduce 2k multi-round dialogues in our training dataset. To get diverse questions, we use GPT-4V [15] to generate questions and corresponding responses, where the prompt is as Fig. 6.

However, the responses generated by GPT-4V still contain hallucinations. To get the high-quality response, we further employ Qwen-VL-Chat [6], LLaVA-v1.5 13B [3], LLaVA-RLHF 13B [11], RLHF-V [12] to provide three extra responses. Thus, together with the response of GPT-4V, we get 5 responses in total. Then, we send these responses to human annotators and ask them to find the optimal response. If the optimal response still exhibits hallucinations, this image-text pair will be deprecated. Through these steps, we get the final 2k high-quality multi-round dialogues.

A.3 Details of Human-free Multi-level Preference Dataset Generation

A.3.1 Multi-size Expert Dataset Generation

To make the language style more consistent, we use models from the same family, including LLaVA-2B, LLaVA-7B, LLaVA-13B, and LLaVA-34B. We leverage the greedy decoding strategy with specific parameters: beams (4), temperature (0.7), repetition penalty (1.1), and max tokens (512).

A.3.2 Incremental Generation

In the Incremental Generation Generation, we obtain MLLMs with varying capabilities by using train datasets of different sizes. Specifically, we use the 7B version of LLaVA-V1.5 as the pre-trained model and follow the training detail of [3] and further fine-tune it on 30k/60k/90k high-quality image-text pairs. The whole dataset contains 10k ShareGPT4V [16], 20k Flickr30k [49], 30k VQAv2 [50], and 30k LRV [26]. Finally, we get 5 different responses, including the Ground Truth (GT), responses generated by 3 fine-tuned MLLMs and LLaVA-V1.5.

A.4 Multi-round Dialogue Hallucination Benchmark (MRHal-Bench)

To evaluate the hallucinations in multi-round dialogue, we build a Multi-round Dialogue Hallucination Benchmark, simplified by "MRHal-Bench". Specifically, MRHal-Bench contains 105 multi-round dialogues, where the length of rounds ranges from 2 to 5, with an average length of 2.99. The questions in MRHal involve five categories where MLLMs tend to generate hallucinations:

- Attribute: Visual characteristics of objects, including color, shape, state, type, *etc*.
- Description: Detailed descriptions of objects, behaviors, environments, background, foreground, *etc*.

https://doi.org/10.52202/079017-0823

You are an image content annotation expert, and you are seeing a single image. Design a multi-round conversation between you and a person asking about this photo.

The conversation format is
User: ********
GPT: ********
where the user is the role of the person, and GPT is the role of you.

The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers. Only include questions that have definite answers:
(1) Ask the questions that can be answered confidently. Do not ask ambiguous questions. Do not give any ambiguous answers.
(2) Your answers may involve some facts about the image, such as news, social, science, etc.

First, your questions may include the characteristics of all the objects or elements in the image, their type, color, and style, the number of objects, the movement of the characters, the precise location, text, etc. Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss events happening in the image, etc. Again, do not ask about uncertain details, do not imagine anything. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary. Your output should only contain the conversations without any other things.

Figure 6: The text prompt for GPT-4V in the annotated process of multi-round dialogue.

- Existence: Questions with absolute answers, *e.g.*, yes/no.
- Counting: The number of specific objects.
- Reasoning: Questions that require the model to integrate analysis based on image content or other knowledge to generate final responses.
- Spatial Relation: The relative or absolute spatial relationships of objects in given images.

To obtain a fair, objective, and trustable evaluation result, we employ GPT-4 via a meticulous template. Since GPT-4V also suffers from hallucinations, we use GPT-4 API and replace visual contents with category names and a standard human-generated answer, which is similar with [11]. The template will be released in our code.

We design two types of metrics for evaluating MRHal-Bench, *i.e.*, cumulative and mean metrics (denoted by c/m in Section 4). The cumulative ($Metric_{\mathrm{c}}$) and mean scores ($Metric_{\mathrm{m}}$) are calculated by,

$$\begin{aligned} Metric_{\mathrm{c}} &= \sum_{i=1}^{105} \left[\mathrm{Sum}\left(Metric_i\right)\right] / \sum_{i=1}^{105} \left[\mathrm{Len}\left(Metric_i\right)\right], \\ Metric_{\mathrm{m}} &= \left\{ \sum_{i=1}^{105} \left[\mathrm{Sum}\left(Metric_i\right) / \mathrm{Len}\left(Metric_i\right)\right] \right\} / 105, \end{aligned} \tag{8}$$

where $Metric_i \in \mathbb{R}^j$ symbolizes the Metric of i-th dialogue with j rounds.

MRHal-Bench will be available at this link.

Figure 7: Case studies in terms of short-form Question Answering (QA), long-form QA, Existence Hallucinations, and Image Caption, including our AMP-MEG, LLaVA-V1.5 [3], and LLaVA-RLHF [11]. Hallucinations, correct responses are highlighted in different colors.

A.5 Case Studies

In Fig. 7, we present some examples of our AMP alongside other MLLMs, *i.e.*, LLaVA-V1.5 [3], and LLaVA-RLHF [11] for an intuitive comparison. We focus on four typical scenarios. 1) Short-form

https://doi.org/10.52202/079017-0823

Question Answering (QA). Our AMP generally provides accurate responses, such as the counting (*e.g.*, "four bicycles") and character (*e.g.*, "Proverbial Philosophy"). 2) Long-form QA. Our AMP outperforms other MLLMs in terms of helpfulness and faithfulness. Specifically, our AMP accurately interprets all the valuable information in the given image, including the time, date, and temperature. In contrast, other MLLMs make a wrong judgment about or neglect the information in this image. 3) Existence Hallucinations. Compared with responses from other MLLMs, our AMP is not misdirected by "identify a person wearing a bright yellow helmet" and predicts the non-existence of this person. 4) Image Caption. For the detailed caption, our AMP captures all the significant visual components correctly and infers the location ("Swiss Alps") from the Swiss national flag and mountains. However, other MLLMs overlook this flag and generate some hallucinations about people and cars. These qualitative results verify the superiority of our AMP framework.

A.6 Implementation Details of Optimal-Level Experiments

In Section 4.4, we report the performance of K-level preferences, where K ranges from 2 to 5. When K is equal to 4, the refined MLLM gets the best performance. However, the performance of optimized MLLM varies from different model pools. Take $K = 3$ as the example, the model pool may be 'Response S&34B&13B', 'Response S&13B&7B', *etc*. The performance of optimized MLLM varies from different model pools. Therefore, we only report the optimal performance under each level, where the details of model pools are reported in Table 7.

Table 7: The model pools for each level preference.

Settings	GT	LLaVA-34B	LLaVA-13B	LLaVA-7B	LLaVA-2B
2-level preference	✓	✓			
3-level preference	✓	✓	✓		
4-level preference	✓		✓	✓	✓
5-level preference	✓	✓	✓	✓	✓

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide our contributions both in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We express our limitations in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/079017-0823

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, we claim our experimental settings. Our code is available at this link.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code and data is available at this link.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The training and test details are described in the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because it would be too computationally expensive.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

https://doi.org/10.52202/079017-0823

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide these information in the implementation details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper for assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

https://doi.org/10.52202/079017-0823

Answer: [Yes]

Justification: MRHal-Bench is available at this link.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.