Calibrated Self-Rewarding Vision Language Models

Yiyang Zhou[1]*, **Zhiyuan Fan**[5]*, **Dongjie Cheng**[6]*, **Sihan Yang**[7], **Zhaorun Chen**[2]
Chenhang Cui[8], **Xiyao Wang**[3], **Yun Li**[1], **Linjun Zhang**[4], **Huaxiu Yao**[1]
[1]UNC-Chapel Hill, [2]University of Chicago, [3]University of Maryland
[4]Rutgers University, [5]HKUST , [6]PolyU, [7]NTU, [8]NUS
yiyangai@cs.unc.edu, huaxiu@cs.unc.edu

Abstract

Large Vision-Language Models (LVLMs) have made substantial progress by integrating pre-trained large language models (LLMs) and vision models through instruction tuning. Despite these advancements, LVLMs often exhibit the hallucination phenomenon, where generated text responses appear linguistically plausible but contradict the input image, indicating a misalignment between image and text pairs. This misalignment arises because the model tends to prioritize textual information over visual input, even when both the language model and visual representations are of high quality. Existing methods leverage additional models or human annotations to curate preference data and enhance modality alignment through preference optimization. These approaches are resource-intensive and may not effectively reflect the target LVLM's preferences, making the curated preferences easily distinguishable. Our work addresses these challenges by proposing the Calibrated Self-Rewarding (CSR) approach, which enables the model to self-improve by iteratively generating candidate responses, evaluating the reward for each response, and curating preference data for fine-tuning. In the reward modeling, we employ a step-wise strategy and incorporate visual constraints into the self-rewarding process to place greater emphasis on visual input. Empirical results demonstrate that CSR significantly enhances performance and reduces hallucinations across ten benchmarks and tasks, achieving substantial improvements over existing methods by 7.62%. Our empirical results are further supported by rigorous theoretical analysis, under mild assumptions, verifying the effectiveness of introducing visual constraints into the self-rewarding paradigm. Additionally, CSR shows compatibility with different vision-language models and the ability to incrementally improve performance through iterative fine-tuning. Our data and code are available at https://github.com/YiyangZhou/CSR.

1 Introduction

Large Vision-Language Models (LVLMs) [1–4] have achieved significant success by incorporating pre-trained large language models (LLMs) and vision models through instruction tuning. However, these LVLMs suffer from the hallucination phenomenon [5], which generates text responses that are linguistically plausible but contradict the visual information in the accompanying image. For instance, the description generated by LVLMs may include visual elements that are not depicted in the image. This issue can also occur when the LLM is highly factual and the visual backbone is capable of producing sufficiently high-quality representations. As indicated in Cui et al. [6], Guan et al. [7], the potential reason for this lies in the misalignment problem between image and text modalities in LVLMs, which causes the model to prioritize the text knowledge present in the training language data while ignoring the actual visual input information.

*Equal contribution

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

Several works have been proposed to enhance modality alignment capability in LVLMs through preference fine-tuning techniques, such as reinforcement learning from human feedback (RLHF) [8] and direct preference optimization (DPO) [9, 10]. However, these methods often either introduce additional models, such as GPT-4, or depend on human annotation to generate preference data. This data generation process is not only resource-intensive but, more critically, fails to capture the inherent preferences of the target LVLM. Consequently, the target LVLM may easily discern preferences from such curated data, making them less effective (detailed analysis provided in Appendix A.4). Recently, self-rewarding approaches have emerged, utilizing a single LLM for both response generation and preference modeling, showing promising results in LLM alignment [11, 12]. Unlike LLMs, LVLMs face modality misalignment issues in both response generation and preference modeling stages, potentially resulting in self-generated preferences overlooking visual input information. Directly applying these self-rewarding approaches to LVLMs is not capable of addressing the modality alignment problem and redirecting LVLM's attention towards emphasizing input image information.

To tackle these challenges, our work introduces the **C**alibrated **S**elf-**R**ewarding (**CSR**) approach, aimed at calibrating the self-rewarding paradigm by incorporating visual constraints into the preference modeling process. Specifically, we train the target LVLM using an iterative preference optimization framework that continuously generates preferences and optimizes the target LVLM over multiple iterations. Starting with a seed model, each iteration employs sentence-level beam search [13, 14] to produce fine-grained candidate responses for each image and text prompt. During the beam search, for each generated sentence, we first utilize the language decoder to establish an initial reward (i.e., sentence-level cumulative probabilities). Subsequently, we calibrate this initial reward by incorporating an image-response relevance score, resulting in the calibrated reward score. These calibrated reward scores are utilized to guide the generation of the next batch of candidate sentences. Finally, responses with the highest and lowest cumulative calibrated reward scores are identified as preferred and dispreferred responses, respectively, for preference fine-tuning in the subsequent iteration.

The primary contribution of this paper is CSR, a novel calibrated self-rewarding paradigm for improving modality alignment in LVLMs. Theoretically, with mild assumptions, we show that introducing visual constraints in the self-rewarding paradigm can improve performance. Empirically, when compared with other competitive approaches (see Figure 1 for some representative methods), the results demonstrate that CSR is capable of improving performance on comprehensive LVLM evaluation benchmarks, VQA tasks, and reducing hallucination, achieving up to a 7.62% improvement on average. Additionally, we demonstrate CSR is capable of continuously improving performance over iterations, compatible with different large vision-language backbone models, and redirecting the attention of LVLMs toward the visual modality to achieve stronger modality alignment.

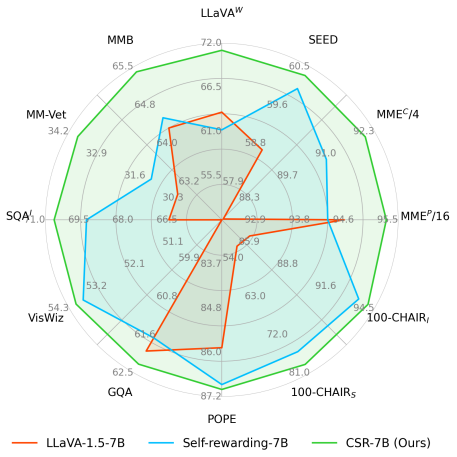

Figure 1: Benchmark performance comparison.

2 Preliminaries

In this section, we will provide a brief overview of LVLM and preference optimization.

Large Vision Language Models. LVLMs extend LLMs to multimodal scenario, which progressively predict the probability distribution of the next token for each input prompt. Given an <image x_v, text x_t> pair as input prompt x, LVLM outputs a text response y.

Preference Optimization. Preference optimization has shown promise in fine-tuning language models and aligning their behavior with desired outcomes. Given an input prompt x , a language model with policy π_θ can produce a conditional distribution $\pi_\theta(y \mid x)$ with y as the output text response. The preference data is defined as $\mathcal{D} = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^N$, where $y_w^{(i)}$ and $y_l^{(i)}$ denote the preferred

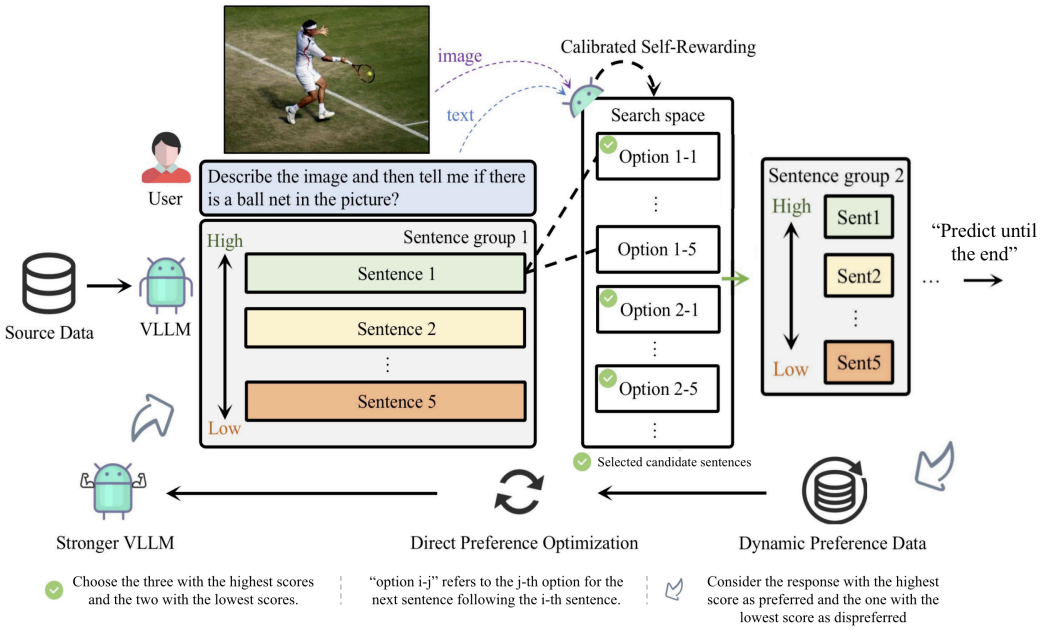

Figure 2: The CSR framework operates an iterative process of preference data generation and learning. During preference data generation, CSR utilizes a sentence-level beam search approach to construct responses sentence by sentence, assigning a reward to each sentence. This reward, initially generated by the model itself, is then calibrated using image-relevance information. Preferences are determined based on the cumulative reward for each response. In each iteration, CSR generates new preference data and performs preference learning based on this data, continuously enhancing the model's performance.

and dispreferred responses for the input prompt $x^{(i)}$. Preference optimization leverage the preference data to optimize language models. Taking DPO [15] as a representative example,, it formulates the probability of obtaining each preference pair as $p(y_w \succ y_l) = \sigma(r(x, y_w) - r(x, y_l))$, where $\sigma(\cdot)$ is the sigmoid function. DPO optimizes the language models with the following classification loss:

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\alpha\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \alpha\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right], \tag{1}$$

where $\pi_{\text{ref}}(y|x)$ represents the reference policy, i.e., the language model after performing supervised fine-tuning.

3 Calibrated Self-Rewarding Vision Language Models

To address this challenge, we propose **C**alibrated **S**elf-**R**ewarding (**CSR**), a novel approach aimed at improving modality alignment in LVLMs by integrating visual constraints into the self-rewarding paradigm. As illustrated in Figure 2, CSR trains the target LVLM by alternately performing two stages: candidate response generation and preference curation and fine-tuning. In the candidate response generation stage, we employ sentence-level beam search for each input prompt to produce fine-grained candidate responses. During this process, the language decoder determines the initial reward for each generated sentence, which is then calibrated by incorporating an image-response relevance score. This calibrated reward score guides the generation of subsequent sentences and finally generate the entire response. Moving on to the preference curation and fine-tuning stage, we use the responses with the highest and lowest cumulative calibrated rewards to construct the preferred and dispreferred responses, and utilize the constructed preference pairs for fine-tuning. In the remaining of this section, we will provide detailed explanations of CSR.

3.1 Step-Level Reward Modeling and Calibration

Before delving into how to generate candidate response and construct preference data, in this section, we first discuss how to formulate the reward within CSR. The ideal reward in the LVLM fulfills two specific criteria:

- Vision-Constrained Reward: This aspect aims to integrate image-relevance information into the reward definition of LVLMs. By doing so, we address the limitation of LVLM in overlooking image input data when generating preferences.
- Step-Wise Reward: Instead of assigning a single reward for the entire response, we opt for a step-wise approach. This involves assigning rewards at each step of response generation. Compared to a single reward, this finer-grained reward offers more detailed guidance and is more robust.

To fulfill these criteria, we propose a step-wise calibrated reward modeling strategy. Inspired by Process-Supervised Reward Models [16], we assign a reward score, $R(s)$, to each generated sentence s during the sentence-level beam search. This score is a combination of two components: the self-generated instruction-following score, $R_T(s)$, and the image-response relevance score, $R_I(s)$.

Specifically, the self-generated instruction-following score, $R_T(s)$, is calculated using the language decoder of the LVLM. It represents the sentence-level cumulative probability of generating sentence s, formulated as:

$$R_T(s) = \prod_{t=1}^{N_o} P(r_o \mid x, r_1, r_2, \ldots, r_{o-1}), \tag{2}$$

where N_o is the number of tokens in sentence s and r_o represents token o in sentence s. A higher self-generated instruction-following score indicates a stronger capability of the generated response to follow instructions.

While the self-generated instruction-following score partially reflects the LVLM's preference, it still suffers from modality misalignment, potentially overlooking visual input information. To address this, we introduce an image-response relevance score, $R_I(s)$, to calibrate the reward score $R_T(s)$. This score depicts the relevance between the generated sentence s and input image x_v. We leverage CLIP-score [17] for this calculation, where the vision encoder in the CLIP model aligns with the vision encoder in the target LVLM. The image-response relevance score $R_I(s)$ is defined as:

$$R_I(s) = \max(100 * \cos(\mathcal{F}_I(x_v), \mathcal{F}_T(s)), 0), \tag{3}$$

where the $\mathcal{F}_I(x_v)$ and $\mathcal{F}_T(s)$ represent the visual CLIP embedding and textual CLIP embedding, respectively. Finally, the calibrated reward score $R(s)$ for the generated sentence s is defined as:

$$R(s) = \lambda \cdot R_I(s) + (1 - \lambda) \cdot R_T(s), \tag{4}$$

where λ is a hyperparameter used to balance the language instruction-following and image-response relevance scores. By combining both scores, we aim to redirect the attention of LVLM towards the input visual information, thus enhancing its modality alignment ability.

3.2 Iterative Fine-Tuning

After establishing the reward framework in CSR, we next discuss our iterative fine-tuning process. Within this framework, we iteratively perform two essential steps, namely candidate response generation and preference data curation and optimization. These steps are elaborated upon as follows:

3.2.1 Step-Level Candidate Response Generation

In candidate response generation, our objective is to generate responses to build preference data. To accomplish this, we employ a sentence-level beam search strategy. Initially, we concurrently sample multiple candidate sentences, utilizing the "end of sub-sentence" marker (e.g., "." in English) as the delimiter. Subsequently, for each sentence s, we compute its reward score $R(s)$ using Eqn. (4). From these scores, we then select the top-k and bottom-k sentences with the highest and lowest reward scores, respectively, to proceed to the subsequent round of sentence-level beam search. This iterative process continues until reaching the "end of response," conventionally represented as $\langle$eos$\rangle$. Once all sentences for a response $y = \{s_1, \cdots, s_{N_y}\}$ are generated, we calculate the cumulative reward score for the response as the sum of the reward scores for each sentence within it. This is

Algorithm 1 Calibrated Self-Rewarding

Require: Dataset: $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$; Reference model:$\pi_{\text{ref}}$; Number of iterations: T

for $t = 1, \ldots, T$ **do**
  for each $x \in \mathcal{D}$ **do**
    while not reach the end of response **do**
      Generate a bunch of candidate sentences from last-round sentences
      for each candidate sentence s **do**
        Compute the self-generated instruction-following score $R_T(s)$ by Eqn. (2)
        Calculate the image representation $\mathcal{F}_I(x_v)$ and sentence representation $\mathcal{F}_T(s)$
        Compute the image-response relevance score $R_I(s)$ by Eqn. (3)
        Compute the calibrated reward score $R(s)$ by Eqn. (4)
      Select top-k and bottom-k sentences with the highest and lowest reward scores
    Select the preferred response $y_{w,t}$ and dispreferred response $y_{l,t}$
  Update $\pi_\theta \leftarrow \arg\min_\theta \mathcal{L}_t(\pi_\theta; \pi_{\text{ref}})$, $\pi_{\text{ref}} \leftarrow \pi_\theta$.

defined as: $R(y) = \sum_{i=1}^{N_y} R(s_i)$, where N_y is the number of sentences in response y. The detailed algorithm for candidate response generation is outlined in Algorithm 1.

3.2.2 Preference Curation and Optimization

After generating candidate responses with their reward scores, our next step is to curate preference dataset. Here, for each input prompt, we select the responses with the highest and lowest cumulative calibrated reward scores as the preferred and dispreferred responses, respectively, to construct the preference dataset for fine-tuning. For each iteration t, we denote the constructed preference data as: $\mathcal{D}_t = \{(x^{(i)}, y_{w,t}^{(i)}, y_{l,t}^{(i)})\}_{i=1}^{N}$. After obtaining the preference data, we fine-tune the target LVLM using DPO. At iteration t, we use the last iteration fine-tuned model $\pi_{\theta_{t-1}}$ as the reference model. Following Eqn (1), the loss at iteration t of CSR is defined as:

$$\mathcal{L}_t = -\mathbb{E}_{(x, y_{w,t}, y_{l,t}) \sim \mathcal{D}} \left[\log \sigma \left(\alpha \log \frac{\pi_\theta(y_{w,t}|x)}{\pi_{\theta_{t-1}}(y_{w,t}|x)} - \alpha \log \frac{\pi_\theta(y_{l,t}|x)}{\pi_{\theta_{t-1}}(y_{l,t}|x)} \right) \right]. \tag{5}$$

The training process of CSR is detailed in Algorithm 1.

4 Experiment

In this section, we empirically investigate CSR in addressing the modality misalignment problem of LVLMs, focusing on the following questions: (1) Can CSR help improve the performance of models on both comprehensive benchmarks and hallucination benchmarks? (2) Can CSR iteratively improve multimodal alignment progressively in LVLMs and lead to more factual LVLMs? (3) Is CSR compatible with different open-sourced LVLMs? (4) How does CSR change attention weights and preference pairs to align image and text modalities?

4.1 Experimental Setups

Implementation Details. We utilize LLaVA-1.5 7B and 13B [1] as the backbone models. During the preference learning process, we adapt LoRA fine-tuning [18]. The images and prompts used to construct the preference data are randomly sampled from the detailed description and complex reasoning subclasses of the LLaVA150k dataset, totaling approximately 13,000 samples [19]. It is worth noting that each iteration uses the same prompt and image as the previous round. Overall, the iterative training is conducted over three iterations, completed on one A100 80GB GPU. It takes roughly 3.5 and 5 hours for fine-tuning LLaVA-1.5 7B and LLaVA-1.5 13B, respectively. For more detailed information on training hyperparameters and training data, please refer to Appendix A.1.

Evaluation Benchmarks. We conducted evaluations on three types of benchmarks: comprehensive benchmarks, general VQA and hallucination benchmarks. Specifically, this includes: (1) Comprehensive benchmarks (MME [20], SEEDbench [21], LLaVA$^{\text{W}}$ [19], MMbench [22], MM-Vet [23]); (2) General VQA (ScienceQA (SQA) [24], VisWiz [25], GQA [26]); (3) Hallucination benchmark (POPE [27], CHAIR [28]). More detailed description are discussed in Appendix A.1.

Table 1: The performance of CSR on LLaVA-1.5 across all benchmarks is presented. Most baseline results, except those for self-rewarding, are sourced from Zhou et al. [10].

	Comprehensive Benchmark						General VQA			Hallucination Benchmark		
Method	MME^P	MME^C	SEED	$LLaVA^W$	MMB	MM-Vet	SQA^I	VisWiz	GQA	POPE	$CHAIR_S$	$CHAIR_I$
LLaVA-1.5-7B	1510.7	348.2	58.6	63.4	64.3	30.5	66.8	50.0	62.0	85.90	48.8	14.9
+ Vlfeedback	1432.7	321.8	59.3	62.1	64.0	31.2	66.2	52.6	**63.2**	83.72	40.3	13.2
+ Human-Prefer	1490.6	335.0	58.1	63.7	63.4	31.1	65.8	51.7	61.3	81.50	38.7	11.3
+ POVID	1452.8	325.3	60.2	68.7	64.9	31.8	68.8	53.6	61.7	86.90	35.2	8.3
+ RLHF-V	1489.2	349.4	60.1	65.4	63.6	30.9	67.1	**54.2**	62.1	86.20	29.7	7.5
+ Self-rewarding	1505.6	362.5	60.0	61.2	64.5	31.4	69.6	53.9	61.7	86.88	24.0	6.7
+ CSR (Ours)	**1524.2**	**367.9**	**60.3**	**71.1**	**65.4**	**33.9**	**70.7**	54.1	62.3	**87.01**	**21.0**	**6.0**
LLaVA-1.5-13B	**1531.3**	295.4	61.6	70.7	67.7	35.4	71.6	53.6	63.3	85.90	48.3	14.1
+ Self-rewarding	1529.0	300.1	62.8	65.6	64.5	35.3	74.3	56.1	63.2	86.58	37.0	8.8
+ CSR (Ours)	1530.6	**303.9**	**62.9**	**74.7**	**68.8**	**37.8**	**75.1**	**56.8**	**63.7**	**87.30**	**28.0**	**7.3**

Baselines. We will first compare CSR with the self-rewarding approach described by Yuan et al. [29]. Here, we directly apply self-rewarding to LVLM, using the prompts and experimental settings outlined in Yuan et al. [29] (see detailed settings in Appendix A.1 and Table 3). We also compared CSR with several data-driven preference learning methods, including Silkie (Vlfeedback) [9], LLaVA-RLHF (Human-preference) [8], POVID [10], and RLHF-V [30]. Furthermore, we compared the performance of the optimized LLaVA-1.5 via CSR with other state-of-the-art open-source LVLMs, including InstructBLIP [31], Qwen-VL-Chat [32], mPLUG-Owl2 [33], BLIP-2 [34], and IDEFICS [35], after the final rounds of training (CSR with iteration = 3). Additionally, to evaluate the effectiveness of CSR on other LVLMs, we applied CSR to a recent LVLM called Vila [36]. For more information on these baselines, please refer to Appendix A.1.

4.2 Results

CSR Continuously Improves Model Performance over Iterations. In Figure 3, we report the average performance of LLaVA-1.5 7B and 13B models concerning the number of training iterations on comprehensive benchmarks, general VQA tasks, and hallucination benchmarks. To facilitate score calculation, we first calculated an average score on a 100-point scale by adjusting the original values: MME^P was divided by 16, and MME^C was divided by 4, corresponding to the number of categories in MME. Additionally, since a lower CHAIR value indicates better performance, we standardized all metrics to follow a higher is better approach by transforming the $CHAIR_S$ and $CHAIR_I$ metrics into 100 - $CHAIR_S$ and 100 - $CHAIR_I$. We then calculated the average score by averaging these standardized values, which were used to compute the average percentage increase. In the experiment, the 7B model achieved an improvement of approximately 7.62% across all benchmarks through online iterative updates, while the 13B model saw an improvement of approximately 5.25%. According to the full results in Table 6 and Table 7 of Appendix A.5, the improvement is particularly significant on the $LLaVA^W$ and CHAIR benchmarks, with improvements of 8.9% and 49.50%, respectively. The results indicate that CSR is capable of incrementally improving model performance over iterations, demonstrating its effectiveness in self-improving the quality of generated preference data and leading to stronger modality alignment. The degree of improvement gradually becomes smaller, which is not surprising, indicating that the model is gradually converging.

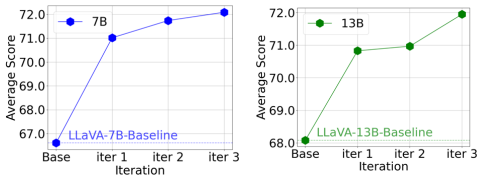

Figure 3: Average scores of CSR at different iterations over all benchmarks (see Table 6 and Table 7 in Appendix A.5 for full results).

CSR Outperforms Competitive Preference Fine-Tuning Baselines. Compared to preference data curation approaches (e.g., POVID, RHLF-V) that generate preference data from either additional models or human annotations, the superiority of CSR indicates that adapting a self-rewarding paradigm better captures the inherent preferences of the target LVLMs, achieving stronger modality alignment. Furthermore, CSR outperforms existing self-rewarding methods, with an average performance improvement of

Table 2: Ablation study of vision-text reward score.

Method	7B	13B
Base	66.61	68.08
Only R_T	68.46	68.12
Only R_I	67.49	69.23
CSR (Ours)	**72.39**	**71.95**

2.43%, demonstrating its effectiveness in calibrating the reward model by incorporating image-response relevance scores. This mitigates the potential issue of overlooking visual input information when estimating self-generated preferences.

In addition, we compare the performance of LLaVA-1.5 after three rounds of online CSR with other state-of-the-art open-sourced VLLMs and report the results in Table 5 of Appendix A.5. Although different open-sourced VLLMs utilize various image and text encoders, CSR still outperforms other open-sourced VLLMs in 9 out of 10 benchmarks, further corroborating the effectiveness of CSR in improving modality alignment.

4.3 Analysis

Ablation Study. To validate the effectiveness of using the image-response relevance score (R_I) to complement the self-generated instruction following score (R_T), we specifically compare CSR with three variants: (1) without applying CSR on LLaVA 1.5 (Base); (2) using CSR with only the self-generated instruction following score (Only R_T); and (3) using CSR with only the image-response relevance score (Only R_I). The results are reported in Table 2. We first observe that CSR improves performance by jointly considering both the self-generated instruction following and image-response relevance scores. This verifies its effectiveness in enhancing modality alignment by calibrating the language-driven self-rewarding paradigm with visual constraints. Additionally, we further conduct the analysis on the change of λ in Eqn (4) and found that incorporating external visual scores to calibrate the models rewarding process effectively enhances performance (see detailed results in Appendix A.5.)

Compatibility Analysis. To validate CSR for its applicability to other LVLMs, we deployed CSR on Vila 7B and conducted three rounds of online iterations. We conducted experiments on all ten evaluation benchmarks and tasks, and the results are shown in Figure 4. Similar to the findings in Figure 3, Vila demonstrates a similar phenomenon during the online iterations of CSR, where it can self-correct preferences, leading to gradual improvements in all benchmarks. For Vila, the overall performance improved by 3.37% after three rounds of CSR iterations, with particularly notable increases of 8.48% on VisWiz and 14.0% on MM-Vet. The compatibility analysis further corroborates the generalizability and effectiveness of CSR in enhancing the performance of LVLMs.

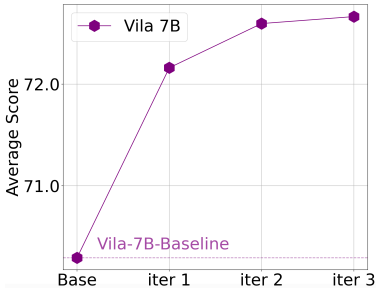

Figure 4: Average scores of CSR in Vila 7B at different iterations over all benchmarks (see Table 8 in Appendix A.5 for full results).

How Does CSR Change the Image-Response Relevance Over Iterations? To investigate how CSR gradually improve the performance over iterations, we analyzed the change of self-generated preference data with the LLaVA-1.5 7B model. In Figure 5, we illustrated the distribution of image-response relevance scores of three iterations over 500 examples from LLaVA-150k [19]. We first observe that both the chosen (preferred) and rejected (dispreferred) responses achieve higher image-response relevance scores after the model undergoes CSR online iterations. This indicates that, following CSR, the responses generated by LVLMs are more closely aligned with the image information. Secondly, it can be observed that after multiple rounds of online iterations with CSR, the average image-response relevance scores for the rejected and chosen responses become closer to each other. This makes the self-generated preference data during CSR iterations more challenging to distinguish, while further strengthening the learning process.

How Does CSR Improve Modality Alignment? To further understand how CSR affects modality alignment, in Figure 6, we present the changes in image and text attention maps for three models: the original LLaVA-1.5 7B model, the self-rewarding approach, and CSR. These attention maps illustrate the distribution of attention scores over image and text tokens. We observe that applying CSR strengthens the model's attention to certain visual tokens. Simultaneously, the change of attention values of the text tokens indicates that CSR is capable of alleviating

Figure 5: Image relevance scores before and after employing CSR.

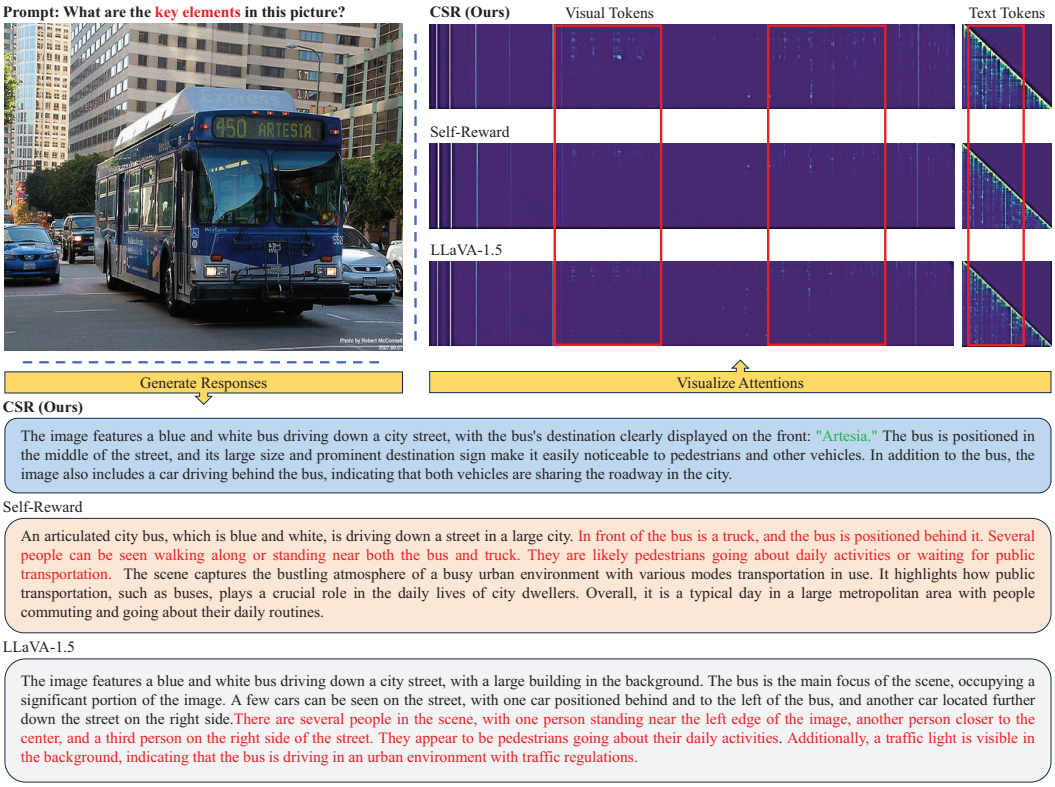

Figure 6: Comparison of attention maps. After optimizing the model with CSR, the attention scores allocated to visual tokens increase, indicating that CSR effectively redirects the model's attention toward the input visual information during the response generation process.

the issue of over-reliance on context mentioned in Huang et al. [37]. Additionally, compared with the self-rewarding approach, CSR shows a more effective distribution of attention between image and text tokens. These findings indicate that with CSR, LVLMs can better align different modalities through a calibrated self-rewarding strategy, focusing more on the visual modality rather than over-relying on contextual text.

4.4 Case Study

In this section, we use LLaVA-1.5 13B as an example to illustrate changes in the models own responses during CSR iterations and the preference data sampled in the CSR learning process, with hallucinations and errors highlighted in red. The results are shown in Figures 7 and 9, respectively. As shown in Figures 7, with each iteration of CSR, hallucinations in the models responses noticeably decrease. This indicates that CSR effectively refines the models preferences through iterative preference learning, making the models responses more accurate. In Figure 9 of the Appendix, we present a sampled preference data pair from the model during the CSR learning process. It can be seen that through CSR, the model not only gradually refines its own preferences but also obtains high-quality preference data pairs without human annotation.

5 Theoretical Explanation

In this section, we present a theoretical framework to explain the empirical phenomenon that incorporating an image-response relevance score can calibrate the self-rewarding procedure, ultimately improving generation accuracy.

As we consider an LVLM, to facilitate the analysis, we decompose the input prompt into $x = (x_v, x_t) \in \mathbb{R}^{d_v} \times \mathbb{R}^{d_t}$, representing the image and text prompts respectively. Although text data

Figure 7: Two cases selected from the CSR-generated datasets.

typically comprises discrete tokens, we follow the CLIP theory literature [38–40] in modeling them as continuous-value random vectors in this section to elucidate the rationale behind our proposed method. More specifically, we assume the following data generative model for x_v and x_t:

$$x_v = U_1 z_1 + \xi_1, \text{ and } x_t = U_2 z_2 + \xi_2,$$

where $U_1 \in \mathbb{O}^{d_v \times r}$ and $U_2 \in \mathbb{O}^{d_t \times r}$ are two orthonormal matrixces, representing decoders that transform the latent (low-dimensional) signals $z_1, z_2 \in \mathbb{R}^r$ to images and text respectively. We assume the covariance matrices of z_1, z_2 are identity matrices. $\xi_1 \in \mathbb{R}^{d_v}$ and $\xi_2 \in \mathbb{R}^{d_t}$ are noise vectors, and we assume they follow sub-gaussian distributions with well-conditioned covariance matrices and sub-gaussian norms upper bounded by a universal constant. We consider the infinite data setting. This is a widely used simplification to avoid the influence of sample randomness [41–43]. According to [38], with an abundance of image-text pairs, the learned visual CLIP embedding $\mathcal{F}_I(x_v)$ and textual CLIP embedding $\mathcal{F}_T(x_t)$ converge to $U_1^\top x_v$ and $U_2^\top x_t$ respectively. To simplify our analysis without loss of generality, we consider a single score for each response y and define the image-response relevance score $R_I(y) = \langle U_1^\top x_v, U_2^\top y \rangle$.

We assume the ground truth $y_{truth} = V_1^* x_v + V_2^* x_t + \epsilon_y$ with weights $V_1^* \in \mathbb{R}^{d_v \times d_v}$ and $V_2^* \in \mathbb{R}^{d_v \times d_t}$. In CSR, we assume the conditional distribution at iteration t, $\pi_{\theta_t}(y \mid x)$ with $\theta_t = (V_1, V_2)$, follows a Gaussian distribution $\pi_{\theta_t}(y \mid x) \propto \exp(-\|y - (V_1 x_v + V_2 x_t)\|^2/\sigma^2)$, where $V_1 \in \mathbb{R}^{d_v \times d_v}$ and $V_2 \in \mathbb{R}^{d_v \times d_t}$ are the weights matrices for the image and text inputs respectively, and $\sigma > 0$ is the standard deviation. As the likelihood is monotonically decreasing with respect to $\|y - (V_1 x_v + V_2 x_t)\|^2$, we consider the self-generated instruction-following score $R_T(y) = -\|y - (V_1 x_v + V_2 x_t)\|^2$. Then the calibrated reward score becomes $R(y) = \lambda \cdot R_I(y) + (1 - \lambda) \cdot R_T(y)$, for some $\lambda \in [0, 1]$. In theoretical analysis, we consider a simpler version of CSR, where we assume $y_w = \arg\max_y R(y)$ (whose distribution is denoted by $p^*_{\theta_t}(y \mid x)$), and y_l is the text output generated by $\pi_{\theta_t}(y \mid x)$. As $R(y)$ depends on λ, we denote the solution θ_{t+1} by $\theta_{t+1}(\lambda)$. In the special case where $\lambda = 1$, this corresponds to the setting where we do not use the image-response relevance score at all.

To evaluate the quality of the text output y, we consider a regression problem where there is an outcome z associated with the ground-truth text output y_{truth}: $z = \beta^{*\top} y_{truth}$ with $\beta^* \in \mathbb{R}^{d_t}$. We evaluate the quality of y by considering the loss function $L(y) = \min_{\beta \in \mathbb{R}^{d_t}} \mathbb{E}[(z - \beta^\top y)^2]$. We then have the following theorem.

Theorem 5.1. *Suppose that $\pi^*_{\theta_t}(y \mid x)$ lies in the LLM space $\{\pi_\theta(y \mid x) : \theta \in \Theta\}$, $\|\beta^{*\top} V_1^{*\top} \beta^*\| \gg \|\beta^{*\top} V_2^{*\top} \beta^*\|$ and $\|\beta^{*\top} V_1^\top \beta^*\| \ll \|\beta^{*\top} V_2^\top \beta^*\|$, then there exists $\lambda < 1$, such that*

$$\mathbb{E}_{\pi_{\theta_{t+1}(\lambda)}(y|x)}[L(y)] < \mathbb{E}_{\pi_{\theta_{t+1}(1)}(y|x)}[L(y)].$$

Our theoretical analysis implies that as long as $\|\beta^{*\top} V_1^\top \beta^*\| \ll \|\beta^{*\top} V_2^\top \beta^*\|$, which happens when the model tends to prioritize textual information over visual input. By incorporating the image-response relevance score (corresponding to $\lambda < 1$), CSR is able to increase the attention on image signals in generating y. As a result, the solution produced by CSR will be better than the method without using the image-response relevance score (corresponding to $\lambda = 1$).

6 Related Work

Large Visual-Language Model Hallucination. Recently, the rapid development of visual-language alignment methods [19, 44–49] and LLMs [50–54] has significantly accelerated the progress of LVLMs, which extend LLMs with visual modalities and demonstrate impressive visual understanding by unifying the encoding of visual and text tokens [34, 55–57]. However, LVLMs still face the problem of hallucination [58, 59], where generated text descriptions contradict the visual modality information. Various approaches have been proposed to address hallucination in LVLMs, including enhancing dataset quality for fine-tuning [60, 8, 61, 9], manipulating the decoding process [37, 62–66], and leveraging external closed-source models to facilitate post-hoc mitigation of hallucination [58, 67–70]. Though these approaches alleviate hallucination to some extent, they do not focus directly on improving modality alignment.

Preference and Modality Alignment. In large models, alignment is necessary to ensure their behavior aligns with human preferences [71, 15, 72]. In LVLMs, alignment manifests as modality misalignment, where the generated textual responses are supposed to follow the input visual information. Recently, preference optimization has been used to address the modality misalignment problem. These optimizations involve preference data curated by human annotators [8, 60, 30] and additional models (e.g., GPT-4) [9, 10]. While these methods improve the ability of LVLMs to align modalities, their reliance on human annotation or additional models is resource-intensive and may introduce additional biases. Furthermore, these models cannot fully capture the inherent preferences of LVLMs, making the curated preference data less effective. Instead, CSR leverages a calibrated self-rewarding strategy, aiming to stimulate the LVLMs' self-correction and enhancement capabilities, thereby further improving modality alignment.

Self-Improvement in Large Language Models. Self-improvement emerges as a powerful paradigm for LLMs to enhance themselves without significant external intervention. For example, self-rewarding and online alignment propose a method that selects consistent answers generated by the model to fine-tune itself [73, 74] , thereby improving its reasoning ability. Similarly, Chen et al. [12] utilizes self-play to enhance the model's performance by distinguishing its self-generated responses from those in human-annotated training data. Unlike prior methods that primarily target LLMs, CSR addresses the modality misalignment issue in LVLMs during the preference modeling process by introducing visual constraints, making it particularly well-suited for LVLMs.

7 Conclusion

In this paper, we investigate the challenge of enhancing modality alignment in LVLMs by introducing a calibrated self-rewarding approach, which integrates visual constraints into the preference modeling process of the self-rewarding paradigm. Empirically, CSR enhances the alignment between image and text modalities, significantly reducing hallucination and improving performance on various LVLM evaluation benchmarks. These empirical results are further supported by rigorous theoretical findings. Additionally, CSR is capable of continuously enhancing LVLM capabilities over iterations, leading to better utilization of visual information.

Acknowledgement

We thank the Center for AI Safety for supporting our computing needs. This research was supported by Cisco Faculty Research Award.

References

[1] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024.

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

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

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[5] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

[6] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023.

[7] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023.

[8] 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.

[9] 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.

[10] 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.

[11] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models, 2024.

[12] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

[13] Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

[14] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[15] 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 *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://arxiv.org/abs/2305.18290`.

[16] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *CoRR*, abs/2104.08718, 2021. URL `https://arxiv.org/abs/2104.08718`.

[18] 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, 2021.

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

[20] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

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

[22] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024.

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

[24] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[25] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people, 2018.

[26] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[27] 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.

[28] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning, 2019.

[29] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

[30] 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.

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

[32] 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, 2023.

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

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[35] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023.

[36] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023.

[37] 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.

[38] Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In *International Conference on Artificial Intelligence and Statistics*, pages 4348–4380. PMLR, 2023.

[39] Zixiang Chen, Yihe Deng, Yuanzhi Li, and Quanquan Gu. Understanding transferable representation learning and zero-shot transfer in clip. *arXiv preprint arXiv:2310.00927*, 2023.

[40] Yuhang Liu, Zhen Zhang, Dong Gong, Biwei Huang, Mingming Gong, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Revealing multimodal contrastive representation learning through latent partial causal models. *arXiv preprint arXiv:2402.06223*, 2024.

[41] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

[42] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. 2021.

[43] Haotian Ye, James Zou, and Linjun Zhang. Freeze then train: Towards provable representation learning under spurious correlations and feature noise. In *International Conference on Artificial Intelligence and Statistics*, pages 8968–8990. PMLR, 2023.

[44] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. 2022.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[46] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024.

[47] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[48] 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. *arXiv preprint arXiv:2311.12793*, 2023.

[49] Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, et al. Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models. *arXiv preprint arXiv:2410.10139*, 2024.

[50] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

[51] 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.

[52] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

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

[54] 01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2024.

[55] 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.

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

[57] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL `https://www.adept.ai/blog/fuyu-8b`.

[58] 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. *arXiv preprint arXiv:2310.00754*, 2023.

[59] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024.

[60] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models, 2024.

[61] 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.

[62] 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.

[63] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data, 2024.

[64] Zongbo Han, Zechen Bai, Haiyang Mei, Qianli Xu, Changqing Zhang, and Mike Zheng Shou. Skip: A simple method to reduce hallucination in large vision-language models. *arXiv preprint arXiv:2402.01345*, 2024.

[65] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024.

[66] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Li Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *ArXiv*, abs/2311.16922, 2023. URL `https://api.semanticscholar.org/CorpusID:265466833`.

[67] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

[68] Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. Rule: Reliable multimodal rag for factuality in medical vision language models. *arXiv preprint arXiv:2407.05131*, 2024.

[69] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*, 2024.

[70] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024.

[71] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020.

[72] Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of human preferences in dialog, 2020. URL `https://openreview.net/forum?id=rJl5rRVFvH`.

[73] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve, 2022.

[74] Zhaoyang Wang, Weilei He, Zhiyuan Liang, Xuchao Zhang, Chetan Bansal, Ying Wei, Weitong Zhang, and Huaxiu Yao. Cream: Consistency regularized self-rewarding language models. *arXiv preprint arXiv:2410.12735*, 2024.

[75] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

A Additional Results

A.1 Experimental Setup

A.1.1 Hyperparameter Settings

Sentence-Level Beam Search. We configure our parameters as follows to ensure both diversity and quality in the sampled data. The num_beamsparameter, set to 5, determines the capacity of input at each search layer. Additionally, num_token_beams, also set to 5, ensures that each beam search returns 5 token-level search results. The eos_token_id is set to the token for a period, effectively controlling the sentence-by-sentence generation process. The max_length parameter, set to 1024, prevents truncation errors and infinite repetitions by controlling the maximum length, while max_new_tokens, set to 74, limits the maximum length of newly generated content to avoid exceeding the CLIP encoding limit.

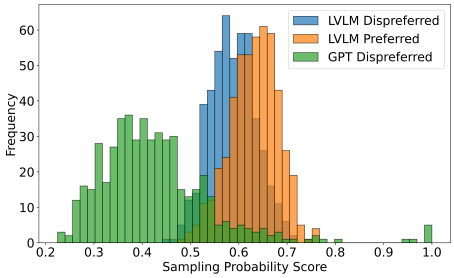

Figure 8: Distribution of preferred responses and dispreferred responses based on the sampling probability scores generated by LVLMs' language models.

To further enhance data diversity, we utilize group beam search by setting the num_beam_group parameter to 5. This approach, when matched with token-level search, significantly boosts the diversity of each data point. The diversity_penalty parameter, set to a value of 3.0, effectively controls the diversity and quality of the sampled data among different beam groups.

Calibrated Rewarding. We set the clip score weight to 0.9 and the language score weight to 0.1 when calculating the scores, giving greater emphasis to visual calibration.

A.2 Evaluation Metrics and Benchmarks

- MME [20] is a comprehensive benchmark for assessing the capabilities of LVLMs in multimodal tasks. It systematically evaluates models across two primary dimensions: perception and cognition, through 14 meticulously designed subtasks that challenge the models' interpretative and analytical skills.
- SEED-Bench [21] is designed to evaluate the generative comprehension capabilities of LVLMs. It features an extensive dataset of 19K multiple-choice questions with precise human annotations, covering 12 distinct evaluation dimensions that assess both spatial and temporal understanding across image and video modalities.
- LLaVA$^{\text{W}}$ [19] is a comprehensive benchmark for evaluating visual reasoning models. It comprises 24 diverse images with a total of 60 questions, covering a range of scenarios from indoor and outdoor settings to abstract art.
- MMBench [22] introduces a dual-pronged approach: a meticulously curated dataset that significantly expands the scope and diversity of evaluation questions, and a pioneering CircularEval strategy that leverages ChatGPT to transform free-form predictions into structured choices.
- MM-Vet [23] is an evaluation benchmark tailored for assessing the multifaceted competencies of LVLMs. It systematically structures complex multimodal tasks into 16 distinct integrations derived from a combination of 6 core vision-language capabilities, providing a granular analysis of model performance across diverse question types and answer styles.
- ScienceQA [24] is a multimodal benchmark designed to evaluate and diagnose the multi-hop reasoning ability and interpretability of AI systems within the domain of science. It offers an expansive dataset of approximately 21k multiple-choice questions across a broad spectrum of scientific topics, complemented by detailed answer annotations, associated lectures, and explanations.
- VizWiz [25] is a dataset in the field of visual question answering (VQA), derived from a naturalistic setting with over 31,000 visual questions. It is distinguished by its goal-oriented approach, featuring images captured by blind individuals and accompanied by their spoken queries, along with crowdsourced answers.

- GQA [26] is a dataset engineered for advanced real-world visual reasoning, utilizing scene graph-based structures to generate 22 million diverse, semantically-programmed questions. It incorporates a novel evaluation metrics suite focused on consistency, grounding, and plausibility, establishing a rigorous standard for assessing in vision-language tasks.
- POPE [27] is an assessment methodology designed to scrutinize object hallucination in LVLMs. It reformulates the evaluation into a binary classification task, prompting LVLMs with straightforward Yes-or-No queries to identify hallucinated objects. POPE offers a stable and adaptable approach, utilizing various object sampling strategies to reveal model tendencies towards hallucination.
- CHAIR [28] is a widely-recognized tool for evaluating the incidence of object hallucination in image captioning tasks, which has two variants: CHAIR_I and CHAIR_S, which assess object hallucination at the instance and sentence levels, respectively. Formulated as:

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|} \quad \text{CHAIR}_S = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}$$

  Specifically, we randomly sampled 500 images from the COCO [75] validation set and evaluated object hallucination using the CHAIR metric.

A.3 Overview of the Baselines

- LLaVA-1.5 [1] is an improvement based on the original LLaVA [19] model demonstrating exceptional performance and data efficiency through visual instruction tuning. It enhanced with a CLIP-ViT-L-336px visual backbone and MLP projection. By incorporating academic-task-oriented VQA data and simple response formatting prompts, LLaVA-1.5 achieves the state-of-the-art results at that time with a remarkably modest dataset of just 1.2 million public images.
- InstructBLIP [31] leverages instruction tuning on pretrained BLIP-2 models, integrating an instruction-aware Query Transformer to enhance feature extraction for diverse vision-language tasks. It achieved state-of-the-art zero-shot performance at the time across 13 datasets and excelled in fine-tuned downstream tasks, such as ScienceQA, showcasing its advantage over contemporaneous multimodal models.
- Qwen-VL-Chat [32] is built upon the Qwen-LM [4] with a specialized visual receptor and input-output interface. It is trained through a 3-stage process and enhanced with a multilingual multimodal corpus, enabling advanced grounding and text-reading capabilities.
- mPLUG-Owl2 [33] employs a modular network design with a language decoder interface for unified modality management. It integrates shared modules for cross-modal collaboration and modality-adaptive components for feature retention, enhancing generalization in both text-only and multimodal tasks.
- BLIP-2 [34] is a vision-language pre-training framework that efficiently leverages off-the-shelf frozen image encoders and LLMs. Employing a two-stage pre-training strategy with a lightweight Querying Transformer, BLIP-2 bridges the modality gap between vision and language, enabling zero-shot image-to-text generation that adheres to natural language instructions while maintaining high compute-efficiency.
- IDEFICS [35] is an open-access visual language model that expands upon the Flamingo [44] architecture, offering both base and instructed variants with 9 billion and 80 billion parameter sizes. It is developed using solely publicly available data and models.
- POVID [10] is a novel training paradigm aligns the preferences of VLLMs through external preference data from GPT4 and the inherent hallucination patterns within the model triggered by noisy images.
- RLHF-V [30] collected fine-grained paragraph-level corrections from humans on hallucinations and performing dense direct preference optimization on the human feedback.
- Silkie [9] constructed a VLFeedback dataset using VLLMs annotation. Specifically, the responses were generated by 12 LVLMs models conditioned on multimodal instructions extracted from different datasets. The entire dataset was evaluated using GPT-4V to assess the generated outputs in terms of helpfulness, visual faithfulness, and ethical considerations. In this paper, the VLFeedback dataset was utilized to perform one round of DPO on LLaVA-1.5.

- LLaVA-RLHF [8] proposes a novel alignment algorithm called Factually Augmented RLHF, which enhances the reward model by incorporating additional factual information such as image captions and ground-truth multi-choice options. In this paper, the annotated preference data is used to conduct one round of preference learning on LLaVA1.5.
- Self-rewarding [29] introduces a method for self-feedback learning in LLMs and serves as a baseline for our approach, referred to as CSR. Specifically, for each input image and prompt, two outputs are sampled from LLaVA-1.5. The model is provided with the prompt mentioned in Table 3 and is tasked with determining which output is better. Finally, LLaVA-1.5 is fine-tuned using the collected preference data, with the entire setup and the images and prompts used for inference matching those of CSR.

A.4 Do Different Sources of Preference Data Have Different Impacts?

The sources of preference data generally fall into two main categories: external preference data and self-generated data. External preference data typically represent preferences obtained from human annotations or GPT-4. Although external preference data generally have higher quality compared to self-generated data, are they really more effective? We conducted an analysis using 500 samples obtained from the original LLaVA-1.5 7B model. Following the same pipeline as CSR, we selected samples with the highest and lowest rewards as preferred (chosen) and dispreferred (rejected) responses. We further employed the GPT-4 API to transform preferred responses into dispreferred ones, with specific prompts referenced in Table 4.

In Figure 8, we present the distribution based on both the sampling probabilities score generated by the target LVLM, which describes the probability of the LVLM generating this response. Clearly, compared to the model's own generated dispreferred responses, the dispreferred responses modified by GPT-4V are not as easily confusable for the model. This result partially supports the idea that dispreferred responses generated by external models are more easily distinguishable by the target LVLM, making them less effective.

Table 3: Prompt for self-reward: utilizing the model itself as a judge to determine whether the corresponding response is a chosen response or a reject response.

Now you act as a judge, helping me determine which of the two texts I provide is closer to the given image and has fewer errors. ********** **Response 1**: {response 1} **Response 2**: {response 2} ********** Please strictly follow the following format requirements when outputting, and don't have any other unnecessary words. **Output Format**: response 1 or response 2.

A.5 Additional Experiments

In this subsection, we provide a comparison of CSR with other state-of-the-art models, a performance comparison of different CSR iterations, a comparison of hallucinations in different CSR iterations, validation experiments of CSR on other models, ablation study on λ in Eqn (4), and the relationship between reward score and average performance score. Experiments strongly demonstrate the effectiveness of CSR.

For the ablation study on λ in Eqn (4), our training settings are consistent with Table 1, with three rounds of iteration. The experimental results in Table 9 show that as the value of λ increases, the models performance on various benchmarks improves. This suggests that calibrating the models

Table 4: Prompt for GPT-4 API: transform the provided response into negative ones based on the provided image.

Transform the provided response into negative ones based on the provided image.

****

Response:

{chosen response from another LVLM or ground truth}

Requirements:

(1) Revise the response while maintaining its original format and order as much as possible.

(2) Based on the provided image, primarily add, replace, or modify entities in the input response to make them related to the image but incorrect. Adjust their attributes and logical relationships accordingly.

(3) The modifications in (2) must align with the image information, making the revised result difficult to discern.

****

Please strictly follow the following format requirements when outputting, and don't have any other unnecessary words.

Output Format:

negative response

Table 5: Comparison of LLaVA-1.5 with CSR and other open-sourced state-of-the-art LVLMs.

	Comprehensive Benchmark						General VQA		
Method	MME^P	MME^C	SEED	$LLaVA^W$	MMB	MM-Vet	SQA^I	VisWiz	GQA
BLIP-2	1293.8	290.0	46.4	38.1	-	22.4	61.0	19.6	41.0
InstructBLIP	1212.8	291.8	53.4	60.9	36.0	26.2	60.5	34.5	49.2
IDEFICS	1177.3	-	45.0	45.0	48.2	30.0	-	35.5	38.4
Qwen-VL-Chat	1487.6	360.7	58.2	67.7	60.6	**47.3**	68.2	38.9	57.5
mPLUG-Owl2	1450.2	313.2	57.8	59.9	64.5	36.2	68.7	54.5	56.1
CSR iter-3 7B	1524.2	**367.9**	60.3	71.1	65.4	33.9	70.7	54.1	62.3
CSR iter-3 13B	**1530.6**	303.9	**62.9**	**74.7**	**68.8**	37.8	**75.1**	**56.8**	**63.7**

Table 6: The performance of CSR online iteration with LLaVA-1.5 as the backbone on comprehensive benchmarks and general VQA.

	Comprehensive Benchmark						General VQA		
Method	MME^P	MME^C	SEED	$LLaVA^W$	MMB	MM-Vet	SQA^I	VisWiz	GQA
LLaVA-1.5-7B	1510.7	348.2	58.6	63.4	64.3	30.5	66.8	50.0	62.0
+ **CSR iter-1**	1500.6	367.5	60.4	69.7	64.7	32.2	70.3	54.0	62.1
+ **CSR iter-2**	1519.0	**368.9**	60.3	70.4	65.2	33.7	70.1	54.0	62.3
+ **CSR iter-3**	1524.2	367.9	60.3	71.1	65.4	**33.9**	70.7	54.1	62.3
+ **CSR iter-4**	**1524.6**	368.8	60.4	71.0	65.3	33.9	70.4	54.0	62.2
+ **CSR iter-5**	1520.1	367.2	**60.5**	**71.3**	**65.4**	33.8	**70.8**	**54.2**	**62.4**
LLaVA-1.5-13B	1531.3	295.4	61.6	70.7	67.7	35.4	71.6	53.6	63.3
+ **CSR iter-1**	**1533.1**	303.6	63.0	74.4	68.4	37.4	74.8	56.8	63.2
+ **CSR iter-2**	1530.4	301.1	**63.0**	74.3	68.5	37.2	75.0	56.0	63.2
+ **CSR iter-3**	1530.6	**303.9**	62.9	**74.7**	**68.8**	**37.8**	75.1	**56.8**	63.7
+ **CSR iter-4**	1530.4	301.4	63.0	74.2	68.3	37.3	**75.2**	56.6	63.4
+ **CSR iter-5**	1531.1	302.2	62.8	74.0	68.2	37.4	74.8	56.7	**63.7**

rewarding process using the visual score can enhance the preference learning process, thereby boosting performance.

Table 7: The performance of CSR online iteration with LLaVA-1.5 as the backbone on hallucination benchmarks.

	Hallucination Benchmark				
Method	POPE_{acc}	POPE_{f1}	CHAIR_{S}	CHAIR_{I}	Avg Length
LLaVA-1.5-7B	85.90	84.29	48.8	14.9	89.03
+ **CSR iter-1**	86.94	85.80	26.6	7.2	80.59
+ **CSR iter-2**	86.82	85.62	23.0	6.1	82.62
+ **CSR iter-3**	87.01	85.93	21.0	6.0	83.29
+ **CSR iter-4**	87.05	85.95	19.0	5.9	81.34
+ **CSR iter-5**	**87.16**	**85.98**	**18.3**	**5.4**	82.07
LLaVA-1.5-13B	85.90	84.87	48.3	14.1	89.73
+ **CSR iter-1**	87.28	86.29	36.0	9.0	98.85
+ **CSR iter-2**	**87.33**	86.36	36.0	7.8	105.0
+ **CSR iter-3**	87.30	86.31	28.0	7.3	107.8
+ **CSR iter-4**	87.20	**86.58**	**27.4**	7.4	112.3
+ **CSR iter-5**	87.18	86.51	28.0	**7.3**	102.4

Table 8: The performance of CSR online iteration with Vila 7B as the backbone.

	Comprehensive Benchmark						General VQA			Hallucination Benchmark		
Method	MME^{P}	MME^{C}	SEED	LLaVA^{W}	MMB	MM-Vet	SQA^{I}	VisWiz	GQA	POPE	CHAIR_{S}	CHAIR_{I}
Vila	1533.0	316.4	61.1	69.7	68.9	34.9	68.2	57.8	62.3	85.50	31.0	8.8
+ **CSR iter-1**	1520.6	321.9	63.2	73.5	**69.3**	38.3	71.9	62.3	62.2	86.82	29.2	**7.9**
+ **CSR iter-2**	1536.0	**322.6**	**63.4**	74.2	69.1	39.7	**72.3**	62.6	62.1	87.30	28.2	8.0
+ **CSR iter-3**	**1542.2**	321.5	**63.4**	**74.3**	**69.3**	**39.8**	72.2	**62.7**	**62.4**	**87.31**	**28.0**	8.2

Table 9: Performance comparison of CSR on LLaVA-1.5 7B with different λ values on various benchmarks.

Method	MME^{P}	MME^{C}	SEED	LLaVA^{W}	MMB	MM-Vet	SQA^{I}	VisWiz	GQA	POPE	CHAIR_{S}	CHAIR_{I}
($\lambda = 0.1$)	1508.6	**369.3**	60.0	66.7	64.9	31.6	70.0	54.0	62.0	86.90	40.8	10.2
($\lambda = 0.5$)	1515.4	364.5	60.1	68.2	64.9	32.4	69.7	54.0	62.1	86.90	28.2	6.7
($\lambda = 0.9$)	**1524.2**	367.9	**60.3**	**71.1**	**65.4**	**33.9**	**70.7**	**54.1**	**62.3**	**87.01**	**21.0**	**6.0**

Table 10: Reward score and average performance score across multiple iterations of CSR on LLaVA-1.5 7B.

Iteration	Iter-1	Iter-2	Iter-3	Iter-4	Iter-5
Chosen reward	0.4885	0.5040	0.5052	0.5055	0.5066
Rejected reward	0.4551	0.4788	0.4789	0.4794	0.4799
Avg performance score	66.61	71.02	71.74	72.09	72.24

B Proofs

Proof of Theorem 5.1. Let us first denote the distribution of y_w by $\pi^*_{\theta_t}(y \mid x)$. As we take $y_w = \arg\max_y R(y)$, this distribution is a point mass. As a result, the global minimizer to (5) will then converge to $\pi^*_{\theta_t}(y \mid x)$.

In the following, we analyze how $\pi^*_{\theta_t}(y \mid x)$ is shaped.

By the CSR procedure, we have

$$y_w = \arg\max_y (1-\lambda))\langle U_1^\top x_v, U_1^\top y\rangle - \lambda\|y - V_1 x_v + V_2 x_t\|^2 = \frac{1-\lambda}{\lambda} U_1 U_1^\top x_v + V_1 x_v + V_2 x_t.$$

We can see that CSR up-weights the signal of the image input.

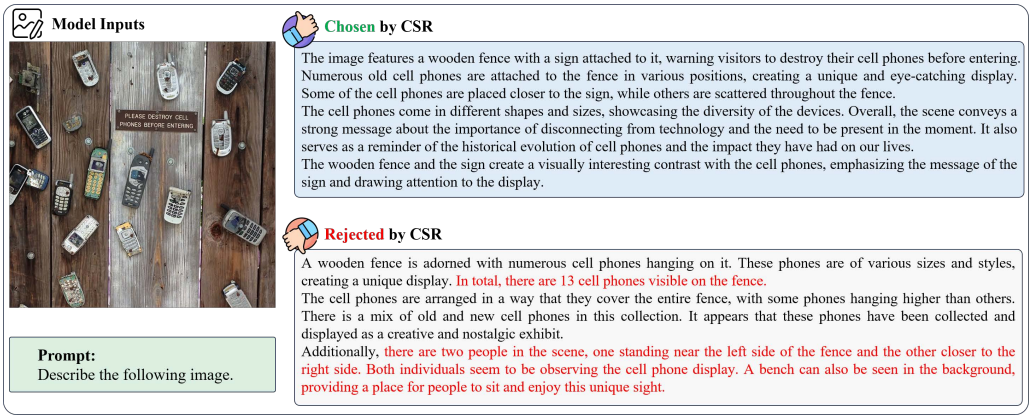

Figure 9: A case including both self-generated preferred and dispreferred responses.

Then

$$L(y) = \min_{\beta \in \mathbb{R}^{d_t}} \mathbb{E}[(z - \beta^\top y)^2] = \min_{\beta \in \mathbb{R}^{d_t}} \mathbb{E}[(\beta^{*\top} y_{truth} - \beta^\top y)^2]$$
$$= \min_{\beta \in \mathbb{R}^{d_t}} \mathbb{E}[(\beta^{*\top}(V_1^* x_v + V_2^* x_t)) - \beta^\top y)^2] + Var(\epsilon_y)\|\beta^*\|^2$$

We have

$$\mathbb{E}[(\beta^{*\top}(V_1^* x_v + V_2^* x_t)) - \beta^\top y)^2] = \mathbb{E}[(\beta^{*\top}(V_1^* x_v + V_2^* x_t)) - \beta^\top \left((\frac{1-\lambda}{\lambda} U_1 U_1^\top + V_1) x_v + V_2 x_t \right))^2]$$

As we assume $\frac{\|V_1\|}{\|\beta^{*\top} V_1^*\|} \ll \frac{\|V_2\|}{\|\beta^{*\top} V_2^*\|}$ and due to the smoothness over parameters. Without loss of generality, we prove the claim for the case where $\|V_1\| = 0$, that is V_1=0.

In this case, we want to show that there exists $\lambda \in (0, 1)$, such that

$$\min_{\beta \in \mathbb{R}^{d_t}} \mathbb{E}[(\beta^{*\top}(V_1^* x_v + V_2^* x_t)) - \beta^\top \left((\frac{1-\lambda}{\lambda} U_1 U_1^\top) x_v + V_2 x_t \right))^2]$$
$$< \min_{\beta \in \mathbb{R}^{d_t}} \mathbb{E}[(\beta^{*\top}(V_1^* x_v + V_2^* x_t)) - \beta^\top (V_2 x_t))^2]$$

Due to the independence between x_v and x_t, the right-hand sides is lower bounded by $\beta^{*\top} V_1^* Cov(x_t) V_1^{*\top} \beta^*$.

The left-hand side, on the other hand, can be upper bounded by the value when we take β_0 such that $\frac{1-\lambda}{\lambda} U_1 U_1^\top \beta_0 = U_1 U_1^\top V_1^{*\top} \beta^*$, which equals to $\beta^{*\top} V_1^* (I - U_1 U_1^\top) Cov(x_t) (I - U_1 U_1^\top) V_1^{*\top} \beta^*$.

As we assume $\|\beta^{*\top} V_1^{*\top} \beta^*\| \gg \|\beta^{*\top} V_2^{*\top} \beta^*\|$, this is a dominating term when the left-hand side is evaluated at β_0.

In addition, we assume $Cov(\xi_1)$ is well-conditioned, implying $Cov(x_t)$ is well-conditioned, and therefore

$$\beta^{*\top} V_1^* (I - U_1 U_1^\top) Cov(x_t) (I - U_1 U_1^\top) V_1^{*\top} \beta^* < \beta^{*\top} V_1^* Cov(x_t) V_1^{*\top} \beta^*.$$

We complete the proof.

□

C Limitations

Due to limitations in computing resources, we conducted only three iterations of CSR. Additionally, our experiments were confined to 7B and 13B models. This restriction prevents us from determiing

if our method adheres to a scaling law. We hope to continue iterative training in the future and to train larger models, given access to more computing resources, to explore the upper limits of our method.

D Broader Impacts

Our approach requires no additional human annotations and significantly enhances model performance using the model itself. Technically, our method may inspire more researchers to explore how multimodal models can learn from themselves. From a societal impact perspective, our method significantly reduces hallucinations in LVLMs, a major factor affecting the application of AI in real-world scenarios. Our approach promotes more responsible use of LVLMs. However, it is important to note that while our method significantly reduces hallucinations, they still occur. Therefore, it is crucial to employ various measures to ensure safety and stability when applying this approach in real-world scenarios.

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: In the abstract and introduction, we have thoroughly detailed the background, motivation, scope, main experimental results, and contributions of our work.

   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: In Section C, we discussed that due to a lack of computing resources, our method was not trained for more iterations and was not tested on larger models.

   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: [Yes]

Justification: In Section 5 and Appendix B, we provide a complete theoretical analysis and proof.

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: In our paper, we provide comprehensive details about the backbone, dataset, algorithm, training hyperparameters, and hardware platform used to ensure experiment reproducibility. Additionally, we will open-source all training code and logs.

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: The data we used is publicly accessible, and we will release the code upon acceptance of the paper.

   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: We provide detailed experimental setups and specifics in Section 4.1 and Appendix A.1 of the paper.

   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: [Yes]

   Justification: We adopt standard evaluation benchmarks and metrics, which are accompanied by statistical significance.

   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.
- 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: The experiments are sufficiently discussed to be run by others.

   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: The conducted research conforms in every respect with 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: [Yes]

    Justification: We have explained the impact of our work from both technical and societal perspectives in Section D.

    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 does not involve any high risks or potential misuse.

    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: All the mentioned previous work are properly cited.

    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?

    Answer: [NA]

    Justification: The code will be made public upon acceptance.

    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: This 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: This 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.