Calibrated Self-Rewarding Vision Language Models

Anonymous Authors[1]

1. Introduction

Large Vision-Language Models (LVLMs) (Liu et al., 2024a; Dai et al., 2023b; Ye et al., 2023b; Bai et al., 2023a) 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 (Rohrbach et al., 2018), 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. (2023); Guan et al. (2023), 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.

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) (Sun et al., 2023) and direct preference optimization (DPO) (Li et al., 2023d; Zhou et al., 2024). 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 C.1). Recently, self-rewarding approaches have emerged, utilizing a single LLM for both response generation and preference modeling, showing promising results in LLM alignment (Yuan et al., 2024a; Chen et al., 2024a). 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 (Graves, 2012; Sutskever et al., 2014) 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 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, 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.

2. Calibrated Self-Rewarding

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 1, CSR trains the target LVLM by alternately

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the ICML 2024 Workshop on Foundation Models in the Wild. Do not distribute.

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.

2.1. 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 into the reward definition of LVLMs. By doing so, we address the limitation of LVLM in overlooking image input data.
- Step-Wise Reward: Instead of assigning a single reward for the entire response, we opt for a step-wise approach, involving 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 PRM (Lightman et al., 2023), 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:

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

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 (Hessel et al., 2021) 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:

$$R_I(s) = \max(100 * \cos(\mathcal{F}_I(x_v), \mathcal{F}_T(s)), 0), \tag{2}$$

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:

$$R(s) = \lambda \cdot R_I(s) + (1 - \lambda) \cdot R_T(s), \tag{3}$$

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.

2.2. Iterative Fine-Tuning

After establishing the reward framework, 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:

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. (3). 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 $\langle \text{eos} \rangle$. 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 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.

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

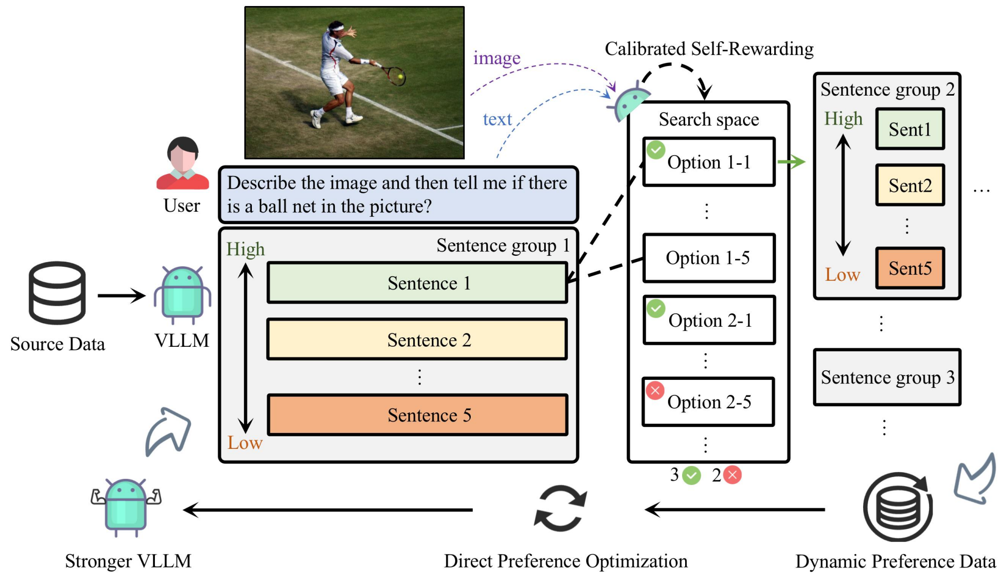

Figure 1: The CSR framework operates an iterative process of preference data generation and learning.

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 (5), the loss at iteration t of CSR is defined as:

$$\mathcal{L}_t = -\mathbb{E}_{(x,y_{w,t},y_{l,t})\sim\mathcal{D}} \left[\log\sigma\left(\alpha\log\frac{\pi_\theta(y_{w,t}|x)}{\pi_{\theta_{t-1}}(y_{w,t}|x)} -\alpha\log\frac{\pi_\theta(y_{l,t}|x)}{\pi_{\theta_{t-1}}(y_{l,t}|x)}\right)\right]. \tag{4}$$

The training process of CSR is detailed in Algorithm 1.

3. Experiment

In this section, we empirically investigate CSR in addressing the modality misalignment problem (see additional experiments in Appendix C.2). We provide the theoretical analysis to understand the empirical phenomena in Appendix D.

Implementation Details. We utilize LLaVA-1.5 7B and 13B (Liu et al., 2024a) as the backbone models. During the preference learning process, we adapt LoRA fine-tuning (Hu et al., 2021). 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 (Liu et al., 2023b). It is worth noting that each iteration uses the same prompt and image as the previous round. For more detailed information on training hyperparameters and training data, please refer to Appendix A.1.

Evaluation Benchmarks and Baselines. We conducted evaluations on three types of benchmarks: comprehensive benchmarks, general VQA and hallucination benchmarks. In terms of baselines, we will first compare CSR with the self-rewarding approach described by Yuan et al. (2024b). Here, we directly apply self-rewarding to LVLM, using the prompts and experimental settings outlined in Yuan et al. (2024b). We also compared CSR with several data-driven preference learning methods (e.g., POVID (Zhou et al., 2024)). More detailed descriptions are discussed in Appendices A.2 and A.3.

3.1. Results

CSR Continuously Improves Model Performance over Iterations. In Figure 2 ((see Table 6 and Table 7 in Appendix C.2 for full results)), we report the average performance of LLaVA-1.5 7B and 13B models concerning the number of training iterations on all baselines. 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%. 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.

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 anno-

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. (2024).

	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**

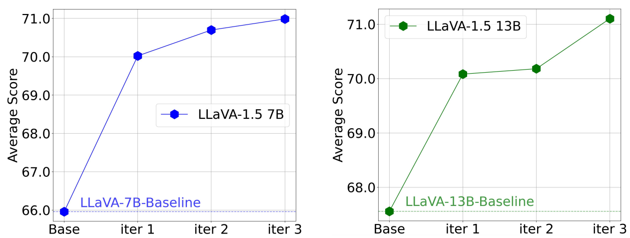

Figure 2: Average scores of CSR at different iterations.

tations, 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 2.43% improvements, 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 C.2. 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.

Table 2: Ablation study.

Method	7B	13B
Base	65.96	67.56
Only R_T	67.66	68.70
Only R_I	66.77	68.23
CSR (Ours)	**70.99**	**71.10**

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.

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 3, we illustrated the distribution of image-response relevance scores of three iterations (Liu et al., 2023b). 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.

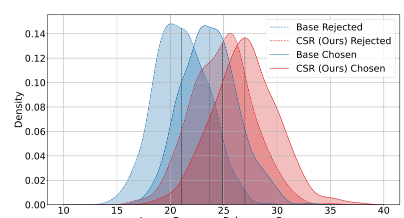

Figure 3: The change in image relevance scores before and after employing CSR.

4. 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 improving performance on various LVLM evaluation benchmarks.

References

AI, ., :, Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., Yu, K., Liu, P., Liu, Q., Yue, S., Yang, S., Yang, S., Yu, T., Xie, W., Huang, W., Hu, X., Ren, X., Niu, X., Nie, P., Xu, Y., Liu, Y., Wang, Y., Cai, Y., Gu, Z., Liu, Z., and Dai, Z. Yi: Open foundation models by 01.ai, 2024.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. 2022.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023a.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023b.

Bavishi, R., Elsen, E., Hawthorne, C., Nye, M., Odena, A., Somani, A., and Taşırlar, S. Introducing our multimodal models, 2023. URL https://www.adept.ai/blog/fuyu-8b.

Chen, Z., Deng, Y., Li, Y., and Gu, Q. Understanding transferable representation learning and zero-shot transfer in clip. *arXiv preprint arXiv:2310.00927*, 2023.

Chen, Z., Deng, Y., Yuan, H., Ji, K., and Gu, Q. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024a.

Chen, Z., Zhao, Z., Luo, H., Yao, H., Li, B., and Zhou, J. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024b.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J., and Yao, H. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023a.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023b.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P. N., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., and Ji, R. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Linearized two-layers neural networks in high dimension. 2021.

Graves, A. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023.

Gunjal, A., Yin, J., and Bas, E. Detecting and preventing hallucinations in large vision language models, 2024.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. Vizwiz grand challenge: Answering visual questions from blind people, 2018.

Han, Z., Bai, Z., Mei, H., Xu, Q., Zhang, C., and Shou, M. Z. Skip: A simple method to reduce hallucination in large vision-language models. *arXiv preprint arXiv:2402.01345*, 2024.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. *CoRR*, abs/2104.08718, 2021. URL https://arxiv.org/abs/2104.08718.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021.

Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., and Han, J. Large language models can self-improve, 2022.

Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., and Yu, N. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*, 2023.

Hudson, D. A. and Manning, C. D. 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*, pp. 6700–6709, 2019.

Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., and Picard, R. Way off-policy batch deep reinforcement learning of human preferences in dialog, 2020. URL https://openreview.net/forum?id=rJl5rRVFvH.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023.

Kim, M. P., Ghorbani, A., and Zou, J. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A. M., Kiela, D., Cord, M., and Sanh, V. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., and Bing, L. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *ArXiv*, abs/2311.16922, 2023a. URL https://api.semanticscholar.org/CorpusID:265466833.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., and Bing, L. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023b.

Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., and Shan, Y. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023a.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023c.

Li, L., Xie, Z., Li, M., Chen, S., Wang, P., Chen, L., Yang, Y., Wang, B., and Kong, L. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023d.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023e.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. Microsoft coco: Common objects in context, 2015.

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning, 2024a.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., and Lin, D. Mmbench: Is your multi-modal model an all-around player?, 2024b.

Liu, Y., Zhang, Z., Gong, D., Huang, B., Gong, M., Hengel, A. v. d., Zhang, K., and Shi, J. Q. Revealing multimodal contrastive representation learning through latent partial causal models. *arXiv preprint arXiv:2402.06223*, 2024c.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Nakada, R., Gulluk, H. I., Deng, Z., Ji, W., Zou, J., and Zhang, L. Understanding multimodal contrastive learning and incorporating unpaired data. In *International*

Conference on Artificial Intelligence and Statistics, pp. 4348–4380. PMLR, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. 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.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. Object hallucination in image captioning, 2019.

Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.-Y., Wang, Y.-X., Yang, Y., et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Team, C. Chameleon: Mixed-modal early-fusion foundation models, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., and Wolf, T. Zephyr: Direct distillation of lm alignment, 2023.

Ye, H., Zou, J., and Zhang, L. Freeze then train: Towards provable representation learning under spurious correlations and feature noise. In *International Conference on Artificial Intelligence and Statistics*, pp. 8968–8990. PMLR, 2023a.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023b.

Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., Huang, F., and Zhou, J. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023c.

Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., and Chen, E. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

Yu, Q., Li, J., Wei, L., Pang, L., Ye, W., Qin, B., Tang, S., Tian, Q., and Zhuang, Y. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data, 2024.

Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Sun, M., et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2023a.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2023b.

Yuan, W., Pang, R. Y., Cho, K., Li, X., Sukhbaatar, S., Xu, J., and Weston, J. Self-rewarding language models, 2024a.

Yuan, W., Pang, R. Y., Cho, K., Sukhbaatar, S., Xu, J., and Weston, J. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024b.

Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., and Yao, H. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

Zhou, Y., Cui, C., Rafailov, R., Finn, C., and Yao, H. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences, 2020.

Algorithm 1 Calibrated Self-Rewarding

Require: Dataset: $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$; Reference model: π_{ref}; Number of iterations: T
1: **for** $t = 1, \ldots, T$ **do**
2:    **for** each x in $\mathcal{D}$ **do**
3:       **while** not reach the end of response **do**
4:          Generate a bunch of candidate sentences from last-round sentences
5:          **for** each candidate sentence s **do**
6:             Compute the self-generated instruction-following score $R_T(s)$ by Eqn. (1)
7:             Calculate the image representation $\mathcal{F}_I(x_v)$ and sentence representation $\mathcal{F}_T(s)$
8:             Compute the image-response relevance score $R_I(s)$ by Eqn. (2)
9:             Compute the calibrated reward score $R(s)$ by Eqn. (3)
10:         **end for**
11:         Select top-k and bottom-k sentences with the highest and lowest reward scores
12:      **end while**
13:      Select the preferred response $y_{w,t}$ and dispreferred response $y_{l,t}$
14:   **end for**
15:   Update $\pi_\theta \leftarrow \arg\min_\theta \mathcal{L}_t(\pi_\theta; \pi_{\text{ref}})$, $\pi_{\text{ref}} \leftarrow \pi_\theta$
16: **end for**

A. Experimental Setups

A.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.

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 (Fu et al., 2024) 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 (Li et al., 2023a) 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$^{\text{W}}$ (Liu et al., 2023b) 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 (Liu et al., 2024b) 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 (Yu et al., 2023b) 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 (Lu et al., 2022) 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 (Gurari et al., 2018) 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 (Hudson & Manning, 2019) 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 (Li et al., 2023e) 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 (Rohrbach et al., 2019) 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 (Lin et al., 2015) validation set and evaluated object hallucination using the CHAIR metric.

A.3. Overview of the Baselines

- LLaVA-1.5 (Liu et al., 2024a) is an improvement based on the original LLaVA (Liu et al., 2023b) 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 (Dai et al., 2023a) 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 (Bai et al., 2023b) is built upon the Qwen-LM (Bai et al., 2023a) 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 (Ye et al., 2023c) 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 (Li et al., 2023c) 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 (Laurençon et al., 2023) is an open-access visual language model that expands upon the Flamingo (Alayrac et al., 2022) 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 (Zhou et al., 2024) 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 (Yu et al., 2023a) collected fine-grained paragraph-level corrections from humans on hallucinations and performing dense direct preference optimization on the human feedback.
- Silkie (Li et al., 2023d) 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 (Sun et al., 2023) 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 (Yuan et al., 2024b) 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.

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.

B. 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 and dispreferred responses for the input prompt $x^{(i)}$. Preference optimization leverage the preference data to optimize language models. Taking DPO (Rafailov et al., 2023) as a representative example,, it formulates

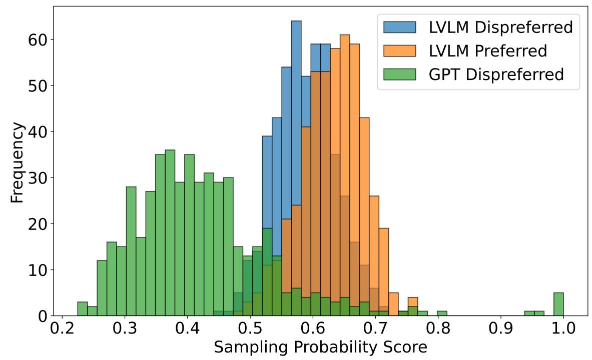

Figure 4: Distribution of preferred responses and dispreferred responses based on the sampling probability scores generated by LVLMs' language models.

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:

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\alpha\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \alpha\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right], \tag{5}$$

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

C. Additional Results

C.1. 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 4, 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.

C.2. Additional Experiments

In this subsection, we provide a additional results and analysis of CSR. All experiments demonstrate the effectiveness of CSR.

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 5. Similar to the findings in Figure 2, 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.

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

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

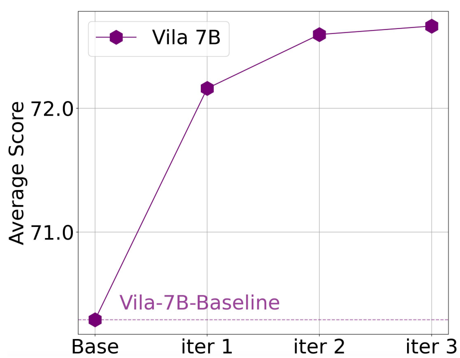

Figure 5: Average scores of CSR in Vila 7B at different iterations over all benchmarks (see Table 8 in Appendix ?? for full results).

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 the issue of over-reliance on context mentioned in (Huang et al., 2023). 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.

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**
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**

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
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

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

D. 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 typically comprises discrete tokens, we follow the CLIP theory literature (Nakada et al., 2023; Chen et al., 2023; Liu et al., 2024c) 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 (Kim et al., 2019; Ghorbani et al., 2021; Ye et al., 2023a). According to (Nakada et al., 2023), 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

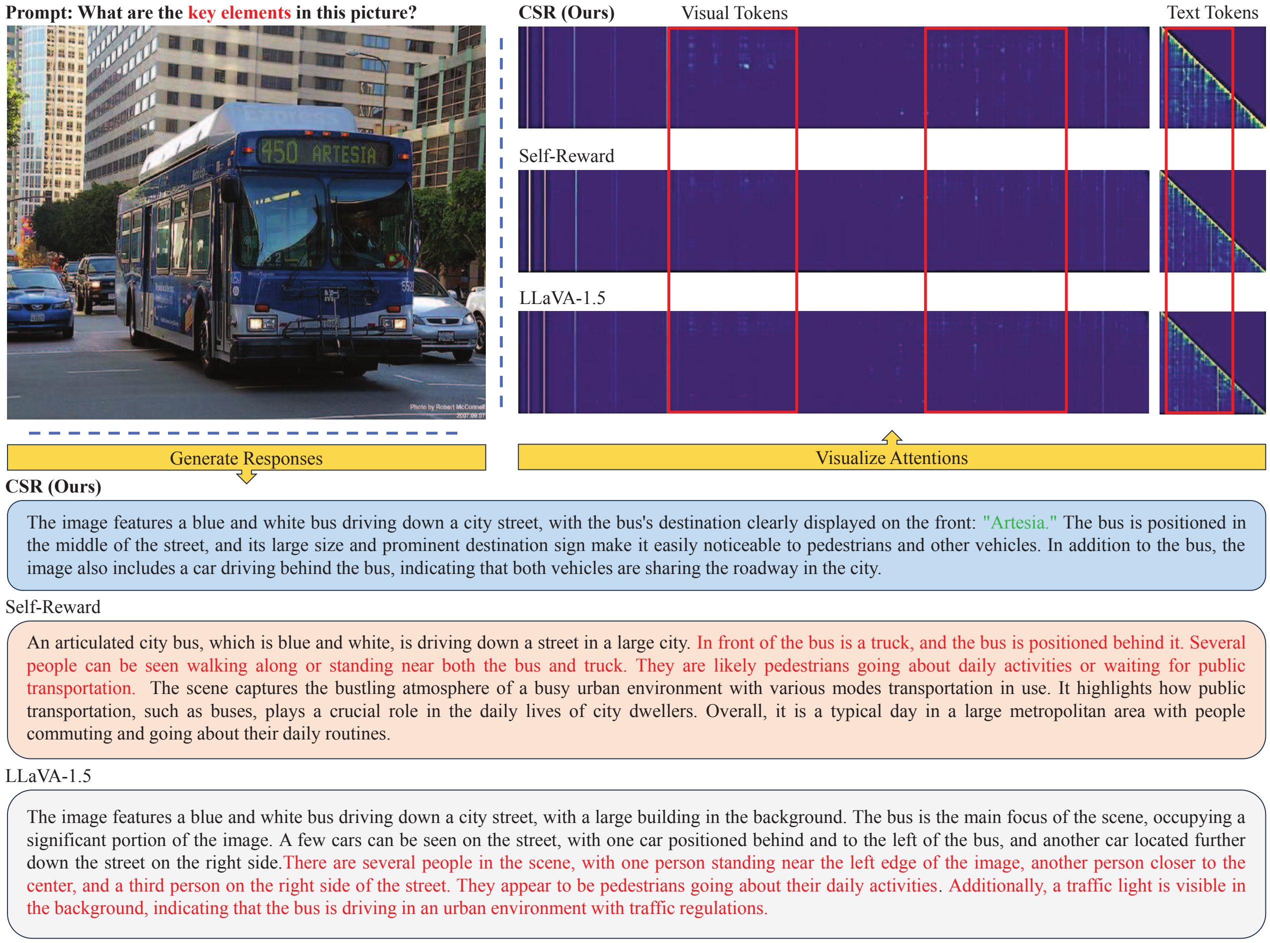

Figure 6: Comparison of attention maps. After optimizing the model with CSR, the attention scores allocated to visual tokens increase, indicating that CSR effectively redirects the model's attention toward the input visual information during the response generation process.

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 D.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 $\lambda = 1$).

D.1. Proofs

Proof of Theorem D.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 (4) 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.

Then

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

We have

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

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

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

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.

□

E. Related Work

Large Visual-Language Model Hallucination. Recently, the rapid development of visual-language alignment methods (Liu et al., 2023b; Alayrac et al., 2022; Radford et al., 2021; Team, 2024) and LLMs (Chiang et al., 2023; Touvron et al., 2023;

Jiang et al., 2023; Tunstall et al., 2023; AI et al., 2024) 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 (Li et al., 2023b; Dai et al., 2024; Zhu et al., 2023; Bavishi et al., 2023). However, LVLMs still face the problem of hallucination (Zhou et al., 2023), 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 (Gunjal et al., 2024; Sun et al., 2023; Liu et al., 2023a; Li et al., 2023d), manipulating the decoding process (Huang et al., 2023; Leng et al., 2023b; Yu et al., 2024; Han et al., 2024; Chen et al., 2024b; Leng et al., 2023a), and leveraging external closed-source models to facilitate post-hoc mitigation of hallucination (Zhou et al., 2023; Yin et al., 2023). 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 (Ziegler et al., 2020; Rafailov et al., 2023; Jaques et al., 2020). 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 (Sun et al., 2023; Gunjal et al., 2024; Yu et al., 2023a) and additional models (e.g., GPT-4) (Li et al., 2023d; Zhou et al., 2024). 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 (Huang et al., 2022) propose a method that selects consistent answers generated by the model to fine-tune itself, thereby improving its reasoning ability. Similarly, (Chen et al., 2024a) 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.