Enhancing Visual-Language Modality Alignment in Large Vision Language Models via Self-Improvement

Anonymous ACL submission

Abstract

Large vision-language models (LVLMs) have achieved impressive results in visual question-answering and reasoning tasks through vision instruction tuning on specific datasets. However, there remains significant room for improvement in aligning visual and language modalities. Existing methods often depend on external models or data, leading to uncontrollable and unstable alignment results. In this paper, we propose SIMA, a self-improvement framework that enhances visual and language modality alignment without external dependencies. SIMA leverages existing vision instruction tuning datasets to self-generate responses, incorporating an in-context self-critic mechanism that constructs preference pairs for tuning. Crucially, our approach allows LVLMs to act as critics by designing effective critic prompts, eliminating the need for additional fine-tuning with external instruction data. We introduce three novel visual metrics within the self-critic process to guide judgement, significantly improving the accuracy of self-critic. Through extensive experiments across 14 hallucination and comprehensive benchmarks, we demonstrate that SIMA significantly improves LVLM's performance and outperforms previous approaches, achieving superior modality alignment.

1 Introduction

Large Language Models (LLMs) (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023) have significantly advanced the development of Large Vision Language Models (LVLMs) through pre-training on image-text pairs (Alayrac et al., 2022; Xiao et al., 2023) or fine-tuning on specialized vision instruction datasets (Liu et al., 2023a, 2024; Zhu et al., 2023). Despite these advancements, effectively aligning visual and language modalities remains a critical challenge in LVLMs.

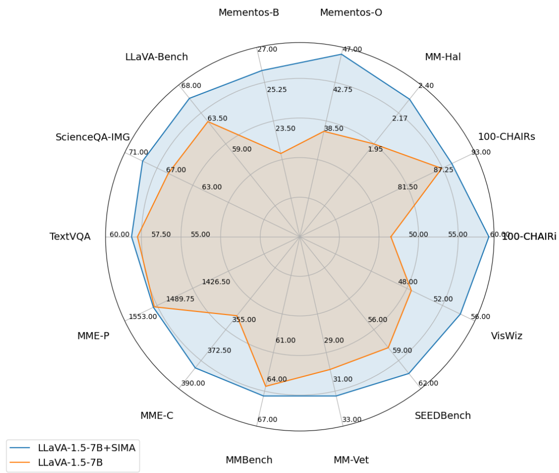


Figure 1: Performance comparison between our propose framework SIMA and LLaVA-1.5-7B on 14 hallucination and comprehensive benchmarks. After applying SIMA, LLaVA's performance is improved significantly across all benchmarks, with an average performance increase of 7.5%.

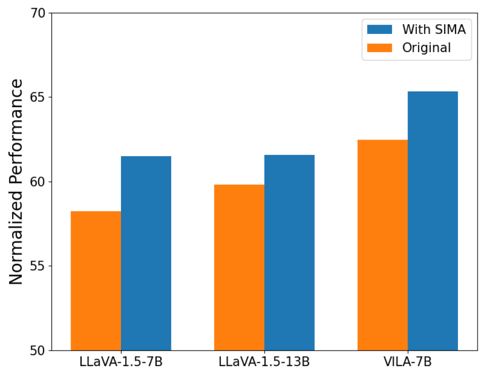


Figure 2: Normalized average performance across 14 hallucination and comprehensive benchmarks of three different LVLMs before and after using SIMA. SIMA demonstrates significant improvement on all three LVLMs.

Recent works (Sun et al., 2023; Zhao et al., 2023; Zhou et al., 2024) have attempted to enhance this alignment through preference tuning methods such

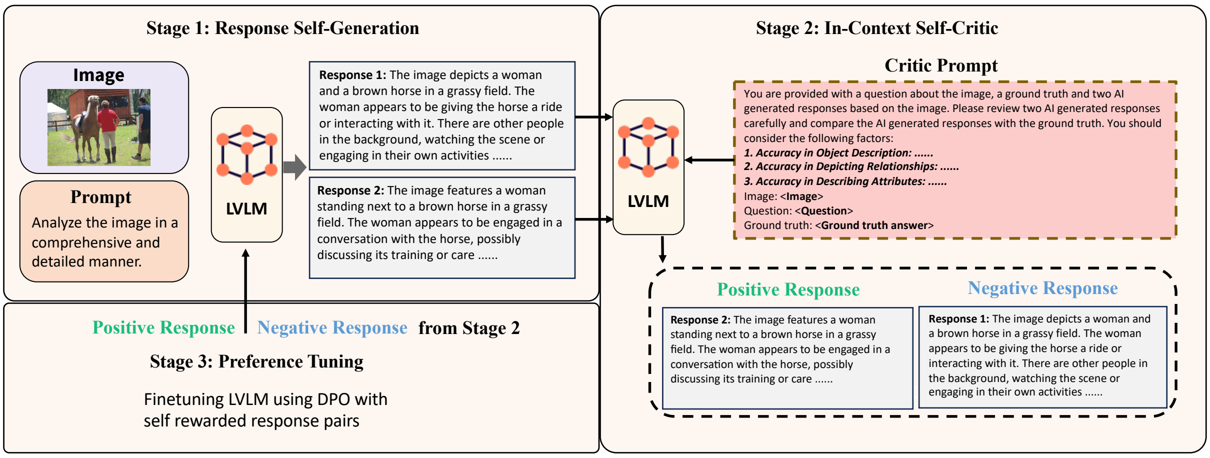


Figure 3: Flowchart of the SIMA framework. SIMA consists of three parts: Response Self-Generation, In-Context Self-Critic, and Preference Tuning.

as reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) and Direct Preference Optimization (DPO) (Rafailov et al., 2024). However, these methods often rely on external models or human-labeled data, introducing issues of uncontrollable and unstable alignment results. Specifically, they face two major challenges: **(1) Distribution Shifts:** Utilizing external LVLMs to generate preference pairs can introduce hallucinations from external models that are not representative of the current model's inference behavior (Li et al., 2023c; Zhao et al., 2023; Zhou et al., 2024). This discrepancy can lead to instability in the optimization process and potentially degrade performance. **(2) High Costs:** Dependence on human-labeled datasets or feedback from third-party AI models incurs significant annotation or API costs, making it difficult to scale high-quality preference datasets in resource-constrained environments (Sun et al., 2023; Yu et al., 2024a,b).

To address these challenges, we propose the **Self-Improvement Modality Alignment (SIMA)** framework, designed to enhance the alignment between visual and language modalities within LVLMs through a self-improvement mechanism. SIMA eliminates the need for external data or models by leveraging the intrinsic capabilities of the model itself to generate diverse responses. Moreover, it utilizes the model's own judgment for evaluating response quality, thus avoiding the high costs associated with external feedback and scaling up preference datasets efficiently.

SIMA consists of three stages: **response self-generation**, **in-context self-critic**, and **preference tuning**. In the response self-generation stage, we sample prompts from the current LVLM's visual instruction tuning dataset to generate diverse responses without introducing external data or models. During the in-context self-critic stage, a carefully designed critic prompt allows the LVLM to evaluate all self-generated responses and form preference pairs. Finally, preference tuning is applied to update the LVLM based on these pairs.

The core innovation of SIMA lies in the in-context self-critic process, which offers several key advantages: **(1) Self-Critic without Fine-Tuning:** Unlike previous self-rewarding methods in LLMs that require additional instruction tuning before the critic step (Yuan et al., 2024; Pang et al., 2024; Wu et al., 2024), our approach shows that by properly configuring the critic prompt, the LVLM can accurately evaluate responses without fine-tuning. **(2) Visual Critic Metrics:** To ensure accurate evaluation of self-generated responses, we introduce three visual critic metrics within the prompt—Accuracy in Object Description, Accuracy in Depicting Relationships, and Accuracy in Describing Attributes—each contributing to a more precise evaluation of visual content.

We apply SIMA to LLaVA-1.5 (Liu et al., 2023a) and VILA (Lin et al., 2024), evaluating it across 14 hallucination and comprehensive benchmarks. The experimental results show that SIMA not only mitigates hallucinations but also significantly enhances comprehension capabilities in LVLMs. As illustrated in Figure 1, the performance of LLaVA-1.5-7B, LLaVA-1.5-13B, and VILA-7B improved by 7.5%, 4.5%, and 5.3%, respectively. Additionally, our method outperforms other preference-tuning approaches that rely on external models and data.

The contribution of this paper can be summarized as follows: **(1)** We introduce Self-Improvement Modality Alignment SIMA, a novel framework designed to enhance alignment between visual and language modalities in LVLMs. To the best of our knowledge, SIMA is the first to achieve self-improvement in LVLMs without external data or third-party AI models. **(2)** We propose the in-context self-critic method, enabling LVLMs to accurately evaluate responses without instruction tuning, significantly improving judgment accuracy through three visual critic metrics. **(3)** SIMA demonstrates significant performance improvements in LLaVA-1.5-7B, LLaVA-1.5-13B and VILA-7B on 14 hallucination and comprehensive benchmarks, validating the effectiveness of our approach.

2 Self-Improvement Modality Alignment

In this section, we introduce the proposed **S**elf-**I**mprovement **M**odality **A**lignment (SIMA) framework. SIMA is consisted of three stages: response self-generation, in-context self-critic, and preference tuning. We will first explain how to obtain self-generated response candidates in Sec 2.1, then discuss how to use model itself π_θ to critique the response candidates in Sec 2.2. Finally, we will introduce how to use self-rewarded responses to update the π_θ in Sec 2.3. The pseudo-code of SIMA is provided in Algorithm 1.

Algorithm 1 SIMA

Require: Prompt Dataset $\{x_i, I_i\}_{i\in[N]}$, Preference dataset $\mathcal{D}_p = \{\}$, Current optimized LVLM π_θ, Reference model π_{ref},

1: **for** $i = 1, \ldots, N$ **do**
2: *// Stage 1: Response self-generation*
3: Generate one response using greedy decoding with π_θ,
4: Generate one response using temperature sampling with π_θ,
5: *// Stage 2: In-context self-critic*
6: Criticizing two generated responses with π_θ,
7: Add preference pair $\{y_w, y_l\}$ into $\mathcal{D}_p$,
8: *// Stage 3: Preference tuning*
9: Update π_θ using Eq 1 with π_{ref}

2.1 Response self-generation

Previous works often require the introduction of external models to generate preference dataset to improve current LVLM (Sun et al., 2023; Zhou et al., 2024). However, due to the significant distribution shift between the external models and the currently optimized LVLM, the generated dataset by these approaches may not be helpful to the LVLM. For example, a common method to obtain negative responses is to use external models to deliberately modify the ground truth and inject object hallucinations (Zhou et al., 2024), while the hallucinations generated by external models do not necessarily indicate that the currently optimized model would produce. In this case, using such data for learning can not enhance LVLM.

Based on our goal to identify and correct the potential misunderstandings the current LVLM may have about images and improve the modality alignment, we propose using the currently optimized LVLM to self-generate responses. This approach avoids the potential distribution shift introduced by external models. As shown in Stage 1 of Figure 3, given an image and its corresponding prompt, we use the currently optimized model to generate two different response candidates for subsequent ranking and preference tuning. Specifically, the two responses are generated using **greedy decoding** and **temperature sampling** to ensure diversity between the responses.

2.2 In-context self-critic

The core part of SIMA is criticizing the self-generated responses without introducing an additional reward model. As shown in Stage 2 of Figure 3, we directly input the self-generated responses and the critic prompt into the currently optimized LVLM. The LVLM then selects the better response as the positive response and the other one as the negative response. The most critical part of this stage is designing an appropriate critic prompt, since the quality of the critic directly determines the performance of the LVLM optimized using the response pairs. If the worse response is selected as the positive response, it will harm the training of the LVLM.

Our critic prompt consists of the following parts:

- **Image, Question, and Ground Truth Response:** Unlike LLMs, which primarily focus on aspects such as the format, helpfulness, and harmlessness of the textual response,

You are provided with a question about the image, a ground truth and two AI generated responses based on the image. \
Please review two AI generated responses carefully and compare the AI generated responses with the ground truth. \
You should consider the following factors: \
1. **Accuracy in Object Description**: Evaluate the accuracy of the descriptions concerning the objects mentioned in the ground truth answer. Responses should minimize the mention of objects not present in the ground truth answer, and inaccuracies in the description of existing objects. \
2. **Accuracy in Depicting Relationships**: Consider how accurately the relationships between objects are described compared to the ground truth answer. Rank higher the responses that least misrepresent these relationships. \
3. **Accuracy in Describing Attributes**: Assess the accuracy in the depiction of objects' attributes compared to the ground truth answer. Responses should avoid inaccuracies in describing the characteristics of the objects present. \
You need to choose one AI generated response better aligned with the ground truth. \
You must not choose the ground truth itself. \n\
The final output format is: The reason: <compare response 1 with the ground truth>. <compare response 2 with the ground truth>. \
The better AI generated response: [[<AI generated response number>]].\n\
Here are some examples: \n\
{**Demonstration 1**}\n\
{**Demonstration 2**}\n\
Question: [{}]\n\
Ground truth: [{}]\n\
AI generated response 1: [{}]\n\
AI generated response 2: [{}]\n\
ASSISTANT:\n

Figure 4: Critic prompt structure used for in-context self-critic.

LVLMs primarily focus on the accuracy of the response's understanding of the image content. This means there is a quantifiable accuracy metric to measure the quality of the response. Therefore, during in-context self-critic, we must provide the ground truth response as a reference to choose the positive response. It is worth noting that since the prompts used to generate responses are sampled from the training data of the visual instruction tuning stage, the corresponding ground truth responses have all been used for visual instruction tuning. Hence, using the ground truth in the in-context self-critic stage is reasonable.

- **Three critic metrics:** Although we provide the ground truth response as a reference, without proper guidance, the LVLM might still choose a response that aligns more with the ground truth in terms of output format or harmlessness rather than focusing on the accuracy of visual comprehension. Therefore, we propose three metrics to guide LVLM ranking, ensuring it select the positive response from the visual comprehension perspective. The three critic metrics are: **Accuracy in Object Description**, **Accuracy in Depicting Relationships**, and **Accuracy in Describing Attributes**.

  Accuracy in Object Description aims to guide current LVLM in evaluating the accuracy of the descriptions concerning the objects mentioned in the ground truth answer. The responses should minimize the mention of objects not present in the ground truth answer and inaccuracies in the description of existing objects. Accuracy in Depicting Relationships considers how accurately the relationships between objects are described compared to the ground truth answer and aims to let LVLM rank higher the responses that least misrepresent these relationships. Accuracy in Describing Attributes assesses the accuracy in depicting objects' attributes compared to the ground truth answer. The responses should avoid inaccuracies in describing the characteristics of the objects present.

- **Demonstrations:** To ensure the correct format of the ranking output, we also leverage in-context learning by providing two ranking demonstrations in the designed ranking prompt for the LVLM to imitate.

In Figure 4, we provide the structure of the critic prompt. For the detailed critic prompt, please refer to the Appendix A.

2.3 Preference tuning

After obtaining the preference pairs through self-ranking, we use these preference pairs to perform preference tuning on the current LVLM. We choose direct preference optimization (DPO) (Rafailov et al., 2024) as the preference tuning method. The preference dataset is denoted as $\mathcal{D}_p = \{(I, x, y_w, y_l)\}$, where I is the image, x is the corresponding question, y_w is the positive response and y_l is the negative response, the DPO objective is defined as below:

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

where π_θ is the current optimized LVLM and π_{ref} is the base reference model, both models are initialized with visual instruction tuning weights. σ is the logistic function.

3 Experiment

In this section, we conduct experiments and aim to answer the following questions: 1. How much does SIMA improve baseline performance? 2. How significant are the three ranking metrics in the ranking prompt?

3.1 Benchmark evaluation

Implementation details Since LLaVA (Liu et al., 2024) is the most widely used open-source LVLM and following recent LVLM preference tuning studies (Sun et al., 2023; Zhou et al., 2024; Yu et al., 2023a; Xiao et al., 2024), we select LLaVA-1.5-7B (Liu et al., 2023a) and LLaVA-1.5-13B (Liu et al., 2023a) as the backbone models and apply SIMA on them. The prompts used to generate preference data are randomly sampled from two categories, 'complex_reasoning_77k' and 'detail_23k', in LLaVA's visual instruction tuning dataset, LLaVA-Instruct-150K, thus avoiding introducing additional data. We sample a total of 17k prompts. To demonstrate the generalizability of SIMA, we also choose VILA-7B (Lin et al., 2024), a recent LVLM, as the base model for our experiments. Similar to the LLAVA experimental setting, we randomly sample 17k prompts from the VILA visual instruction tuning dataset to generate preference pairs for training. After obtaining the preference pairs, we finetune LLaVA and VILA with SIMA on this data using LoRA (Hu et al., 2021) for three epochs on LLaVA-1.5-7B, one epoch on LLaVA-1.5-13B, and one epoch on VILA-7B since we find that LLaVA-1.5-13B and VILA-7B is prone to overfitting on the sampled dataset. All experiments are conducted on one A100 80GB GPU with 15 gpu hours for three epochs training on LLaVA-1.5-7B, 7 gpu hours for one epoch training on LLaVA-1.5-13B, and 6 gpu hours for one epoch training on VILA-7B.

Baselines For the baselines, we compare with three previous methods that use preference optimization to improve LVLM performance: LLaVA-RLHF (Sun et al., 2023), HA-DPO (Zhao et al., 2023), and POVID (Zhou et al., 2024). LLaVA-RLHF trains a reward model by incorporating additional human-annotated preference data and then finetunes LLaVA using PPO. HA-DPO uses GPT to rewrite AI-generated responses for hallucination mitigation and data augmentation and then apply DPO to fine-tune the LVLM. POVID introduces GPT to inject hallucinations into the ground truth answers and add noise to images to induce hallucinations in the LVLM to obtain negative samples and also uses DPO to finetune the LVLM. These three methods are all based on LLaVA-1.5-7B. Besides, we compare the method of using the ground truth answer as the positive sample and the LVLM-generated response as the negative sample for DPO finetuning, which we refer to as GT-DPO. We also report comparison with other popular open-source LVLMs as a reference to demonstrate the superiority of our experimental results in Appendix B.3.

Benchmarks We select 14 hallucination and comprehensive benchmarks for evaluation. For the hallucination benchmark, we randomly sample 5000 images from the COCO (Lin et al., 2014) validation set and randomly pair them with 5 questions, resulting in 5000 <image, question> pairs. We then evaluate the object hallucination rate on these 5000 pairs using the CHAIR (Rohrbach et al., 2018) metric , calculated as follows:

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|},$$
$$\text{CHAIR}_S = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}. \tag{2}$$

We also use MM-Hal (Sun et al., 2023) and Mementos (Wang et al., 2024) as benchmarks for evaluating hallucination. In Mementos, we use F1 score as the metric to assess the LVLM's object hallucination and behavior hallucination when understanding multi-image inputs. For the comprehensive benchmark, we select nine commonly used comprehensive benchmarks and general VQA tasks: LLaVA in the Wild (Liu et al., 2024), ScienceQA (Lu et al., 2022), TextVQA (Singh et al., 2019), MME Perception (Fu et al., 2024), MME Cognition (Fu et al., 2024), MMBench (Liu et al., 2023b), MM-Vet (Yu et al., 2023b), SeedBench (Li et al., 2023a), and VisWiz (Gurari et al., 2018). For details on these benchmarks, please refer to the Appendix C.

Experiment results **(a) SIMA can significantly reduce hallucinations of LVLMs.** As shown in Table 1, SIMA significantly improves the performance of all three LVLMs on five hallucina-

Table 1: Performance comparison between SIMA and other baselines on hallucination benchmarks

LVLMs	CHAIRs ↓	CHAIRi ↓	MM-Hal ↑	Mementos^{O} ↑	Mementos^{B} ↑
LLaVA-1.5-7B	50.8	11.7	2.04	39.29%	23.02%
+ RLHF	45.3	11.1	2.11	40.53%	22.71%
+ GT-DPO	47.3	11.2	2.00	43.67%	24.35%
+ HA-DPO	46.5	10.7	1.97	41.07%	23.58%
+ POVID	48.4	11.3	2.28	42.95%	23.84%
+ **SIMA** (ours)	**40.9**	**10.4**	**2.30**	**46.08%**	**26.03%**
LLaVA-1.5-13B	48.6	10.8	2.19	40.37%	24.65%
+ GT-DPO	47.2	10.8	2.27	42.59%	25.84%
+ **SIMA** (ours)	**45.8**	**10.6**	**2.41**	**45.84%**	**27.17%**
VILA-7B	34.7	9.2	2.53	41.96%	25.88%
+ GT-DPO	32.4	8.9	2.61	44.25%	26.91%
+ **SIMA** (ours)	**28.4**	**8.4**	**2.66**	**48.15%**	**27.04%**

Table 2: Performance comparison between SIMA and other baselines on comprehensive benchmarks

LVLMs	LLaVA^{W} ↑	SQA^{I} ↑	VQA^{T} ↑	MME^{P} ↑	MME^{C} ↑	MMB ↑	MM-Vet ↑	SEED ↑	VisWiz ↑
LLaVA-1.5-7B	63.4	66.8	58.2	1506.4	355.7	64.3	30.5	58.6	50.0
+ RLHF	63.7	65.8	58.3	1508.2	360.2	60.4	31.1	60.0	52.2
+ GT-DPO	64.7	67.4	58.1	**1510.8**	365.0	64.6	31.2	60.4	53.8
+ HA-DPO	64.2	68.1	58.0	1507.2	362.3	63.9	30.9	60.2	53.9
+ POVID	65.3	**69.2**	58.1	1493.5	363.5	64.1	31.3	60.3	54.0
+ **SIMA** (ours)	**66.1**	69.1	**58.5**	1507.7	**379.3**	**64.9**	**31.6**	**60.6**	**54.4**
LLaVA-1.5-13B	66.5	71.6	**61.3**	1531.1	296.1	67.7	36.1	61.6	53.6
+ GT-DPO	66.9	72.3	61.2	1532.6	296.7	68.0	36.3	62.2	54.4
+ **SIMA** (ours)	**67.4**	**72.5**	61.2	**1538.1**	**298.6**	**68.4**	**38.3**	**63.0**	**55.5**
VILA-7B	69.7	68.2	64.4	1533.0	316.4	68.9	34.9	61.1	57.8
+ GT-DPO	71.4	70.6	65.9	1547.8	325.7	69.0	37.1	61.9	60.3
+ **SIMA** (ours)	**73.5**	**72.2**	**66.1**	**1559.6**	**326.8**	**69.2**	**38.4**	**62.5**	**62.1**

tion benchmarks. On the CHAIRs, CHAIRi, and Mementos-Object benchmarks, which test object hallucination, SIMA improves he performance of LLaVA-1.5-7B, LLaVA-1.5-13B, and VILA-7B by an average of 16.1%, 7.1%, and 8.4%, respectively. On the MM-Hal benchmark, which uses GPT as an evaluator for a more comprehensive assessment of hallucinations, SIMA achieves 12.7%, 10.1%, and 5.1% performance improvement compared with LLaVA-1.5-7B, LLaVA-1.5-13B, and VILA-7B. Notably, despite our three critic metrics focusing primarily on object hallucination, SIMA also achieves the greatest improvement of 13.1% on the Mementos-Behavior benchmark based on LLaVA-1.5-7B model, which tests behavior hallucination arising from understanding sequential image inputs. This improvement is significant because there is a correlation between object hallucination and behavior hallucination in sequential image understanding (Wang et al., 2024); reducing object hallucination increases the likelihood of correctly inferring the corresponding behavior. **(b) SIMA also enhances the comprehension capabilities of LVLMs.** As shown in Table 2, on the nine comprehensive and VQA benchmarks, although the improvements are not as pronounced as on the hallucination benchmarks, SIMA still achieves an average improvement of 3.5%, 2.1%, and 4.4% compared to LLaVA-1.5-7B, LLaVA-1.5-13B, and VILA-7B. This is superior to other preference tuning methods.

3.2 Importance of our critic metric

In this section, our main objective is to demonstrate the importance of the three critic metrics in the in-context self-critic stage through experiments and case studies. We use LLaVA-1.5-7B as base model to conduct experiments. As in the experimental setup described in Section 3.1, the prompts used to generate response candidates are sampled from LLaVA's visual instruction tuning dataset. After self-generating the response candidates, we keep these candidates unchanged and use LLaVA to evaluate them with both metric-inclusive and metric-

Table 3: The performance comparison between training LLaVA with preference pairs obtained using metric-inclusive and metric-free critic prompts in the in-context self-critic process.

	Hallucination Benchmark					Comprehensive Benchmark								
	CHAIRs	CHAIRi	MM-Hal	Mem^{O}	Mem^{B}	LLaVA^{W}	SQA^{I}	VQA^{T}	MME^{P}	MME^{C}	MMB	MM-Vet	SEED	VisWiz
LLaVA-1.5-7B	50.8	11.7	2.04	39.29%	23.02%	63.4	66.8	58.2	1506.4	355.7	64.3	30.5	58.6	50.0
+ SIMA w/o metrics	41.5	10.8	2.12	41.55%	23.92%	63.3	68.9	58.3	1504.6	371.7	64.0	31.5	60.4	53.7
+ **SIMA** (ours)	40.9	10.4	2.30	46.08%	26.03%	66.1	69.1	58.5	1507.7	379.3	64.9	31.6	60.6	54.4

free critic prompts, resulting in preference pairs that are then used to update the LLaVA. We test the performance of both methods on 14 benchmarks, with the results shown in Table 3. Upon comparison, we find that removing the critic metrics still improved performance compared to the original LLaVA, but there remained a significant gap compared to SIMA with metrics. This disparity is particularly notable in more challenging tasks like MM-Hal and Mementos, where the improvement from SIMA without critic metrics is quite limited. This demonstrates that with the correct design of critic prompts, LVLMs can gain critic capabilities and improve model performance without requiring instruction fine-tuning. Moreover, the three visual critic metrics are crucial for further enhancing performance.

Table 4: Comparison of response critic results with human judgment.

	Select 1	Select 2	Align w. human
Human	183	317	-
GPT-4v	198	302	95.6%
SIMA	215	285	89.8%
SIMA w/o metrics	246	254	78.2%

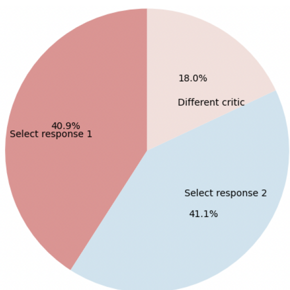
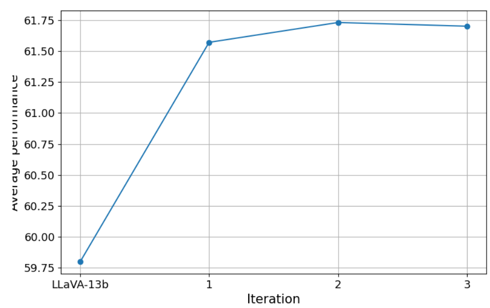


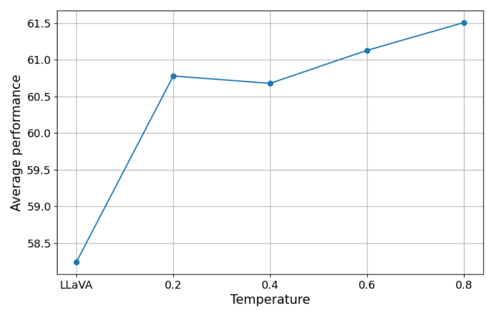
Figure 5: Comparison of critic results with and without critic metrics in SIMA.

We compare the evaluation results distribution of response candidates with and without using metrics, as shown in Figure 5. It can be seen that approximately 20% of the response candidates have inconsistent evaluations between the two methods. Additionally, we randomly sample 500 response candidates and evaluate them both manually by the authors of this paper and with GPT-4v. For human evaluation, we provide 500 response pairs and asked individuals to directly select the better one. For GPT-4v, we use the same critic prompt with metrics as SIMA for the evaluation. Comparing these evaluations with SIMA's results in Table 4, we find that without the critic metrics, SIMA's evaluations are only 78% consistent with human evaluations. After incorporating metrics, this consistency improved by 11.2% to 89.8%, which is very close to the evaluation results of GPT-4v and human. In Appendix E.2, we also present an example of evaluation results with and without metrics to further illustrate the magic of these three visual metrics.

3.3 Ablation studies



(a) Average performance of LLaVA-1.5-13B with SIMA at different iterations.



(b) Average performance of SIMA on LLaVA-1.5-7B with different decoding temperature.

Figure 6: Ablation studies of SIMA.

In this section, we conduct ablation studies on SIMA from two aspects: the performance variation of SIMA under multi-iteration finetuning and the impact of different decoding temperatures on performance when generating response candidates.

Performance of multi-iteration finetuning Figure 6(a) shows the average performance of the model on benchmarks across different training iterations on LLaVA-1.5-13B. For detailed performance on each benchmark, please refer to Table 8 in the Appendix B. In each iteration, we randomly resample 17k prompts from LLaVA's visual instruction tuning dataset for self-generation. We observe that the performance improvement is most noticeable in the first iteration compared to the base model. In the second iteration, there is an improvement, but it is not as pronounced. Although the average performance saturates in the third iteration, performance in some benchmarks continues to improve, as seen in Table 8.

Different decoding temperatures We also conduct an ablation study on the temperature used in temperature decoding during the response self-generation phase. The experimental results are shown in Figure 6(b). We find that as the temperature increases, the performance of SIMA also improves. We believe this is because, as the temperature increases, the responses generated by LVLM become more diverse and are more likely to exhibit hallucination. This increases the distribution shift between the responses generated by greedy decoding and those generated with higher temperature, leading to better performance improvements for LVLM during the preference tuning phase.

4 Related Work

Vision-Language Models Vision-Language Models (VLMs) (Li et al., 2019, 2020; Wang et al., 2021; Radford et al., 2021; Li et al., 2022) have emerged as critical tools in bridging visual and textual modalities, enabling advancements in multimodal understanding and reasoning tasks. Recent developments have been driven by the integration of large language models (LLMs) (Touvron et al., 2023; Jiang et al., 2023; Chiang et al., 2023) and sophisticated image encoders, leading to more robust and versatile Large Vision-Language Models (LVLMs) (Bai et al., 2023; Zhu et al., 2023; Chen et al., 2023; Dai et al., 2024; Lin et al., 2024; Yao et al., 2024; Liu et al., 2024). For instance, models like LLaVA (Liu et al., 2024) and InstructBLIP (Dai et al., 2024) combine advanced vision encoders with LLMs, enhancing their ability to follow vision-language instructions. In this work, we focus on further enhancing LVLM's visual understanding and reasoning abilities based on LVLM's visual instruction tuning data through self-improvement.

Modality Alignment Vision-language modality misalignment is a key challenge in LVLMs, where the generated textual outputs may not fully correspond to the input visual data. Preference learning (Rafailov et al., 2024; Azar et al., 2024; Ethayarajh et al., 2024) is generaly used to improve modality alignment in LVLMs. Some methods, such as using human annotation (Sun et al., 2023; Yu et al., 2024a) and third-party AI model feedback (Li et al., 2023c; Zhao et al., 2023; Zhou et al., 2024; Yu et al., 2024b; Jing and Du, 2024) for preference learning, have been proposed. However, these methods are resource-intensive and may introduce additional external hallucinations, leading to LVLM performance that is uncontrollable and unstable after optimization. In this paper, we addresses both issues through a self-improvement approach, significantly enhancing modality alignment without introducing any external models or data.

Self-Improvement in Large Language Models Self-improvement is proposed in LLM to improve LLM itself with self-generated data. Several papers have explored self-improvement in LLM (Yuan et al., 2024; Pang et al., 2024; Wu et al., 2024). To the best of our knowledge, this paper is the first to explore self-improvement in LVLMs. Different from previous self-improvement methods in LLM which need to finetune the LLM with additional instruction tuning data before critic, our method demonstrate that LVLM can acquire the ability to act as a critic by properly configuring critic prompt without fine-tuning.

5 Conclusion

In this paper, we introduce SIMA framework in enhancing the alignment between visual and language modalities in LVLMs through self-improvement. This is achieved through self-generated responses, evaluating them via in-context self-critic, and employing preference tuning. SIMA bypasses the need for the third-party AI model for data generation and response evaluation, making it more scalable and cost-effective. This approach not only improves the modality alignment but also significantly enhances the model's comprehension abilities and reduces hallucinations across various benchmarks.

Limitations

One limitation of this paper is that the reliance on self-generated responses and self-critic inherently ties the SIMA's performance to the current capabilities of LVLM and does not address the inherent potential biases caused by the vision instruction tuning dataset. This can result in SIMA providing less significant improvements for LVLMs on certain benchmarks, such as LLaVA-1.5-7B and LLaVA-1.5-13B on TextVQA. In future work, we will further explore this issue.

Broader Impacts

To the best of our knowledge, we are the first to apply self-rewarding in LVLMs. This approach avoids the introduction of external models and data, enhancing the alignment between visual and language modalities through the model itself. This significantly reduces hallucinations and improves reasoning capabilities, greatly increasing the reliability of LVLMs.

From a societal impact perspective, while SIMA has made substantial progress, it has not entirely eliminated potential risks within LVLMs. For example, reliance on self-generated and self-critic data may unintentionally reinforce biases caused by distribution shifts in the training data. Therefore, despite SIMA's significant advancements, it is crucial to implement ethical guidelines and safeguards to mitigate these risks and ensure responsible use of this technology.

References

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

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

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. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

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

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

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

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023c. Silkie: Preference distillation for large visual language models.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

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

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023b. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

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

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

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

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

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. 2024.

Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2023. Florence-2: Advancing a unified representation for a variety of vision tasks. *arXiv preprint arXiv:2311.06242*.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Hao Jiang, Fei Wu, and Linchao Zhu. 2024. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. *arXiv preprint arXiv:2404.14233*.

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

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

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

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024a. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023b. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

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

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

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

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

A Detailed critic prompt

In this section, we provide detailed critic prompt used during the in-context self-critic phase, as shown in Figure 7.

B Detailed ablation studies

B.1 LLaVA-1.5-7B

In this section, we provide the performance of LLaVA-1.5-7B-7B across all benchmarks at different training epochs, as shown in Table 5. Additionally, in Table 6, we present the performance of SIMA across all benchmarks when using different temperature coefficients for temperature decoding.

B.2 LLaVA-1.5-13B

In this section, we present the detailed performance of LLaVA-1.5-13B across all benchmarks at different epochs and iterations in Tables 7 and Tables 8, respectively. It can be observed that the best results for LLaVA-1.5-13B are achieved after just one epoch. During multiple iteration training, performance on some benchmarks continues to improve in the third iteration while some declines due to overfitting.

B.3 Comparison with other open-source LVLMs

In this section, we report the performance of five other popular open-source LVLMs (BLIP-2 (Li et al., 2023b), InstructBLIP (Dai et al., 2024), IDEFICS (Laurençon et al., 2024), Qwen-VL-Chat (Bai et al., 2023), and mPLUG-Owl2 (Ye et al., 2023)) as a reference to demonstrate the superiority of our experimental results in Table 9.

```
You are provided with a question about the image, a ground truth and two AI generated responses based on the image. \
Please review two AI generated responses carefully and compare the AI generated responses with the ground truth. \
You should consider the following factors: \
1. Accuracy in Object Description: Evaluate the accuracy of the descriptions concerning the objects mentioned in the ground truth answer. Responses should minimize the mention of objects not present in the ground truth answer, and inaccuracies in the description of existing objects. \
2. Accuracy in Depicting Relationships: Consider how accurately the relationships between objects are described compared to the ground truth answer. Rank higher the responses that least misrepresent these relationships. \
3. Accuracy in Describing Attributes: Assess the accuracy in the depiction of objects' attributes compared to the ground truth answer. Responses should avoid inaccuracies in describing the characteristics of the objects present. \
You need to choose one AI generated response better aligned with the ground truth. \
You must not choose the ground truth itself. \n\
The final output format is: The reason: <compare response 1 with the ground truth>. <compare response 2 with the ground truth>. \
The better AI generated response: [[<AI generated response number>]].\n\
Here are some examples: \n\
Question: [Analyze the image in a comprehensive and detailed manner.]\n\
Ground truth: [The image depicts a large horse standing in a field, accompanied by two main individuals. One person, dressed in English-style riding pants and boots, is standing close to the horse, while the other is further back to the right. There are several other people in the backdrop, making the scene appear lively and social. A potted plant can also be found towards the right side of the image, adding a touch of greenery to the scene.]\n\
AI generated response 1: [The image features a woman standing next to a brown horse in a grassy field. The woman appears to be engaged in a conversation with the horse, possibly discussing its training or care. There are several other people in the background, some of whom might be observing the interaction between the woman and the horse. In addition to the people and the horse, there are two potted plants visible in the scene, one located near the left edge of the image and the other towards the right side. A book can also be seen on the ground, possibly belonging to one of the people present in the field.]\n\
AI generated response 2: [The image features a woman standing next to a brown horse in a grassy field. The woman appears to be engaged in a conversation with the horse, possibly discussing its training or care. The horse is wearing a bridle, indicating that it is well-trained and responsive to commands. In the background, there are several other people present, possibly observing the interaction between the woman and the horse or engaging in their own activities. The scene captures a moment of connection between the woman and the horse, showcasing the bond between humans and animals.]\n\
ASSISTANT:\n\
The reason: AI generated response 1 incorrectly mentions two potted plants and a book on the ground, none of which are described in the ground truth. There's only one potted plant mentioned in the ground truth, and no book is noted. AI generated response 2 does not mention the potted plants or a book, thus avoiding the inaccuracies present in response 1. \
The better AI generated response: [[2]].\n\n\
Question: [What could be a possible reason for the man wearing a number one on his chest while riding the horse?]\n\
Ground truth: [While riding the chestnut-colored horse in the competition, such as a polo match or equestrian event, the man is wearing a number one on his chest to identify his team or individual position. In these types of events, each participant or team member is assigned a unique number for easy identification and tracking during the competition. This allows organizers and spectators to differentiate between participants and teams, helping them follow the progress and performance of each individual or team member throughout the event.]\n\
AI generated response 1: [A possible reason for the man wearing a number one on his chest while riding the horse could be that he is participating in a competition or event. In such events, riders often wear numbers to identify themselves and their horses. The number one on his chest might indicate that he is the first rider in the competition or event, or it could simply be a unique identifier assigned to him by the organizers.]\n\
AI generated response 2: [A possible reason for the man wearing a number one on his chest while riding the horse could be that he is participating in a competitive event, such as a horse show, equestrian competition, or even a casual game with friends. In such events, participants often wear numbers to identify themselves and keep track of their performance, placing in the competition, or personal achievements. The number one on his chest could indicate that he is the top-ranked participant, or he is simply using it to distinguish himself from the other competitors.]\n\
ASSISTANT:\n\
The reason: AI generated response 2 includes additional speculation about the man being the top-ranked participant or using the number to distinguish himself, which is not mentioned in the ground truth. It also suggests that the event could be a casual game with friends, which is not supported by the ground truth focusing on a competitive context. AI generated response 1 is more aligned with the ground truth as it strictly suggests participation in a competition or event and does not add unrelated possibilities. Both responses mention the use of numbers for identification, which aligns with the ground truth.
\
The better AI generated response: [[1]].\n\n\
Question: [{}]\n\
Ground truth: [{}]\n\
AI generated response 1: [{}]\n\
AI generated response 2: [{}]\n\
ASSISTANT:\n
```

Figure 7: Critic prompt used for in-context self-critic.

Table 5: Performance of different epochs on LLaVA-1.5-7B.

	Hallucination Benchmark					Comprehensive Benchmark									
	CHAIRs	CHAIRi	MM-Hal	Mem^{O}	Mem^{B}	LLaVA^{W}	SQA^{I}	VQA^{T}	MME^{P}	MME^{C}	MMB	MM-Vet	SEED	VisWiz
LLaVA-1.5-7B	50.8	11.7	2.04	39.29%	23.02%	63.4	66.8	58.2	1506.4	355.7	64.3	30.5	58.6	50.0
+ **SIMA** Epoch 1	43.9	10.8	2.17	42.39%	23.88%	65.3	68.9	58.2	1511.9	369.6	64.9	30.5	60.1	53.7
+ **SIMA** Epoch 2	41.6	10.4	2.28	45.71%	24.93%	66.1	69.2	58.2	1514.8	371.8	65.0	31.5	60.4	54.0
+ **SIMA** Epoch 3	40.9	10.4	2.30	46.08%	26.03%	66.1	69.1	58.5	1507.7	379.3	64.9	31.6	60.6	54.4

Table 6: Performance of different decoding temperature.

	Hallucination Benchmark					Comprehensive Benchmark									
	CHAIRs	CHAIRi	MM-Hal	Mem^{O}	Mem^{B}	LLaVA^{W}	SQA^{I}	VQA^{T}	MME^{P}	MME^{C}	MMB	MM-Vet	SEED	VisWiz
T=0.2	40.2	10.1	2.11	45.42%	24.99%	65.2	68.5	58.3	1505.0	371.8	64.7	31.1	60.1	53.7
T=0.4	40.7	10.2	2.19	45.93%	25.37%	64.9	68.9	58.3	1506.4	355.7	65.0	31.1	60.3	53.8
T=0.6	40.9	10.3	2.23	45.71%	25.61%	65.7	69.2	58.2	1504.8	371.8	64.9	31.3	60.3	54.1
T=0.8	40.9	10.4	2.30	46.08%	26.03%	66.1	69.1	58.5	1507.7	379.3	64.9	31.6	60.6	54.4

Table 7: Performance of different epochs on LLaVA-1.5-13B.

	Hallucination Benchmark					Comprehensive Benchmark									
	CHAIRs	CHAIRi	MM-Hal	Mem^{O}	Mem^{B}	LLaVA^{W}	SQA^{I}	VQA^{T}	MME^{P}	MME^{C}	MMB	MM-Vet	SEED	VisWiz
LLaVA-1.5-13B	48.6	10.8	2.19	40.37%	24.65%	66.5	71.6	61.3	1531.1	296.1	67.7	36.1	61.6	53.6
+ **SIMA** Epoch 1	45.8	10.6	2.41	45.84%	27.17%	67.4	72.5	61.2	1538.1	298.6	68.4	38.3	63.0	55.5
+ **SIMA** Epoch 2	46.1	10.6	2.26	45.53%	26.99%	67.2	72.4	61.2	1537.5	291.1	68.5	37.6	63.0	55.0
+ **SIMA** Epoch 3	45.9	10.6	2.21	45.61%	26.74%	66.0	72.4	61.1	1529.2	291.4	68.3	35.9	63.0	54.9

Compared to other open-source LVLMs, SIMA also significantly outperforms all except for Qwen-VL-Chat on MM-Vet.

Table 8: Performance of different iterations on LLaVA-1.5-13B.

	Hallucination Benchmark					Comprehensive Benchmark								
	CHAIRs	CHAIRi	MM-Hal	Mem^{O}	Mem^{B}	LLaVA^{W}	SQA^{I}	VQA^{T}	MME^{P}	MME^{C}	MMB	MM-Vet	SEED	VisWiz
LLaVA-1.5-13B	48.6	10.8	2.19	40.37%	24.65%	66.5	71.6	61.3	1531.1	296.1	67.7	36.1	61.6	53.6
+ **SIMA** Iter 1	45.8	10.6	2.41	45.84%	27.17%	67.4	72.5	61.2	1538.1	298.6	68.4	38.3	63.0	55.5
+ **SIMA** Iter 2	45.3	10.6	2.46	46.02%	27.58%	67.5	72.7	61.2	1528.9	298.6	68.5	38.3	62.9	55.9
+ **SIMA** Iter 3	45.4	10.6	2.42	46.91%	27.63%	67.3	72.6	61.1	1529.8	298.6	68.6	37.9	63.0	56.0

Table 9: Performance comparison between SIMA and other open-source LVLMs on comprehensive benchmarks

LVLMs	LLaVA^{W} ↑	SQA^{I} ↑	VQA^{T} ↑	MME^{P} ↑	MME^{C} ↑	MMB ↑	MM-Vet ↑	SEED ↑	VisWiz ↑
BLIP-2	38.1	61.0	42.5	1293.8	290.0	-	22.4	46.4	19.6
InstructBLIP	60.9	60.5	50.1	1212.8	291.8	36.0	26.2	53.4	34.5
IDEFICS	45.0	-	25.9	1177.3	-	30.0	30.0	45.0	35.5
Qwen-VL-Chat	67.7	68.2	61.5	1487.6	360.7	60.6	47.3	58.2	38.9
mPLUG-Owl2	59.9	68.7	58.2	1450.2	313.2	64.5	36.2	57.8	54.5
LLaVA-1.5-7B	63.4	66.8	58.2	1506.4	355.7	64.3	30.5	58.6	50.0
+ **SIMA** (ours)	**66.1**	69.1	**58.5**	1507.7	**379.3**	**64.9**	**31.6**	**60.6**	**54.4**
VILA-7B	69.7	68.2	64.4	1533.0	316.4	68.9	34.9	61.1	57.8
+ **SIMA** (ours)	**73.5**	**72.2**	**66.1**	**1559.6**	**326.8**	**69.2**	**38.4**	**62.5**	**62.1**

C Benchmark details

LLaVA^{W} is an extensive benchmark for assessing visual reasoning models. It includes 24 varied images accompanied by a total of 60 questions, encompassing scenarios from indoor and outdoor settings to abstract art.

ScienceQA is a multi-modal benchmark designed to evaluate and diagnose the multi-hop reasoning capabilities and interpretability of artificial intelligence systems in science. It provides an extensive data set of approximately 21,000 multiple-choice questions covering a wide range of scientific topics, supported by detailed answer notes, relevant lectures and explanations.

TextVQA is a dataset that benchmarks visual reasoning based on text in images. TextVQA requires models to read and reason about text in images to answer questions about them. Specifically, the model needs to incorporate a new form of text into the image and reason about it to answer the TextVQA question.

MME serves as a comprehensive benchmark for evaluating the capabilities of LVLMs in multimodal tasks. It evaluates models systematically across two main dimensions: perception and cognition, using 14 carefully designed subtasks that test the models' interpretative and analytical abilities.

MMBench introduces a two-pronged approach: a carefully curated dataset that significantly expands the scope and diversity of evaluation questions, and a groundbreaking CircularEval strategy that leverages ChatGPT to transform free-form predictions for structured choices.

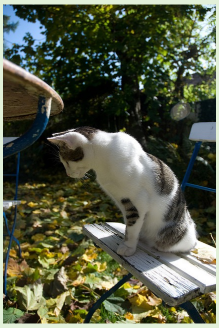
MM-Vet is an evaluation benchmark specially designed to evaluate the multi-faceted capabilities of LVLM. It systematically builds complex multimodal tasks into 16 different ensembles derived from combinations of 6 core visual language features, providing granular analysis of model performance across different question types and answer styles.

SEEDBench is intended to rigorously assess the generative comprehension capabilities of LVLMs. It includes a comprehensive dataset of 19K multiple-choice questions with accurate human annotations, spanning 12 distinct evaluation dimensions that test both spatial and temporal understanding across image and video modalities.

VizWiz is a dataset in the field of visual question answering (VQA) derived from a naturalistic setting containing over 31,000 visual questions. It features a goal-oriented approach, featuring images taken by blind people, accompanied by their verbal queries, and crowdsourced answers.

D Hyperparameters

In this section, we provide the hyperparameters used during training, as well as the GPT version utilized during evaluation, as shown in Table 10.



Prompt: Write a detailed description for the given image

LLaVA: The image features a white and black cat sitting on a wooden bench in a park. The bench is located near a dining table, which is surrounded by several chairs. The cat appears to be looking at the table, possibly intrigued by something on it. The scene is set in a garden, with a chair placed close to the bench and another chair further away. The overall atmosphere is calm and serene, with the cat enjoying its time outdoors.

SIMA: The image features a cat sitting on a wooden bench in a park. The bench is located near a dining table, which is surrounded by several chairs. The cat appears to be looking at the ground, possibly exploring the area or observing something of interest. The scene captures a peaceful and relaxing moment in the park.

Prompt: Describe the given image in detail

LLaVA: The image features a large brown cow lying down on the ground, taking up a significant portion of the scene. The cow appears to be resting or relaxing in the outdoor area. In the background, there are two motorcycles parked, one on the left side and the other on the right side of the cow. Additionally, there is a scooter parked further back on the right side of the scene. The presence of these vehicles suggests that the location might be a busy street or a parking area.

SIMA: The image features a large brown cow lying down on the ground, taking up a significant portion of the scene. The cow appears to be relaxed and comfortable in its position. In the background, there are two motorcycles parked nearby. One motorcycle is located on the left side of the cow, while the other is on the right side. The presence of the motorcycles suggests that the location might be a public space or a place where people gather.

Figure 8: Examples to illustrate the effect of SIMA on LVLM. SIMA is presented to be less hallucinated compared with LLaVA.

Table 10: Hyperparameters of SIMA during training and evaluation.

Parameter	Value
lora r	128
lora alpha	256
mm projector lr	2e-5
learning rate	1e-7
model max length	2048
batch size	1
decoding temperature	0.8
GPT api version (Eval)	gpt-4-turbo

E Case Study

E.1 SIMA case study

In Figure 8, we compare the detailed captions generated from LLaVA and our method SIMA. For instance, in the above demonstration in Figure 8, LLaVA inaccurately describes the scene 'with a chair close to the bench and another chair further away', details that are not present in the provided image, while SIMA accurately captures the details in images. Both demonstrations in Figure 8 indicate that SIMA is capable of generating more accurate descriptions with fewer hallucinations.

E.2 Critic prompt case study

In Figure 9, we present an example of evaluation results with and without metrics. The example shows that after using the critic metrics, the SIMA's evaluation is guided to focus more on the details of the visual content, leading to correct judgments. Therefore, based on the analysis and results above, it is evident that critic metrics are crucial for improving the accuracy of response evaluations during in-context self-critic.

F Status of Exemption from Institutional Review Board

Before starting any segments of the study involving human evaluation, the research team completed and submitted a "Human Subjects Research Determination" form to the appropriate Institutional Review Board (IRB). We obtained a determination letter from the IRB before any human study activities commenced, indicating that our project proposal had been granted 'Exempt' status. This classifica-

Prompt: Can you describe the main features of this image for me?

Ground truth: The image captures a lively outdoor market, bustling with people browsing through the fruits and vegetables on display. Various individuals are present in the scene, with a man standing in front of a fruit stand close to the center, while another man wearing a backpack is nearby. Additional people can be seen in the background, possibly exploring the market or conversing with one another. The fruit stand is mainly filled with bananas, which come in different shapes and sizes, indicating a vast selection for shoppers. There are also apples available, located towards the bottom of the scene. The market provides a vibrant atmosphere, with its inviting array of fresh produce and the interactions of customers and vendors.

Self-Generation

Temperature decoding: The image depicts a lively market scene with several people shopping for various fruits and vegetables. There are at least four people visible in the scene, engaged in different activities related to the market. Among the produce, there are numerous apples and oranges, with several apples scattered throughout the scene. Oranges can also be seen, placed in different locations within the market. In addition to the fresh fruits, there is a backpack on the right side of the image, possibly belonging to one of the shoppers. A cell phone is also visible, likely being used by someone in the market.

Greedy decode: The image depicts a lively outdoor market scene with a variety of fresh fruits and vegetables on display. There are numerous bananas, apples, and oranges spread across the market, attracting the attention of shoppers. Several people are present in the scene, browsing and shopping for the fresh produce. Some are standing close to the fruit stands, while others are scattered throughout the market area. A backpack can be seen in the background, likely belonging to one of the shoppers. The market appears to be a popular destination for people looking to purchase fresh, healthy food options.

Self-Critic

SIMA without metrics: The reason: AI generated response 1 includes additional details about a cell phone and a backpack, which are not present in the ground truth. AI generated response 2 is more accurate in describing the market as lively and filled with fresh produce, but it does not provide specific details about the people or the backpack. AI generated response 1 is more aligned with the ground truth, as it describes the market as lively and filled with people shopping for fruits and vegetables, and it provides specific details about the produce and the people present in the scene. The better AI generated response: [[1]].

SIMA with metrics : The reason: AI generated response 1 includes additional details about a cell phone and a backpack, which are not present in the ground truth. AI generated response 2 is more accurate in describing the market as lively and bustling with people, as well as the variety of fruits and vegetables available. It also provides more specific details about the produce, such as the presence of bananas, apples, and oranges. AI generated response 1 focuses more on the presence of a backpack and a cell phone, which are not relevant to the main features of the image. The better AI generated response: [[2]].

Figure 9: An example of the critic results with and without using critic metrics in the in-context self-critic stage. The results show that after using metrics, the LVLM's evaluation is more focused on the details of the visual content, leading to a correct judgment.

tion implies that the proposed research was deemed 'Not Human Subjects Research'.