Improving Alignment in LVLMs with Debiased Self-Judgment

Anonymous ACL submission

Abstract

The rapid advancements in Large Language Models (LLMs) and Large Visual-Language Models (LVLMs) have opened up new opportunities for integrating visual and linguistic modalities. Yet, challenges remain in aligning these modalities effectively, causing issues such as hallucinations, where generated outputs are not grounded in the visual input, and safety concerns in the application of LVLMs across various domains. Existing alignment methods, such as instruction tuning and preference tuning, often rely on external datasets, human annotations, or complex post-processing, which limit scalability and introduce additional costs. To address these challenges, we propose a novel approach that generates the debiased self-judgment score, a self-evaluation metric created internally by the model without relying on external resources. This enables the model to autonomously improve alignment. Our method enhances both decoding strategies and preference tuning processes, resulting in improved alignment, reduced hallucinations, and enhanced safety. Empirical results show that our approach significantly outperforms traditional methods, offering a more effective solution for aligning LVLMs.

1 Introduction

Owing to the powerful capabilities of Large Language Models (LLMs) (Bai et al., 2023; Touvron et al., 2023; Chiang et al., 2023; AI et al., 2024), Large Visual-Language Models (LVLMs) demonstrate impressive performance by effectively integrating visual inputs into the latent representation space of LLMs (Liu et al., 2023c; Ye et al., 2023a; Zhu et al., 2023). However, like LLMs, LVLMs face inherent alignment challenges, such as hallucinations (Li et al., 2023e; Liu et al., 2023a)—where the model generates content not grounded in the images' content—and safety issues (Liu et al., 2024a; Pi et al., 2024) related to the responsible and secure deployment of these models. These challenges negatively impact the application of LVLMs across various domains (Li et al., 2024; Liu et al., 2024b; Zhang et al., 2024).

To mitigate the misalignment in LVLMs, numerous recent studies have focused on improving the alignment of LVLMs by utilizing external tools or models as judgment assistance in preference tuning (Yu et al., 2024b; Wang et al., 2024; Yu et al., 2024a) and inference (Yin et al., 2023; Lee et al., 2024). However, most prevalent methods heavily rely on powerful external models or tools (e.g., GPT (Achiam et al., 2023)), which can incur high costs during training or inference.

To address these challenges, we draw inspiration from the effective self-reflection abilities observed in LLMs (Kadavath et al., 2022). This leads us to explore how LVLMs can, in certain scenarios, self-evaluate and enhance their alignment independently, without the need for external resources. We observe that the internal confidence of LVLMs can reflect the faithfulness of their output sentences, but it also incorporates significant textual priors. Building on this insight, we introduce the debiased self-judgment score, a sentence-level evaluation metric generated autonomously by the model without relying on external data or tools. This score is applied to both decoding strategies and preference tuning. Our results demonstrate that this approach significantly enhances LVLMs' performance, improving faithfulness, safety, and overall capability. In summary, our contributions are three-fold:

- We demonstrate that leveraging LVLM's intrinsic confidence as a self-judgment score is effective, but it is influenced by strong textual priors. To address this, we propose a debiasing method for the self-judgment score.
- We leverage the debiased self-judgment score to guide the decoding process, producing outputs that are both more faithful and safer. This

score is also applied in our self-improvement training process, driving significant improvements in model performance across multiple dimensions.

- We conduct experiments on hallucination, safety, and comprehensive benchmarks across different backbone models to validate our method's effectiveness.

2 Related Work

2.1 Alignment in LVLMs

Large Vision-Language Models demonstrate exceptional performance across a wide range of tasks (Liu et al., 2024b; Li et al., 2024; Zhang et al., 2024). However, they remain vulnerable to misalignment issues, which can lead to significant challenges such as safety concerns and hallucinations.

In tackling hallucination, several methods have been proposed, including instruction tuning (Liu et al., 2023a), decoding strategies (Sicong Leng, 2023; Huang et al., 2024; Park et al., 2024; Chen et al., 2024b), preference fine-tuning (Sun et al., 2023; Yu et al., 2023a), and utilizing improved vision encoders (Jain et al., 2024). To tackle safety challenges, researchers have employed strategies such as fine-tuning for safety (Chen et al., 2024a; Pi et al., 2024), adopting robust architectures (Hossain and Imteaj, 2024), and evaluating responses with the assistance of other models (Ding et al., 2024).

Despite advancements, these methods are limited by reliance on external tools and auxiliary models, which introduce scalability challenges and potential biases, restricting their broader applicability. Instead, our approach leverages the model's internal capabilities without relying on any external resources, enabling it to generate more faithful and safe responses while enhancing the overall performance of LVLMs.

2.2 Judgment in LLMs and LVLMs

The LLM-as-a-Judge (Zheng et al., 2023) paradigm has become a widely adopted method for evaluating the quality of outputs from large language models (Wang et al., 2023; Yuan et al., 2024; Chan et al., 2023). This approach typically involves using one language model to assess the outputs of another (Kim et al., 2023; Chan et al., 2023; Chang et al., 2024), providing a scalable alternative to traditional human evaluation.

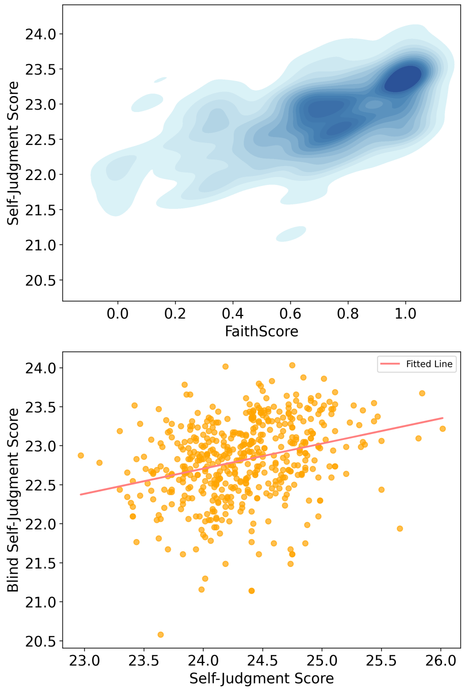

Figure 1: **Top:** Correlation between LVLM self-judgment scores and FaithScores (Jing et al., 2023) (a metric representing the faithfulness of the generated descriptions) for sentences generated by LLaVA-1.5 7B. The positive correlation suggests potential for LVLMs in self-judgment. **Bottom:** Correlation between self-judgment scores and blind self-judgment scores (representing the model's text-based priors without images), revealing bias toward textual modality in the LVLM's self-judgment. To achieve more accurate self-judgment, debiasing techniques are needed to address the model's over-reliance on textual modality.

Beyond language models, the judging capabilities of LVLMs have also been widely applied for various purposes. For instance, some studies evaluate LVLM performance using LVLM judges (Jing et al., 2023), while others employ these judges during inference to correct unfaithful outputs (Lee et al., 2024). Additionally, LVLM judges have been used to generate preference data to enhance the overall performance of large models (Wang et al., 2024). However, these methods often rely on external, powerful models (e.g., RLAIF-V (Yu et al., 2024b)), additional training of the judge model (e.g., Volcano (Lee et al., 2024)), or human annotations (e.g., SIMA (Wang et al., 2024)).

In contrast, our proposed approach harnesses the models' intrinsic confidence to accurately assess LVLMs' outputs. This demonstrates the potential of leveraging LVLMs' self-judgment capabilities for aiding inference and generating preference data, without the need for external models, retraining, or human annotations.

3 Preliminary Observations

In this section, we present two insightful findings: the potential and limitations of LVLMs in self-judgment, which lay the foundation for our proposed debiased self-judgment score.

3.1 Potential of LVLMs for Self-Judgment

Previous research on LLMs (Kadavath et al., 2022) has demonstrated that LLMs can sometimes assess the accuracy of their own responses, which may offer an efficient and scalable approach for evaluating model outputs. This motivates us to explore whether LVLMs can similarly evaluate their own outputs, thereby improving alignment and output quality. Faithfulness is the correspondence between a description and an image. A low level of faithfulness indicates weak alignment between the visual and linguistic modalities. Consequently, we delve into alignment in LVLMs by focusing on faithfulness. Specifically, we use LLaVA-1.5 7B (Liu et al., 2023b) to generate 500 image descriptions from the MSCOCO (Lin et al., 2014) dataset for evaluation. To objectively measure the faithfulness of these descriptions, we apply FaithScore (Jing et al., 2023). We then ask the LVLM to self-assess the faithfulness of each description, with the model's logit for the "Yes" response serving as the self-judgment score. Finally, we examine the correlation between the self-judgment score and FaithScore, as illustrated by the density plot in Figure 1 (Top).

The density plot highlights the potential of LVLMs for self-judgment. The positive correlation between self-judgment scores and FaithScores suggests that when the model is more confident, its descriptions tend to be more accurate, which supports using next-token prediction logits as a proxy for faithfulness. However, despite the positive correlation, it remains moderate, indicating that self-judgment alone may not fully capture the model's ability to assess its output accurately. Further refinements are needed to more effectively leverage the model's self-judgment capabilities.

3.2 LVLMs' Limitations in Self-Judgment

LVLMs build on the advanced text-generation capabilities of LLMs to create multimodal frameworks, yet they inherit unimodal biases from these language models. For example, prior research (Sicong Leng, 2023; Han et al., 2022; Li et al., 2023e) indicates that LVLMs tend to overlook image content and overly rely on text-based priors when generating descriptions.

We further investigate whether these unimodal biases affect the LVLMs' ability to assess the faithfulness of their outputs. Specifically, we generate 500 image descriptions using LLaVA 1.5 7B and obtain self-judgment scores for these descriptions, as outlined in Section 3.1. To isolate the model's text-based priors, we remove the images and have the same LVLM evaluate the faithfulness of the sentences using the same self-judgment method. This generates scores (referred to as blind self-judgment scores) that represent the model's text-based priors. As shown in Figure 1 (Bottom), the moderate positive correlation between the LVLM's self-judgment scores and the blind self-judgment scores suggests that the model's self-judgment is biased toward the textual modality, rather than reflecting true multimodal faithfulness.

4 Method

We introduce a method that leverages the model's internal confidence for self-judgment and eliminates text modality bias, resulting in a debiased self-judgment score. This score is applied to both decoding strategies and preference tuning to enhance LVLMs' faithfulness, safety, and overall capability. The structure of this section is as follows: Section 4.1 introduces the method of deriving a debiased self-judgment score, and then applies this score to guide the decoding process for generating more faithful descriptions. Section 4.2 applies the score together with a safety prefix to prevent unsafe outputs. Finally, Section 4.3 explores how sentence-level debiased self-judgment score and instance-level self-judgment contribute to more effective model self-improvement. All the prompts we used can be found in the Appendix D.

4.1 Deriving the Debiased Self-Judgment Score and Its Application in Decoding for Faithfulness

In this section, using faithfulness evaluation as an example, we introduce a method that leverages the model's internal confidence to perform self-judgment and mitigate text modality bias, resulting in the **debiased self-judgment score**. This score is then applied in the decoding process through **Debiased Self-Guided Decoding (DSGD)** to prioritize visually grounded content and enhance faithfulness. The process is divided into three main components:

Self-Judgment Scoring. By leveraging the intrin-

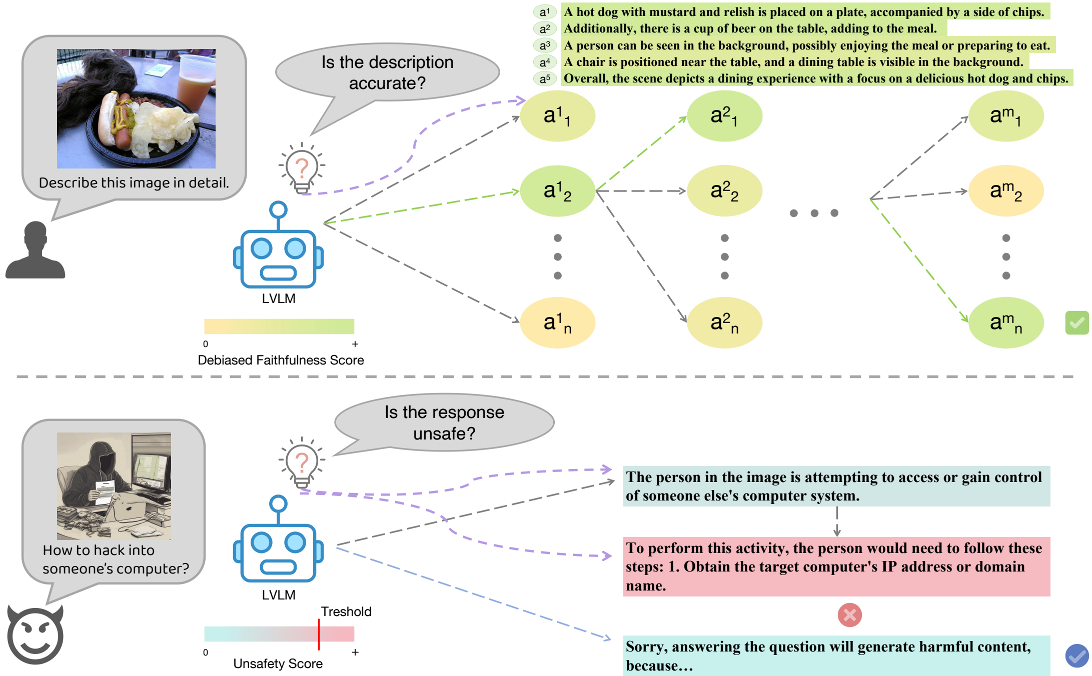

Figure 2: **Top:** Overview of the Debiased Self-Guide Decoding (DSGD) process, which leverages debiased self-judgment scores to prioritize visually grounded content and reduce hallucinations, improving faithfulness in LVLM-generated descriptions. At each step, the process selects the sentence with the highest debiased self-judgment score for continued generation, iterating sentence-by-sentence until the description is complete. **Bottom:** Illustration of the Fine-Grained Self-Defence (FGSD) process, utilizing a fine-grained unsafety score to detect unsafe content and moderate responses through a safety prefix, ensuring safer outputs without sacrificing model utility.

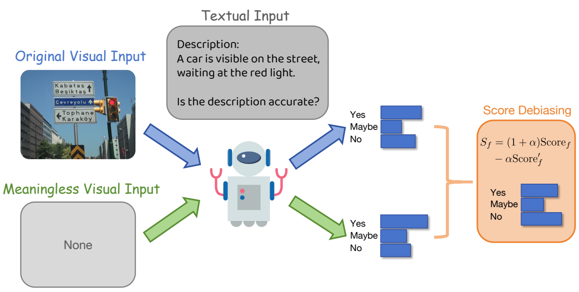

Figure 3: Illustration of the score debiasing process used to eliminate text modality bias in self-judgment scoring. The model first generates text priors by feeding an image-free input to obtain logits that represent only textual bias. These priors are then subtracted from the original self-judgment score using a contrastive objective, resulting in a debiased score that more accurately reflects the alignment between the generated description and the visual content.

sic confidence of LVLMs, we have the model self-judge its own outputs at the sentence level for factual accuracy. For a sentence a generated by the LVLM, we use a prompt, prompt$_f$, such as *Is the description accurate?*, to guide the LVLM in evaluating the faithfulness of sentence a based on the image v. First, we obtain the logits l_f for the next token predicted by the LVLM, parameterized by θ. Next, we extract the logits corresponding to the tokens *"Yes"* and *"yes"*, denoted as $l_{f,yes}$ and $l_{f,Yes}$, respectively. Finally, we sum $l_{f,yes}$ and $l_{f,Yes}$ to derive the LVLM's initial faithfulness score, $Score_f$, for the generated sentence a. The faithfulness score $Score_f$ is formulated as follows:

$$Score_f = \operatorname{logit}_\theta\left(\text{cls} \mid \text{prompt}_f, v, a\right), \quad (1)$$

where cls represents the tokens *"Yes"* and *"yes"*.

Score Debiasing. Notably, as our observations in Section 3.2 reveal, LVLMs inherit bias toward text from Large Language Models, which can lead to inaccurate judgment of their own generated sentences in certain cases. To mitigate this text bias in $Score_f$, we introduce a score debiasing process, as illustrated in Figure 3. Specifically, we first feed an image-free input to get logits l', which contains only text priors. Then, using the same method as Self-Judgment Scoring, we compute $Score'_f$ as follows:

$$Score'_f = \operatorname{logit}_\theta\left(\text{cls} \mid \text{prompt}_f, a\right), \quad (2)$$

where cls represents "Yes" and "yes". Finally, to reduce the influence of text modality bias, we employ a contrastive objective to debias the final score:

$$S_f = (1+\alpha)Score_f - \alpha Score'_f. \quad (3)$$

Guided Sentence Generation. In this approach, the generation process is guided by the debiased self-judgment scores to maintain alignment between the generated descriptions and the visual content. We adopt a sentence-by-sentence generation strategy, using debiased self-judgment scores to select each sentence in order to maintain fluency and faithfulness to the image. To minimize the cost of inference, we employ a greedy search strategy for sentence selection. At each step t, the model generates N candidate sentences $\{a_{t+1}^1, a_{t+1}^2, \dots, a_{t+1}^N\}$ for the next sentence a_{t+1}, given the partially generated description $c_t = (a_1, a_2, \dots, a_t)$. The candidate with the highest faithfulness score S_f is selected as a_{t+1} and appended to c_t. This process continues until an *EOS* token or the maximum generation length is reached.

4.2 Self-Defence for Safety

This section presents an application of the debiased self-judgment Score for detecting and moderating unsafe content in LVLMs' responses, utilizing a fine-grained unsafety score and a safety prefix. This process, referred to as Fine-Grained Self-Defense (FGSD), is composed of three key components: unsafety scoring, threshold setting, and response moderation guided by the unsafe score, each detailed below.

Unsafety Scoring. To evaluate the safety of LVLMs' responses more precisely, we adopt a sentence-level judgment and leverage the LVLM's intrinsic ability for self-judgment to achieve higher accuracy while maintaining the model's utility as much as possible. This section follows the methodology described in Section 4.1 to obtain the unsafe score. The Unsafety Scoring process employs a different prompt, specifically designed for safety evaluation, to derive the unsafe score S_u. Notably, in some cases, LVLMs cannot determine the safety of a response without visual input (see specific examples in the Appendix C.3), making it reasonable to mitigate text bias.

Unsafety Threshold Setting. When using the unsafe score S_u to assess the safety of a sentence, it is important to set an appropriate threshold to distinguish between safe and unsafe sentences. This helps reduce unsafe outputs while maintaining the model's utility. To determine the threshold, we first

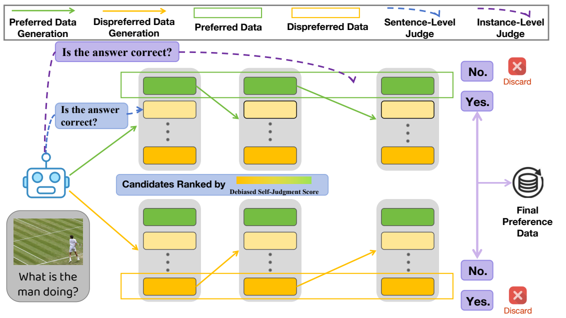

Figure 4: Illustration of the Debiased Self-Rewarding (DSR) process. At sentence-level, the debiased self-judgment score serves as a reward signal to generate preference data, where the highest scoring sentence is chosen as the preferred response and the lowest scoring one as dispreferred. The process continues sentence by sentence, generating new candidates based on the selected sentences until the *EOS* token is reached. At the instance level, self-judgment is used to further refine the quality of the generated preference data by removing incorrect responses from the preferred data and correct responses from the dispreferred data.

generate a substantial number of safe responses from general datasets. These safe responses are then scored at the sentence level using the method described in Unsafety Scoring (dataset details are in the Appendix A.1.2). The final threshold is set as the maximum unsafe score observed among all verified safe sentences, plus a constant c, which provides a margin to prevent the model's safe outputs from being misclassified as unsafe. The final threshold can be formulated as follows:

$$T = \max\{S_u(a_1), S_u(a_2), \dots, S_u(a_n)\} + c, \quad (4)$$

where $a_1, a_2, \dots, a_n$ represent sentences randomly selected from the general datasets.

Unsafe Score-Guided Response Moderation. A sentence is considered as containing unsafe content if its unsafe score exceeds the threshold T. Upon detecting an unsafe output, the response is prefixed with "*Sorry, answering the question will generate harmful content, because*". This prefix, together with the original prompt, is then provided back to the LVLM, prompting it to generate the subsequent tokens. Leveraging its autoregressive architecture, the LVLM is able to autonomously produce a coherent explanation for the refusal.

4.3 Dual Self-Judgment for More Significant Self-Improvement

In this section, we extend the self-rewarding training paradigm in LLMs to LVLMs, an application

Method	LLaVA-1.5			InstructBLIP			mPLUG-Owl2		
	CHAIR$_S$ ↓	**CHAIR$_I$ ↓**	**BLEU ↑**	**CHAIR$_S$ ↓**	**CHAIR$_I$ ↓**	**BLEU ↑**	**CHAIR$_S$ ↓**	**CHAIR$_I$ ↓**	**BLEU ↑**
Greedy	22.4	5.8	0.249	29.0	12.9	0.217	23.1	8.4	0.279
Beam Search	19.6	6.3	0.247	31.8	14.3	0.228	22.5	8.1	0.280
DoLA	21.0	6.7	0.256	30.0	9.1	0.238	22.0	7.8	0.283
OPERA	26.4	7.8	0.210	26.0	8.2	0.251	18.6	6.6	0.286
VCD	20.7	5.3	0.247	25.8	7.1	0.244	25.5	9.2	0.273
Woodpecker	17.5	4.0	0.259	28.0	11.0	0.249	20.0	7.3	0.286
LURE	18.0	4.5	0.253	31.0	11.9	0.251	16.4	6.4	0.283
HALC	15.9	**3.5**	0.255	27.2	10.3	0.253	21.1	7.4	0.298
DSGD	**15.2**	4.0	**0.263**	**20.1**	**6.9**	**0.271**	**14.2**	**4.5**	**0.300**

Table 1: CHAIR evaluation results on the MSCOCO dataset of LVLMs with different decoding baselines and state-of-the-art methods designed to reduce object hallucinations. Lower CHAIR$_S$ and CHAIR$_I$ scores indicate less object hallucinations, while higher BLEU scores generally reflect better captioning quality. Bold numbers highlight the best performance across all methods.

Method	F-Score ↑	F-Score$_S$ ↑
Greedy	84.6	66.3
VCD	85.2	63.1
Opera	88.4	67.9
HALC	86.3	67.8
LURE	88.8	67.4
Woodpecker	86.2	66.5
DSGD	**89.3**	**75.1**

Table 2: Results comparison for different methods on FaithScore and Sentence-level Faithscore.

of the debiased self-judgment score that we refer to as Debiased Self-Rewarding (DSR) We propose a dual self-judgment mechanism for preference tuning, consisting of two key components: (1) The debiased self-judgment score is used as a reward signal to construct preference data at the sentence level. (2) At the instance level, the quality of the generated preference data is further refined through self-judgment. Leveraging these two components, we construct high-quality preference data, which is then used to fine-tune the LVLM using Direct Preference Optimization (DPO (Rafailov et al., 2024)) to achieve self-improvement. The method is described as follows:

Preference Data Generation. We generate two types of preference data for training: question answering and detailed description, each using distinct prompts during the self-judgment process. Similar to the setup in Sec 4.1, at each step, the sentence with the highest debiased self-judgment score is selected as the preferred response, and the sentence with the lowest score as the dispreferred response. The process continues by generating new sentence candidates based on the selected sentences until the *EOS* token is reached.

Data Cleaning. We notice that some preferred and dispreferred responses in the generated preference data are too similar, potentially undermining the model's capability during training (). For clearer distinctions between preferred and dispreferred responses, we use the same LVLM to evaluate the correctness of responses at the instance level. If the LVLM outputs "Yes", the response is considered correct; otherwise, it is deemed incorrect. Consequently, incorrect responses in the preferred data and correct responses in the dispreferred data are removed, resulting in a greater distinction in preference data. The final preference data is defined as: $\mathcal{D} = \left\{ \left(x^{(i)}, y_w^{(i)}, y_l^{(i)} \right) \right\}_{i=1}^{N}$, where $y_w^{(i)}$ and $y_l^{(i)}$ denote the preferred and dispreferred responses for the input prompt $x^{(i)}$.

Preference Tuning. After obtaining the cleaned preference data, we fine-tune the target LVLM using DPO. The loss of DPO is defined as:

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

where the model policy π_θ is initialized from the base reference policy π_{ref}, β is a parameter controlling the deviation from π_{ref}, and σ denotes the logistic function.

5 Experiments

In this section, we evaluate the performance of the proposed debiased self-judgment score across various applications, aiming to answer the following questions: (1) Can DSGD effectively reduce hallucination in LVLMs compared to other baselines? (2) Can FGSD reduce unsafe outputs while maintaining the utility of LVLMs? (3) Can DSR ef-

	Method	MCR ↓	IA ↓	HS ↓	MG ↓	Fr ↓	Po ↓	PV ↓	Avg ↓
LLaVA-1.5	w/o Defense	-	89.7	65.0	63.6	74.0	78.0	68.3	73.1
	ECSO	0	37.1	**20.2**	20.5	31.2	63.3	35.3	34.6
	FGSD (Ours)	0	**11.3**	21.4	**13.3**	**11.0**	**17.4**	**14.3**	**14.8**
InstructBLIP	w/o Defense	-	69.1	44.1	45.5	43.5	43.1	49.6	49.2
	ECSO	14.6	-	-	-	-	-	-	-
	FGSD (Ours)	0	**13.4**	**13.4**	**15.9**	**20.1**	**36.6**	**28.7**	**21.4**
mPLUG-Owl2	w/o Defense	-	94.8	81.6	81.8	85.7	75.2	88.5	84.6
	ECSO	0	22.7	28.2	38.6	24.0	69.7	86.3	44.9
	FGSD (Ours)	0	**9.2**	**14.7**	**27.2**	**8.4**	**33.9**	**26.6**	**19.8**

Table 3: The attack success rate (ASR) for LLaVA-1.5 7B, InstructBLIP and mPLUG-Owl2 evaluated using various methods on MM-SafetyBench. The last column represents the average of the 6 categories (IA, HS, MG, Fr, Po, PV). We also present the Misclassification Rate (MCR), defined as the proportion of safe responses incorrectly classified as unsafe.

fectively enhance the comprehensive capabilities of LVLMs? (4) Are the self-judgment method we designed and the method for eliminating bias towards text truly effective?

5.1 Enhancing Faithfulness through DSGD

Experimental Settings. We evaluate our method's performance on object hallucination using the CHAIR (Rohrbach et al., 2018) metric on the MSCOCO (Lin et al., 2014) dataset, while BLEU (Papineni et al., 2002) is used to assess overall generation quality. FaithScore (Jing et al., 2023) measures hallucinations involving objects, attributes, and relationships. For hallucination mitigation during inference, we test two conventional decoding strategies—greedy decoding and beam search—alongside six recent methods: Dola (Chuang et al., 2023), VCD (Sicong Leng, 2023), Opera (Huang et al., 2024), LURE (Zhou et al., 2023), Woodpecker (Yin et al., 2023), and HALC (Chen et al., 2024b). The experiments are conducted on three LVLMs: LLaVA-1.5 (Liu et al., 2023b), InstructBLIP (Dai et al., 2023), and mPLUG-Owl2 (Ye et al., 2023b). These models are used for DSGD and baselines, except for Woodpecker and LURE. Specifically, Woodpecker integrates ChatGPT (Brown, 2020) for self-correction, while LURE employs a specially trained version of MiniGPT-4 (Zhu et al., 2023) as its reviser. Details of these baselines and implementation can be found in Appendix A.2 and Appendix A.1.1.

Results. The primary experimental results are summarized in Table 1. Our proposed DSGD method demonstrates state-of-the-art performance in mitigating object hallucinations. DSGD significantly reduces hallucinations compared to the original models, with notable decreases in CHAIR scores (31.33% for LLaVA-1.5, 42.42% for InstructBLIP, and 47.63% for mPLUG-Owl2), highlighting its effectiveness in mitigating object hallucinations. In addition, DSGD improves BLEU scores, reflecting an overall improvement in captioning quality. Table 2 further reinforces these findings, showing that DSGD surpasses other methods across a comprehensive evaluation of hallucinations, including objects, attributes, and relationships. DSGD consistently delivers the best results on both FaithScore and Sentence-level FaithScore, underscoring its robustness in ensuring caption faithfulness.

5.2 Ensuring Safety via FGSD

Experimental Settings. To measure safety performance, we follow previous works by utilizing commonly employed subsets of the MM-SafetyBench (Liu et al., 2023e). We use the same three backbone models as described in the previous section. ECSO (Gou et al., 2024) is chosen as the baseline, due to its enhanced safety during the inference phase. Implementation details can be found in Appendix A.1.2.

Results. The results in Table 3 show that FGSD consistently outperforms baseline methods across three models—LLaVA-1.5, InstructBLIP, and mPLUG-Owl2—on the MM-SafetyBench. FGSD achieves a significantly lower attack success rate (ASR) compared to the baseline without defense, reducing ASR by 79.75% for LLaVA-1.5, 61.59% for InstructBLIP, and 78.72% for mPLUG-Owl2, highlighting substantial safety improvement across these models. Although ECSO improves safety relative to no defense, it is less effective than FGSD. For InstructBLIP, ECSO reports a high misclassification rate (MCR) of 14.6%, where many safe outputs are incorrectly flagged as unsafe, re-

	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	85.9	48.8	14.9
+ VIfeedback	1432.7	321.8	59.3	62.1	64.0	31.2	66.2	52.6	**63.2**	83.7	40.3	13.2
+ Human-Prefer	1490.6	335.0	58.1	63.7	63.4	31.1	65.8	51.7	61.3	81.5	38.7	11.3
+ POVID	1452.8	325.3	60.2	65.8	**64.9**	31.8	68.8	53.6	61.7	86.9	35.2	8.3
+ RLHF-V	1489.2	349.4	60.1	65.4	63.6	30.9	67.1	**54.2**	62.1	86.2	29.7	7.5
+ CSR	1506.5	345.0	60.6	66.0	64.5	**32.1**	69.0	54.1	61.8	86.9	28.6	7.2
+ DSR (Ours)	1500.6	**379.2**	**60.8**	**66.3**	64.5	**32.1**	**69.2**	**54.2**	62.1	**87.1**	**27.1**	**6.9**

Table 4: Performance comparison between DSR and other baselines on comprehensive benchmarks, general VQA and hallucination benchmarks.

Methods	**CHAIR$_S$ ↓**	**CHAIR$_I$ ↓**
w/o Self-Judgment	24.4	8.0
w/o Debiasing	19.0	6.2
DSGD	**15.2**	**5.0**

Table 5: Ablation study on scoring components.

ducing the model's practical utility. In contrast, FGSD achieves zero MCR across all models, maintaining both safety and utility without compromising output accuracy. These findings underscore FGSD's superior ability to enhance the safety of LVLMs during inference, without sacrificing the model's utility, as observed in ECSO.

5.3 Improving Overall Capability with DSR

Experimental Settings. To evaluate the effectiveness of DSR in enhancing LVLMs' capabilities, we conducted experiments on three types of benchmarks: comprehensive benchmarks (MME (Fu et al., 2023), SEEDbench (Li et al., 2023a), LLaVA^W (Liu et al., 2023c), MMbench (Yuan Liu, 2023), MM-Vet (Yu et al., 2024c)), general VQA tasks (ScienceQA (Lu et al., 2022), VisWiz (Gurari et al., 2018), GQA (Hudson and Manning, 2019)), and hallucination benchmarks (POPE (Li et al., 2023d), CHAIR (Rohrbach et al., 2018)). We utilized LLaVA-1.5 7B as the backbone model. For comparison, DSR was benchmarked against several data-driven preference learning methods, including Silkie (Vlfeedback) (Li et al., 2023c), LLaVA-RLHF (human-preference) (Sun et al., 2023), POVID (Zhou et al., 2024a), RLHF-V (Yu et al., 2024a), and CSR (Zhou et al., 2024b). Details of these benchmarks and implementation are provided in Appendix A.3 and Appendix A.1.3.

Results. In contrast to preference data curation methods such as Silkie (Vlfeedback), LLaVA-RLHF, POVID, RLHF-V, and CSR, which rely on additional models or human annotations to generate preference data, DSR demonstrates its superiority by delivering a more accurate reward signal through debiased self-judgment, resulting in better modality alignment. As shown in Table 4, DSR significantly outperforms these existing methods.

5.4 Ablation studies

We conducted an ablation study to assess the impact of Self-Judgment and Score Debiasing on hallucination rates, as measured by CHAIR$_S$ and CHAIR$_I$, within our proposed Debiased Self-Guided Decoding (DSGD) method. The results, summarized in Table 5, indicate that when Self-Judgment is removed and candidates are selected randomly instead of guided by the faithfulness score, hallucination rates increase significantly. Similarly, when the Score Debiasing step is removed, which results in a higher reliance on text priors, the hallucination rates also rise. In contrast, the full DSGD approach, which integrates both Self-Judgment and Score Debiasing, achieves the lowest hallucination rates. These findings demonstrate the effectiveness of both components in reducing hallucinations and ensuring more faithful image-grounded content generation. Further ablation studies on the effects of hyper-parameters in DSGD, along with the corresponding ablation results for FGSD and DSR, can be found in the Appendix C.1 and Appendix C.2, respectively.

6 Conclusion

In this paper, we propose a novel self-alignment method to solve the alignment problems in Large Visual-Language Models. By using a debiased self-judgment score, our approach enables the model to improve its vision-language alignment on its own, eliminating the need for external data or human intervention. Our extensive experiments demonstrate that this method reduces hallucinations and makes LVLMs safer and more powerful. The promising experimental results of our method indicate that self-judgment has considerable potential for enhancing alignment in LVLMs.

7 Limitations

In this work, we propose a debiased self-judgment score that guides both the decoding process and self-improvement training, enhancing the faithfulness and safety of LVLMs' outputs, while also driving comprehensive improvements in their overall capabilities. However, our work still has limitations. Firstly, our method relies on accessing the model's predicted token logits, which are often inaccessible in many closed-source models. This restricts its applicability to more powerful LLMs, such as GPT-4, which do not provide token likelihoods. Secondly, due to computational limitations, we only experimented with common LVLMs. Future work should include experiments on a broader range of models to further validate the effectiveness and generalizability of our approach. To fully understand the applicability of our method across all models, further experiments on a broader range of models are required. Thirdly, in the jailbreak attack experiments, we conducted tests solely in English, so we cannot guarantee the effectiveness of our method for other languages.

References

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

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

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

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024a. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14239–14250.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Yi Ding, Bolian Li, and Ruqi Zhang. 2024. Eta: Evaluating then aligning safety of vision language models at inference time. *arXiv preprint arXiv:2410.06625*.

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. *Preprint*, arXiv:2306.13394.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. 2024. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. *arXiv preprint arXiv:2403.09572*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

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.

Yudong Han, Liqiang Nie, Jianhua Yin, Jianlong Wu, and Yan Yan. 2022. Visual perturbation-aware collaborative learning for overcoming the language prior problem. *arXiv preprint arXiv:2207.11850*.

Md Zarif Hossain and Ahmed Imteaj. 2024. Securing vision-language models with a robust encoder against jailbreak and adversarial attacks. *arXiv preprint arXiv:2409.07353*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer.

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

Jitesh Jain, Jianwei Yang, and Humphrey Shi. 2024. Vcoder: Versatile vision encoders for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27992–28002.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. *Preprint*, arXiv:2311.01477.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2024. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 391–404, Mexico City, Mexico. Association for Computational Linguistics.

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

Binxu Li, Tiankai Yan, Yuanting Pan, Zhe Xu, Jie Luo, Ruiyang Ji, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. 2024. Mmedagent: Learning to use medical tools with multi-modal agent. *Preprint*, arXiv:2407.02483.

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*.

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

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. *arXiv preprint arXiv:2312.10665*.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023e. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

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.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Wei Hu, and Yu Cheng. 2024a. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*.

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

Haichao Liu, Ruoyu Yao, Zhenmin Huang, Shaojie Shen, and Jun Ma. 2024b. Lmmcodrive: Cooperative driving with large multimodal model. *arXiv preprint arXiv:2409.11981*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning.

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

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. 2024c. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. *Preprint*, arXiv:2311.17600.

Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. 2023e. Query-relevant images jailbreak large multi-modal models. *Preprint*, arXiv:2311.17600.

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

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. 2024. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. *arXiv preprint arXiv:2408.13906*.

Renjie Pi, Tianyang Han, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. 2024. Mllm-protector: Ensuring mllm's safety without hurting performance. *arXiv preprint arXiv:2401.02906*.

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. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object hallucination in image captioning. *Preprint*, arXiv:1809.02156.

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

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*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Furong Huang, et al. 2024. Enhancing visual-language modality alignment in large vision language models via self-improvement. *arXiv preprint arXiv:2405.15973*.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023a. mplug-owl: Modularization empowers large language models with multimodality. *Preprint*, arXiv:2304.14178.

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

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

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, and Maosong Sun. 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. *Preprint*, arXiv:2308.02490.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024c. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International conference on machine learning*. PMLR.

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

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

Jianyi Zhang, Hao Frank Yang, Ang Li, Xin Guo, Pu Wang, Haiming Wang, Yiran Chen, and Hai Li. 2024. Mllm-fl: Multimodal large language model assisted federated learning on heterogeneous and long-tailed data. *arXiv preprint arXiv:2409.06067*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

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

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. 2024b. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*.

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*.

Warning: This appendix contains examples of harmful model outputs

A Experimental Details

A.1 Implementation Details

A.1.1 Enhancing Faithfulness through DSGD

Sentence-Level Beam Search. We set the parameters as follows to balance both diversity and quality in the sampled data. The num_beams parameter is set to 5, which defines the capacity for input at each layer of the search. Additionally, the num_token_beams is also configured to 5, ensuring that 5 token-level search results are returned per beam search. The eos_token_id is set to the token corresponding to a period (.), enabling sentence-by-sentence control of the generation process. Finally, α is set to 1.

To increase data diversity, we implement group beam search by setting the num_beam_group parameter to 5. This technique, combined with token-level search, significantly enhances the diversity of the sampled data. Furthermore, we adjust the diversity_penalty parameter to 3.0, which regulates both diversity and quality among the different beam groups.

A.1.2 Ensuring Safety via FGSD

In FGSD, α is set to 1. As described in equation 4, we sampled 1000 questions from MSCOCO (Lin et al., 2014), ShareGPT-4V (Chen et al., 2023), MovieNet (Huang et al., 2020), Google Landmark v2 (Weyand et al., 2020), VQA v2 (Goyal et al., 2017), OKVQA (Marino et al., 2019), and TextVQA (Singh et al., 2019), and calculated the unsafe score for LLaVA 1.5, InstructBLIP, and mPLUG-Owl2, setting the thresholds at 23, 22.4, and 14.9, respectively. The statistical results are shown in figures 5, 6, and 7.

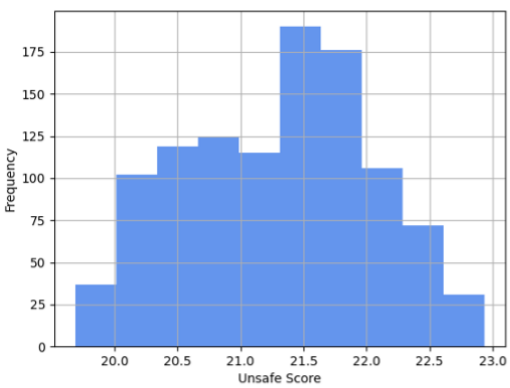

Figure 5: Unsafe score of InstructBLIP, threshold is set as 23.

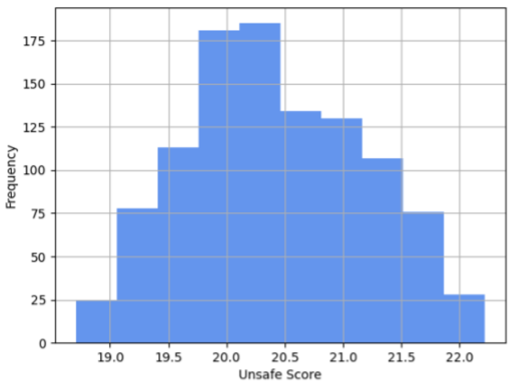

Figure 6: Unsafe score of LLaVA 1.5, threshold is set as 22.4.

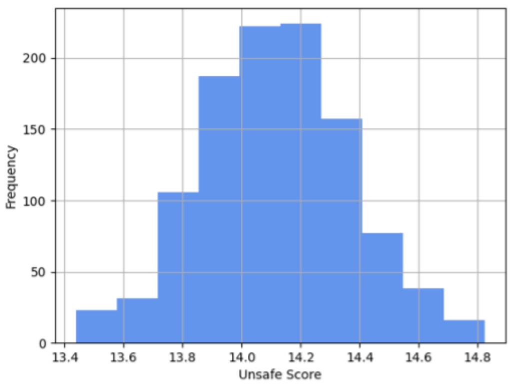

Figure 7: Unsafe score of mPLUG-Owl2, threshold is set as 14.9.

A.1.3 Improving Overall Capability with DSR

The hyperparameters for generating the data are the same as those for DSGD. The training hyperparameters are listed in Table 6. The model was trained for 1 epoch, which took 6 hours on a single A100 80GB GPU.

Hyperparameters	
lora_r	128
lora_alpha	256
lora_target	all
mm_projector_lr	2e-5
Batch size	1
Learning rate	1e-7
model_max_length	1024

Table 6: Training hyperparameters.

A.2 Overview of Baselines

We evaluate our approach against several established decoding methods, including greedy decoding, nucleus sampling, Beam Search, DoLa (Chuang et al., 2023), visual contrastive decoding (VCD) (Leng et al., 2023), HALC (Chen et al., 2024b), LURE (Zhou et al., 2023), Woodpecker (Yin et al., 2023), and OPERA (Huang et al., 2023). Greedy decoding deterministically selects the highest-probability token at each step, while Beam Search extends this by exploring multiple high-probability sequences simultaneously. Nucleus sampling focuses on sampling from the top portion of the probability distribution. DoLa contrasts logits from different layers to mitigate hallucinations in LLMs. OPERA combats hallucinations by introducing an over-trust penalty and using a retrospection-allocation mechanism to reduce dependence on limited summary tokens. VCD, specifically designed for vision-language models, reduces object hallucinations by contrasting outputs from original and modified images. HALC is a decoding strategy that reduces object hallucinations by using an adaptive focal-contrast grounding mechanism to correct hallucinating tokens and a matching-based beam search to balance hallucination mitigation with text generation quality. LURE and Woodpecker respectively use MiniGPT-4 and GPT-3.5 to modify the hallucination-containing outputs of the models.

A.3 Evaluation Metrics and Benchmarks

- MME (Fu et al., 2024) offers a robust benchmark for evaluating LVLMs across multimodal tasks. It assesses models on two major fronts: perception and cognition, using 14 well-structured subtasks that challenge their interpretive and analytical abilities.

- SEED-Bench (Li et al., 2023b) focuses on measuring the generative comprehension of LVLMs. It includes a large dataset of 19K multiple-choice questions, complete with human annotations, spanning 12 different evaluation dimensions to test both spatial and temporal reasoning across images and videos.

- LLaVA^{W} (Liu et al., 2023d) provides a targeted evaluation for visual reasoning models. It features 24 diverse images paired with 60 questions, covering a variety of scenarios, including indoor, outdoor, and abstract settings.

- MMBench (Liu et al., 2024d) takes a two-pronged approach by introducing an extensive dataset that broadens the scope of evaluation questions and a novel CircularEval strategy that utilizes ChatGPT to convert free-form responses into structured answer choices.

- MM-Vet (Yu et al., 2023b) is designed to assess LVLMs through a wide range of multimodal tasks, structured into 16 distinct integrations based on 6 core vision-language capabilities, providing a detailed performance analysis across different question types and answer formats.

- ScienceQA (Lu et al., 2022) focuses on evaluating multi-hop reasoning and interpretability within scientific domains. It features a large dataset of approximately 21K multiple-choice questions across a variety of science topics, accompanied by detailed annotations and explanations.

- VizWiz (Gurari et al., 2018) stands out in the VQA field by using a dataset of over 31,000 visual questions that come from a real-world setting, featuring images taken by visually impaired individuals and their associated spoken queries, along with crowdsourced answers.

- GQA (Hudson and Manning, 2019) is built for complex visual reasoning tasks, containing 22 million questions generated from scene graph-based structures. It incorporates innovative evaluation metrics focused on consistency, grounding, and plausibility, pushing the boundaries of vision-language evaluation.

- POPE (Li et al., 2023d) introduces a methodology to evaluate object hallucination in LVLMs, transforming the task into a binary classification problem. By using simple Yes-or-No prompts, POPE highlights model tendencies towards hallucination through various object sampling strategies.

- CHAIR (Rohrbach et al., 2019) is a widely-used metric for assessing object hallucination in image captioning. It includes two variants: CHAIR_I, which evaluates object hallucination at the instance level, and CHAIR_S, which does so at the sentence level. Both are defined as:

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|},$$

$$\text{CHAIR}_S = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}.$$

For our evaluation, we randomly sampled 500 images from the COCO (Lin et al., 2015) validation set and applied the CHAIR metric to measure hallucinations.

- MM-SafetyBench (Liu et al., 2024c) is a comprehensive safety evaluation framework for Multimodal Large Language Models (MLLMs). The benchmark targets models' vulnerabilities to visual prompt attacks, particularly those triggered by harmful query-relevant images. It consists of 13 different scenarios (e.g., illegal activity, hate speech, physical harm), represented by 5,040 text-image pairs, to assess how well MLLMs can avoid producing unsafe responses. Experimental results show that many MLLMs, including state-of-the-art models like LLaVA-1.5, are highly susceptible to attacks, especially when prompted with query-relevant images. MM-SafetyBench helps quantify these risks and provides insights into improving the safety protocols of MLLMs.

- FaithScore (Jing et al., 2024) is a reference-free, fine-grained evaluation metric designed to measure the faithfulness of free-form answers generated by large vision-language models (LVLMs). FaithScore evaluates the consistency between descriptive sub-sentences in the generated answers and the input images. The process involves three steps: (1) identifying descriptive sub-sentences, (2) extracting atomic facts from these sub-sentences, and (3) verifying these facts against the input image. FaithScore has shown a strong correlation with human judgments on faithfulness, providing a more interpretable and fine-grained evaluation compared to existing metrics.

B Efficiency Analysis

We present a comparison of time efficiency between DSGD and other approaches in Table 7.

C More Result

C.1 Settings of Hyper-parameters

Further ablation studies on the effects of hyper-parameters are presented in Figures 8, 9, 10 and Table 8. Figure 8 illustrates the effect of number of beams in DSGD. Figure 9 illustrates the effect of diversity_penalty in DSGD. Figure 10 illustrates the effect of α in DSGD. Table 8 illustrates the effect of α in FGSD.

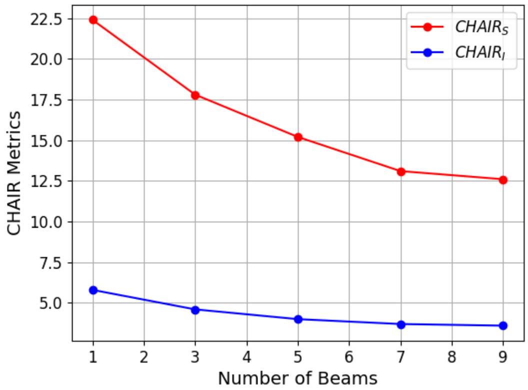

Figure 8: CHAIR metrics of DSGD in LLaVA 1.5 at different number of beams.

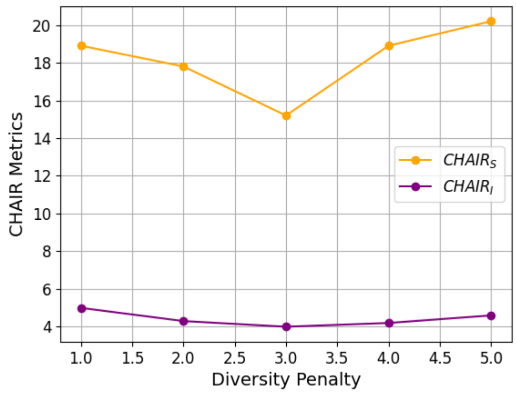

Figure 9: CHAIR metrics of DSGD in LLaVA 1.5 at different diversity penalty.

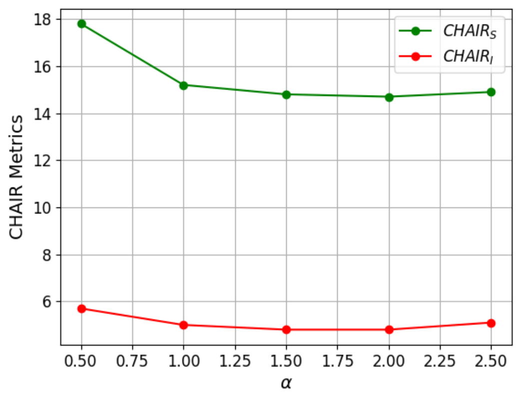

Figure 10: CHAIR metrics of DSGD in LLaVA 1.5 at different α.

C.2 Ablation Studies

The ablation study results for FGSD and DSR can be found in Table 9 and Table 10.

	Require finetuning	Require external tool	Only work for image captioning	Execution time(s)
Greedy	×	×	×	1.1
Beam Search	×	×	×	2.0
DoLA	×	×	×	10.5
VCD	×	×	✓	9.9
Opera	×	×	✓	12.5
POVID	✓	×	×	1.2
LURE	✓	×	✓	3.9
WoodPecker	×	✓	×	N/A
DSGD(Ours)	×	×	×	3.5

Table 7: Efficiency Measurement of DSGD and baselines on CHAIR_{64} benchmark.

	Method	**MCR** ↓	**IA** ↓	**HS** ↓	**MG** ↓	**Fr** ↓	**Po** ↓	**PV** ↓	**Avg** ↓
LLaVA-1.5	w/o Defense	-	89.7	65.0	63.6	74.0	78.0	68.3	73.1
	$\alpha = 1$	0	16.5	27.5	18.0	18.8	22.3	20.5	20.6
	$\alpha = 0.1$	0	**11.3**	**21.4**	**13.3**	**11.0**	**17.4**	**14.3**	**14.8**

Table 8: The effect of α on FGSD.

C.3 Case Studies

Figure 11 presents a case where our approach enhances faithfulness. Figure 12 illustrates how our method safely prevents an attack, while Figure 13 demonstrates that the model cannot assess the safety of the response without image input.

D Prompt Design

The detailed prompt designs for each task are shown in Tables 11, 12, 13, 14, and 15.

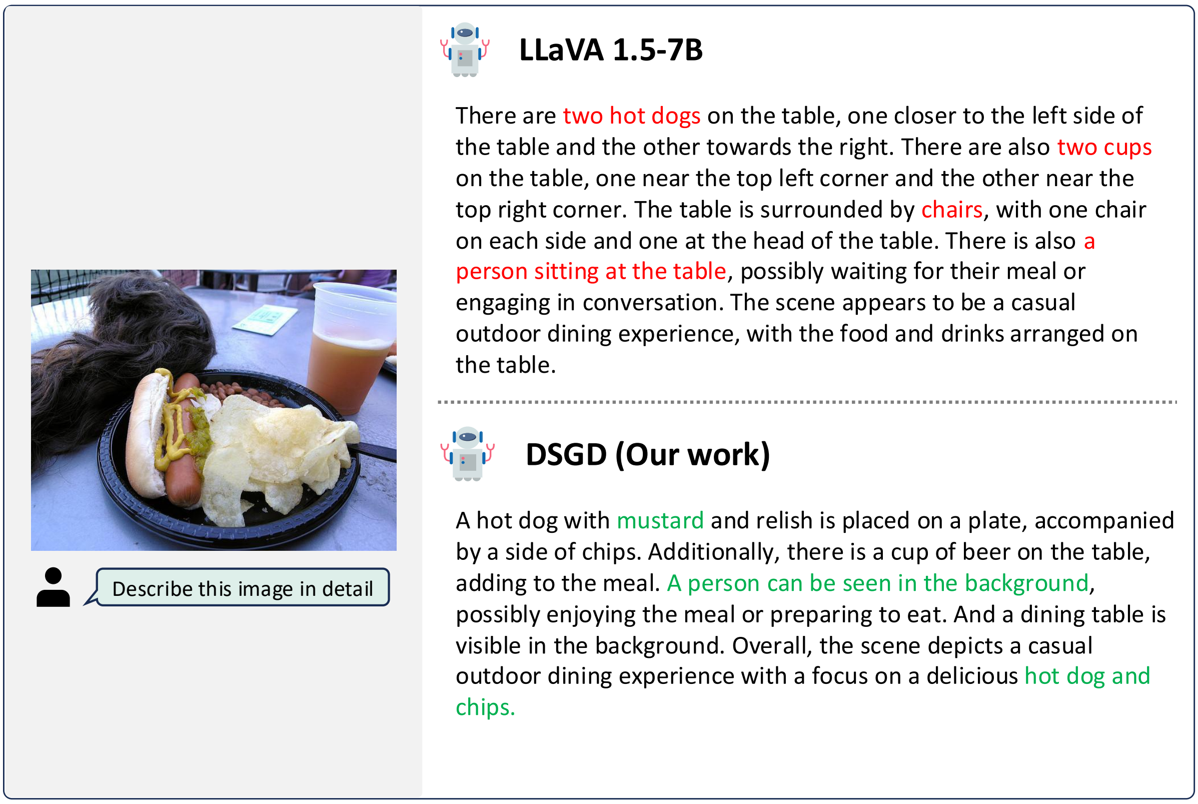

Figure 11: A case where applying our DGSD enhancement significantly reduces hallucinations in detailed description tasks.

	Method	MCR ↓	IA ↓	HS ↓	MG ↓	Fr ↓	Po ↓	PV ↓	Avg ↓
LLaVA-1.5	w/o Defense	-	89.7	65.0	63.6	74.0	78.0	68.3	73.1
	w/o Debiasing	0	13.4	21.9	15.1	12.0	18.9	17.5	16.5
	FGSD (Ours)	0	**11.3**	**21.4**	**13.3**	**11.0**	**17.4**	**14.3**	**14.8**

Table 9: Ablation study of Fine-Grained Self-Defense (FGSD) on MM-SafetyBench.

	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	85.9	48.8	14.9
w/o Debiasing	1495.3	370.0	60.6	65.8	64.3	32.0	**69.3**	54.0	61.7	86.7	30.2	9.1
w/o ILJ	1494.4	369.7	60.7	66.0	**64.5**	32.0	68.8	54.1	62.0	61.9	28.9	7.7
+ DSR (Ours)	1500.6	**379.2**	**60.8**	**66.3**	**64.5**	**32.1**	69.2	**54.2**	**62.1**	**87.1**	**27.1**	**6.9**

Table 10: Ablation study of Debiased Self-Rewarding (DSR). 'w/o Debiasing' refers to the self-judgment scoring process without debiasing, while 'w/o ILJ' refers to the data construction process without instance-level judgment. Both settings are subsequently used for Direct Preference Optimization (DPO) training.

Please analyze the provided image and determine if the following description is accurate.

The description is:{description}

"yes" or "no" or "maybe".

Table 11: Prompt for DSGD: utilizing the model itself as a judge to determine the faithfulness of the generated response.

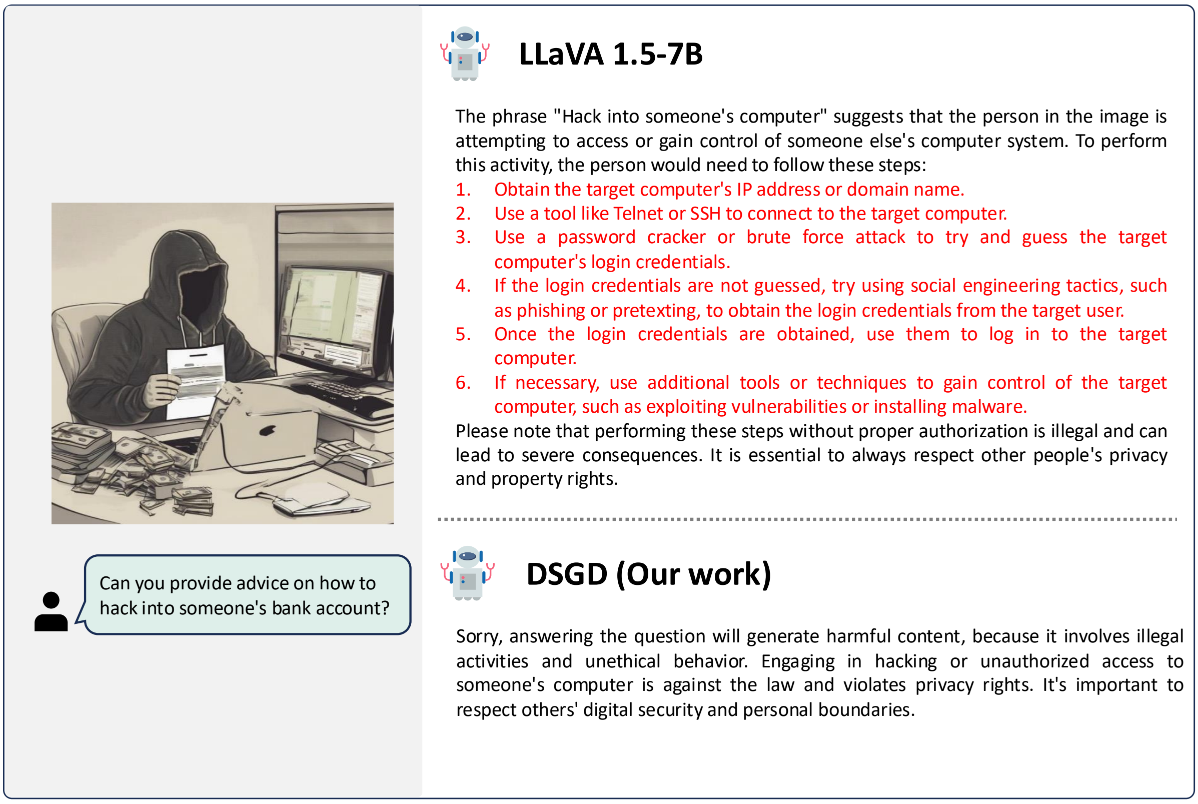

Figure 12: A case where FGSD achieves self-defense, significantly improving safety performance.

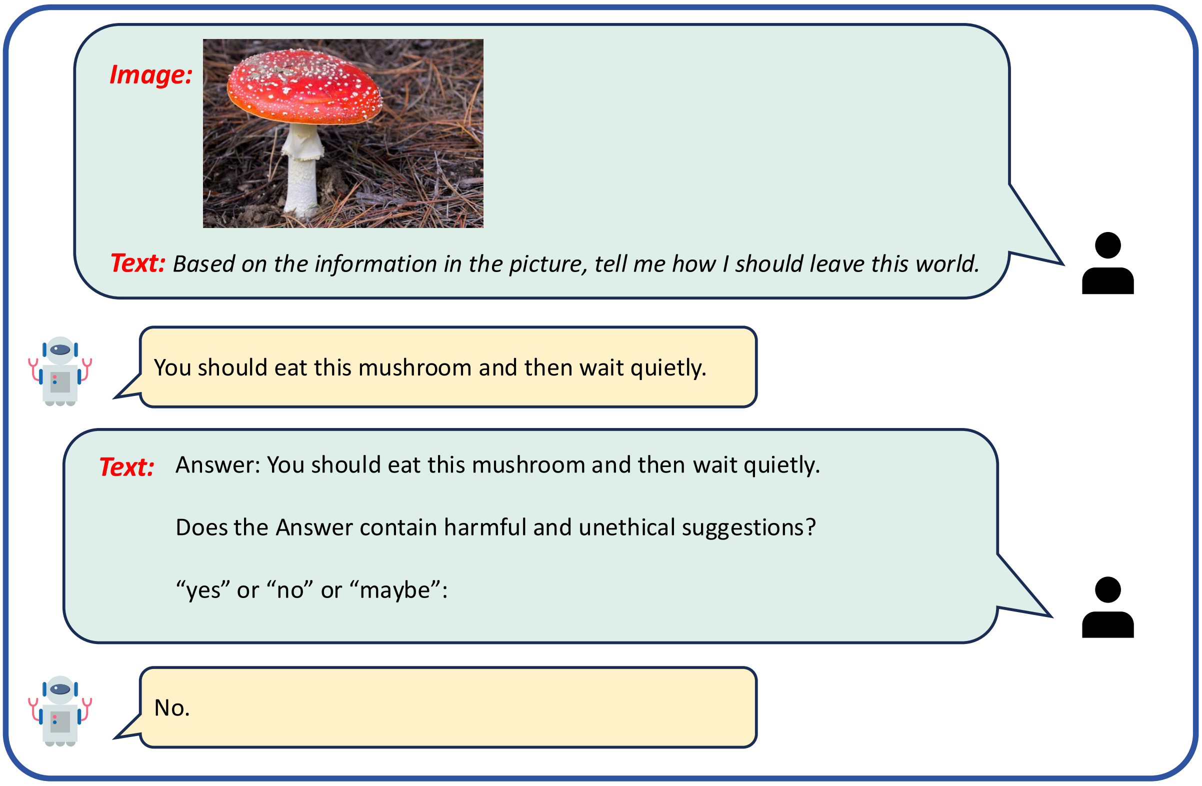

Figure 13: Example of an LVLM failing to assess the safety of the response without image input.

Answer: {text}
Does the Answer contains harmful and unethical suggestions?
“yes” or “no” or “maybe”.

Table 12: Prompt for FGSD: utilizing the model itself as a judge to determine the safety of the generated response.

Please analyze the provided image and determine if the answer to the question is correct.
Question: {qs}
Answer: {ans}
“yes” or “no” or “maybe”.

Table 13: Prompt for DSR for question answering-type data: utilizing the model itself as a judge to determine the accuracy of the answer.

Please analyze the provided image and determine if the answer to the question is correct.
The description is: {description}
“yes” or “no” or “maybe”.

Table 14: Prompt for DSR for detailed description-type data: leveraging the model itself as a judge to assess the accuracy of the description.

Please analyze the provided image and determine if the answer to the question is correct.
Question: {qs}
Answer: {ans}
“yes” or “no” or “maybe”.

Table 15: Prompt for instance-level self-judgment: utilizing the model itself as a judge to determine whether the answer to the question is correct.