Aligning Modalities in Large Vision Language Models via Preference Fine-tuning

Anonymous ACL submission

Abstract

Instruction-following Large Vision Language Models (LVLMs) have achieved significant progress recently on a variety of tasks. These approaches merge strong pre-trained vision models and large language models (LLMs). Since these components are trained separately, the learned representations need to be aligned with joint training on additional image-language pairs. This procedure is not perfect and can cause the model to hallucinate - provide answers that do not accurately reflect the image, even when the core LLM is highly factual and the vision backbone has sufficiently complete representations. In this work, we frame the hallucination problem as an alignment issue, tackle it with preference tuning. Specifically, we propose POVID to generate feedback data with AI models. We use ground-truth instructions as the preferred response and a two-stage approach to generate dispreferred data. First, we prompt GPT-4V to inject plausible hallucinations into the correct answer. Second, we distort the image to trigger the inherent hallucination behavior of the LVLM. This is an automated approach, which does not rely on human data generation or require a perfect expert, which makes it easily scalable. Finally, both of these generation strategies are integrated into an preference optimization pipeline. In experiments across broad benchmarks, we show that we can not only reduce hallucinations, but improve model performance across standard benchmarks, outperforming prior approaches.

1 Introduction

Large Vision Language Models (LVLMs) have achieved significant success in various vision understanding tasks, such as image captioning (Vinyals et al., 2015; Li et al., 2022, 2023c) and vision question answering (Ye et al., 2023; Antol et al., 2015). These LVLM models fuse larger-scale pre-trained vision models into the representation space of a large language models (LLM), allowing the LLM access to the visual representations. However, such LVLMs are not perfect and even suffer from "hallucinations", a phenomenon in which the language model generates content that is not grounded in the image, such as imagined objects and even scenes, wrong spatial relationships or categories, etc. Such artifacts are present even when both the vision backbone produces high-quality visual features and the language model itself is factual and accurate. These issues can pose significant risks when LVLMs are deployed in high-stakes scenarios, such as medical domains (Li et al., 2023b) or autonomous driving (Dewangan et al., 2023).

As discussed by Cui et al. (2023), the potential reason for hallucinations in LVLMs lies in their tendency to prioritize common sense present in the training language data, often disregarding the actual visual input information. In this paper, we attribute this issue to the lack of alignment between the image and text modalities, resulting in a reduced focus on input image information. Recent research efforts have sought to enhance the alignment between modalities through preference fine-tuning techniques, such as reinforcement learning from human feedback (RLHF) (Sun et al., 2023). Concurrent works (Li et al., 2023d; Zhao et al., 2023b) also use the Direct Preference Optimization (DPO) framework, but they rely on the traditional preference data generation process in LLMs, where both preferred and dispreferred responses may potentially be incorrect. However, in LVLMs, the produced responses are centered around the image data rather than being generated freely like in LLMs. When comparing two responses, both of which may be incorrect for the given task, the model may struggle to accurately align the image with the correct generated response. In (Yu et al., 2023a) the authors propose to solve this issue by collection corrective feedback, which shows strong results, but relies on costly human data gathering.

Unlike prior works that generate both preferred and dispreferred data, we propose **P**reference **O**ptimization in L**V**LM with A**I**-Generated **D**ispreferences (**POVID**) framework, aiming to exclusively generate dispreferred feedback data using AI models. In POVID we employ a high-quality ground truth multi-modal instruction as the preferred answer and employ two strategies to generate dispreferred responses. *First*, we utilize GPT-4V to introduce plausible hallucinations into the answer, which we then use as the dispreferred response. *Second*, we aim to provoke inherent hallucination patterns and subsequently correct them within the target LVLM that requires fine-tuning. We achieve this goal by introducing noise, triggering inherent hallucination patterns within the LVLMs. The introduction of noise disrupts the LVLM's comprehension of the image, leading it to generate uncertain responses that rely more on textual context or the knowledge it has acquired from the training data. Given that the inherent hallucination patterns of the target LVLM evolve during the training process, the response generation with the noisy image occurs in real-time during training, and this is treated as dispreference. Finally, we integrate both forms of dispreference into the DPO optimization framework, specifically targeting the alignment of language generation with the image.

The primary contribution of this paper is POVID, which aligns the image and text modalities in LVLMs. This approach explicitly contrasts a hallucinatory answer with a truthful one, eliminating the need for gathering human feedback and making it easily deployable at scale. Our empirical results demonstrate the promise of our framework in reducing hallucinations and enhancing other LVLM-related tasks. In particular, our approach significantly improves performance compared to other preference tuning methods in LVLMs. Additionally, we demonstrate that POVID can redirect the attention of LVLMs towards the image modality, resulting in better modality alignment.

2 Preliminaries

Our approach aims to fine-tune LVLMs for better aligning the image and text modalities uses the framework of preference tuning from preferences over responses. In this section, we will provide some notations of LVLMs and an overview of direct preference optimization (Rafailov et al., 2023).

Vision Large Language Models. LVLMs is an multimodal extension of large language models, which can generate sentences in an autoregressive manner, aiming to progressively predict the probability distribution of the next token. Here, the input prompt x contains both images and text prompts, and the output contains text response y. A typical application scenario for LVLMs is image captioning and Vision Question Answering (VQA).

Direct Preference Optimization. Direct preference optimization (DPO) (Rafailov et al., 2023) leverages preference data for preference optimization in language models. Here, the preference data is defined as $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$, where $y_w^{(i)}$ and $y_l^{(i)}$ represent preferred and dispreferred responses given an input prompt x. $r(x, y)$ is defined as the reward function. Following a Bradley-Terry model (Bradley and Terry, 1952), the probably of obtaining each preference pair is:

$$p(y_w \succ y_l) = \sigma(r(x, y_w) - r(x, y_l)), \quad (1)$$

where we omit the superscript (i) for simplicity and $\sigma(\cdot)$ is defined as a sigmoid function. The DPO loss can be formulated as classification loss over the preference data as:

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log\sigma\left(\alpha\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \alpha\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]. \quad (2)$$

DPO enables learning π_θ from a fixed dataset of preferences, which is lightweight. However, the key challenge lies in generating effective preference data for fine-tuning and aligning image and text modalities in LVLMs.

3 Constructing Preferences to Aligning Modalities in LVLMs

While preference learning approaches (e.g., DPO) facilitate the lightweight training of LVLMs, they require data in the form of preferences. In contrast to LLMs, which support more freestyle generation in many scenarios, LVLMs used in various applications, such as VQA or image captioning, produce responses linked to input images. This inherent image-centricity presents distinct challenges in the preference data generation process for LVLMs, setting it apart from the process in LLMs. Specifically, in LVLMs, when comparing two responses, neither of which is correct for the required task (e.g., image captioning), the model may not be able to accurately align the image with the response.

To address this challenge, we propose **P**reference **O**ptimization in L**V**LM with A**I**-Generated

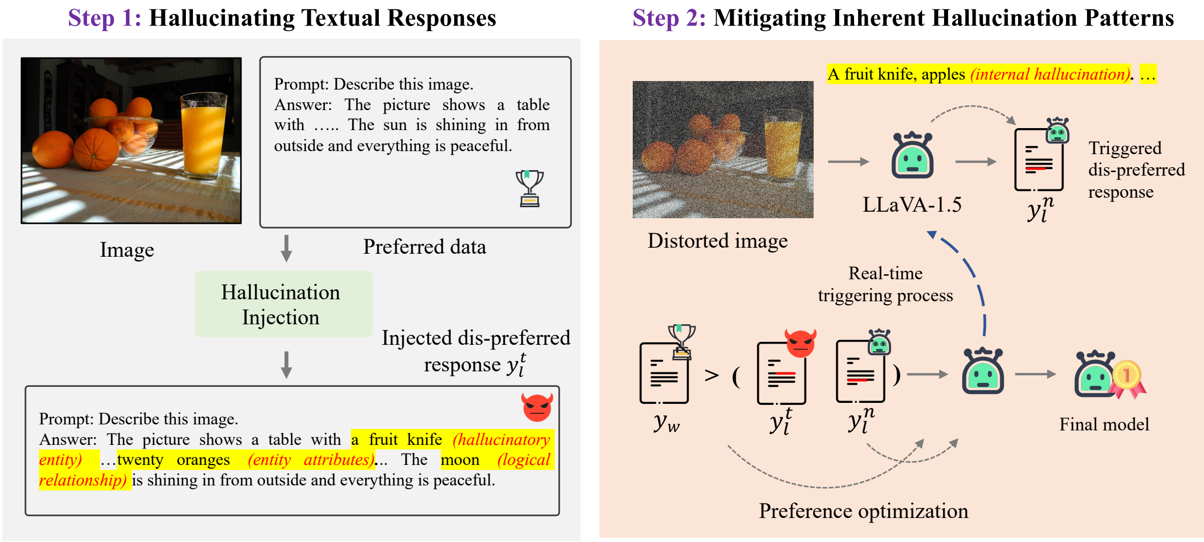

Figure 1: The framework of POVID. The preference generation process is divided into two steps: hallucinating textual responses and trigger dispreference during training. Here, different types of triggered hallucinations are labeled in *(types of hallucinations)*.

Dispreferences (**POVID**), a novel approach aimed at better aligning image and text modalities. As illustrated in Figure 1, POVID leverages AI models to generate dispreferred responses without the need for human labeling efforts. These generated dispreferred responses, when combined with groundtruth image descriptions (treated as preferred responses), form the preference data pairs. Specifically, we employ two strategies to generate the dispreferred response: (1) Firstly, we manipulate the groundtruth response by transforming the groundtruth response into hallucinated response, which serves as the dispreferred response; (2) Secondly, we introduce distortion to the image input during the training process, intending to trigger inherent hallucination patterns within the LVLMs. These patterns are then formalized as the dispreferred response, motivating the model to correct its inherent dispreferred patterns. In the remainder of this section, we will provide detailed explanations of both strategies and demonstrate how to integrate them into the preference training framework.

3.1 Hallucinating Textual Responses

In our first strategy, we aim to generate dispreferred hallucinatory responses by hallucinating the groundtruth correct response. We construct the hallucinatory response based on a subset with 17K examples that are randomly sampled from LLaVA-Instruct-150K (Liu et al., 2023b) dataset. Here, the LLaVA-Instruct-150K datasets is used to train LLaVA LLaVA with supervised fine-tuning. The 17K examples includes various task types, including image captioning, VQA and logical reasoning.

To construct the preferences, we treat the original answers in the 17K examples as preferred responses. In terms of constructing dispreferred responses, we hallucinate the original answers using GPT-4V (OpenAI, 2023). Here, we adopt two hallucinating approaches tailored to different tasks:

I. Hallucinating Image Captioning Tasks. First, we hallucinate the image captioning tasks by considering three fundamental causes of hallucination in LVLMs: (1) *Object Co-occurrence*: This phenomenon arises when the training data contains spurious co-occurring patterns between objects, leading LVLMs to generate objects based on these learned spurious correlations. In this context, we aim to leverage GPT-4V to deduce object co-occurrence within the given image and subsequently revise the original responses accordingly; (2) *Logical Relationships Between Entities*: This involves using GPT-4V to modify the relationships between the original objects; (3) *Incorrect Attributes*: In this case, we employ GPT-4V to alter the attributes of various objects, such as changing their colors. We illustrate these three distinct hallucination scenarios with an example provided in Figure 2(a). The prompt we used to generate the dispreferred response is in Appendix A.2.

II. Hallucinating Reasoning Tasks. Secondly, when dealing with tasks involving reasoning, such as VQA and logical reasoning, we task GPT-4V with modifying the reasoning process. This entails introducing errors related to logical relationships,

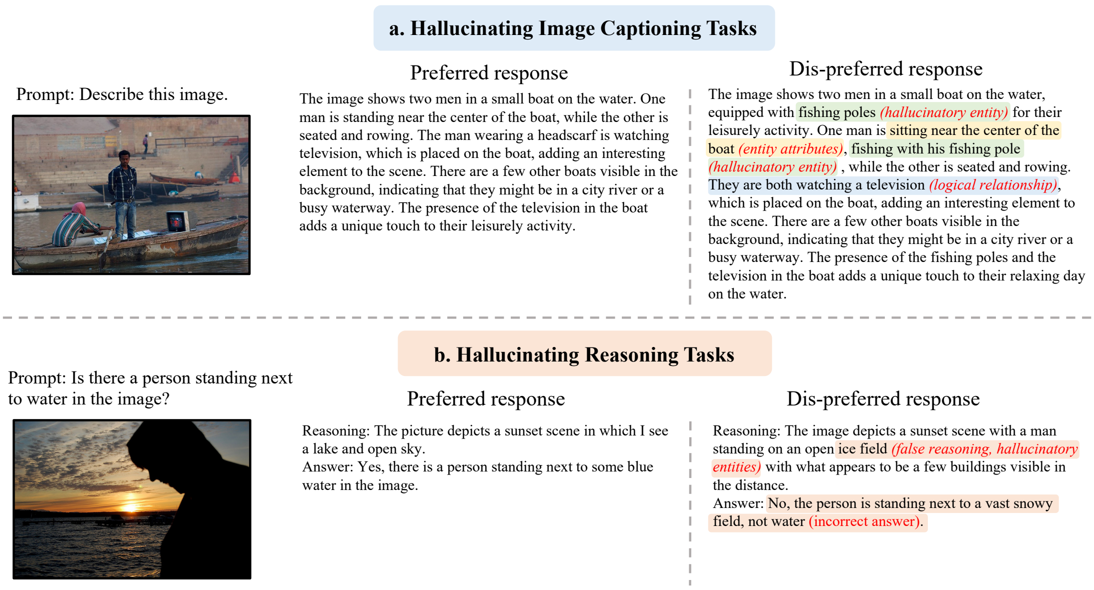

Figure 2: Two examples extracted from hallucinated image captioning tasks and reasoning tasks. Different types of hallucinations are labeled in *(types of hallucinations)*.

entity information, entity attributes, and more. Additionally, we recommend that GPT-4V attempts to make subtle changes to the reasoning process, ensuring it remains independent of factual reasoning results, meaning that an incorrect reasoning process may still yield correct results. However, if the introduction of errors necessitates alterations to the reasoning results, we instruct GPT-4V to adjust the results accordingly. Likewise, in Figure 2(b), we provide an example to demonstrate both the original and the generated dispreferred responses. The prompt we used is detailed in Appendix A.2.

3.2 Mitigating Inherent Hallucination Patterns

In addition to generating the dispreferred response using powerful external models like GPT-4V, we also aim to provoke inherent hallucination patterns to be finetuned. Our second strategy introduces noise into the image to trigger inherent hallucination patterns. This noise disrupts the LVLM's understanding of the image, leading it to produce uncertain responses that rely more on textual context or acquired knowledge from the training data. This occurs because, in the presence of noisy images, the model tends to prioritize inherent object associations over visual information. Notably, the noise step should remain within a reasonable range, ensuring that the image remains easily recognizable by humans. For example, as depicted in Figure 3, when presented with the context "There are a knife and _", under specific noisy conditions, the likelihood of "fork" surpasses that of "plate" (ground truth). This may occur because "knife" is more likely to co-occur with "fork" in the training data. With an increase in noise steps, the term "pixel" becomes predominant, owing to the noticeable noise patterns within the image. We further demonstrate the generalizability of this phenomenon through experiments on multiple models and different images in Appendix C.1. Consequently, establishing an appropriate noise step to trigger inherent hallucination patterns is a reasonable approach.

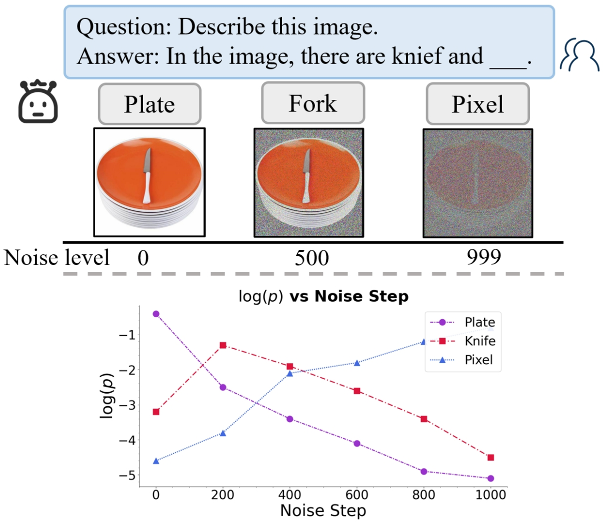

Figure 3: Illustration of logits for the next token generation with "In the image, there are knife and _". This figure shows the predictive uncertainty in token generation, emphasizing the influence of visual cues from objects identified as "knife" and "plate" (see Appendix C.1 for more detailed discussion).

To achieve this goal, we introduce diffusion noise into the original image. We define the noise step as k, and the noised image with step k can be expressed as follows:

$$x(k) = \sqrt{\bar{\xi}_k} \cdot x + \sqrt{1 - \bar{\xi}_k} \cdot \epsilon, \tag{3}$$

where $\bar{\xi}_t = \prod_{i=0}^{k} \xi_i$ and $\xi_k \in (0, 1)$ is a hyperparameter chosen prior to model training. Detailed settings can be found in Appendix A.1. After obtaining the noised image, in order to more effectively capture changes in inherent hallucination patterns during the fine-tuning process of the LVLM, we integrate the image noising process into the DPO fine-tuning process. Specifically, for each input prompt x, we take into account the dispreferred responses from both the hallucinated text responses discussed in Section 3.1 and the responses triggered by distorted images. We then reformulate the DPO loss as follows:

$$\begin{aligned}\mathcal{L}_{POVID} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \Bigg[\log \sigma \Bigg(\alpha \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} \\ - \left(\beta_1 \log \frac{\pi_\theta(y_l^t|x)}{\pi_{\text{ref}}(y_l^t|x)} + \beta_2 \log \frac{\pi_\theta(y_l^n|x^n)}{\pi_{\text{ref}}(y_l^n|x^n)} \right) \Bigg) \Bigg],\end{aligned} \tag{4}$$

where α, β_1 and β_2 are coefficients that balance preferred and dispreferred terms. y_l^g represents the dispreferred response generated using the approach outlined in Section 3.1. Additionally, x^n represents the noisy image, which triggers the generation of the dispreferred response y_l^n. It's important to note that for each token i in the sequence y_l^n, the value of $y_{l,i}^n$ is determined by selecting the maximum probability from the set $\pi_\theta(\cdot \mid x^n, y_{w,<i})$. Here, each generated token in the dispreferred response $y_{l,i}^n$ is conditioned on the prior tokens from the preferred response $y_{w,<i}$. This conditioning allows us to control the reliability of the triggered dispreferred response. As a result, we aim to capture the most significant changes between the preferred and dispreferred responses, since a substantial portion of dispreferred response overlaps with preferred response. The training process of our method is detailed in Algorithm 1.

Algorithm 1 POVID Training Process

Require: $\mathcal{D}$: Dataset of paired images and text context. π_θ: Parameters of the LVLM. π_{ref}: Parameters of the reference model. α, β_1, β_2: Hyperparameters. ξ_k: Noise hyperparameter for each timestep. T: Noise Steps.

AddNoiseToImage(x_0, k)
  $\epsilon \sim \text{N}(0, 1)$
  $x(k) \leftarrow \sqrt{\bar{\xi}_k} \cdot x_0 + \sqrt{1 - \bar{\xi}_k} \cdot \epsilon$
Generate dispreferred data and place it in $\mathcal{D}$
Initialize reference policy π_θ
for epochs **do**
  for $(x, y_w, y_l^t) \in \mathcal{D}$ **do**
    for $k = 0$ to T **do**
      $x(k) \leftarrow$ AddNoiseToImage(x, k)
    end for
    Update π_θ through Eq. (4)
  end for
end for

4 Experiment

In this section, we empirically investigate the effectiveness of POVID in aligning image and text modalities in LVLMs and reducing hallucination. We aim to answer the following questions: (1) Can POVID effectively reduce hallucination in LVLMs compared to other preference fine-tuning strategies? (2) Can hallucinating textual responses and image distortion benefit performance? (3) How does POVID change attention weights to align image and text modalities?

4.1 Experimental Setups

In this section, we briefly introduce the implementation details, baselines, and evaluation settings.

Implementation Details. Following concurrent LVLM preference tuning studies Yu et al. (2023b); Li et al. (2023d), we choose LLaVA-1.5 (7B) as our backbone model for all experiments and have applied POVID to fine-tune LLaVA-1.5 (7B), including both LoRA fine-tuning and full fine-tuning. The training process is divided into two stages. In the first stage, we exclusively utilize the preferences generated through the hallucinating textual responses, as discussed in Section 3.1, to fine-tune LLaVA-1.5 using DPO. In the second stage, we employ image distortion to rectify the model's inherent hallucinatory behaviors using the loss defined in Eqn. (4). The first stage involves training for 3 epochs, and the second stage for 1 epoch. Please refer to Appendix A.1 for more details.

Baseline Approaches. We first compare the proposed approach with other LVLM preference tuning methods, which include Silkie (Li et al., 2023d), LLaVA-RLHF (Sun et al., 2023), and RLHF-V (Yu et al., 2023b). These methods enhance model performance by creating curated datasets and subsequently applying preference tuning techniques to fine-tune the model based on these datasets. To ensure a fair and equitable comparison, we utilize the same curated datasets employed by these approaches and apply DPO to fine-tune LLaVA-1.5 (7B)'s LoRA parameters for the same number of training epochs as in the first stage of POVID. Furthermore, we compare the performance with other open source LVLMs, including InstructBLIP (Dai et al., 2023), Qwen-VL-Chat (Bai et al., 2023) and mPLUG-Owl2 (Ye et al., 2023).

Evaluation Benchmark. To evaluate the performance, we first adopt LVLM hallucination benchmarks, including CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023f), and MMHal (Sun et al., 2023). In addition, we evaluate all approaches on comprehensive LVLM evaluation benchmarks, including SQA$^{\mathrm{I}}$ (Lu et al., 2022), VQA$^{\mathrm{v2}}$ (Goyal et al., 2017), GQA (Hudson and Manning, 2019), VQA$^{\mathrm{T}}$ (Singh et al., 2019), MME (Fu et al., 2023), MMB (Liu et al., 2023c), MM-Vet (Yu et al., 2023c) and LLaVA$^{\mathrm{W}}$ (Liu et al., 2023b). Detailed descriptions of all benchmarks are in Appendix B.

4.2 Results

Comparison with Different Preferences in LVLMs. In Table 1, we present the results of a comparison between various LVLM preferences, evaluating both hallucination and comprehensive benchmarks. Firstly, in the hallucination benchmarks, POVID effectively enhances performance by creating dispreferred preferences through textual data manipulation and image distortion. We achieve a significant improvement of 17.08% across all hallucination benchmarks, effectively reducing hallucinations in the generated responses. This outcome aligns with our expectations, as constructing dispreferences from the ground-truth correct responses maximally enables the model to discern differences between correct and incorrect responses while optimizing alignment between the image and text modalities within the model. Moreover, in more comprehensive evaluation benchmarks, which encompass not only factuality and hallucination assessment but also other aspects, POVID continues to demonstrate superior performance when compared to other preference data collection methods. This further indicates our model's capacity to enhance LVLM performance through improved modality alignment.

Comparison with Open-Sourced LVLMs Models. We present a comparison between POVID and other open-sourced LVLMs in Table 6 of Appendix. Although various approaches utilize different image and text encoders, POVID outperforms other popular LVLMs in eight out of twelve benchmarks. In contrast, the second-best baseline, Qwen-VL-Chat, achieves the best performance in only three out of twelve benchmarks. This underscores the superiority of POVID and further corroborates its effectiveness in aligning image and text modalities to improve the performance of LVLMs.

4.3 Analysis

In this section, we provide a comprehensive analysis to demonstrate how different components contribute to the performance of POVID and illustrate how POVID enhances overall performance. We further conduct fine-grained analysis of different preference collection strategies in Appendix D. In addition, we discuss the compatibility of POVID on other state-of-the-art open-source LVLMs.

Ablation Studies. To further demonstrate the essential role of the key components of POVID in contributing to performance, we conduct ablation experiments on POVID (Full), and present the results in Table 2. In this ablation study, we evaluate the effectiveness of two aspects: (1) hallucinating groundtruth responses and (2) image distortion. According to the results, we initially observe that image distortion can enhance performance across all benchmarks. This indicates its effectiveness in aligning multimodalities by compelling the model to rectify inherent hallucination patterns. Additionally, generating dispreference from groundtruth responses significantly enhances performance, underscoring the effectiveness of the AI-generated dispreference strategy. Finally, when combining both strategies, POVID achieves the best performance, further affirming its effectiveness in enhancing LVLMs through improved modality alignment.

Compatibility Analysis. To verify the compatibility of POVID we have migrated POVID to two state-of-the-art LVLMs - SVIT (Zhao et al., 2023a) and Vila (Lin et al., 2023), to validate its compatibility. For the experiments in this section, we only fine-tuned the LoRA parameters of the language models, with SVIT using a 13B-parameter language model and Vila using a 7B-parameter language model. The training setup is same as the training of LLaVA shown in Appendix A.1. We

Table 1: Comparison between POVID and other preferences construction approaches in both hallucination and comprehensive benchmarks. We bold the best and underline the second best results.

	Hallucination Benchmark				Comprehensive Benchmark								
Method	C_S	C_i	POPE	MMHal	SQA^I	MM-Vet	MMB	$LLaVA^W$	MME	VQA^{v2}	VQA^T	GQA	Avg rank
LLaVA-1.5	66.8	12.7	85.90	2.42	66.8	30.5	64.3	63.4	**1510.7**	78.5	58.2	62.0	4.3
+ Vlfeedback	56.3	11.4	83.72	2.62	66.2	31.2	63.9	62.1	1432.7	77.3	57.5	**63.2**	4.6
+ Human-Preference	54.0	9.3	81.50	2.53	65.8	31.1	60.4	63.7	1490.6	78.4	58.6	61.3	4.4
+ RLHF-V	44.6	7.9	86.20	2.59	67.1	30.9	63.6	65.4	1489.2	78.2	58.3	62.1	3.5
+ POVID (LoRA)	**31.8**	**5.4**	86.90	2.69	68.8	31.8	64.9	68.7	1452.8	**78.7**	**58.9**	61.7	2.1
+ POVID (Full)	33.5	5.7	**87.12**	**3.08**	**70.0**	**36.4**	**65.6**	**69.9**	1449.1	78.6	57.8	62.0	**2.0**

Table 2: Results of ablation study. Text disprefer (Txt) indicates solely training with hallucinated responses. Image distortion (Img) means that we use distorted images to trigger inherent hallucination patterns.

		Hallucination Benchmarks				Comprehensive Benchmarks							
Txt	Img	C_S	C_i	POPE	MMHal	MME	VQA^T	SQA^I	GQA	MM-Vet	MMB	$LLaVA^W$	VQA^{v2}
×	×	66.8	12.7	85.90	2.42	**1510.7**	78.5	**58.2**	62.0	30.5	64.3	63.4	78.5
✓	×	35.0	9.9	87.01	2.67	1445.4	78.5	57.6	**62.2**	34.2	65.4	64.2	78.5
×	✓	45.0	10.7	85.91	2.52	1440.7	78.2	54.1	59.9	31.8	63.4	66.0	78.2
✓	✓	**33.5**	**5.7**	**87.12**	**3.08**	1449.1	**78.6**	57.8	62.0	**36.4**	**65.6**	**68.7**	**78.6**

present the results in Table 3. POVID improves the performance of both SVIT and Vila across several benchmarks. For SVIT, POVID significantly reduce the C_S and C_i scores, indicating better performance in captioning and the reliability of its responses to images. Similarly, Vila also saw reductions in C_S and C_i scores, along with improvements in other key benchmarks, demonstrating the effectiveness and compatibility of POVID when integrated into these LVLMs. The results from Table 3 demonstrate the robustness and utility of POVIDin enhancing performance and dependability across various open-sourced LVLMs.

Modality Alignment Analysis. We assess the impact of POVID on modality alignment by comparing the attention maps generated by POVID with those of the original LLaVA-1.5 model, with a specific focus on image captioning and VQA tasks. We illustrate two cases in Figure 4, where these attention maps reveal the distribution of attention scores assigned to generated textual tokens within the input image-text sequence throughout the LVLM's output generation phase. Our findings reveal that the original LLaVA-1.5 model tends to overemphasize the context of the text, which can result in hallucinations. In contrast, POVID increasingly prioritizes attention towards the image, indicating a strong alignment between image and text modalities. One potential explanation for this phenomenon is that, through a comparison between the ground truth and the generated dispreferred data, along with the mitigation of internal hallucination patterns, POVID redirects the LVLM's attention, leading to a greater focus on the image tokens.

5 Related Work

LVLMs and LVLM Hallucination. The advent of autoregressive large-scale language models (LLMs), highlighted in works by (Touvron et al., 2023a,b; Taori et al., 2023), has led to the development of Vision-Large Language Models (LVLMs). To align the image and text modalities, recent research has concentrated on instruction tuning (Li et al., 2023a), scaling up training dataset (Jia et al., 2021), and better alignment between image and text with local feature enhancement (Cha et al., 2023). These advancements have successfully combined LLMs with image inputs and excel in image comprehension. However, such LVLMs are not perfect and even suffer from "hallucinations", generating outputs that may not accurately or faithfully represent the content of a user-provided image. There are various sources of hallucinations in LVLMs, including biased data (Chuang et al., 2023; Tu et al., 2023), insufficient training (Chen et al., 2023), and imperfect inference (Huang et al., 2023). Recently, addressing hallucination in LVLMs is primarily achieved through various techniques such as decoding approaches (Leng et al., 2023; Huang et al., 2023), post-processing (Zhou et al., 2023; Yin et al.,

Table 3: The performance of POVID when migrated to other open-source LVLMs on comprehensive benchmarks.

Method	C_S	C_i	POPE	MMHal	VQA^{v2}	VQA^{T}	SQA^{I}	GQA	MM-Vet	MMB	LLaVA^{W}	MME
SVIT	48.9	4.6	86.25	2.71	**80.3**	60.8	70.0	**64.1**	34.2	68.6	67.4	**1565.8**
SVIT + **POVID**	**42.4**	**4.3**	**86.30**	**2.76**	80.2	**60.9**	**70.1**	63.9	**35.4**	**69.1**	**70.2**	1560.2
Method	C_S	C_i	POPE	MMHal	VQA^{v2}	VQA^{T}	SQA^{I}	GQA	MM-Vet	MMB	LLaVA^{W}	MME
Vila	26.3	6.6	85.5	2.56	79.9	64.4	68.2	**62.3**	34.9	68.9	69.7	**1533.0**
Vila + **POVID**	**23.4**	**6.1**	**86.1**	**2.61**	**81.2**	**64.4**	**68.7**	62.1	**36.3**	**69.2**	**69.9**	1529.7

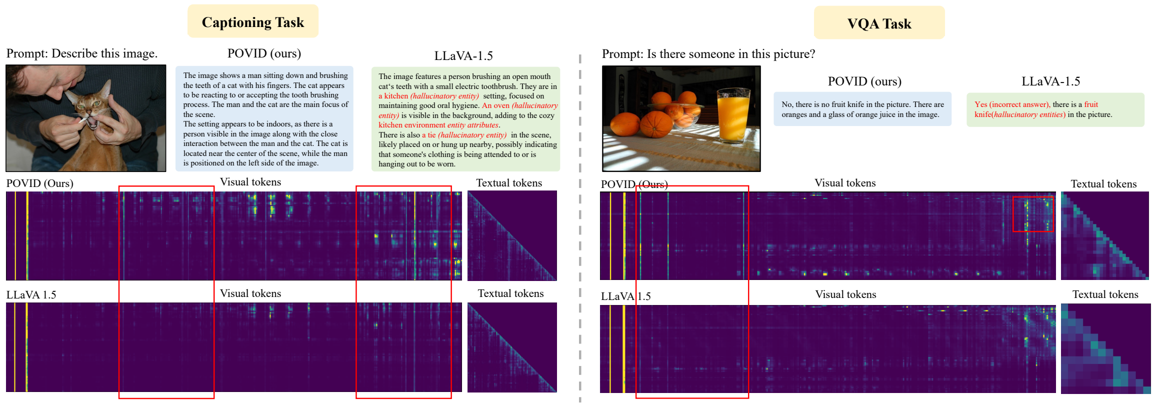

Figure 4: Comparison of attention map between POVID and LLaVA-1.5 at different tasks. The red box region is labeled with the image attentions that can be significantly improved by POVID.

2023) and the construction of higher-quality dataset (Liu et al., 2023a; Li et al., 2023e). While these approaches can mitigate hallucination to some extent, they often fail to directly guide LVLMs to align image and text modalities.

Preference Alignment. Aligning with human preferences for large models has emerged as a critical issue due to the limitations imposed by safety and ethical considerations in real-world applications. Preference alignment can be broadly categorized into two main approaches: alignment through feedback, which encompasses both human (Bai et al., 2022; Rafailov et al., 2023) and AI-generated feedback (Lee et al., 2023) and alignment via prompt guidance (Wei et al., 2022). Initial investigations into preference alignment for LVLMs have recently been conducted. Sun et al. (2023) introduced LLaVA-RLHF, which utilizes a preference dataset annotated by humans to decrease hallucinations in LLaVA. Li et al. (2023d) proposed a method for distilling preferences into LVLMs to enhance their ability to generate relevant and accurate responses based on visual context. Yu et al. (2023b) collected human preferences in the form of segment-level corrections to hallucinatory content and optimizing the model's behavior based on dense, direct feedback. While these initial results are promising, these works heavily rely on the traditional preference data generation process in LLMs, which generate both preferred and dispreferred responses, but none of them are guaranteed to be correct. In LVLMs, when both responses prove incorrect for the given task, accurately aligning the image with the correct generated response becomes challenging. In contrast, POVID directly generates dispreferred responses, effectively addressing this challenge.

6 Conclusion

In this work, we introduce a novel approach, Preference Optimization in LVLM with AI-Generated Dispreferences (POVID) to address the challenges in modality alignment for large vision-language models. In POVID, we adopt two strategies to generate dispreferred responses: first, we use synthetic data from GPT-4V to inject plausible hallucinations into the correct answer. Second, we use distorted images to trigger the inherent hallucination behavior of the LVLM. Then both of these answers are integrated into an RLHF framework via Direct Preference Optimization. Empirical evaluations across multiple benchmarks reveal that POVID not only mitigates hallucination effectively but boosts the overall performance of model.

7 Limitation

While our results provide significant insights into the behavior of LVLMs under varying conditions, several limitations of our study need to be addressed. The training and evaluation of the models were conducted using high-performance hardware, such as multiple A100 80G GPUs. This setup may not be feasible for all research teams or practical applications, potentially limiting the reproducibility and accessibility of our findings. Additionally, the specific formula used to adjust the diffusion noise level is manually designed rather than automatically generated.

References

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

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*.

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

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2023. Honeybee: Locality-enhanced projector for multimodal llm. *arXiv preprint arXiv:2312.06742*.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. 2023. Lion: Empowering multi-modal large language model with dual-level visual knowledge. *arXiv preprint arXiv:2311.11860*.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. 2023. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*.

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

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.

Vikrant Dewangan, Tushar Choudhary, Shivam Chandhok, Shubham Priyadarshan, Anushka Jain, Arun K Singh, Siddharth Srivastava, Krishna Murthy Jatavallabhula, and K Madhava Krishna. 2023. Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving. *arXiv preprint arXiv:2310.02251*.

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*.

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.

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*.

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.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

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*.

Chen Li, Yixiao Ge, Dian Li, and Ying Shan. 2023a. Vision-language instruction tuning: A review and analysis.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023b.

Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

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

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

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

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

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*.

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

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

Pan Lu, Swaroop Mishra, Tony 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. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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

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

Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. 2023. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023.

mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

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, and Tat-Seng Chua. 2023a. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arxiv*.

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

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

Bo Zhao, Boya Wu, and Tiejun Huang. 2023a. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*.

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

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*.

A Experimental Setup and Prompt Design

A.1 Training Setup

Training hyperparameters are shown in Table 4. For the first phase, we trained for 3 epochs, and for the second phase, the training was conducted for 1 epoch. Under the setup of DeepSpeed ZeRO2, for POVID LoRA, we utilize a single A100 80G during the training process, which takes approximately 6 hours. For POVID full, our first stage employs four A100 80G, taking approximately 2.5 hours, while the second stage utilizes eight A100 80G, taking approximately 1 hour. For the second phase, we adjust the diffusion noise level, symbolized by ξ through a specific formula: $\xi = \text{Sigmoid}(l_t) \times (0.5 \times 10^{-2} - 10^{-5}) + 10^{-5}$, where ϵ is drawn from a normal distribution.

A.2 Construction of the Dispreference Dataset

This section details the prompts utilized to compile the dataset focusing on dispreferences, specifically within the realms of image captioning and reasoning tasks. The prompts are designed to elicit responses that reveal dispreference patterns, categorized into two main types: image captioning tasks intended to provoke imaginative descriptions, and reasoning tasks aimed at stimulating inferential thought processes. These prompts, central to our methodology, are enumerated in Table 7, offering a comprehensive view of the data generation framework.

B Details about Baselines and Benchmark

This section provides a detailed introduction to the benchmarks used in the experimental part of this paper.

- CHAIR, including CHAIR_S (C_S) and CHAIR_i (C_i), is a metric used in image captioning tasks to evaluate the accuracy of object descriptions in captions. It compares the objects mentioned in a caption with those present in the image.

- MMHal (Sun et al., 2023) assesses hallucinations and response informativeness by utilizing GPT-4V to compare model output with human responses and various object labels, determining the scores accordingly.

- POPE (Li et al., 2023f) uses a set of binary classification tasks, prompting LVLMs with simple Yes-or-No questions about the existence of certain objects in images.

- MME (Fu et al., 2023) is a comprehensive evaluation tool designed to measure both perception and cognition abilities across 14 sub-tasks for LVLMs.

- MMB: MMbench (Liu et al., 2023c) is known for its approach to assessing both perception and reasoning abilities, categorized into top-level dimensions in the ability taxonomy. This benchmark includes different levels of abilities, each encompassing specific aspects of perception and reasoning.

- MM-Vet (Yu et al., 2023c) focuses on evaluating six core capabilities: recognition, knowledge,

Table 4: Training hyperparameters.

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
noise_step (only for internal preference optimization)	500

Table 5: Fine-grained performance comparison of various models on LLaVA$^{\mathrm{W}}$, where we adopt the following abbreviation: Convo for Conversation, Captioning for Detail description, Reasoning for Complex reasoning.

Method	Convo	Captioning	Reasoning	Overall
LLaVA-1.5	53.3	53.4	79.6	63.4
+ Vlfeedback	51.3	49.3	78.5	62.1
+ Human-Preference	49.6	43.3	81.3	63.7
+ RLHF-V	55.8	56.1	80.3	65.4
+ **POVID (LoRA)**	55.9	60.1	81.5	68.7
+ **POVID (Full)**	**56.5**	**67.2**	**81.7**	**69.9**

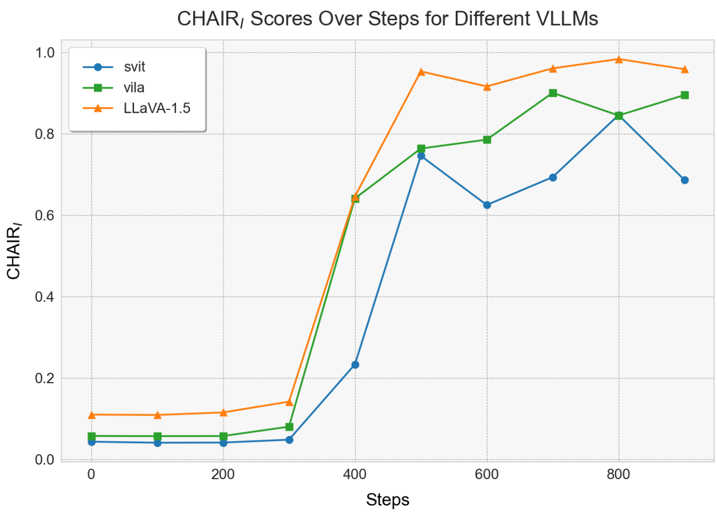

Figure 5: Comparison of CHAIR$_I$ scores on different LVLMs across various noise levels.

OCR, spatial awareness, language generation, and math. These capabilities cover a wide range of functions, from general visual recognition to specific tasks like arithmetic problem-solving.

- LLaVA$^{\mathrm{W}}$: LLaVA-bench (Liu et al., 2023b) assesses models in more complex tasks and their adaptability to new domains. It consists of 24 diverse images, encompassing a variety of scenes such as indoor and outdoor settings, memes, paintings, and sketches. Each image in LLaVA$^{\mathrm{W}}$ is paired with a detailed, manually crafted description and a carefully chosen set of questions, totaling 60 questions. This setup aims to provide a thorough and varied evaluation of the models' capabilities.

- VQA$^{\mathrm{v2}}$ (Goyal et al., 2017) is a dataset comprising open-ended questions related to images, demanding comprehension of vision, language, and commonsense knowledge for answers.

Table 6: Comparison between POVID and other state-of-the-art LVLMs across both hallucination and comprehensive benchmarks. We bold the best results and underline the second best results. Notably, when LLaVA-1.5 7B utilizes POVID for preference learning, it can achieves an average rank at 2.0 over other open-source models across all benchmarks.

Method	Vision Encoder	Language Model	C_S	C_i	POPE	MMHal	VQA^{v2}	VQA^{T}
InstructBLIP	ViT-g (1.3B)	Vicuna (7B)	40.0	8.0	77.83	2.10	70.1	50.1
Qwen-VL-Chat	ViT-G (1.9B)	Qwen (7B)	48.2	9.1	87.07	2.89	78.2	**61.5**
mPLUG-Owl2	ViT-L (0.3B)	LLaMA (7B)	54.4	12.0	86.20	2.17	**79.4**	58.2
LLaVA-1.5 + **POVID (LoRA)**	ViT-L (0.3B)	Vicuna (7B)	**31.8**	**5.4**	86.90	2.69	78.7	58.9
LLaVA-1.5 + **POVID (Full)**	ViT-L (0.3B)	Vicuna (7B)	33.5	5.7	**87.12**	**3.08**	78.6	57.8
Method	Vision Encoder	Language Model	SQA^{I}	GQA	MM-Vet	MMB	$LLaVA^{W}$	MME
InstructBLIP	ViT-g (1.3B)	Vicuna (7B)	60.5	49.2	26.2	36.0	60.9	1212.8
Qwen-VL-Chat	ViT-G (1.9B)	Qwen (7B)	68.2	57.5	**41.2**	60.6	67.7	**1487.5**
mPLUG-Owl2	ViT-L (0.3B)	LLaMA (7B)	64.5	56.1	36.2	64.5	59.9	1450.2
LLaVA-1.5 + **POVID (LoRA)**	ViT-L (0.3B)	Vicuna (7B)	68.8	61.7	31.8	64.9	68.7	1452.8
LLaVA-1.5 + **POVID (Full)**	ViT-L (0.3B)	Vicuna (7B)	**70.0**	**62.0**	36.4	**65.6**	**69.9**	1449.1

- GQA (Hudson and Manning, 2019) is a novel dataset tailored for real-world visual reasoning and compositional question answering. It addresses shortcomings of previous VQA datasets by leveraging scene graph structures and a robust question engine to generate 22 million diverse reasoning questions, each paired with functional programs representing their semantics.

- VQA^{T}: TextVQA (Singh et al., 2019) is a dataset aimed at addressing the significant challenge of visually impaired users reading text in images of their surroundings. It consists of 45,336 questions and 28,408 images, requiring reasoning about text in the images to answer. SQA^{I}: SciQA-IMG (Lu et al., 2022) is a new benchmark dataset designed to assess the multi-hop reasoning capability and interpretability of artificial intelligence systems on multimodal multiple-choice scientific questions. It consists of approximately 21,000 diverse science-themed questions, along with annotated answers and corresponding lecture and explanation annotations.

- SQA^{I} (Lu et al., 2022): ScienceQA is a new benchmark dataset designed to evaluate the multi-hop reasoning ability and interpretability of AI systems. ScienceQA consists of approximately 21,000 multimodal multiple-choice science questions, covering a variety of scientific topics, and provides annotations of the answers along with corresponding lectures and explanations.

C Experimental Supplement for Inherent Hallucination Pattern

C.1 The Impact of Noise Levels on Inherent Hallucination Pattern in LVLMs

To further demonstrate that noise in the image contributes to activating inherent hallucination patterns, we compare $CHAIR_I$ scores on LLaVA, svit and Vila across different noise levels. The experimental settings align with the hallucination evaluation benchmark CHAIR. As illustrated in Figure 5, it is evident that as noise levels increase, the $CHAIR_I$ scores also tend to rise, indicating a higher occurrence of hallucinations.

D Fine-grained Performance Analysis

Table 5 presents a fine-grained performance analysis of different preference collection strategies on the LLaVA-Bench benchmark. This analysis encompasses a spectrum of multi-modal reasoning and perception dimensions, such as Conversation, Detail Description, and Complex Reasoning. According to Table 5, it is evident that, when compared with other preference data collection approaches, POVID excels in image captioning and providing detailed descriptions for a given image. This outcome aligns with our expectations, as our training data includes various long-form captions, and such comprehensive preference comparisons result in improved alignment and stronger image captioning results.

Table 7: Two types of prompts to GPT4V (The format of the obtained data is {image, prefer data, disprefer data}).

Prompts for hallucinating image captioning tasks: Help me generate one highly confusing response based on the image and the standard caption in the Question-Answer Pair. ** Question-answer Pair: Q: {question} A: {answer} Requirements: (1) The generated caption is generally similar to the given A, with the same main meaning; (2) You can refer to the following errors to generate the wrong caption (1. The wrong caption can contain some co-occurring objects, which are prone to appear in such scenarios but do not appear in the image; 2. The wrong caption can be an error in the number of entities or the logical relationships between entities; 3. The attributes of entities in the caption can also be modified, such as color, appearance, etc.) (3) Compared to the original caption A, the caption you modified is incorrect based on the provided image. ** Output Format: Answer: your answer
Prompts for hallucinating reasoning tasks: Now, please help me generate new answers with hallucination errors based on the image, question, and answer provided. There are two cases now: 1. If the given question and answer are short and do not require logical reasoning, then modify the answer to a hallucination error answer, such as some quantity errors or entity and property errors. 2. If the entire question requires logical reasoning, then help me reorganize the answers based on the given image, questions, and answers into the format “Reason: xxx, Result: xxx" (Answer 1). Modify the reasons by introducing errors related to logical relationships, entity information, entity attributes, etc. If the error in the reason would lead to a new result, modify the result accordingly. If the error does not lead to a new result, keep the original result. Similarly, organize it in the format “Reason: xxx, Result: xxx" (Answer 2). ** Question-answer Pair: Q: {question} A: {answer} Requirements: (1) The generated wrong answer and reasoning process should be combined with the image and be misleading.. ** Output Format: Answer: your answer