AdPO: Enhancing the Adversarial Robustness of Large Vision-Language Models with Preference Optimization

Anonymous authors
Paper under double-blind review

Abstract

Large Vision-Language Models (LVLMs), such as GPT-4o and LLaVA, have recently witnessed remarkable advancements and are increasingly being deployed in real-world applications. However, inheriting the sensitivity of visual neural networks, LVLMs remain vulnerable to adversarial attacks, which can result in erroneous or malicious outputs. While existing efforts utilize adversarial fine-tuning to enhance robustness, they often suffer from significant performance degradation on clean inputs. In this paper, we propose AdPO, a novel adversarial defense strategy for LVLMs based on preference optimization. For the first time, we reframe adversarial training as a preference optimization problem, aiming to enhance the model's preference for generating normal outputs on clean inputs while rejecting the potential misleading outputs for adversarial examples. Notably, AdPO achieves this by solely modifying the image encoder, e.g., CLIP ViT, resulting in superior clean and adversarial performance in a variety of downstream tasks. Due to the computational cost of training large language models, we show that training on smaller LVLMs and transferring to larger ones achieves state-of-the-art performance with efficiency comparable to previous methods. Our comprehensive experiments confirm the effectiveness of the proposed AdPO which highlights the potential of preference-based learning in adversarially robust multimodal systems.

1 Introduction

The emergence of Large Vision-Language models (LVLMs) has substantially propelled the development of general artificial intelligence, attracting considerable attention from the AI community (Yin et al., 2023; Cui et al., 2024; Liu et al., 2024c). These models generally consist of two key components: visual modules and Large Language Models (LLMs) (Zhao et al., 2023a). The visual modules, frequently utilizing pre-trained image encoders like CLIP's ViT (Radford et al., 2021), are responsible for extracting salient visual features from images and projecting them onto the input space of the language model. This alignment facilitates the next-token prediction in an autoregressive manner within the framework of the language model. Cutting-edge LVLMs, such as Qwen2.5-VL (Bai et al., 2025) and LLaVA (Liu et al., 2023), have demonstrated outstanding capabilities in understanding and reasoning with both visual and textual information. These models have delivered exceptional performance across a broad range of tasks, such as image captioning (Nguyen et al., 2023), visual question answering (Liu et al., 2024b), and text recognition (Liu et al., 2024a; Cao et al., 2023).

Given their transformative potential in multimodal learning and understanding, LVLMs are increasingly being deployed across a diverse range of real-world applications. However, this widespread deployment raises significant security concerns, as malicious adversaries can exploit vulnerabilities in LVLMs to induce undesirable outputs and hallucinations (Schlarmann & Hein, 2023; Shayegani et al., 2024; Wang et al., 2024e). Consequently, it is imperative to rigorously test and improve the robustness of these models prior to deployment. Recent research has identified a critical vulnerability in LVLMs to adversarial attacks targeting both textual and visual inputs (Zhao et al., 2023b). Notably, the continuous nature of the visual modality renders it more susceptible to manipulation via numerical optimization techniques (Wang et al., 2024c; Carlini et al., 2023; Qi et al., 2024b; Luo et al., 2024a). Researchers disrupt the understanding of LVLMs by injecting imperceptible noise into images, thereby enabling both targeted and untargeted adversarial attacks.

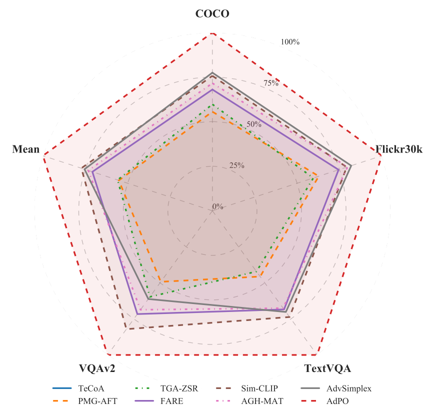

Figure 1: AdPO achieves a significant improvement in clean performance compared with previous methods.

To improve the adversarial robustness of LVLMs, two main training paradigms have been explored: multimodal contrastive learning and generative pre-training. Multimodal contrastive learning methods (e.g., FARE (Schlarmann et al., 2024) and TeCoA (Mao et al., 2023)) align the features of adversarial images with those of text to obtain a robust image encoder, which can then be transferred to LVLMs. This approach is computationally efficient but often fails to achieve fine-grained alignment. In contrast, generative pre-training leverages the full LVLM, enabling finer-grained alignment, but generally suffers from poor generalization, which in turn degrades clean performance (Chu et al., 2025).

Inspired by the significant success of preference optimization in the LLM community (Wang et al., 2024g; Ouyang et al., 2022), we identify that applying preference optimization to adversarial training is highly promising, given the alignment between their objectives. More specifically, adversarial training aims to enhance model robustness against adversarial attacks while preserving performance on clean data. Preference optimization, such as DPO (Rafailov et al., 2023), aligns LLMs with human values by increasing the probability of preferred outputs while decreasing the likelihood of non-preferred ones. Building on this insight, we propose **AdPO**, a novel **A**dversarial **d**efense strategy based on **P**reference **O**ptimization, which enables LVLMs to generate correct outputs from clean image inputs while rejecting misleading outputs from adversarial images.

However, applying DPO to adversarial training presents non-trivial challenges. In comparison to standard offline DPO, we introduce key improvements: (1) We extend DPO from an offline to an online setting to eliminate the reliance on image annotations. In this framework, the policy model generates interpretations for both clean and adversarial images, which are then used as sources of positive and negative samples. (2) We propose **Preferred Image Optimization** (PIO), which simultaneously increases the probability of producing correct outputs under clean inputs while reducing erroneous outputs under adversarial images. This leads to a significant improvement in clean performance, as illustrated in Figure 1. (3) We propose **Adversarial Image Optimization** (AIO), which leverages dynamic fine-tuning to explicitly optimize the probability of producing correct responses under adversarial inputs, thereby mitigating the potential multimodal unconditional preference issue (Wang et al., 2024a).

Another potential concern is computational efficiency. Directly training a commonly used LVLM model, such as LLaVA-7B Liu et al. (2024b), may be prohibitively expensive in resource-constrained scenarios. In this paper, **we explore fine-tuning the image encoder of a smaller LVLM and subsequently transferring it to a larger LVLM model.** This strategy not only achieves high computational efficiency and mitigates the risk of potential overfitting during evaluation, but also enables a fair comparison with prior CLIP-based approaches.

By constraining our adversarial training to modifying only the CLIP ViT parameters on the ImageNet dataset (Deng et al., 2009), extensive evaluations demonstrate that our proposed AdPO produces a more robust image encoder while maintaining almost intact clean performance. These findings not only validate the effectiveness of our approach but also extend the applicability of preference optimization techniques beyond their traditional use in language models.

In summary, our contributions are as follows:

- We introduce **AdPO** (Adversarial defense based on Preference Optimization), which, to the best of our knowledge, is the first attempt to explore the application of preference optimization for adversarial training.

- We propose a dual strategy combining **Preferred Image Optimization (PIO)** and **Adversarial Image Optimization (AIO)** to preserve the model's clean performance while enhancing its adversarial robustness. This serves as a general adversarial training framework that is not restricted to any specific algorithm or model.

- We validate the feasibility of conducting adversarial **training on smaller LVLMs and subsequently transferring it to larger models**, which reduces computational costs and mitigates potential overfitting during evaluation.
- We conduct extensive experiments on multiple vision-language tasks and datasets using various models and the results show that our method consistently achieves state-of-the-art performance.

2 Related Work

In this section, we primarily review the related studies on large vision-language models, adversarial attacks, adversarial defenses, and preference optimization methods.

Large Vision-Language Models. Recently, large multimodal models have emerged, including LLaVA 1.5 (Liu et al., 2024b), OpenFlamingo (OF) (Awadalla et al., 2023), BLIP-2 (Li et al., 2023b), MiniGPT-4 (Zhu et al., 2024), Otter (Li et al., 2023a), mPLUG-Owl (Ye et al., 2023), Qwen-VL (Bai et al., 2023), MiniCPM-V (Yao et al., 2024), DeepSeek-VL (Lu et al., 2024), InternVL (Chen et al., 2024), and Idefics2 (Laurençon et al., 2024). These models typically use pre-trained image encoders (e.g., CLIP or SigCLIP) to extract image features, which are then aligned with text embedding spaces (Radford et al., 2021; Zhai et al., 2023). The visual and textual embeddings are then fed into LLMs for autoregressive generation. This approach allows the model to simultaneously understand and generate content related to both images and text. To mitigate computational load, a practical strategy is to freeze the image encoder and train only the projection layer, which not only simplifies the training process but also enhances efficiency (Liu et al., 2023; Awadalla et al., 2023). Therefore, image encoders can significantly impact the performance of LVLMs, receiving significant attention from the multimodal community (Cao et al., 2023; Zhou et al., 2024). We mainly focus on evaluating the performance of LLaVA-1.5 and OpenFlamingo, as both adopt CLIP ViT-L/14 (Radford et al., 2021) as their image encoder, while additionally assessing our method on Qwen-2.5-VL (Bai et al., 2025), a non-CLIP-based model, for further validation.

Adversarial attacks. The vulnerability of visual neural network models to adversarial attacks is well-established and has been extensively investigated (Szegedy et al., 2014; Goodfellow et al., 2015; Madry et al., 2018; Brown et al., 2017; Zhang et al., 2023; 2024; Zhou et al., 2023). By introducing carefully crafted noise into images, adversaries can cause the victim model to generate incorrect outputs with high confidence. Capitalizing on this vulnerability, recent studies have shown that LVLMs are also vulnerable to attacks targeting visual inputs (Schlarmann & Hein, 2023; Shayegani et al., 2024; Luo et al., 2024a; Gao et al., 2024; Dong et al., 2023b). Zhao *et al.* (Zhao et al., 2023b) showed that transferable black-box attacks could be generated using text-to-image models and other work (Carlini et al., 2023) demonstrated how adding adversarial noise to images can circumvent safety constraints of LLMs. Qi *et al.* (Qi et al., 2024a) explored how adversarial attacks embedding deceptive information into images can mislead LVLMs and deceive users. The widespread deployment of LVLMs has raised urgent security concerns due to the threat of adversarial attacks.

Adversarial defenses. Adversarial defenses in machine learning safeguard models from malicious inputs to ensure their integrity and reliability, especially in security-sensitive contexts (Madry et al., 2018; Fares et al., 2024; Papernot et al., 2016; Zhou & Patel, 2022; Luo et al., 2024b; Ledda et al., 2024; Debbi, 2024; Xue et al., 2024; Zhao et al., 2024; Liang et al., 2024; Li et al., 2024; Li & Li, 2024; Hotegni & Peitz, 2024; Jiang et al., 2024). For example, Detectors (Huang et al., 2024; Mumcu & Yilmaz, 2024; Mavali et al., 2024; Roth et al., 2019; Xu et al., 2018; Meng & Chen, 2017; Metzen et al., 2017) identify and filter out adversarial examples, but these external modules can introduce additional inference time and may also obstruct normal inputs. Purification methods (Samangouei et al., 2018; Nie et al., 2022; Ho & Vasconcelos, 2022; Das et al., 2018) use techniques such as diffusion models to eliminate adversarial perturbations in input data, and this can also modify the input, thus affecting performance. Adversarial training (Kurakin et al., 2017b; Tramèr et al., 2018; Dong et al., 2023a; Liu & Chen, 2024; Jia et al., 2024b; Lv et al., 2024; Palma et al., 2024; Dong et al., 2024; RIbeiro et al., 2024; Jia et al., 2022) is a foundational method for enhancing a model's inherent robustness by integrating adversarial examples into the training dataset. In the multimodal field (Wang et al., 2024b), recent research has predominantly concentrated on enhancing the adversarial robustness of CLIP-based models in zero-shot classification tasks. For example, TeCoA (Mao et al., 2023) applies text-guided adversarial training, while AdvXL (Wang et al., 2024f) leverages large-scale training data. TGA-ZSR (Yu et al., 2024a) introduces a text-guided attention mechanism to further strengthen robustness under zero-shot settings. FARE (Schlarmann et al., 2024) enhances the robustness of LVLMs by minimizing the representation distance between clean and adversarial images in CLIP, and transferring the CLIP image encoder to models such as LLaVA. Despite these advances, a persistent challenge remains: the clean performance of LVLMs still suffers a significant drop.

Preference optimization. Preference optimization has emerged as a novel training paradigm for aligning LLMs with human values and has garnered significant attention in recent research (Ouali et al., 2024; Yu et al., 2023; 2024b; Wang et al., 2024a;d). Reinforcement Learning from Human Feedback (RLHF) utilizes human preferences as a reward model and applies reinforcement learning to guide model training (Bai et al., 2022; Ouyang et al., 2022) Direct Preference Optimization (DPO) streamlines the training process by increasing the log probability of preferred samples while reducing that of non-preferred samples, enabling broader applications (Rafailov et al., 2023). Subsequent advancements, such as StepDPO (Lai et al., 2024), SimPO (Meng et al., 2025), and IPO (Azar et al., 2024), have further improved DPO's performance. Considering its stability and efficiency in training, we also adopt DPO for adversarial training of LVLMs in this work.

3 Method

This section provides a detailed introduction to our AdPO, with its overall framework illustrated in Figure 2. First, Section 3.1 outlines the basics of the DPO algorithm, and Section 3.2 discusses adversarial example generation, which forms the preference sample pairs required for DPO. Sections 3.3 and 3.4 introduce preferred image optimization and adversarial image optimization, respectively.

3.1 Preliminaries

DPO has emerged as a prominent method in the domain of offline preference optimization. This method provides a novel framework for optimizing language models in accordance with human preferences. In a typical setup, given an input x and an output text y, a language model (i.e., policy model) π_θ generates a conditional distribution $\pi_\theta(y|x)$. Unlike RLHF, which employs an explicit reward model, DPO reformulates the reward function using a closed-form expression with respect to the optimal policy. The main objective of DPO is to maximize the expected reward of the outputs generated by this policy, with the reward function defined as $r(x, y)$:

$$r(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x) \quad (1)$$

where β is a constant, π_{ref} is the reference policy model (identical to the original π_θ), and $Z(x)$ is the partition function.

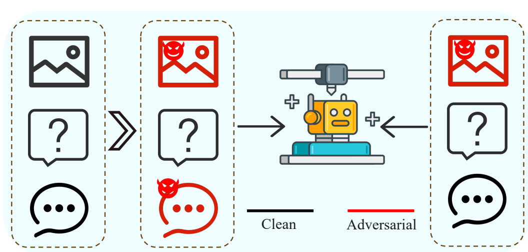

Figure 2: The architecture of our proposed AdPO. AdPO mainly consists of two parts: (**left**) preferred image optimization and (**right**) adversarial image optimization. Preferred image optimization incorporates both clean and adversarial images into adversarial training while maintaining the model's performance on clean inputs, and adversarial image optimization can significantly enhance the model's adversarial robustness.

Given a preference dataset $\mathcal{D} = \{x, y_w, y_l\}$, where y_w and y_l represent the winning and losing responses respectively, DPO employs a Bradley-Terry model (Bradley & Terry, 1952) to express the probability for each preference pair:

$$p(y_w \succ y_l) = \sigma(r(x, y_w) - r(x, y_l)) \quad (2)$$

where $\sigma(\cdot)$ is typically defined as a sigmoid function. The key innovation of DPO is its formulation of the likelihood of preference data using the policy model, as opposed to relying on an explicit reward model. This leads to the formulation of the DPO objective:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right] \quad (3)$$

This formulation captures the core principles of DPO, providing a robust framework for optimizing language models in alignment with human preferences.

3.2 Adversarial Example Generation

In the context of LVLMs, the input to the model comprises $x = \{x_m, x_{text}\}$, where x_m denotes the image input and x_{text} represents the text input. This section outlines the principles behind generating adversarial images.

Adversarial images are generated by introducing small, nearly imperceptible perturbations to original images, with the goal of deceiving machine learning models and inducing incorrect predictions. Although adversarial images appear nearly identical to the original images to humans, they can drastically alter the model's output, exposing its vulnerability to malicious inputs (Kurakin et al., 2017a). Adversarial attacks can be broadly categorized into targeted and untargeted attacks: targeted attacks compel the model to produce specific outputs (Luo et al., 2024a), whereas untargeted attacks merely lead the model to generate incorrect outputs (Wang et al., 2024e; Gao et al., 2024). In this study, we employ untargeted attack methods to generate adversarial images for three reasons: (1) They eliminate dependence on specifically labeled datasets and do not rely on the text encoder, enabling our method to generalize to unseen datasets (Schlarmann et al., 2024; Yu et al., 2024a). (2) Untargeted attacks typically achieve a high success rate, allowing the stable generation of negative adversarial samples during training (Cui et al., 2023). (3) Their broader attack capability enhances the model's resilience against various types of adversarial attack methods (Wang et al., 2024e).

Given an image encoder ϕ (e.g., CLIP ViT) and a clean image x_m, adversarial examples are generated by optimizing to maximize the discrepancy between the encoded features of the adversarial image and the clean image:

$$x_{adv} = \underset{\|x_{adv}-x_m\|_\infty \le \varepsilon}{\arg\max} \|\phi(x_{adv}) - \phi_{\text{org}}(x_m)\|_2^2 \tag{4}$$

where x_{adv} is the adversarial image obtained through iterative optimization like PGD (Madry et al., 2018), ϕ_{org} is the original image encoder and ϵ is the image perturbation magnitude. This approach has been widely adopted in prior work, such as FARE and TGA-ZSR, and we also employ it to ensure a fair comparison. Note that in subsequent adversarial training, the parameters of ϕ will be updated.

3.3 Preferred Image Optimization

This section primarily delineates the methodology for constructing pairs of preferred and non-preferred samples from unlabeled image data, a fundamental step in the DPO training pipeline.

Model Selection. Compared to previous methods (Mao et al., 2023; Yu et al., 2024a; Wang et al., 2024b) that rely solely on CLIP's image and text encoders, AdPO utilizes the entire LVLM model. Using a commonly adopted model such as LLaVA-7B would result in high computational costs. To address this, we construct TinyLLaVA[1], which leverages OpenELM-450M-Instruct (Mehta et al., 2024) as its language model. This lightweight LVLM not only achieves training efficiency comparable to previous approaches but also mitigates potential overfitting during evaluation.

Given a clean image x_m and its adversarial image x_{adv}, we employ an online approach to directly prompt the model (e.g., *"What is the content of the image?"*) to generate interpretations, thereby obtaining the preferred response y_w and the non-preferred response y_l. Complete prompts are provided in Appendix B. Accordingly, in the setting of multimodal adversarial training, our preferred image optimization can be formulated as:

$$\mathcal{L}_{\text{PIO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x_m, x_{text})\sim\mathcal{D}} \\ \log\sigma\left(\beta \log \frac{\pi_\theta(y_w \mid x_m, x_{text})}{\pi_{\text{ref}}(y_w \mid x_m, x_{text})} - \beta \log \frac{\pi_\theta(y_l \mid x_{adv}, x_{text})}{\pi_{\text{ref}}(y_l \mid x_{adv}, x_{text})}\right) \tag{5}$$

This straightforward approach presents several advantages. First, it removes the need for data annotation, thus facilitating its application to previously unseen image data. Second, this method resembles semi-supervised learning, especially as LVLMs now possess advanced capabilities, enabling them to incorporate labeled images into their training data. Moreover, allowing the model to generate its own text as labels effectively mitigates distribution shift issues, thus concentrating attention on the adversarial images themselves (Li et al., 2023c).

[1]https://github.com/TinyLLaVA/TinyLLaVA_Factory

Notably, this optimization does not presuppose that negative samples are always incorrect. The core idea of DPO is its relative objective: encouraging the model to prefer certain responses over others based on comparative judgments (Rafailov et al., 2023). In the extreme case where positive and negative samples are indistinguishable, no relative preference exists, and thus no model update is applied. Given the rapid development of preference optimization algorithms, we will evaluate the performance of DPO variants in experiments to assess the adaptability of AdPO.

3.4 Adversarial Image Optimization

While Preferred Image Optimization can maintain the performance of LVLMs on clean inputs, it remains insufficient to reach the optimal adversarial robustness. We identify two fundamental limitations. First, recent work has shown that multimodal DPO can be dominated by language-only preferences, causing the model to disregard visual conditions, a failure mode termed "unconditional preferences" that leads to hallucinations and suboptimal performance (Wang et al., 2024a). Second, as formulated in Eq. 5, the optimization objective focuses on maintaining clean outputs under clean inputs and rejecting harmful responses under adversarial inputs, yet fails to explicitly encourage the generation of correct outputs when adversarial perturbations are present. These limitations hinder the attainment of truly robust performance.

To address this gap, we introduce **Adversarial Image Optimization (AIO)**, which explicitly encourages the model to produce correct outputs under adversarial inputs. The most straightforward approach is to apply Supervised Fine-Tuning (SFT) to optimize the objective:

$$\mathcal{L}_{\mathrm{SFT}}(\pi_\theta) = -\mathbb{E}_{(x_m, x_{text})\sim\mathcal{D}}\left[\log \pi_\theta\left(y_w \mid x_{adv}, x_{text}\right)\right] \tag{6}$$

However, a growing body of recent work shows that SFT tends to overfit the objective, thereby significantly reducing the model's generalization ability (Chu et al., 2025; Wu et al., 2025). To mitigate this issue, we employ dynamic fine-tuning, whose core idea is to adjust the token-level loss based on the model's confidence (Wu et al., 2025):

$$\begin{aligned} \mathcal{L}_{\mathrm{AIO}}(\pi_\theta) &= -\mathbb{E}_{(x_m, x_{text})\sim\mathcal{D}}\left[\mathrm{sg}\left(\pi_\theta(y_w \mid x_{adv}, x_{text})\right)\log \pi_\theta(y_w \mid x_{adv}, x_{text})\right] \\ &= -\mathbb{E}_{(x_m, x_{text})\sim\mathcal{D}}\left[\sum_{t=1}^{|y_l|}\mathrm{sg}\left(\pi_\theta(y_w^t \mid y_w^{<t}, x_{adv}, x_{text})\right)\log \pi_\theta(y_w^t \mid y_w^{<t}, x_{adv}, x_{text})\right] \end{aligned} \tag{7}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator and y_w^t denotes the t-th token of y_w. By increasing the weight on high-confidence predictions, AIO explicitly enhances adversarial robustness while minimally affecting generalization.

Based on the analysis above, the final objective of AdPO is a combination of preferred image optimization and adversarial image optimization:

$$\mathcal{L}_{\mathrm{AdPO}} = \mathcal{L}_{\mathrm{PIO}} + \lambda\mathcal{L}_{\mathrm{AIO}}, \tag{8}$$

where λ is the scaling factor that balances the two loss terms. By leveraging joint optimization, AdPO attains enhanced adversarial robustness while maintaining its performance on clean samples.

4 Experiments

In this section, we conduct extensive experiments to evaluate the performance of AdPO on various LVLM tasks. For a more comprehensive evaluation, please refer to the Appendix.

Models. To facilitate a thorough comparison with prior work, we focus on CLIP-based models in the main text. For training, we adopt TinyLLaVA (Jia et al., 2024a), which pairs CLIP's ViT-L/14 image encoder with the OpenELM-450M-Instruct language model. This lightweight setup maintains computational efficiency comparable to prior methods while mitigating potential overfitting during evaluation. For evaluation, we primarily use LLaVA-1.5-7B(Liu et al., 2024b), a model widely adopted in the multimodal community. To show that our approach generalizes beyond CLIP-based models, we also evaluate **Qwen2.5-VL-7B**(Bai et al., 2025) (improved self-attention encoder), **InternVL3.5-8B**(Wang et al., 2025) (InternViT encoder), and **BLIP-2-12B**(Li et al., 2023b) (EVA-CLIP ViT encoder (Sun et al., 2023)). Results for these models are provided in Appendix E.

Table 1: Comparison of our proposed AdPO with prior methods under untargeted attacks. We evaluate the clean performance and adversarial robustness of various methods across multiple tasks. The results indicate that AdPO significantly exceeds our baseline methods, attaining outstanding robustness along with exceptional clean performance. The best results are shown in **bold**.

Method	COCO			Flickr30k			TextVQA			VQAv2		
	clean	ℓ_∞		clean	ℓ_∞		clean	ℓ_∞		clean	ℓ_∞	
		$2/255$	$4/255$		$2/255$	$4/255$		$2/255$	$4/255$		$2/255$	$4/255$
CLIP	115.5	4.0	3.1	77.5	1.6	1.0	37.1	0.5	0.0	74.5	2.9	0.0
TeCoA	98.4	44.2	30.3	57.1	23.2	15.3	24.1	12.1	8.8	66.9	33.8	21.8
PMG-AFT	107.8	56.1	30.5	68.9	28.1	18.2	29.3	14.9	8.5	70.2	34.5	23.9
TGA-ZSR	108.5	55.6	31.1	68.3	28.6	17.7	28.9	14.5	8.7	70.9	35.1	23.1
FARE	109.9	53.6	31.0	71.1	29.5	17.5	31.9	14.7	9.1	71.7	34.9	23.0
Sim-CLIP	111.2	54.5	31.8	72.0	30.1	18.2	32.5	15.3	9.6	72.4	35.5	23.8
AGH-MAT	110.5	57.2	29.9	72.1	29.4	19.5	31.8	16.1	9.2	71.5	36.2	24.5
AdvSimplex	111.5	55.8	32.6	72.5	31.2	18.9	32.1	15.9	10.0	71.0	38.4	26.1
AdPO	**115.3**	**68.9**	**47.6**	**75.9**	**38.6**	**27.9**	**35.5**	**24.2**	**17.6**	**73.6**	**52.3**	**37.6**

Adversarial training settings. For fair comparison, we train on ImageNet (Deng et al., 2009) using an online learning approach that relies solely on images without category labels. Adversarial perturbations are generated via 10-step PGD under the ℓ_∞ norm by optimizing Equation 4. To balance robustness and clean accuracy, we apply perturbation radii $\epsilon = 2/255$. λ is set to 1 by default. We use the AdamW optimizer with a weight decay of 1e-4 and a learning rate of 1e-5. We conduct training for two epochs with a batch size of 128. The preference parameter β is set to 0.1.

Baseline methods. Given the limited prior work on enhancing adversarial robustness of LVLMs, and to fully demonstrate the advantages of our proposed method, we conduct extensive comparisons in the main text against CLIP-based adversarial training approaches, including TeCoA (Mao et al., 2023), FARE (Schlarmann et al., 2024), Sim-CLIP (Hossain & Imteaj, 2024), PMG-AFT (Wang et al., 2024b), TGA-ZSR (Yu et al., 2024a), AGH-MAT (Chen et al., 2025), and AdvSimplex (Dong et al., 2025). To ensure fair comparison, we use adversarial images with the same noise radius for training. Note that AdPO does not benefit from broader optimization, allowing for a fair comparison with previous methods, as neither has been exposed to the final language model.

4.1 Evaluation of Untargeted Attacks on LVLMs

Attack setup. We utilize the approach outlined in Schlarmann & Hein (2023) to perform untargeted attacks aimed at degrading the model's performance. Given that attacks on LVLMs often demand more iterations, we employ a 100-step APGD attack (Croce & Hein, 2020), which utilizes ground-truth captions as labels. After each attack, we discard samples with scores below a specified threshold to ensure that computationally expensive attacks are only performed when necessary, following Schlarmann et al. (2024). Further details are provided in Appendix A.

Datasets and metrics. We utilize a variety of datasets for image captioning tasks, including COCO (Lin et al., 2014) and Flickr30k (Plummer et al., 2015), as well as for visual question answering tasks, such as VQAv2 (Goyal et al., 2017) and TextVQA (Singh et al., 2019). Considering that adversarial attacks are time-consuming and costly, we randomly selected 500 images for evaluation. We employ the CIDEr score (Vedantam et al., 2015) for image captioning and VQA accuracy (Antol et al., 2015) for visual question answering tasks to present our results.

As presented in Table 1, our proposed AdPO method sets a new state-of-the-art in adversarial robustness across all evaluated tasks. It consistently and substantially outperforms all baseline methods, marking a significant advance in adversarial defense. Crucially, AdPO achieves this enhanced robustness while preserving performance on benign inputs. The method incurs only a negligible performance drop on clean data compared to the original CLIP baseline, effectively avoiding the typical trade-off between robustness and clean accuracy. Furthermore, AdPO shows excellent generalization from weaker to stronger attacks. Despite being trained only on perturbations with an budget of $2/255$, it maintains superior robustness against larger, unseen perturbations of $\epsilon = 4/255$. To further underscore its resilience, we include evaluations against even larger perturbations ($8/255$ and $16/255$) in Appendix F. We also find that evaluating directly on TinyLLaVA yields even

Table 2: Quantitative evaluation of targeted attacks at $\epsilon = {}^4/_{255}$ radii. We assess the Attack Success Rate (ASR) for each setup.

Target	CLIP	TeCoA	PMG	TGA	FARE	AGH	Adv.	AdPO
`A group of people ...`	20/20	1/20	1/20	2/20	1/20	0/20	0/20	0/20
`a bunch of people ...`	20/20	1/20	0/20	1/20	1/20	1/20	1/20	0/20
`The pizza on the table...`	20/20	2/20	2/20	0/20	0/20	0/20	0/20	0/20
`An earthquake is about...`	20/20	2/20	1/20	1/20	1/20	1/20	0/20	0/20
`This patient needs ...`	20/20	0/20	1/20	1/20	0/20	1/20	2/20	0/20
Mean ASR:	100%	4%	5%	5%	3%	3%	3%	**0%**

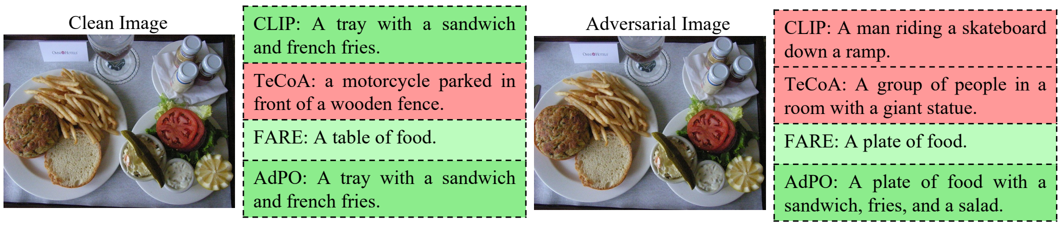

Figure 3: Qualitative assessment of targeted attacks on LLaVA. (**Left**) When encountering clean images, TeCoA may exhibit noticeable errors, which is undesirable in adversarial defense, while FARE and AdPO demonstrate better clean performance. (**Right**) When faced with adversarial images, the original LLaVA is easily compromised, FARE shows some adversarial robustness but loses more details or makes subtle errors, whereas AdPO performs better.

more significant improvements (Appendix G), which can be attributed to the direct alignment between visual and language representations.

4.2 Evaluation of Targeted Attacks on LVLMs

In contrast to the untargeted attacks discussed in Section 4.1, targeted attacks on LVLMs pose a significantly greater threat. Targeted attacks aim to compel the model to produce specific outputs, with the added noise in the image remaining imperceptible to the user. Through image manipulation, attackers can circumvent the model's security mechanisms, leading it to generate malicious content (Carlini et al., 2023; Niu et al., 2024; Qi et al., 2024b). Additionally, attackers can embed phishing links into images through adversarial attacks to deceive users (Bagdasaryan et al., 2023).

Attack setup. We perform targeted attack experiments on LLaVA-1.5-7B, using the attack success rate (ASR) as the primary evaluation metric. A sample is deemed successfully attacked if the model's output contains the target string. Targeted attacks on LVLMs generally require more iterations, prompting us to execute APGD attacks for 10,000 iterations. Given that larger image perturbations pose more significant threats, we employ ℓ_∞ threat models with a radius of $\epsilon = {}^4/_{255}$. We evaluate five target strings incorporating errors such as incorrect medical diagnoses and fake news, sampling 20 images for each string.

The quantitative evaluation results are presented in Table 2. The attack success rate for the clean version of the CLIP model reaches 100%, underscoring the vulnerability of current vision-language models to visual input and the substantial security risks posed. Although baseline methods exhibit a certain degree of robustness, they still expose considerable vulnerabilities. In contrast, AdPO achieves the strongest robustness, effectively safeguarding the model against malicious attacks. Additional details are provided in Appendix C.

4.3 Further Evaluation

Although we conduct extensive quantitative evaluations above, they are still insufficient for a comprehensive assessment of LVLMs. In this section, we first present a qualitative evaluation, followed by an analysis of other vision-language tasks and the training efficiency.

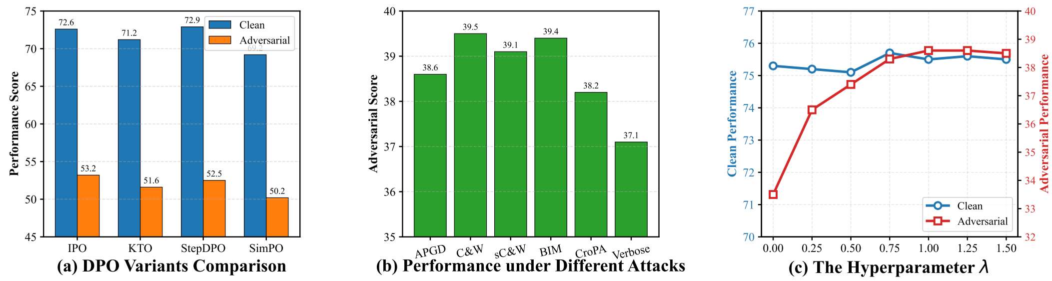

Figure 4: Ablation experiments on AdPO. **(a)** The performance of DPO variants. **(b)** The evaluation of attack types. **(c)** The impact of the parameter λ.

Qualitative evaluation. As depicted in Figure 3, the LLaVA model, using the original CLIP as the encoder, provides the most accurate and detailed understanding of clean images. However, when faced with adversarial images, they are completely vulnerable to successful attacks. TeCoA fails to exhibit robust performance against both clean and adversarial images, whereas FARE experiences a loss of detail or minor errors in image understanding, ultimately falling short of optimal performance. In the absence of adversarial defenses, LLaVA is susceptible to manipulation, resulting in biased outputs that can mislead users and have detrimental effects.

Recent work has shown that LVLMs are prone to hallucinations and are more susceptible to jailbreak attacks compared to purely language models (Qi et al., 2024a; Li et al., 2023d). Additional experimental evaluations presented in Appendix D demonstrate that our method exhibits better performance in both hallucination reduction and jailbreak prevention. We quantitatively evaluate the average runtime per batch across different methods in Appendix I, demonstrating that AdPO achieves comparable efficiency to previous approaches when trained on a lightweight LVLM.

4.4 Ablation Study

The impact of DPO variants. In Figure 4 (a), we evaluate four commonly used DPO variants to analyze the effectiveness of AdPO. The results show that IPO (Azar et al., 2024), KTO (Ethayarajh et al., 2024), and StepDPO (Lai et al., 2024) perform well, while SimPO (Meng et al., 2025) performs relatively poorly, possibly due to the removal of the reference model. This experiment also demonstrates that AdPO serves as **a general preference framework** for enhancing model robustness, rather than being restricted to a specific algorithm.

Analysis of attack types. In addition to APGD, we further evaluate the impact of other attack methods, including C&W (Carlini & Wagner, 2017), sC&W (Zhang et al., 2020), BIM (Kurakin et al., 2016), CroPA (Luo et al., 2024a), and Verbose (Gao et al., 2024). As shown in Figure 4 (b), our method remains robust even against attacks specifically designed for LVLMs.

The impact of λ. We perform untargeted attacks to evaluate the effectiveness of AdPO trained with diffenrent λ on the Flickr30K dataset, with experimental results shown in Figure 4 (c). We find that the clean performance is largely insensitive to AIO, whereas increasing λ significantly improves adversarial robustness, with the best empirical results achieved around $\lambda = 1$.

We also provide additional ablation studies, including direct adversarial training, SFT-based AIO, and full fine-tuning, in Appendix H.

5 Conclusion

We propose AdPO, the first adversarial defense strategy based on preference optimization. It jointly optimizes the model's outputs on both clean and adversarial images, thereby better preserving clean performance under adversarial training. Unlike previous adversarial fine-tuning methods, which typically only impose single-target constraints to improve adversarial robustness, leading to a loss of clean performance. Moreover, AdPO does not require labeled image data or the CLIP text encoder, offering greater flexibility. Both quantitative and qualitative analyses demonstrate that our method achieves state-of-the-art results, with particularly significant improvements on generative vision-language understanding tasks.

Ethics Statement

This research did not involve any human subjects, nor did it collect or process personally identifiable information. The core focus of this paper is on adversarial defense for Large Vision-Language Models. While the broader field of adversarial examples includes the creation of attacks, our primary contribution is the development of a robust defense mechanism (AdPO) designed to enhance model security and reliability. By making models more resilient to malicious manipulation, our work aims to mitigate the potential for these systems to generate erroneous, misleading, or harmful content. We believe this research contributes positively to the development of safer and more trustworthy AI, which is crucial for their responsible deployment in real-world applications.

Reproducibility Statement

All experiments were conducted using publicly available models and datasets (LLaVA, COCO, Flickr30k, etc.) and standard evaluation protocols. We provide detailed hyperparameters for our main experiments in Section 4 and Appendix A, B.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2425–2433. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.279. URL `https://doi.org/10.1109/ICCV.2015.279`.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *CoRR*, abs/2308.01390, 2023. doi: 10.48550/ARXIV.2308.01390. URL `https://doi.org/10.48550/arXiv.2308.01390`.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Rémi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 2024. URL `https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html`.

Eugene Bagdasaryan, Tsung-Yin Hsieh, Ben Nassi, and Vitaly Shmatikov. (ab) using images and sounds for indirect instruction injection in multi-modal llms. *arXiv preprint arXiv:2307.10490*, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL `https://arxiv.org/abs/2308.12966`.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL `https://arxiv.org/abs/2502.13923`.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022. doi: 10.48550/ARXIV.2204.05862. URL https://doi.org/10.48550/arXiv.2204.05862.

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

Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *CoRR*, abs/1712.09665, 2017. URL http://arxiv.org/abs/1712.09665.

Haoyu Cao, Changcun Bao, Chaohu Liu, Huang Chen, Kun Yin, Hao Liu, Yinsong Liu, Deqiang Jiang, and Xing Sun. Attention where it matters: Rethinking visual document understanding with selective region concentration. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 19460–19470. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01788. URL https://doi.org/10.1109/ICCV51070.2023.01788.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei Koh, Daphne Ippolito, Florian Tramèr, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/c1f0b856a35986348ab3414177266f75-Abstract-Conference.html.

Long Chen, Yuling Chen, Yun Luo, Hui Dou, and Xinyang Zhong. Attention-guided hierarchical defense for multimodal attacks in vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1607–1617, 2025.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites, 2024. URL https://arxiv.org/abs/2404.16821.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 2020. URL http://proceedings.mlr.press/v119/croce20b.html.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, and Chao Zheng. A survey on multimodal large language models for autonomous driving. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACVW 2024 - Workshops, Waikoloa, HI, USA, January 1-6, 2024*, pp. 958–979. IEEE, 2024. doi: 10.1109/WACVW60836.2024.00106. URL https://doi.org/10.1109/WACVW60836.2024.00106.

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. *CoRR*, abs/2312.03777, 2023. doi: 10.48550/ARXIV.2312.03777. URL https://doi.org/10.48550/arXiv.2312.03777.

Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E. Kounavis, and Duen Horng Chau. SHIELD: fast, practical defense and vaccination for deep learning using JPEG compression. In Yike Guo and Faisal Farooq (eds.), *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pp. 196–204. ACM, 2018. doi: 10.1145/3219819.3219910. URL https://doi.org/10.1145/3219819.3219910.

Hichem Debbi. Causadv: A causal-based framework for detecting adversarial examples, 2024. URL https://arxiv.org/abs/2411.00839.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Junhao Dong, Seyed-Mohsen Moosavi-Dezfooli, Jianhuang Lai, and Xiaohua Xie. The enemy of my enemy is my friend: Exploring inverse adversaries for improving adversarial training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 24678–24687. IEEE, 2023a. doi: 10.1109/CVPR52729.2023.02364. URL https://doi.org/10.1109/CVPR52729.2023.02364.

Junhao Dong, Xinghua Qu, Z. Jane Wang, and Yew-Soon Ong. Enhancing adversarial robustness via uncertainty-aware distributional adversarial training, 2024. URL https://arxiv.org/abs/2411.02871.

Junhao Dong, Piotr Koniusz, Yifei Zhang, Hao Zhu, Weiming Liu, Xinghua Qu, and Yew-Soon Ong. Improving zero-shot adversarial robustness in vision-language models by closed-form alignment of adversarial path simplices. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=WR0ahlhOoy.

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google's bard to adversarial image attacks? *CoRR*, abs/2309.11751, 2023b. doi: 10.48550/ARXIV.2309.11751. URL https://doi.org/10.48550/arXiv.2309.11751.

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

Samar Fares, Klea Ziu, Toluwani Aremu, Nikita Durasov, Martin Takác, Pascal Fua, Karthik Nandakumar, and Ivan Laptev. Mirrorcheck: Efficient adversarial defense for vision-language models. *CoRR*, abs/2406.09250, 2024. doi: 10.48550/ARXIV.2406.09250. URL https://doi.org/10.48550/arXiv.2406.09250.

Kuofeng Gao, Yang Bai, Jindong Gu, Shu-Tao Xia, Philip Torr, Zhifeng Li, and Wei Liu. Inducing high energy-latency of large vision-language models with verbose images. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=BteuUysuXX.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670. URL https://doi.org/10.1109/CVPR.2017.670.

Chih-Hui Ho and Nuno Vasconcelos. DISCO: adversarial defense with local implicit functions. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/96930636e3fb63935e2af153d1cc40a3-Abstract-Conference.html.

Md Zarif Hossain and Ahmed Imteaj. Sim-clip: Unsupervised siamese adversarial fine-tuning for robust and semantically-rich vision-language models. *arXiv preprint arXiv:2407.14971*, 2024.

Sedjro Salomon Hotegni and Sebastian Peitz. Morel: Enhancing adversarial robustness through multi-objective representation learning, 2024. URL https://arxiv.org/abs/2410.01697.

Youcheng Huang, Fengbin Zhu, Jingkun Tang, Pan Zhou, Wenqiang Lei, Jiancheng Lv, and Tat-Seng Chua. Effective and efficient adversarial detection for vision-language models via a single vector, 2024. URL https://arxiv.org/abs/2410.22888.

Junlong Jia, Ying Hu, Xi Weng, Yiming Shi, Miao Li, Xingjian Zhang, Baichuan Zhou, Ziyu Liu, Jie Luo, Lei Huang, and Ji Wu. Tinyllava factory: A modularized codebase for small-scale large multimodal models. *arXiv preprint arXiv:2405.11788*, 2024a.

Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. LAS-AT: adversarial training with learnable attack strategy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 13388–13398. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01304. URL https://doi.org/10.1109/CVPR52688.2022.01304.

Xiaojun Jia, Yong Zhang, Xingxing Wei, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Improving fast adversarial training with prior-guided knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(9):6367–6383, 2024b. doi: 10.1109/TPAMI.2024.3381180. URL https://doi.org/10.1109/TPAMI.2024.3381180.

Chengze Jiang, Junkai Wang, Minjing Dong, Jie Gui, Xinli Shi, Yuan Cao, Yuan Yan Tang, and James Tin-Yau Kwok. Improving fast adversarial training via self-knowledge guidance, 2024. URL https://arxiv.org/abs/2409.17589.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017a. URL https://openreview.net/forum?id=HJGU3Rodl.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017b. URL https://openreview.net/forum?id=BJm4T4Kgx.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *CoRR*, abs/2406.18629, 2024. doi: 10.48550/ARXIV.2406.18629. URL https://doi.org/10.48550/arXiv.2406.18629.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *CoRR*, abs/2405.02246, 2024. doi: 10.48550/ARXIV.2405.02246. URL https://doi.org/10.48550/arXiv.2405.02246.

Emanuele Ledda, Giovanni Scodeller, Daniele Angioni, Giorgio Piras, Antonio Emanuele Cinà, Giorgio Fumera, Battista Biggio, and Fabio Roli. On the robustness of adversarial training against uncertainty attacks, 2024. URL https://arxiv.org/abs/2410.21952.

Binghui Li and Yuanzhi Li. Adversarial training can provably improve robustness: Theoretical analysis of feature learning process under structured data, 2024. URL https://arxiv.org/abs/2410.08503.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *CoRR*, abs/2305.03726, 2023a. doi: 10.48550/ARXIV.2305.03726. URL https://doi.org/10.48550/arXiv.2305.03726.

Fengpeng Li, Kemou Li, Haiwei Wu, Jinyu Tian, and Jiantao Zhou. Dat: Improving adversarial robustness via generative amplitude mix-up in frequency domain, 2024. URL https://arxiv.org/abs/2410.12307.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023b. URL https://proceedings.mlr.press/v202/li23q.html.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *CoRR*, abs/2312.10665, 2023c. doi: 10.48550/ARXIV.2312.10665. URL https://doi.org/10.48550/arXiv.2312.10665.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 292–305. Association for Computational Linguistics, 2023d. doi: 10.18653/V1/2023.EMNLP-MAIN.20. URL https://doi.org/10.18653/v1/2023.emnlp-main.20.

Yuhan Liang, Yijun Li, Yumeng Niu, Qianhe Shen, and Hangyu Liu. A hybrid defense strategy for boosting adversarial robustness in vision-language models, 2024. URL https://arxiv.org/abs/2410.14911.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Chaohu Liu, Kun Yin, Haoyu Cao, Xinghua Jiang, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, and Linli Xu. HRVDA: high-resolution visual document assistant. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 15534–15545. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.01471. URL https://doi.org/10.1109/CVPR52733.2024.01471.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26286–26296. IEEE, 2024b. doi: 10.1109/CVPR52733.2024.02484. URL https://doi.org/10.1109/CVPR52733.2024.02484.

Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. Safety of multimodal large language models on images and texts, 2024c. URL https://arxiv.org/abs/2402.00357.

Zhen-Ting Liu and Shang-Tse Chen. Trap-mid: Trapdoor-based defense against model inversion attacks, 2024. URL https://arxiv.org/abs/2411.08460.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding. *CoRR*, abs/2403.05525, 2024. doi: 10.48550/ARXIV.2403.05525. URL https://doi.org/10.48550/arXiv.2403.05525.

Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024a. URL https://openreview.net/forum?id=nc5GgFAvtk.

Rui Luo, Jie Bao, Zhixin Zhou, and Chuangyin Dang. Game-theoretic defenses for robust conformal prediction against adversarial attacks in medical imaging, 2024b. URL https://arxiv.org/abs/2411.04376.

Kangtao Lv, Huangsen Cao, Kainan Tu, Yihuai Xu, Zhimeng Zhang, Xin Ding, and Yongwei Wang. Hyper adversarial tuning for boosting adversarial robustness of pretrained large vision models, 2024. URL https://arxiv.org/abs/2410.05951.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=P4bXCawRi5J.

Sina Mavali, Jonas Ricker, David Pape, Yash Sharma, Asja Fischer, and Lea Schönherr. Fake it until you break it: On the adversarial robustness of ai-generated image detectors, 2024. URL https://arxiv.org/abs/2410.01574.

Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Seyed Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, et al. Openelm: An efficient language model family with open training and inference framework. In *Workshop on Efficient Systems for Foundation Models II@ ICML2024*, 2024.

Dongyu Meng and Hao Chen. Magnet: A two-pronged defense against adversarial examples. In Bhavani Thuraisingham, David Evans, Tal Malkin, and Dongyan Xu (eds.), *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*, pp. 135–147. ACM, 2017. doi: 10.1145/3133956.3134057. URL https://doi.org/10.1145/3133956.3134057.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2025.

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJzCSf9xg.

Furkan Mumcu and Yasin Yilmaz. Detecting adversarial examples, 2024. URL https://arxiv.org/abs/2410.17442.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/45e604a3e33d10fba508e755faa72345-Abstract-Datasets_and_Benchmarks.html.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16805–16827. PMLR, 2022. URL https://proceedings.mlr.press/v162/nie22a.html.

Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. Jailbreaking attack against multimodal large language model. *CoRR*, abs/2402.02309, 2024. doi: 10.48550/ARXIV.2402.02309. URL https://doi.org/10.48550/arXiv.2402.02309.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms, 2024. URL https://arxiv.org/abs/2408.10433.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Alessandro De Palma, Serge Durand, Zakaria Chihani, François Terrier, and Caterina Urban. On using certified training towards empirical robustness, 2024. URL https://arxiv.org/abs/2410.01617.

Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pp. 582–597. IEEE Computer Society, 2016. doi: 10.1109/SP.2016.41. URL https://doi.org/10.1109/SP.2016.41.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2641–2649. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.303. URL https://doi.org/10.1109/ICCV.2015.303.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 21527–21536. AAAI Press, 2024a. doi: 10.1609/AAAI.V38I19.30150. URL https://doi.org/10.1609/aaai.v38i19.30150.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 21527–21536. AAAI Press, 2024b. doi: 10.1609/AAAI.V38I19.30150. URL https://doi.org/10.1609/aaai.v38i19.30150.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning*

Research, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Antônio H. RIbeiro, Thomas B. Schön, Dave Zahariah, and Francis Bach. Efficient optimization algorithms for linear adversarial training, 2024. URL https://arxiv.org/abs/2410.12677.

Kevin Roth, Yannic Kilcher, and Thomas Hofmann. The odds are odd: A statistical test for detecting adversarial examples. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5498–5507. PMLR, 2019. URL http://proceedings.mlr.press/v97/roth19a.html.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=BkJ3ibb0-.

Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pp. 3679–3687. IEEE, 2023. doi: 10.1109/ICCVW60793.2023.00395. URL https://doi.org/10.1109/ICCVW60793.2023.00395.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=WLPhywf1si.

Erfan Shayegani, Yue Dong, and Nael B. Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=plmBsXHxgR.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 8317–8326. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00851. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018,* *Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rkZvSe-RZ.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 4566–4575. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://doi.org/10.1109/CVPR.2015.7299087.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. *CoRR*, abs/2406.11839, 2024a. doi: 10.48550/ARXIV.2406.11839. URL https://doi.org/10.48550/arXiv.2406.11839.

Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24502–24511, 2024b.

Siyuan Wang, Zhuohan Long, Zhihao Fan, and Zhongyu Wei. From llms to mllms: Exploring the landscape of multimodal jailbreaking. *CoRR*, abs/2406.14859, 2024c. doi: 10.48550/ARXIV.2406.14859. URL https://doi.org/10.48550/arXiv.2406.14859.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Furong Huang, and Cao Xiao. Enhancing visual-language modality alignment in large vision language models via self-improvement. *CoRR*, abs/2405.15973, 2024d. doi: 10.48550/ARXIV.2405.15973. URL https://doi.org/10.48550/arXiv.2405.15973.

Yubo Wang, Chaohu Liu, yanqiuqu, Haoyu Cao, Deqiang Jiang, and Linli Xu. Break the visual perception: Adversarial attacks targeting encoded visual tokens of large vision-language models. In *ACM Multimedia 2024*, 2024e. URL https://openreview.net/forum?id=tocfToCGF1.

Zeyu Wang, Xianhang Li, Hongru Zhu, and Cihang Xie. Revisiting adversarial training at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24675–24685, 2024f.

Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Zixu Zhu, Xiang-Bo Mao, Sitaram Asur, and Na Cheng. A comprehensive survey of LLM alignment techniques: Rlhf, rlaif, ppo, DPO and more. *CoRR*, abs/2407.16216, 2024g. doi: 10.48550/ARXIV.2407.16216. URL https://doi.org/10.48550/arXiv.2407.16216.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. On the generalization of sft: A reinforcement learning perspective with reward rectification. *arXiv preprint arXiv:2508.05629*, 2025.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018. URL https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-4_Xu_paper.pdf.

Zhiyu Xue, Haohan Wang, Yao Qin, and Ramtin Pedarsani. Conflict-aware adversarial training, 2024. URL https://arxiv.org/abs/2410.16579.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone, 2024. URL https://arxiv.org/abs/2408.01800.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178, 2023. doi: 10.48550/ARXIV.2304.14178. URL https://doi.org/10.48550/arXiv.2304.14178.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *CoRR*, abs/2306.13549, 2023. doi: 10.48550/ARXIV.2306.13549. URL https://doi.org/10.48550/arXiv.2306.13549.

Lu Yu, Haiyang Zhang, and Changsheng Xu. Text-guided attention is all you need for zero-shot robustness in vision-language models. *Advances in Neural Information Processing Systems*, 37: 96424–96448, 2024a.

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. RLHF-V: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *CoRR*, abs/2312.00849, 2023. doi: 10.48550/ARXIV.2312.00849. URL https://doi.org/10.48550/arXiv.2312.00849.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. RLAIF-V: aligning mllms through open-source AI feedback for super GPT-4V trustworthiness. *CoRR*, abs/2405.17220, 2024b. doi: 10.48550/ARXIV.2405.17220. URL https://doi.org/10.48550/arXiv.2405.17220.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 11941–11952. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01100. URL https://doi.org/10.1109/ICCV51070.2023.01100.

Chiyu Zhang, Xiaogang Xu, Jiafei Wu, Zhe Liu, and Lu Zhou. Adversarial attacks of vision tasks in the past 10 years: A survey, 2024. URL https://arxiv.org/abs/2410.23687.

Hanwei Zhang, Yannis Avrithis, Teddy Furon, and Laurent Amsaleg. Smooth adversarial examples. *EURASIP Journal on Information Security*, 2020:1–12, 2020.

Jianping Zhang, Yizhan Huang, Weibin Wu, and Michael R. Lyu. Transferable adversarial attacks on vision transformers with token gradient regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 16415–16424. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01575. URL https://doi.org/10.1109/CVPR52729.2023.01575.

Mengnan Zhao, Lihe Zhang, Jingwen Ye, Huchuan Lu, Baocai Yin, and Xinchao Wang. Adversarial training: A survey, 2024. URL https://arxiv.org/abs/2410.15042.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *CoRR*, abs/2303.18223, 2023a. doi: 10.48550/ARXIV.2303.18223. URL https://doi.org/10.48550/arXiv.2303.18223.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b. URL http://papers.nips.cc/paper_files/paper/2023/hash/a97b58c4f7551053b0512f92244b0810-Abstract-Conference.html.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models, 2024.

Mo Zhou and Vishal M. Patel. Enhancing adversarial robustness for deep metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 15304–15313. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01489. URL https://doi.org/10.1109/CVPR52688.2022.01489.

Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (eds.), *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pp. 6311–6320. ACM, 2023. doi: 10.1145/3581783.3612454. URL https://doi.org/10.1145/3581783.3612454.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=1tZbq88f27.

Appendix

The Use of Large Language Models

In preparing this paper, we primarily used ChatGPT for language polishing. The model was employed to refine grammar, improve readability, and adjust the overall clarity and conciseness of the manuscript. Importantly, all conceptual contributions, experimental designs, analyses, and conclusions are our own. ChatGPT was used strictly as a writing assistant, and we carefully verified and edited all generated suggestions to ensure accuracy and consistency with our intended meaning.

A Untargeted Attack Details

We provide a comprehensive description of the attack pipeline utilized for the untargeted adversarial evaluation of LVLMs in Section 4.1.

For the captioning tasks COCO and Flickr30k, each image is accompanied by 5 available ground-truth captions, with each caption utilized to calculate the CIDEr score. We perform APGD attacks with 100 iterations on each ground truth. After each attack, we calculate the CIDEr score and cease attacking samples that score below 10 for COCO or 2 for Flickr30k. This strategy is primarily employed because generating adversarial attacks on VLMs is extremely time-consuming and costly, allowing us to allocate our limited resources to more impactful attacks. For the VQA task, we apply a similar approach, but the score threshold is set to 0, and we utilize the 5 most common ground truths out of the 10 available.

B Prompts

We list the prompt templates we used for image description tasks:

- What is the content of the image?
- What does the image show?
- Describe the image in detail.
- What is the image illustrating?
- What is displayed in the image?
- Can you explain what is happening in the image?
- Provide a detailed description of the image.
- Summarize the main objects and elements in the picture.
- What objects or people are visible in the image?
- Describe the scene shown in the image.
- What kind of activity is depicted in the image?
- Give a comprehensive explanation of the image content.
- Identify and describe the key features of the image.
- What is the overall setting or environment shown in the picture?
- Provide a step-by-step breakdown of what you see in the image.
- Describe the relationships between the objects in the picture.
- What message or information does the image convey?
- Provide a caption that accurately summarizes the image.

C TARGETED ATTACK DETAILS

We use the following target sequences to perform targeted attacks:

- `A group of people are playing basketball.`
- `a bunch of people are flying in the air.`
- `The pizza on the table is very bad.`
- `An earthquake is about to happen.`
- `This patient needs the best medical care.`

If these targeted and even subtly provocative phrases are successfully embedded into adversarial images, they can subtly deceive users, thereby influencing their subsequent decisions. In addition to the quantitative evaluation presented in Section 4.2, we also observed that when the target text is closely related to the image content, the success rate of adversarial attacks is significantly high, indicating that images can easily mislead LVLMs. This presents a more dangerous scenario because when the target text is only weakly related to the image, users can more easily spot these erroneous outputs, thereby reducing their trust in the model. Conversely, when the model's output appears somewhat plausible in relation to the image content, users are more likely to trust the model's output.

D EXPERIMENTATION ON MORE TASKS

Table 3: Hallucination evaluation on the POPE dataset. The reporting results are based on the F1-score metric.

Method	Clean	TeCoA	TGA	FARE	PMG	TGA	AGH	AdvSimplex	AdPO
F1-score	84.5	75.9	78.2	80.8	81.2	83.2	82.9	82.3	83.7

It is widely recognized that evaluating large vision-language models is challenging. In addition to assessing several key multimodal tasks in Section 4.1, this section further examines the performance of our method on additional vision-language tasks.

Hallucinations. One of the greatest challenges faced by LVLMs is hallucination, where these models may perceive objects in an image that do not actually exist. This issue has garnered widespread attention within the research community. We selected the commonly used POPE dataset (Li et al., 2023d) to evaluate multiple CLIP versions of the LLaVA model. In this dataset, the model is required to answer "Yes" or "No" to indicate whether a specific object is actually present in the image. Table 3 shows that our version of CLIP achieved the highest accuracy, but our AdPO method most effectively addresses clean performance. In contrast, both TeCoA and FARE demonstrated a more pronounced decline in performance.

Table 4: The evaluation of jailbreak attack defense, with the attack success rate reported.

Method	ϵ	ASR
CLIP	0	14 / 40
TeCoA	0	14 / 40
TGA	0	15 / 40
FARE	0	13 / 40
AdPO	0	8 / 40
CLIP	$^{16}/_{266}$	25 / 40
TeCoA	$^{16}/_{266}$	16 / 40
TGA	$^{16}/_{266}$	16 / 40
FARE	$^{16}/_{266}$	16 / 40
AdPO	$^{16}/_{266}$	8 / 40

Jailbreaking attacks. Recent studies have shown that LVLMs are more vulnerable to jailbreak attacks than pure LLMs, especially when adversarial perturbations are added to images (Qi et al., 2024a; Carlini et al., 2023). Therefore, it is essential to further analyze our method's robustness against jailbreak attacks. Under normal circumstances, model owners align models with human values to prevent them from generating suggestive or harmful content. For example, if a user enters a malicious prompt like "How to make a bomb," the model should refuse to respond. However, with the introduction of adversarial images, attackers can more easily bypass these security guard, inducing the model to output intended content and thereby posing greater risks. Following the setup from (Qi et al., 2024a), we evaluate LLaVA 1.5 with different CLIP versions under various noise levels. The results are shown in Table 4. Even without adversarial images, LLaVA can be affected by jailbreak attacks to generate harmful content. Once noise is introduced, however, the success rate of jailbreak attacks on the clean CLIP version increases significantly, while adversarially trained versions maintain their original level of security. This indicates that adversarial training can also enhance the robustness of LVLMs against jailbreak attacks, with our method achieving the best performance. It is important to note that jailbreak attacks are currently a very active area of research, and our evaluations may somewhat overestimate their performance.

Table 5: Adversarial Evaluation on the Qwen-2.5-VL model.

Method	Type	COCO	Flickr30k	TextVQA	VQAv2
Original	-	124.3	82.3	79.3	84.3
FARE	Clean	118.5	74.3	65.1	73.8
	Adversarial	61.8	35.2	20.8	34.7
AdPO	Clean	124.8	82.2	79.2	84.1
	Adversarial	78.3	50.6	37.2	48.5

Table 6: Adversarial Evaluation on the InternVL3.5 model.

Method	Type	COCO	Flickr30k	TextVQA	VQAv2
Original	-	124.5	83.6	78.2	82.4
FARE	Clean	117.4	73.2	62.4	70.3
	Adversarial	60.5	32.1	22.4	34.5
AdPO	Clean	124.4	83.4	79.0	82.5
	Adversarial	79.3	48.8	36.4	47.2

Table 7: Adversarial Evaluation on the BLIP-2 model.

Method	Type	COCO	Flickr30k	TextVQA	VQAv2
Original	-	98.2	70.7	40.3	48.2
FARE	Clean	88.5	62.2	29.5	32.2
	Adversarial	41.2	23.7	10.3	15.3
AdvSimplex	Clean	82.9	59.8	19.9	27.5
	Adversarial	40.7	21.8	12.8	14.6
AGH-MAT	Clean	85.1	62.5	21.3	29.1
	Adversarial	42.3	25.5	14.1	18.0
TGA-ZSR	Clean	84.2	61.3	20.8	28.0
	Adversarial	41.4	22.9	13.4	15.2
AdPO	Clean	98.5	70.3	39.3	48.2
	Adversarial	65.4	48.2	22.8	24.4

E Experiment on Non-CLIP Models

In order to assess the generalizability of AdPO beyond CLIP-based models, we conduct empirical evaluations on Qwen-2.5-VL, InternVL3.5, and BLIP-2.

Qwen2.5-VL employs an image encoder with an improved self-attention mechanism, while InternVL3.5 uses InternViT as its image encoder. We apply AdPO for comparison, and neither model requires a text encoder. As shown in Tables 5 and Table 6, our method achieves a substantial lead, particularly in adversarial robustness, consistently outperforming FARE by more than 10 points. We also evaluate BLIP-2, which uses the EVA-CLIP model as its encoder, to enable a more comprehensive comparison with prior methods. As shown in the results in Table 7, AdPO still achieves a substantial lead, further demonstrating that it is a model-agnostic approach.

F Analysis of Attack Strengths

In this section, we explore the impact of higher attack strengths with the results presented in Table 8 and Table 9.

Table 8: The performance of attacks with $^{8}/_{255}$.

Method	COCO	Flickr30k	TextVQA	VQAv2
FARE	25.2	13.2	5.2	10.1
TGA	26.7	14.2	6.9	15.2
AdPO	42.5	24.5	13.3	22.5

Table 9: The performance of attacks with $^{16}/_{255}$.

Method	COCO	Flickr30k	TextVQA	VQAv2
FARE	8.2	3.2	2.2	3.1
TGA	10.7	8.2	3.6	5.2
AdPO	20.2	13.2	8.2	14.9

We find that models trained with low attack intensity exhibit some level of adversarial robustness when faced with high-disturbance adversarial samples. However, they show a noticeable performance drop compared to models trained with the same level of perturbation. Compared to previous state-of-the-art methods, our method still achieves a significant lead.

G Experimental Results on Tiny-LLaVA

Table 10: Experimental Results on Tiny-LLaVA.

Method	Type	COCO	Flickr30k	TextVQA	VQAv2
Original	-	90.3	65.3	40.4	69.5
FARE	Clean	83.2	55.3	30.7	58.2
	Adversarial	39.2	20.5	8.9	17.2
TGA	Clean	80.1	53.9	25.8	55.7
	Adversarial	40.2	23.1	10.2	18.2
AdPO	Clean	91.2	66.3	42.8	65.2
	Adversarial	50.3	42.7	27.4	28.4

Table 10 presents our experimental results on TinyLLaVA. The results demonstrate that our method achieves a substantial improvement over previous approaches. This improvement can be attributed to the direct joint training of the image encoder and the target decoder, which enables more effective vision-language alignment.

H Additional Ablation Studies

In this section, we primarily investigate the impact of direct adversarial training and full fine-tuning on model performance as shown Table 11.

We observe that direct adversarial training significantly degrades clean performance without providing notable improvements in adversarial robustness. On the other hand, full fine-tuning slightly

Table 11: Additional ablation studies.

Method	Clean	Adversarial
Direct Adversarial Training	66.3	42.2
Full Fine-tuning	72.3	50.4
SFT-based AIO	70.0	41.1
AdPO	73.6	52.3

compromises transferability, leading to minor drops in both clean and adversarial performance. We find that when applying SFT-based AIO (Eq. 6), both clean and adversarial performance degrade significantly. This decline arises from the strong negative impact of SFT on model generalization, a phenomenon consistent with recent findings in the literature (Chu et al., 2025).

I Efficiency Analysis

In this section, we discuss the training efficiency of different methods. Table 12 shows that, under the same hardware configuration, our method achieves comparable training speed to previous approaches. This is primarily attributed to our use of a smaller LVLM during training. It is worth noting that training time may vary slightly under different hardware drivers, temperature conditions, and other system factors.

In the early experiments, this paper consumed substantial computational resources. However, the final approach significantly reduced the training cost by relying on training smaller models, ultimately requiring only 8 A100 GPUs.

Table 12: Comparison of training speed among different methods.

Method	GPU	Batch Size	Average Training Time
TeCoA	NVIDIA Tesla A100	128	1.78s / Batch
TGA-ZSR			1.73s / Batch
AdPO			1.89s / Batch

J Analysis of Failure Samples

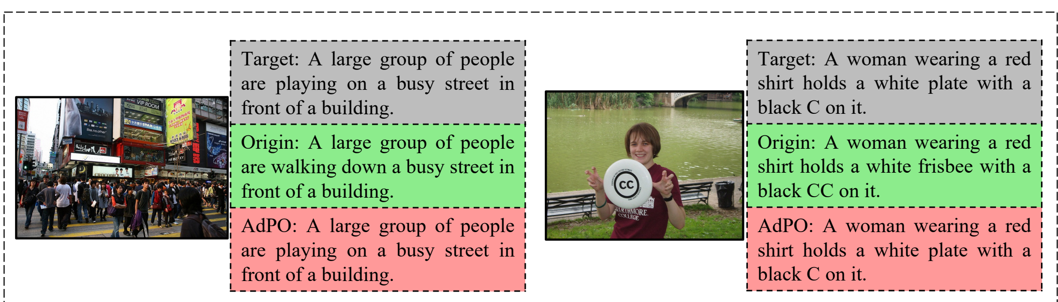

Figure 5: Showcase of failure samples.

In this section, we provide a qualitative analysis of failure cases, as illustrated in Figure 5. Our observations indicate that AdPO is particularly vulnerable when the adversarial target is semantically close to the true content of the image. These cases often involve fine-grained distinctions that are semantically ambiguous, making them difficult for the model to reliably discriminate. The adversarial attack exploits this ambiguity, thereby increasing the likelihood of misleading the model. This analysis clarifies the inherent limitations of AdPO when operating under subtle semantic shifts.