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-4 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 performance degradation on clean inputs. In this paper, we proposes AdPO, a novel adversarial defense strategy for LVLMs based on preference optimization. Preference optimization methods, such as DPO and RLHF, have been widely used to align large language models (LLMs) with human values and preferences. 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 robustness across a range of downstream tasks (including LVLMs and zero-shot classification). Our comprehensive experimental validation confirms the efficacy of the proposed AdPO, which outperforms prior state-of-the-art methods.

1 INTRODUCTION

The emergence of large vision-language models (LVLMs) has substantially propelled the development of general artificial intelligence, attracting considerable attention from the research community (Yin et al., 2023; Cui et al., 2024; Liu et al., 2024b). 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 GPT-4 (OpenAI et al., 2024), LLaVA (Liu et al., 2023b), and OpenFlamingo (Awadalla 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 (Dai et al., 2023; Nguyen et al., 2023), visual question answering (Liu et al., 2023b), and text recognition (Liu et al., 2024a; Li et al., 2023d).

Given their transformative potential for multimodal learning and understanding, LVLMs are positioned for deployment across a growing range of real-world applications. However, this widespread deployment introduces significant security concerns, as malicious attacker could manipulate LVLMs into generating undesirable content and hallucinations (Schlarmann & Hein, 2023; Shayegani et al., 2024). 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., 2024b; Carlini et al., 2023; Qi et al., 2024b; Luo et al., 2024). Researchers have demonstrated both targeted and untargeted attacks by introducing imperceptible noise into images, which consequently alters the model's interpretation and output.

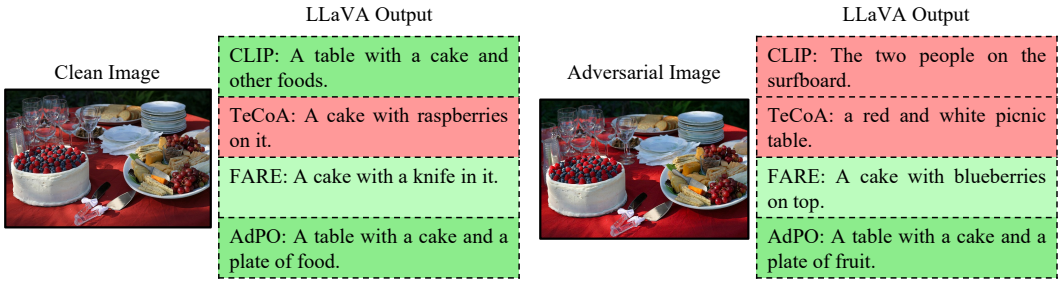

Figure 1: Illustration of f adversarial attacks with $\epsilon = 4/255$ on LLaVA using different CLIP models as encoders. The original model can produce accurate outputs on clean images, but it makes significant errors when faced with adversarial attacks. Although the adversarially trained versions, TeCoA and FARE, have better adversarial robustness, they still tend to hallucinate or fail to fully comprehend the image. Comparatively, our AdPO exhibits strong performance on both clean and adversarially altered images.

To improve the adversarial robustness of LVLMs, existing efforts focus on fine-tuning primarily the image encoder. For example, TeCoA utilizes a text-guided contrastive adversarial training loss, supervising the alignment of text embeddings with adversarial visual features on a limited training dataset (Mao et al., 2023). FARE proposes an unsupervised adversarial fine-tuning scheme to eliminate the dependence on labeled training datasets (Schlarmann et al., 2024). Although these methods have achieved advancements in improving the robustness of CLIP models, they continue to suffer from performance degradation on downstream tasks, including LVLMs and zero-shot classification. As shown in Figure 1, TeCoA generates severe hallucinations with clean samples, whereas FARE tends to lose its fine-grained comprehension of the image.

Inspired by the significant success of preference optimization in the LLM community (Wang et al., 2024e; Ouyang et al., 2022), we find 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. Leveraging 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 unique challenges. In comparison to standard offline DPO, we introduce two key improvements: (1) To remove the reliance on image annotations, we adapt DPO to an online setting. During training, the policy model generates interpretations for both clean and adversarial images, which serve as sources for positive and negative samples. This process is referred to as **preferred image optimization**. (2) Multimodal preference optimization may face an *unconditional preference* issue, where the learning process may neglect image conditions (Wang et al., 2024a). To address this issue, we introduce supplementary **adversarial image optimization** to further improve the adversarial robustness of LVLMs. To ensure consistency with previous research, we confine our adversarial training to adjusting only the parameters of CLIP's ViT on the ImageNet dataset (Deng et al., 2009). Extensive experimental results, including those on LVLMs and zero-shot classification, demonstrate that our proposed AdPO achieves a more robust image encoder, with minimal impact on clean inputs and even shows improvements in certain tasks. These outcomes not only validate the effectiveness of our approach but also expand the potential applications of preference optimization techniques beyond their original scope in language models.

In summary, our contributions can be summarized 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 the dual strategy of preferred image optimization and adversarial image optimization to maintain the model's clean performance while enhancing its adversarial robustness.
- Extensive experiments show that our method achieves state-of-the-art results in improving the adversarial robustness of LVLMs while maintaining the original performance as much as possible.

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., 2023a), 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., 2023b; 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). We focus on the performance evaluation of LLaVA-1.5 and OF, as both use CLIP ViT-L/14 (Radford et al., 2021) as their image encoder.

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). 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. (2024); Gao et al. (2024); Dong et al. (2023b). Zhao et al. (2023b) showed that transferable black-box attacks could be generated using text-to-image models. Carlini et al. (2023) demonstrated how adding adversarial noise to images can circumvent safety constraints of LLMs. 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; Meng & Chen, 2017; Zhou & Patel, 2022). Adversarial training is a foundational method for enhancing a model's inherent robustness by integrating adversarial examples into the training dataset Kurakin et al. (2017b); Tramèr et al. (2018); Dong et al. (2023a). In the multimodal domain, TeCoA improves the adversarial robustness of CLIP's image encoder through text-guided contrastive adversarial training while preserving some of CLIP's zero-shot classification capabilities (Mao et al., 2023). FARE employs unsupervised training by minimizing the distance between adversarial image features and clean image features, maintaining impressive performance on LVLMs (Schlarmann et al., 2024). However, this straightforward adversarial training approach often fails to prevent performance degradation on clean samples. Unlike these fine-tuning strategies, we are the first to frame adversarial training as a preference optimization problem, integrating both clean and adversarial images into the training process to improve robustness while maintaining clean performance.

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; 2024; Wang et al., 2024a;c). Reinforcement Learning from Human Feedback (RLHF) utilizes human preferences as a reward model and applies reinforcement learn-

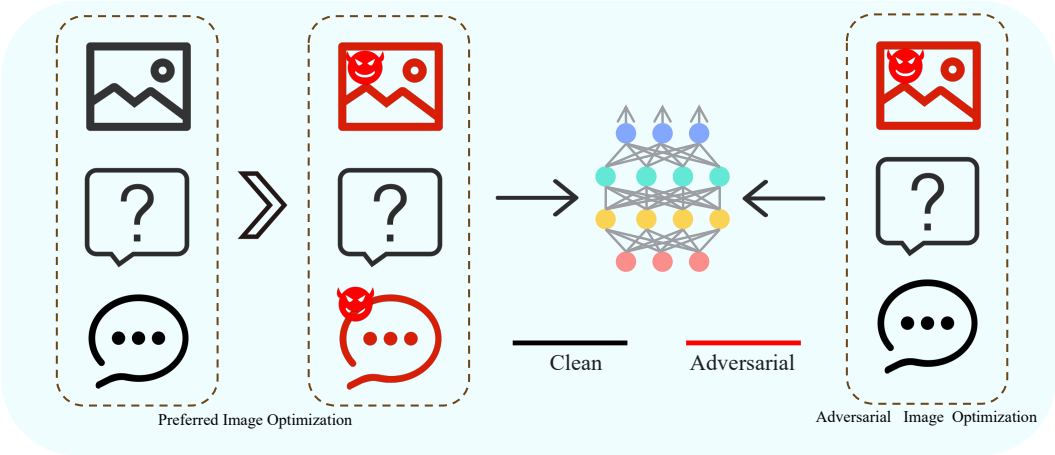

Figure 2: The architecture of our proposed AdPO. AdPO mainly consists of two parts: (**left**) preferred image optimization and (**right**) adversarial image optimization. Preferred image optimization incorporates both clean and adversarial images into adversarial training while maintaining the model's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

ing 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), SmiPO (Meng et al., 2024), 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 $r(x, y)$:

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

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.

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:

$$p(y_w \succ y_l) = \sigma(r(x, y_w) - r(x, y_l)), \tag{2}$$

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:

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

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 large vision-language models, the input to the model comprises $x = \{x_m, x_t\}$, where x_m denotes the image input and x_t 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 (Szegedy et al., 2014; Goodfellow et al., 2015). 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., 2024), whereas untargeted attacks merely lead the model to generate incorrect outputs (Wang et al., 2024d; Gao et al., 2024). In this study, we employ untargeted attack methods to generate adversarial images. This approach eliminates reliance on specific labeled datasets, enabling our method to be extended to unseen datasets.

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:

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

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. Note that in subsequent adversarial training, the parameters of ϕ will be updated.

3.3 Preferred Image Optimization

This section primarily outlines the process of constructing pairs of preferred and non-preferred samples from unlabeled image data, a crucial component of the DPO training pipeline.

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. Accordingly, in the setting of multimodal adversarial training, our preferred image optimization can be formulated as:

$$\mathcal{L}_{\mathrm{P}}(\pi_\theta; \pi_{\mathrm{ref}}) = -\mathbb{E}_{(x_m,x_t,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x_m,x_t)}{\pi_{\mathrm{ref}}(y_w|x_m,x_t)} - \beta\log\frac{\pi_\theta(y_l|x_{adv},x_t)}{\pi_{\mathrm{ref}}(y_l|x_{adv},x_t)}\right)\right], \tag{5}$$

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).

3.4 Adversarial Image Optimization

Although preferred image optimization can maintain the performance of VLMs on clean inputs, it does not significantly enhance adversarial robustness. Recent research indicates that, although multimodal DPO is designed to compute implicit rewards based on all input modalities, it may prioritize

language-only preferences while neglecting image conditions (i.e., unconditional preferences), resulting in suboptimal model performance and increased hallucinations (Wang et al., 2024a).

The issue of unconditional preferences may lead to suboptimal adversarial robustness. To address this, we introduce adversarial image optimization:

$$\mathcal{L}_{\mathrm{A}} = \sum_{t=1}^{T} \log \pi_\theta(y_w^t \mid x_{adv}, x_t^{1:t-1}), \tag{6}$$

where T represents the sequence length of each sample. The objective of AdPO is a combination of preferred image optimization and adversarial image optimization:

$$\mathcal{L}_{\mathrm{AdPO}} = \mathcal{L}_{\mathrm{P}} + \mathcal{L}_{\mathrm{A}}. \tag{7}$$

By leveraging joint optimization, AdPO attains enhanced adversarial robustness while maintaining its performance on clean samples.

4 Experiments

In this section, we evaluate the performance of AdPO on LVLMs and zero-shot classification through extensive experiments. Although we use the complete LVLMs during adversarial training, we modify only the parameters of their image encoders, enabling the robust image encoder to be directly transferred to other LVLMs. All experiments are conducted on 32 Tesla A100 GPUs.

Models. For the LVLM models, we primarily select OpenFlamingo-9B (OF)(Awadalla et al., 2023) and LLaVA 1.5-7B(Liu et al., 2023a), both of which use CLIP's ViT-L/14 as their image encoder (Radford et al., 2021). The two models differ in their language decoders: OF employs MPT-7B (Team et al., 2023), while LLaVA 1.5 uses Vicuna (Chiang et al., 2023). In the subsequent evaluation of OF, we adopt a zero-shot setting, where the model is given textual prompts without any accompanying images (Alayrac et al., 2022). For LLaVA, we employ the default system prompt along with task-specific prompts (Liu et al., 2023b).

Adversarial training settings. In AdPO, we leverage LLaVA 1.5 to fine-tune CLIP's ViT model on the ImageNet dataset (Deng et al., 2009). As we adopt an online learning approach, we do not rely on category labels provided by the dataset, only on the images themselves. By optimizing Equation 4, we generate adversarial perturbations for clean images using a 10-step PGD under the ℓ_∞ norm. It is widely recognized that employing larger image perturbations during adversarial training can significantly improve adversarial robustness, but it often leads to performance degradation on clean data (Madry et al., 2018). To balance robustness and clean accuracy, we apply two perturbation radii: $\epsilon = 2/255$ and $\epsilon = 4/255$. The resulting robust CLIP image encoders are referred as AdPO^2 and AdPO^4, respectively. 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 optimization parameter β is set to 0.1.

Baseline methods. We compare the performance of AdPO with the original CLIP and two state-of-the-art methods, TeCoA (Mao et al., 2023) and FARE (Schlarmann et al., 2024). TeCoA utilizes supervised contrastive learning with image category labels, while FARE performs unsupervised training at the representation level. To ensure fair comparison, we use adversarial images with the same noise radius for training, denoted as TeCoA^2 and FARE^2 for $\epsilon = 2/255$, and TeCoA^4 and FARE^4 for $\epsilon = 4/255$.

4.1 Evaluation of Untargeted Attacks on LVLMs

In this section, we evaluate the clean and robust performance of AdPO in vision-language tasks by replacing the image encoder of LVLMs with robust versions.

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 the Appendix A.1.

Table 1: Evaluation of the adversarial robustness of large vision-language models with different CLIP models. We evaluate the clean performance and adversarial robustness of various methods across multiple tasks and perturbation sizes. 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**.

VLM	Image Encoder	COCO			Flickr30k			TextVQA			VQAv2		
		clean	ℓ_∞		clean	ℓ_∞		clean	ℓ_∞		clean	ℓ_∞	
			$2/255$	$4/255$		$2/255$	$4/255$		$2/255$	$4/255$		$2/255$	$4/255$
OF-9B	CLIP	79.7	1.5	1.1	60.1	0.7	0.4	23.8	0.0	0.0	48.5	1.8	0.0
	TeCoA^2	73.5	31.5	21.2	49.5	14.1	9.5	16.6	3.5	2.1	46.2	23.5	20.5
	FARE^2	79.1	34.2	19.5	57.7	16.4	8.9	21.6	4.1	1.9	47.0	24.0	17.2
	AdPO^2	**84.7**	**34.6**	**25.5**	**57.9**	**18.8**	**12.3**	**22.3**	**6.5**	**3.3**	**48.1**	**26.3**	**22.8**
	TeCoA^4	66.9	28.5	21.6	40.9	12.0	10.3	15.4	2.1	1.8	44.8	23.6	21.3
	FARE^4	74.1	30.9	22.8	51.4	15.7	10.5	18.6	3.4	2.9	46.1	23.6	21.0
	AdPO^4	**75.2**	**33.3**	**25.9**	**54.6**	**17.2**	**12.7**	**20.5**	**5.2**	**3.3**	**46.7**	**24.4**	**21.3**
LLaVA 1.5-7B	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^2	98.4	44.2	30.3	57.1	23.2	15.3	24.1	12.1	8.8	66.9	33.8	21.8
	FARE^2	109.9	53.6	31.0	71.1	29.5	17.5	31.9	14.7	9.1	71.7	**34.9**	23.0
	AdPO^2	**118.3**	**65.3**	**43.9**	**75.4**	**32.5**	**20.1**	**32.4**	**17.8**	**10.5**	**72.9**	34.3	**23.2**
	TeCoA^4	88.3	50.9	35.3	48.6	27.9	19.5	20.7	12.6	9.3	63.2	41.0	31.7
	FARE^4	102.4	57.1	40.9	61.6	31.4	22.8	27.6	15.8	**10.9**	68.3	40.7	30.5
	AdPO^4	**111.5**	**67.2**	**49.3**	**67.0**	**35.3**	**25.4**	**32.3**	**16.1**	10.2	**70.1**	**42.3**	**32.5**

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.

Table 1 summarizes the experimental results. Typically, the original CLIP model achieves optimal clean performance but lacks adversarial robustness, rendering it vulnerable to attacks. When comparing different methods, our AdPO consistently achieves superior clean performance and adversarial robustness compared to baseline methods, emphasizing the significance of including both clean and adversarial images in the training dataset. Across various datasets, our method demonstrates significant improvements in tasks such as COCO image captioning, likely due to the alignment between this task and our adversarial training paradigm, enabling the robust model to potentially outperform the clean model. For different perturbation sizes, $\epsilon = 2/255$ already ensures solid adversarial robustness, while larger perturbations still preserve more clean performance. AdPO^4 exhibits stronger robustness compared to AdPO^2, but at the cost of some clean performance.

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). In this section, we examine the robustness of substituting the CLIP encoder in LLaVA with our adversarially robust variant.

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

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^2	FARE^2	AdPO^2	TeCoA^4	FARE^4	AdPO^4
`A group of people are playing...`	20/20	1/20	1/20	0/20	0/20	0/20	0/20
`A group of people are flying...`	20/20	1/20	1/20	0/20	0/20	0/20	0/20
`The pizza on the table...`	20/20	2/20	0/20	0/20	0/20	0/20	0/20
`An earthquake is about...`	20/20	2/20	1/20	1/20	0/20	0/20	0/20
`This patient needs the best...`	20/20	0/20	0/20	0/20	0/20	0/20	0/20
Mean ASR:	100%	4%	3%	1%	**0%**	**0%**	**0%**

Table 3: Evaluation of clean and adversarial performance on image classification datasets using the CLIP model. We primarily evaluate the performance of the original CLIP model and its adversarially trained versions when faced with clean samples and adversarial samples with a noise $4/255$. Detailed descriptions of the dataset are provided in the appendix.

Eval.	Image Encoder	CalTech	Cars	CIFAR10	CIFAR100	DTD	EuroSAT	FGVC	Flowers	ImageNet-R	ImageNet-S	PCAM
Clean	CLIP	83.3	77.9	95.2	71.1	55.2	62.6	31.8	79.2	87.9	59.6	52.0
	TeCoA^2	80.7	50.1	87.5	60.7	44.4	26.1	14.0	51.8	80.1	58.4	49.9
	FARE^2	84.8	70.5	89.5	69.1	50.0	25.4	26.7	70.6	85.5	59.7	50.0
	AdPO^2	85.1	72.8	91.2	69.5	53.1	35.3	25.9	74.4	87.5	59.6	50.7
	TeCoA^4	78.4	37.9	79.6	50.3	38.0	22.5	11.8	38.4	74.3	54.2	50.0
	FARE^4	84.7	63.8	77.7	56.5	43.8	18.3	22.0	58.1	80.2	56.7	50.0
	AdPO^4	84.9	65.8	80.2	56.6	44.5	21.7	21.4	58.5	82.9	57.8	49.9
$\epsilon = 4/255$	CLIP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	TeCoA^2	57.4	6.5	31.0	17.8	14.7	7.7	1.1	9.8	36.7	32.8	16.0
	FARE^2	46.6	4.8	25.9	13.9	11.7	0.5	0.6	7.1	25.6	22.5	17.2
	AdPO^2	55.3	5.8	28.7	17.5	13.6	5.7	1.0	8.7	33.4	33.1	15.8
	TeCoA^4	60.9	8.4	37.1	21.5	16.4	6.6	2.1	12.4	41.9	34.2	44.0
	FARE^4	64.1	12.7	34.6	20.2	17.3	11.1	2.6	12.5	40.6	30.9	50.2
	AdPO^4	66.8	13.6	36.9	21.7	17.9	9.2	2.6	12.7	42.3	33.3	49.7

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 test five target strings, 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. TeCoA^2, FARE^2, and AdPO^2 demonstrate varying degrees of adversarial robustness, even when subjected to higher levels of adversarial noise. By comparison, the $\epsilon = 4/255$ versions exhibit significantly higher levels of adversarial robustness. Additional details are provided in Appendix A.2.

4.3 Evaluation of Zero-Shot Classification

In this section, we assess the zero-shot classification performance of the robust CLIP image encoder, following the methods of Mao et al. (2023) and Schlarmann et al. (2024). CLIP's zero-shot classi-

Figure 3: Qualitative assessment of targeted attacks on LLaVA. (Left) When encountering clean images, CoTeA 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 CLIP version of LLaVA is easily compromised, FARE shows some adversarial robustness but loses more details or makes subtle errors, whereas AdPO performs better.

fication simultaneously trains both visual and text encoders, enabling the model to project images and textual descriptions into a shared semantic space. For classification, there is no requirement for a specially labeled dataset for each category; instead, CLIP computes the similarity between images and the textual descriptions of categories to classify images into the most relevant category.

Attack setup. To assess the adversarial robustness of the models, we utilize the initial two components of AutoAttack (Croce & Hein, 2020), specifically APGD with cross-entropy loss and APGD with DLR loss, both executed over 100 iterations. In alignment with AutoAttack, we adopt the targeted version of the DLR loss, differing from Mao et al. (2023), where the less effective untargeted variant was applied. We perform the evaluation with a more powerful attack ($\epsilon = 4/255$) in this section and present the $\epsilon = 2/255$ results in Appendix A.3.

As demonstrated in Table 3, similar to evaluations on vision-language tasks, the original CLIP typically achieved the best clean performance but displayed minimal adversarial robustness. Adversarial attacks on the clean CLIP achieved a 100% attack success rate, further confirming CLIP's inherent vulnerability, which introduces several weaknesses in LVLMs. After adversarial training, CLIP exhibits some performance decline on clean samples, but its adversarial robustness significantly improves. In contrast, the AdPO models, particularly AdPO^2, demonstrate substantially higher accuracy on clean data while still preserving robustness.

4.4 Quantitative Evaluation

In addition to quantitative experimental evaluations, we also present a qualitative comparison of different defense methods in this section.

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 generated by targeted attacks, 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. Therefore, it is imperative to enhance the model's adversarial robustness.

4.5 Ablation Study

In this section, we mainly discuss the impact of preferred image optimization (PIO) and adversarial image optimization (AIO) on the final performance.

We use the setup in Section 4.1 to perform untargeted attacks to evaluate the effectiveness of methods trained with a single optimization approach on the COCO dataset, with experimental results shown in Table 4. PIO retains more of the model's clean performance, but only shows a small amount of adversarial robustness. AIO somewhat weakens the model's clean performance, but significantly improves its adversarial robustness. It can also be observed that PIO contributes to enhancing adversarial robustness, indicating the potential of preference optimization in improving adversarial robustness.

Table 4: Ablation study of preferred image optimization and adversarial image optimization.

Metric	Clean	$2/255$	$4/255$
PIO	119.5	35.5	29.7
AIO	102.4	65.8	42.1
AdPO	118.3	65.3	49.9

5 Conclusion

We propose AdPO, the first adversarial defense strategy based on preference optimization. The core idea of preference optimization methods, represented by DPO, is to learn both positive and negative samples simultaneously and optimize the model to better align with user preferences or goals. This is achieved by comparing the differences between positive and negative samples, clarifying the direction in which the model should be optimized. Unlike previous adversarial fine-tuning methods, which typically only impose single-target constraints to improve adversarial robustness, leading to a loss of clean performance. In contrast, AdPO explicitly optimizes two objectives: improving adversarial robustness while maintaining proper understanding of clean images. Both quantitative and qualitative experimental analyses demonstrate the superiority of our proposed method, offering a new perspective for future adversarial defense research. Considering that preference optimization is gaining increasing attention in academia, introducing more refined methods into the adversarial defense field could lead to better outcomes.

Limitations. Although this paper primarily focuses on LVLMs using CLIP ViT as the encoder, other types of models are equally applicable. Considering the computational resources and alignment with previous work, we only adjusted the parameters of the image encoder, but full tuning may yield better results. Carefully crafted malicious prompts also pose significant security risks to the model, and future work needs to address threats from both image and text inputs. While we have performed a significant amount of evaluation, it is clear that evaluating the adversarial robustness of LVLMs in real-world settings is also essential.

Reproducibility

To ensure the reproducibility of our method, we provide a detailed description of our experimental setup in the experiment and appendix sections. The training datasets, evaluation datasets, and involved models are all openly available and accessible.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. 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/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

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.

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, 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.

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.

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

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 3606–3613. IEEE Computer Society, 2014. doi: 10.1109/CVPR.2014.461. URL https://doi.org/10.1109/CVPR.2014.461.

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.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with 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/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html.

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.

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.

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.

Gregory Griffin, Alex Holub, Pietro Perona, et al. Caltech-256 object category dataset. Technical report, Technical Report 7694, California Institute of Technology Pasadena, 2007.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242. URL https://doi.org/10.1109/JSTARS.2019.2918242.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 8320–8329. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00823. URL https://doi.org/10.1109/ICCV48922.2021.00823.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pp. 554–561. IEEE Computer Society, 2013. doi: 10.1109/ICCVW.2013.77. URL https://doi.org/10.1109/ICCVW.2013.77.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

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.

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.

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.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *CoRR*, abs/2311.06607, 2023d. doi: 10.48550/ARXIV.2311.06607. URL https://doi.org/10.48550/arXiv.2311.06607.

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. *CoRR*, abs/2404.06918, 2024a. doi: 10.48550/ARXIV.2404.06918. URL https://doi.org/10.48550/arXiv.2404.06918.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744, 2023a. doi: 10.48550/ARXIV.2310.03744. URL https://doi.org/10.48550/arXiv.2310.03744.

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*, 2023b. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

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

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, 2024. URL https://openreview.net/forum?id=nc5GgFAvtk.

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.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. URL http://arxiv.org/abs/1306.5151.

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.

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. *CoRR*, abs/2405.14734, 2024. doi: 10.48550/ARXIV.2405.14734. URL https://doi.org/10.48550/arXiv.2405.14734.

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.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, pp. 722–729. IEEE Computer Society, 2008. doi: 10.1109/ICVGIP.2008.47. URL https://doi.org/10.1109/ICVGIP.2008.47.

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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan

Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

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.

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.

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.

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.

MosaicML NLP Team et al. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023.

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.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, volume 11071 of *Lecture Notes in Computer Science*, pp. 210–218. Springer, 2018. doi: 10.1007/978-3-030-00934-2_24. URL https://doi.org/10.1007/978-3-030-00934-2_24.

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.

Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 10506–10518, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3eefceb8087e964f89c2d59e8a249915-Abstract.html.

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

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, 2024c. 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*, 2024d. URL https://openreview.net/forum?id=tocfToCGF1.

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, 2024e. doi: 10.48550/ARXIV.2407.16216. URL https://doi.org/10.48550/arXiv.2407.16216.

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.

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, 2024. 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.

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.

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.

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.

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.

A APPENDIX

A.1 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.

A.2 TARGETED ATTACK DETAILS

We use the following target sequences to perform targeted attacks:

- `A group of people are playing basketball.`
- `A group 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.

A.3 ZERO-SHOT EVALUATIONS

We evaluated the model's clean performance and robustness on a series of zero-shot image classification tasks. These datasets include CalTech (Griffin et al., 2007), StanfordCars (Krause et al., 2013), CIFAR10, CIFAR100 (Krizhevsky & Hinton, 2009), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), FGVC Aircrafts (Maji et al., 2013), Flowers (Nilsback & Zisserman, 2008), ImageNet-R (Hendrycks et al., 2021), ImageNet-Sketch (Wang et al., 2019), and PCAM (Veeling et al., 2018). The evaluation protocol is based on the *CLIP Benchmark*[1].

We assess the robustness by evaluating 1000 samples per dataset and reporting the clean accuracy for all samples. We utilize the first two attacks from AutoAttack (Croce & Hein, 2020), specifically, APGD with cross-entropy loss and APGD with targeted DLR loss, each with 100 iterations. Given that the DLR loss is applicable only to multi-class classification, we employ only the first attack on the binary dataset PCAM. We consider ℓ_∞-bounded threat models with radii $\epsilon = 4/255$ and evaluate the robustness on all datasets at a resolution of 224x224, except for CIFAR10, CIFAR100, and STL-10, which are evaluated at their original resolutions.

In Section 4.3, we only presented the performance of different CLIP versions on clean images and adversarial images with noise set to $\epsilon = 4/255$ due to space constraints. In Table 5, we show the evaluation results for an attack noise of $\epsilon = 2/255$. Humans can barely distinguish between images with $2/255$ noise and clean images, yet even such a small amount of noise causes the original CLIP model to nearly lose all its performance. This vulnerability is extremely critical. After adversarial

[1] https://github.com/LAION-AI/CLIP_benchmark

Table 5: **Evaluation of the clean performance and adversarial robustness with a noise $\epsilon = 2/255$ of different CLIP versions.**

Eval.	Image Encoder	CalTech	Cars	CIFAR10	CIFAR100	DTD	EuroSAT	FGVC	Flowers	ImageNet-R	ImageNet-S	PCAM
$\epsilon = 2/255$	CLIP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
	TeCoA^2	70.2	22.2	63.7	35.0	27.0	12.8	5.8	27.6	58.8	45.2	40.0
	FARE^2	73.0	26.0	60.3	35.6	26.7	6.2	5.9	31.2	56.5	38.3	41.9
	AdPO^2	75.1	29.1	64.1	35.4	26.9	10.5	6.4	33.3	59.2	45.7	43.5
	TeCoA^4	69.7	17.9	59.7	33.7	26.5	8.0	5.0	24.1	59.2	43.0	48.8
	FARE^4	76.7	30.0	57.3	36.5	28.3	12.8	8.2	31.3	61.6	41.6	50.2
	AdPO^4	78.1	32.5	64.2	36.1	27.4	13.9	9.3	34.2	62.4	42.5	51.3

training, multiple CLIP versions achieved noticeable adversarial robustness, but at the cost of some clean performance. Overall, AdPO had the least sacrifice in clean performance.