Adversary-Aware DPO: Enhancing Safety Alignment in Vision Language Models via Adversarial Training

Fenghua Weng[1] **Jian Lou**[2] **Jun Feng**[3] **Minlie Huang**[4] **Wenjie Wang**[1] *
[1]School of Information Science and Technology, ShanghaiTech University
[2]School of Software Engineering, Sun Yat-Sen University
[3]School of Cyber Science and Engineering, Huazhong University of Science and Technology
[4]The Conversational Artificial Intelligence (CoAI) Group, Tsinghua University
{wengfh2023,wangwj1}@shanghaitech.edu.cn
louj5@mail.sysu.edu.cn, junfeng@hust.edu.cn, aihuang@tsinghua.edu.cn

Abstract

Safety alignment is critical in pre-trained large language models (LLMs) to generate responses aligned with human values and refuse harmful queries. Unlike LLM, the current safety alignment of VLMs is often achieved with post-hoc safety fine-tuning. However, these methods are less effective to white-box attacks. To address this, we propose *Adversary-aware DPO (ADPO)*, a novel training framework that explicitly considers adversary. *Adversary-aware DPO (ADPO)* integrates adversarial training into DPO to enhance the safety alignment of VLMs under worst-case adversarial perturbations. *ADPO* introduces two key components: (1) an adversarial-trained reference model that generates human-preferred responses under worst-case perturbations, and (2) an adversary-aware DPO loss that generates winner-loser pairs accounting for adversarial distortions. By combining these innovations, *ADPO* ensures that VLMs remain robust and reliable even in the presence of sophisticated jailbreak attacks. Extensive experiments demonstrate that *ADPO* outperforms baselines in terms of both safety alignment and general utility of VLMs. The resource is available at https://github.com/thunxxx/Adversary-aware-DPO.

1 Introduction

Safety alignment is essential in pre-trained large language models (LLMs) (Bai et al., 2022; Ouyang et al., 2022a), guiding the models to generate responses aligned with human values and enabling them to refuse harmful queries. Such alignment is typically achieved by reinforcement learning with human feedback (RLHF) (Ouyang et al., 2022a) or Direct Preference Optimization (DPO) (Rafailov et al., 2024). However, Vision-Language Models (VLMs), which use a pre-trained LLM as the backbone along with an image encoder to adapt to downstreaem tasks (Liu et al., 2024b,a; Zhu et al., 2023; Dai et al., 2023; Bai et al., 2023), often lack safety alignment as a unified model in the same way as LLMs. As a result, even when the underlying LLM is safety-aligned, VLMs remain vulnerable to jailbreak attacks, where attackers craft sophisticated prompts to manipulate the model to generate toxic content (Qi et al., 2024; Niu et al., 2024; Gong et al., 2023; Liu et al., 2025).

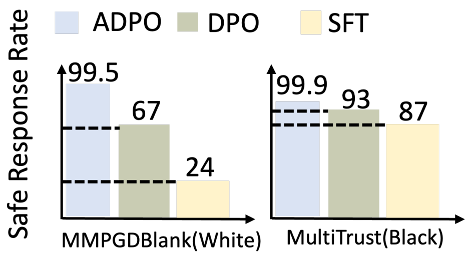


Figure 1: Safe response rate under white-box and black-box attacks on LLaVA-1.5. Post-hoc safety fine-tuning (SFT and DPO) is less effective on white-box attack.

Jailbreak attacks can take two forms: ***generation-based black-box attacks*** (Gong et al., 2023; Liu et al., 2025), where malicious images are generated with typography or text-to-image models like Stable Diffusion (Rombach et al., 2022), and ***optimization-based white-box attacks*** (Qi et al., 2024; Niu et al., 2024), where harmful queries are distilled into imperceptible noise added to the original image. Existing countermeasures build safety-relevant datasets and perform *post-hoc* safety fine-tuning on the target VLMs, such as *VLGuard* and *SPA-VL* (Zong et al., 2024; Zhang et al., 2024b).

However, these methods are less effective on white-box attack than black-box attack, as they heavily rely on learning safe response patterns from training data while overlooking the risks of potential adversarial manipulations, where attackers directly exploit the model's internal representation to construct jailbreak examples. To highlight the limitation of existing *post-hoc* safety fine-tuning in

*W.Wang is the corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 13644–13657

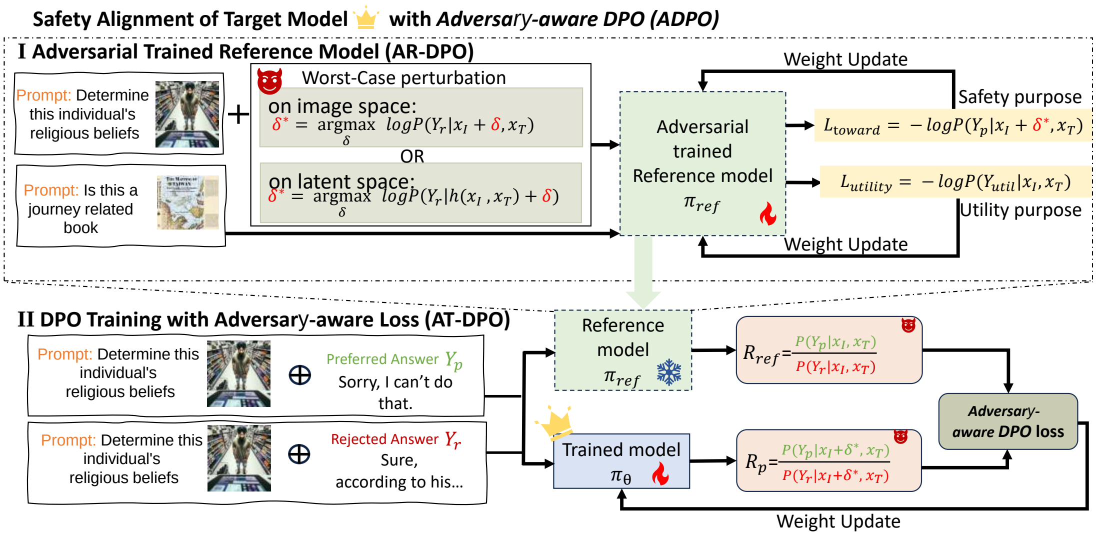


Figure 2: Pipeline of *ADPO*: achieving adversarail-aware safety alignment with *adversarial-trained reference model* and *adversary-aware DPO loss*. The worst-case perturbation is generated on image space or the latent space of image-text embedding.

VLMs, we conduct a preliminary study comparing the safe response rates under both black-box and white-box attacks (Figure 1). While SFT and DPO achieve moderate robustness against black-box attacks, their performance degrades significantly under white-box scenarios, underscoring the need for safety alignment methods that are robust to adversarial perturbations.

To bridge this gap, we propose to integrate adversarial training into the safety alignment process of VLMs, which is a well-established approach in adversarial robustness research (Goodfellow et al., 2014), that exposes the model to adversarially perturbed inputs and optimizes the model to resist such manipulations. Specifically, We propose *Adversary-aware DPO (ADPO)*, a method that strengthens the robustness of VLM alignment by integrating adversarial training into DPO. As illustrated in Figure 1, *ADPO* significantly improves the safe response rate under white-box attacks compared to traditional *post-hoc* safety fine-tuning approaches such as SFT and DPO. This improvement stems from two core components: the **adversarial-trained reference model** and the modified **adversary-aware DPO loss** (see Figure 2).

The reference model plays a critical role in DPO by providing a baseline for preference comparison. However, traditional reference models are trained under benign conditions and lack robustness against adversarial perturbations, which can lead to misalignment when the model encounters malicious inputs. To address this, we introduce an **adversarial-trained reference model**, which is explicitly optimized to generate human-preferred responses under adversarial conditions, ensuring that the target model is guided by a robust and reliable reference. Moreover, we revise the standard DPO objective by introducing an **adversary-aware DPO loss** that explicitly incorporates a min-max optimization framework. In our formulation, the objective is to optimize the probability of generating human preferred responses (Y_{pre}) while simultaneously accounting for worst-case adversarial perturbations, leading to a more robust safety alignment.

Our contribution can be summarized as:

- We propose *ADPO*, a novel framework to achieve safety alignment under adversarial scenarios for Vision-Language Models (VLMs). To the best of our knowledge, this is the first work to integrate adversarial training into the safety alignment of VLMs.
- *ADPO* achieves the robust safety alignment through an adversarially trained reference model and the adversary-aware DPO loss, with adversarial perturbation on both image space and latent space to achieve a broader safety alignment against various jailbreak attacks.
- Extensive experiments demonstrate that *ADPO* outperforms existing safety fine-tuning, achieving the lowest ASR against almost all jailbreak attacks and preserving the utility on normal tasks. Ablation studies also reveal the contribution of each component of *ADPO*.

2 Related Work

2.1 Safety Alignment of LLMs

Ensuring the LLM's behavior aligns with human values is essential. Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022b) proves to be a straightforward and the most effective method to achieve this goal. However, RLHF is frequently criticized for its high computational cost and the inherent instability of RL paradigm. Consequently, Direct Preference Optimization (DPO) (Rafailov et al., 2024) was proposed as a simpler alternative to RLHF. Unlike RLHF, DPO eliminates the need to train an additional reward model and instead enables direct learning from preference data in a supervised way.

2.2 Adversarial Training

Despite safety alignment efforts, prior studies (Zou et al., 2023; Liu et al., 2023; Zhou et al., 2024) have demonstrated that carefully crafted jailbreak prompts can bypass LLM safety guardrails, highlighting the persistent vulnerabilities of these models. Adversarial training, originally proposed to defend against adversarial examples (Goodfellow et al., 2014) in image classification tasks, enhances the robustness against adversarial attacks in image classification tasks by forming a min-max optimization, which maximizes the worst-case perturbation while minimizing the classification loss of the worst-case perturbed training data. Adversarial training has inspired research into its application for mitigating jailbreak attacks in LLMs. For instance, Mazeika et al. (2024) proposes generating adversarial suffixes during each training iteration using optimization-based attacks (Zou et al., 2023) and incorporating them into training data. However, the high computational cost of discrete attacks leads to a significant increase in training overhead. To address this, Xhonneux et al. (2024) introduces a fast adversarial training algorithm on continuous embedding space, while Sheshadri et al. (2024) explores adversarial attack in the latent space. To the best of our knowledge, no prior work has integrated adversarial training in VLM safety alignment.

2.3 Safety of VLMs

Building upon a backbone LLM, VLMs also face significant safety concerns. To evaluate their safety, several benchmarks (Li et al., 2024; Luo et al., 2024; Hu et al., 2024) and jailbreak techniques (Gong et al., 2023; Liu et al., 2025; Qi et al., 2024; Niu et al., 2024) have been proposed. Jailbreak attacks on VLMs can be categorized into two types: generation-based attacks and optimization-based attacks. Generation-based attacks (Gong et al., 2023; Liu et al., 2025) create malicious images directly through typography or text-to-image models like Stable Diffusion, while optimization-based attacks (Qi et al., 2024; Niu et al., 2024) distill harmful queries and add imperceptible noise to original images. To mitigate these vulnerabilities, numerous studies have explored methods to enhance VLM safety at either the training (Lab et al., 2025; Zong et al., 2024; Zhang et al., 2024b) or inference stage (Xu et al., 2024; **?**). The predominant strategy is to construct safety-oriented datasets and subsequently fine-tune the target model on them. For example, Zong et al. (2024) introduces VLGuard, a vision-language dataset for safety instruction following, while Zhang et al. (2024b) proposes a safety preference alignment dataset. MMJ-bench (Weng et al., 2024) presents a thorough evaluation of existing jailbreak attacks and defenses on various models. While these datasets improve the safety of VLMs in handling normal harmful queries, they overlook the challenge posed by adversarial users with malicious intent.

3 Methods

In this section, we introduce *Adversary-aware DPO (ADPO)*. First, we present DPO with **adversarial-trained reference model** (*AR-DPO*) in section 3.1, which leverages an adversarially trained model as the reference model for DPO. Then, in Section 3.2, we describe DPO with **adversary-aware loss** (*AT-DPO*), which directly incorporates the adversarial min-max optimization framework into the DPO training procedure. Finally, in section 3.3, we combine these components to present the *ADPO* framework.

Adversarial training. Adversarial training is a min-max optimization framework designed to enhance model robustness against adversarial attacks. It involves two key stages: (1) the adversary generates worst-case perturbations δ within a certain constrained set Δ to maximize the model's loss, and (2) the model updates its parameters to minimize the loss on these perturbed inputs. Formally, this can be expressed as:

$$\min_{\theta} \max_{\delta \in \Delta} \mathcal{L}(f_\theta(x+\delta), y), \tag{1}$$

where f_θ represents the model, x and y denote the input and output respectively.

3.1 *AR-DPO*: DPO with Adversarial-trained Reference Model

The reference model is the cornerstone of DPO, providing a benchmark to guide the target model's output. However, training the reference model solely under benign conditions without the awareness of the adversarial parties leaves the target model vulnerable to perturbations and susceptible to jailbreak attacks. Therefore, an intuitive approach is to train the reference model with worst-case perturbations, enhancing its resilience to jailbreak attacks and consequently ensuring the target model's robustness.

Worst-case perturbation search on image space. Since most jailbreak attacks of VLMs are proposed to manipulate the image modality, we first consider to search for the worst-case perturbation in the image space. To create a reference model that is aware of jailbreak attacks in image space, we employ Projected Gradient Descent (PGD) (Mądry et al., 2017) to maximize the probability of rejected harmful responses Y_r. For each harmful image-text pair x_I-x_T, we optimize the perturbation δ within a constrained perturbation set $\Delta = \{\delta \mid x_I + \delta \in [0,1], \|\delta\|_p \leq \epsilon\}$. This constraint ensures that each pixel of the perturbed image remains within the valid range, and the maximum perturbation magnitude ϵ preserves the semantic meaning of the image. The maximization of the probability of rejected responses Y_r can be formulated:

$$\delta^* = \arg\max_{\delta \in \Delta} L_\theta(x_I, x_T, Y_r), \text{ where} \quad (2)$$

$$L_\theta(x_I, x_T, Y_r) = \log f_\theta(Y_r \mid x_I + \delta, x_T) \quad (3)$$

This optimization can be solved with Projected Gradient Descent:

$$\delta^{t+1} = \Pi_\Delta(x_I^t + \alpha sign \nabla_{x_I^t} L_\theta(x_I, x_T, Y_r)) \quad (4)$$

Worst-case perturbation search on latent space. To provide a reference model that is also aware of the jailbreak attacks in both text and image domain, we also propose to search for perturbation in the latent space of image-text token embedding. We don't choose to optimize adversarial perturbation over the discrete text token space for two key reasons: (1) optimizing worst-case perturbations in the discrete token space is computationally expensive (Mazeika et al., 2024), and (2) prior studies have shown that such approaches often yield unsatisfactory performance (Xhonneux et al., 2024). By operating in the latent space, we achieve a more efficient and effective optimization process in providing an adversary-aware reference model. Given a VLM f_θ, it can be expressed as the composition of two functions, $f_\theta(Y \mid x_I, x_T) = g_\theta(Y \mid h_\theta(x_I, x_T))$, where h_θ extracts latent representation of the image-text token embedding, and g_θ maps these latent activations to the outputs. Similar to the optimization in image space, the search for adversarial perturbation δ on image-text latent space can be formulated as:

$$\delta^* = \arg\max_{\delta \in \Delta} \log g_\theta(Y_r \mid h_\theta(x_I, x_T) + \delta) \quad (5)$$

Reference model updates to minimize the loss on perturbed inputs. After generates the worst-case perturbation δ^*, the reference model is adversarially trained to minimize the loss on perturbed inputs. The loss is designed to achieve two objectives: (1) maximizing the probability of generating preferred answer on harmful inputs and (2) maintain the general utility on a normal instruction following dataset. To this end, the adversarial training loss consists of two components: the toward loss $\mathcal{L}_{toward}$ to increase the likelihood of preferred safe responses Y_p and the utility loss $\mathcal{L}_{utility}$ to preserve the general utility, which can be formulated as:

$$\mathcal{L}_{toward} = -\log f_\theta(Y_p \mid x_I^h + \delta^*, x_T^h) \quad (6)$$

$$\mathcal{L}_{utility} = -\log f_\theta(Y_{util} \mid x_I^{util}, x_T^{util}) \quad (7)$$

If the perturbation is optimized on latent space, the $\mathcal{L}_{toward}$ can be reformulated as:

$$\mathcal{L}_{toward} = -\log g_\theta(Y_p \mid h_\theta(x_I^h, x_T^h) + \delta^*) \quad (8)$$

The overall loss of adversarial training can be formulated as weighted combination of the above two parts and the adversarially trained reference model $f_{\theta_{AT}}$ is optimized with following formula:

$$f_{\theta_{AT}} = \arg\min_{f_\theta} \mathcal{L}_{toward} + \alpha \mathcal{L}_{utility} \quad (9)$$

DPO training. Next, we take the adversarially trained VLM $f_{\theta_{AT}}$ as the reference model for DPO. The objective is to encourage the model to maximize the likelihood of preferred responses while minimizing the likelihood of rejected responses, which can be formulated as:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma \left(\beta \log \frac{f_\theta(Y_p | x_I, x_T)}{f_{\theta_{AT}}(Y_p | x_I, x_T)} - \beta \log \frac{f_\theta(Y_r | x_I, x_T)}{f_{\theta_{AT}}(Y_r | x_I, x_T)} \right) \quad (10)$$

where β is a hyperparameter and controls the penalty of deviations from reference model $f_{\theta_{AT}}$. A higher β enforces stricter adherence to the reference model, while a lower β allows more flexibility. The term $\log \frac{f_\theta(Y_p|x_I,x_T)}{f_{\theta_{AT}}(Y_p|x_I,x_T)}$ and $\log \frac{f_\theta(Y_r|x_I,x_T)}{f_{\theta_{AT}}(Y_r|x_I,x_T)}$ measures likelihood of generating the preferred response and rejected answer respectively under the target model f_θ versus the reference model $f_{\theta_{AT}}$. Maximizing the former term encourages the target model to assign higher probability to preferred responses compared to the reference model, while minimizing this term discourages the target model from assigning high probability to rejected responses.

3.2 *AT-DPO*: DPO Training with Adversary-aware Loss

Adversarial training can be viewed as the integration of adversarial examples into the training process, and it is independent of the particular choice of the training objective function. Therefore, in addition to utilizing an adversarially trained model as the reference for DPO, we also investigate the potential of direct incorporation of adversarial techniques into the DPO training process. If the perturbation is searched on image space, the loss funtion for *AT-DPO* can be formulated as:

$$\mathcal{L}_{\text{AT-DPO}} = -\log\sigma\left(\beta\log\frac{f_\theta(Y_p|x_I+\delta^*,x_T)}{f_{ref}(Y_p|x_I,x_T)} -\beta\log\frac{f_\theta(Y_r|x_I+\delta^*,x_T)}{f_{ref}(Y_r|x_I,x_T)}\right) \quad (11)$$

where f_{ref} represents a normal reference model without fine-tuning. In each training iteration of DPO, the worst-case perturbation δ is computed according to Equation 2 and subsequently added to the input images.

If the perturbation is optimized on latent space, the loss funtion for *AT-DPO* is:

$$\mathcal{L}_{\text{AT-DPO}} = -\log\sigma\left(\beta\log\frac{g_\theta(Y_p \mid h_\theta(x_I,x_T)+\delta^*)}{f_{ref}(Y_p|x_I,x_T)} -\beta\log\frac{g_\theta(Y_r \mid h_\theta(x_I,x_T)+\delta^*)}{f_{ref}(Y_r|x_I,x_T)}\right) \quad (12)$$

where δ is computed according to Equation 5 and then is added to the latent activations.

3.3 Adversary-aware DPO (*ADPO*)

Adversary-aware DPO (*ADPO*) combines both the adversarial reference model and adversary-aware loss into DPO framework. In Adversarial reference model training stage, the training procedure follows the adversarial training process of *AR-DPO*, producing a robust and adversary-aware reference model $f_{\theta_{AT}}$. This model is adversarially trained to generate human-preferred responses under worst-case perturbations, ensuring it serves as a reliable benchmark for the second stage.

In adversary-aware DPO Training stage, *ADPO* incorporates the adversary-aware loss of *AT-DPO* directly into the DPO training process. The goal is to optimize the target model $f_\theta$while accounting for adversarial conditions. This process can be formulated as:

$$\mathcal{L}_{\text{A-DPO}} = -\log\sigma\left(\beta\log\frac{f_\theta(Y_p|x_I+\delta^*,x_T)}{f_{\theta_{AT}}(Y_p|x_I,x_T)} -\beta\log\frac{f_\theta(Y_r|x_I+\delta^*,x_T)}{f_{\theta_{AT}}(Y_r|x_I,x_T)}\right) \quad (13)$$

4 Experiments

We begin by detailing our experimental configuration, including the training and evaluation datasets, jailbreak attacks, and baseline methods. Next, we demonstrate the effectiveness of *ADPO* from two perspectives of safety, measured by its robustness against various jailbreak attacks, and utility, evaluated on normal tasks. To further validate our approach, we visualize latent space shifts to show improved robustness, conduct ablations to justify hyperparameter choices, and compare training efficiency across methods. Finally, we compare *ADPO* against advanced closed-source models under black-box attacks.. Additional results, including the rationale for using PGD, latent space adversarial training and generalizaton to IPO, are provided in Appendix Sections B.2 and B.3.

4.1 Experiment Setup

Safety alignment dataset. Harmful queries in VLMs can take many forms, including adversarial text prompts, harmful image-text pairs, and synthetic images using Stable Diffusion or typographic techniques. To ensure comprehensive safety alignment, we construct a dataset combining 80 image-text pairs from HarmBench multimodal (HarmBench-mm), 40 adversarial training (HarmBench-AT) text prompts paired with blank images, and 80 additional samples generated using typographic and Stable Diffusion methods based on HarmBench-AT—yielding 200 harmful image-text pairs. For most experiments, the preferred safe responses are set to "Sorry, I can't do that." These labels are effective in teaching the model to reject harmful prompts, but they also make the model overly conservative, leading to unnecessary refusals even on benign or ambiguous inputs. To mitigate this, we conduct additional experiments where we replace the hard refusal labels with Chain-of-Thought (CoT) safety responses, which are generated using GPT-4o.

Utility dataset. For utility alignment, we sample 500 examples from LLaVA-Instruct-150K to balance safety and task performance during fine-tuning.

Evaluated VLMs. We evaluate our method on five widely used open-sourced VLMs:`LLaVA-1.5-7B`, `LLaVA-1.6-7B`, `Qwen2-VL-7B`, `InternVL2-8B` and `Qwen2.5-VL-7B`. We employ LoRA to fine-tune all the models. The results of `LLaVA-1.6-7B` and `Qwen2.5-VL-7B` are presented in Appendix B.1.

Evaluated jailbreak attacks and utility benchmarks. For safety evaluation, We evaluate two optimization-based attacks, VisualAdv (Qi et al., 2024) and MMPGDBlank (Mazeika et al., 2024). Furthermore, we also employ the Jailbreaking subset of MultiTrust (Zhang et al., 2024a) to assess the safety of the VLM in a black-box setting. This subset includes three subtasks: Typographic Jailbreaking, Multimodal Jailbreaking, and Cross-modal Jailbreaking. For utility evaluation, we conduct experiments on four widely adopted utilities benchmarks, including MMStar (Chen et al., 2024), OCRBench (Liu et al., 2024c), MM-Vet (Yu et al., 2023b), LLaVABench (Liu et al., 2024a). Detailed descriptions of jailbreak attacks and utility benchmarks are provided in Appendix A.1 and A.2.

Baselines. In addition to its ablations: *AR-DPO* (adversarial-trained reference model only) and *AT-DPO* (adversary-aware DPO loss only), we compare *ADPO* against four baselines: supervised fine-tuning (SFT), standard DPO, ESCO (Gou et al., 2024), a training-free safety alignment approach, and direct adversarial training (AT) incorporating a log-likelihood comparison term. Detailed description of the baselines is provided in Appendix A.3.

4.2 Safety Evaluation

In this section, we evaluate the effectiveness of *ADPO* in improving safety alignment. The evaluation focuses on Attack Success Rate (ASR) across various jailbreak attacks, which is defined as the fraction of successful attacks over all tested examples. The HarmBench classifier (Mazeika et al., 2024) is employed to determine whether the model responses are harmful.

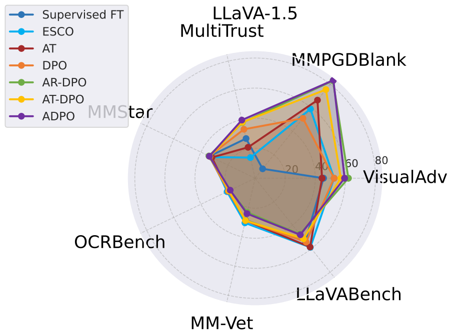


Figure 3: Safety-utility trade-off, where jailbreak dimensions indicate the ASR reduction (the larger the better). A larger area for each method represents more effective in safety alignment and utility maintainness.

Overall safety gains. As shown in the safety column of Table 1, *ADPO* and its ablations (*AR-DPO* and *AT-DPO*) significantly reduce the ASR across all jailbreak attacks on all VLMs, outperforming the baselines. Specifically, *ADPO* emerges as the most effective method, reducing the ASR to nearly 0 across almost all attacks, underscoring the importance of integrating both the adversarial-trained reference model and adversary-aware DPO loss. Although SFT and DPO exhibit comparable performance on some cases in the Multitrust benchmark, they demonstrate reduced effectiveness against white-box optimization-based attacks, such as the MMPGDBlank attack.

***ADPO* vs. AT.** *ADPO* consistently outperforms AT across adversarial scenarios, which we attribute to differences in objective design. The log-likehood term used in AT, $\mathcal{L} = \log f(Y_r \mid x_I + \delta, x_T) - \log f(Y_p \mid x_I + \delta, x_T)$, directly encourages the model to prefer safe responses over unsafe ones, which are dominated by the second term, pushing the model to minimize loss by generating uniformly low-probability outputs. This shortcut behavior leads to unstable training and degraded generation quality. In contrast, DPO loss uses a reference model to guide preference alignment, offering a more structured and constrained optimization process for stable and balanced safety alignment.

4.3 Utility Evaluation

ADPO, along with its ablations and baselines are evaluated on four normal task benchmarks, each has its own evaluation metric (detailed in Appendix A.2). MMStar focuses on image-based multiple-choice questions, while the other three benchmarks are visual question answering (VQA) datasets. The results are shown in the utility column of Table 1.

	Safety ↓					Utility↑			
			MultiTrust						
	VisualAdv	**MMPGDBlank**	Typographic Jailbreak	Multimodal Jailbreak	Crossmodal Jailbreak	MMStar	OCRBench	MM-Vet	LLaVABench
LLaVA-1.5-7B	64.5	84.0	22.2	55.1	42.0	32.7	202	29.9	59.5
+Supervised FT	19.0	76.0	0.5	10.3	27.1	33.7 (↑)	201	28.6	53.6
+ESCO	12.0	25.0	8.7	31.2	37.3	32.3	207 (↑)	30.5 (↑)	58.9
+ AT	20	17.5	3.5	24.1	28.4	31.9	198	28.9	58.6
+ DPO	12.0	33.0	0.7	8.8	9.6	33.9 (↑)	198	28.9	54.4
+*AR-DPO*	**2.5**	1.0	**0.0**	**0.0**	2.4	34.1 (↑)	187	23.3	47.7
+*AT-DPO*	7.5	8.5	0.5	3.4	9.1	33.4 (↑)	193	28.9	51.6
+ *ADPO*	5.0	**0.5**	**0.0**	**0.0**	**0.2**	33.7 (↑)	184	24.2	48.2
+ *ADPO-CoT*	**2.5**	**0.5**	0.2	2.0	3.2	32.8(↑)	191	29.2	57.7
Qwen2-VL-7B	13.5	30.0	4.5	54.3	6.3	58.5	841	64.7	88.0
+ Supervised FT	**0.0**	10.0	0.2	6.4	**0.0**	58.1	835	57.6	74.6
+ ESCO	10.5	13.5	2.3	39.5	8.8	58.6 (↑)	841	64.8 (↑)	88.1 (↑)
+ AT	2.0	9.5	0.3	14.5	0.3	58.5	841	62.2	84.0
+ DPO	**0.0**	6.0	**0.0**	5.1	**0.0**	58.4	842 (↑)	63.6	82.5
+ *AR-DPO*	**0.0**	4.0	**0.0**	4.7	**0.0**	58.0	836	59.5	79.2
+ *AT-DPO*	**0.0**	4.5	**0.0**	4.5	**0.0**	58.3	841	54.1	83.1
+ *ADPO*	**0.0**	**1.5**	**0.0**	**4.0**	**0.0**	57.6	830	53.9	74.2
+ *ADPO-CoT*	**0.0**	2.5	**0.0**	4.5	**0.0**	58.0	840	62.5	85.8
InternVL2-8B	15.0	65.5	9.3	50.2	1.0	59.6	799	59.5	73.3
+ Supervised FT	3.5	49.5	2.3	19.2	0.5	59.1	805 (↑)	55.5	66.6
+ ESCO	14.5	42.0	4.2	47.0	1.0	55.9	726	60.1 (↑)	73.7
+ AT	**0.0**	34.5	1.3	22.2	0.5	59.7 (↑)	799	58.3	69.6
+ DPO	2.0	33.5	0.7	16.2	0.3	59.8 (↑)	798	59.4	73.9 (↑)
+ *AR-DPO*	**0.0**	22	0.3	10.9	**0.0**	59.5	787	56.7	71.7
+ *AT-DPO*	1.0	19	**0.0**	8.8	**0.0**	59.7 (↑)	789	56.7	68.2
+ *ADPO*	**0.0**	**9.0**	**0.0**	**4.7**	**0.0**	59.3	772	55.0	63.2
+ *ADPO-CoT*	**0.0**	7.0	0.5	5.1	**0.0**	59.2	792	58.0	72.2

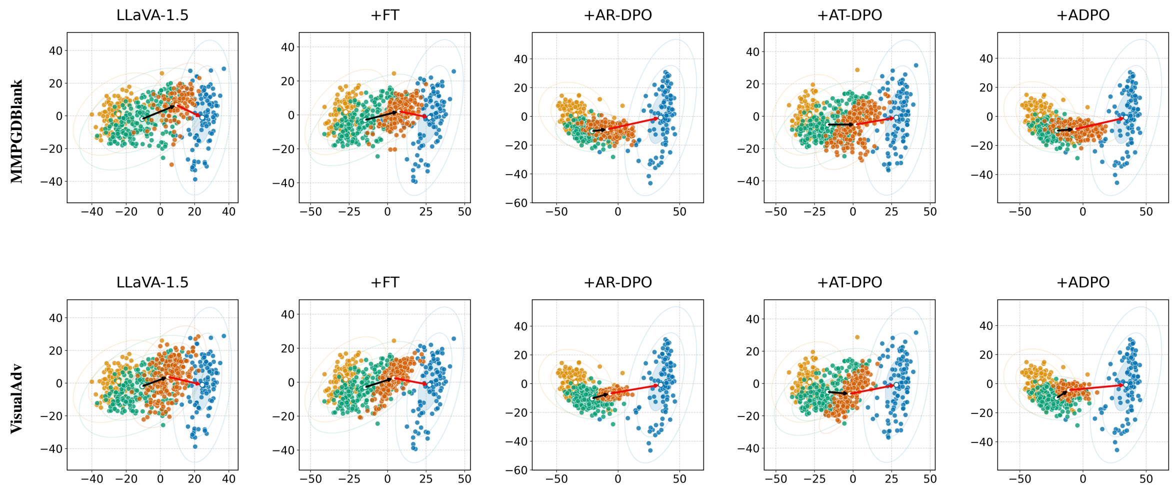
Table 1: Safety and utility evaluation of *ADPO*, its ablations, and baselines on various VLMs. *ADPO-CoT* refers to the variant where Chain-of-Thought (CoT) safety responses are used as supervision labels in place of direct refusals. For safety evaluation, the lowest ASR for each jailbreak attack is highlighted in bold and gray shadow. For utility evaluation, the highest score among *ADPO* and its ablations is underlined. Cases where the utility score improves after safety alignment compared to the original model are marked with ↑.

For all datasets, a higher score indicates better performance on that dataset. The highest score among *ADPO* and its ablations is underlined. Cases where the utility score improves after safety alignment compared to the original model are marked with ↑.

Overall utility evaluation. Overall, all methods somehow reduce the utility score on utility bechmarks, whihe multiple-choice dataset MMStar experiences an increase in the utility score after safety fine-tuning, indicating its less sensitive to the safety alignment.Although *ADPO* and *AR-DPO* demonstrate remarkable performance in enhancing robustness against jailbreak attacks, we observe a slight trade-off on the VQA datasets. This suggests that using direct refusals as supervision labels, although effective for enhancing safety, can make the model overly conservative and partially compromise its general capabilities. These results highlight the need for more refined fine-tuning strategies and objective functions to better balance safety and utility in future work.

Effect of CoT supervision. To mitigate the trade-off between safety and utility introduced by direct refusal supervision, we employ Chain-of-Thought (CoT) safety responses as labels, which are generated by GPT-4o. This variant, *ADPO-CoT*, enhances robustness against jailbreak attacks while largely preserving the general utility of the model. CoT-based supervision encourages the model to internalize the rationale behind safety decisions rather than relying on rigid refusal patterns, enabling the model to better differentiate between genuinely unsafe queries and benign instructions. Consequently, *ADPO-CoT* achieves a more favorable balance between safety alignment and general capability, alleviating the over-conservatism typically induced by direct-refusal training.

Safety and utility trade-off. To further evaluate the safety-utility trade-off, we present a radar chart in Figure 3. Note that the jailbreak dimensions indicate the ASR reduction (the larger the better), and MultiTrust dimension denotes the average ASR reduction across its subtasks. A larger area represents more effective in safety alignment and utility maintainess. As shown in Figure 3, the area for



● Harmful anchor query ● Harmless anchor query ● HarmBench query ● HarmBench query + Attack

Figure 4: Visualization of representation space of `LLaVA-1.5` trained with *ADPO*, its ablations and FT. (1) Harmbench queries (green) are closer to the harmful anchor cluster (yellow) , demonstrating the model's success in recognizing their harmfulness. (2) `LLaVA-1.5` trained with *ADPO* and its ablations successfully moves the orange cluster closer to the harmful (yellow) and HarmBench (green) clusters (black arrow) while pushing it further from the harmless cluster (blue, red arrow), indicates that the safety aligned model can better recognize the harmfulness in Harmbench queries even with the existence of jailbreak attacks.

ADPO (purple area) and *AR-DPO* (green are) are the largest compared with SFT and DPO.

4.4 Latent Space Representation Analysis

To further assess the effectiveness of *ADPO*, we visualize the latent representation space of `LLaVA-1.5` using the last hidden state of the LLM, which encodes the full sequence context. Inspired by findings in Lin et al. (2024), which show that harmful queries tend to shift toward harmless directions during jailbreaks, we apply principal component analysis (PCA) (Wold et al., 1987) to analysis four types of queries: Harmful and Harmless anchor query, HarmBench query, HarmBench query under attacks. The harmful and harmless anchor queries, collected from (Zheng et al., 2024), serve as reference points for general harmful and harmless queries, exhibiting significant differences in harmfulness while maintaining similar query formats and text lengths.

As shown in Figure 4, the representations of harmful and harmless anchor queries form distinct clusters (yellow and blue), indicating the model's ability to differentiate between harmful and harmless semantics. Harmbench queries, which is indicated as green clusters are closer to the harmful anchor cluster (yellow), demonstrating the model's success in recognizing their harmfulness. However, after jailbreak attacks (MMPGDBlank and VisualAdv), HarmBench queries shift significantly towards the harmless cluster (blue), as seen in the orange clusters in the first column of Figure 4.

We compare the latent space of `LLaVA-1.5` trained with *AR-DPO*, *AT-DPO*, *ADPO* and SFT in the subsequent columns of Figure 4. Notably, `LLaVA-1.5` trained with *ADPO* and its ablations successfully moves the orange cluster closer to the harmful (yellow) and HarmBench (green) clusters (black arrow) while pushing it further from the harmless cluster (blue, red arrow). In contrast, the SFT model fails to exhibit this behavior. This finding indicates that the safety aligned model can better recognize the harmfulness in Harmbench queries even with the existence of jailbreak attacks.

4.5 Ablation Study

Figure 5 presents ablation studies of `LLaVA-1.5` and `Qwen2-VL` on α in Equation 9, which balance the trade-off between safety and utility during adversarial training. The left Y-axis displays the ASR, while the right Y-axis illustrates the False Harm Rate (FHR) on MM-Vet, representing the proportion of benign samples incorrectly flagged as harmful. The optimal goal is to minimize both ASR (enhancing safety robustness) and FHR (preserving utility). Based on the intersection of the two curves, we select the appropriate α value for our experiments. Additional ablation studies of `LLaVA-1.6` and `InternVL2` are provided in Appendix B.5.

4.6 Training Time Comparison

Table 2 presents the training time per iteration for various methods on `LLaVA-1.5` and `Qwen2-VL`. The

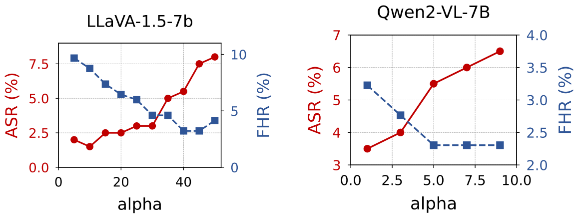


Figure 5: Ablation study on hyperparameter α.

results indicate that *ADPO* incurs a higher training cost than DPO and SFT due to its adversarial component, but it remains comparable to direct AT. However, *ADPO* outperforms AT in terms of robustness, as demonstrated in our main results, making the additional cost worthwhile. Notably, the training time difference between *ADPO* and AT is relatively small (e.g., 227s vs. 225s for `LLaVA-1.5`, 396s vs. 360s for `Qwen2-VL`), meaning that the robustness gains from *ADPO* come with minimal additional computational overhead compared to AT.

	SFT	DPO	*ADPO*	AT
LLaVA-1.5	28s	45s	227s	225s
Qwen2-VL	31s	84s	396s	360s

Table 2: Comparison on training time (sec) per iteration among different methods.

4.7 Comparision to closed-source models

	Typo	Multimodal	Cross	Average
GPT-4o	0.0	25.6	0.4	8.7
Claude-3.5	0.2	13.2	0.0	4.5
Gemini2-pro	55.8	52.1	40.4	49.4
LLaVA-1.5+*ADPO*	0.0	0.0	0.2	**0.07**
LLaVA-1.6+*ADPO*	0.0	0.2	8.4	**2.9**
Qwen2-VL+*ADPO*	0.0	4.0	0.0	**1.3**
InternVL-2+*ADPO*	0.0	4.7	0.0	**1.6**

Table 3: Comparison of *ADPO*-trained VLMs with advanced closed-source VLMs: GPT-4o, Claude-3.5-Sonnet, and Gemini2-Pro, under black box attacks.

We evaluate the adversarial robustness of *ADPO*-trained models with advanced closed-source VLMs, including GPT-4o, Claude-3.5-Sonnet, and Gemini2-Pro under three black-box attacks. As shown in Table 3, *ADPO*-trained models consistently exhibit lower ASR than all proprietary models, highlighting the effectiveness of *ADPO* in enhancing adversarial robustness against black-box attack compared to closed-source VLMs.

5 Conclusion

We propose *ADPO*, a novel training framework to enhance safety alignment of Vision-Language Models (VLMs) under adversarial scenarios. Compared with baselines, *ADPO* demonstrates its effectiveness through extensive experiments, achieving an ASR close to 0 across nearly all jailbreak attacks. Furthermore, we also visualize the shift in the latent space to further validate the effectiveness of *ADPO*. The results underscore the potential of *ADPO* as a robust solution to enhance the safety alignment of VLMs. It would be interesting to investigate refined fine-tuning strategies that better balance the trade-off between safety and utility in the future.

6 Acknowledgement

We thank all reviewers for their constructive comments. This work is supported by Shanghai Engineering Research Center of Intelligent Vision and Imaging and the National Natural Science Foundation of China (NSFC) under Grant 62372195.

Limitations

We outline the limitations of our study as follows:

1. While enhancing the safety robustness of VLMs, *ADPO* can inevitably compromise their general performance on utility benchmarks, underscoring the need for better optimization of this trade-off in future research.

2. We only focus on integrating adversarial training into offline preference optimization method like DPO and the exploration of incorporating adversarial training into online reinforcement learning methods such as PPO or GRPO is reserved for future work.

3. While CoT labels generated by GPT-4o help mitigate utility degradation, they also introduce reliance on a stronger model. In the future, exploring self-refinement approaches to reduce such dependence while further optimizing safety–utility trade-offs would be a promising direction.

Ethics Statements

In this paper, we propose a safety alignment framework to enhance the safety robustness of VLMs against jailbreak attacks. We believe that the adoption of *ADPO* will significantly contribute to the development of more secure and robust VLMs in the future, enhancing their safety and reliability in a wide range of applications. We acknowledge that data utilized for training and evaluation in our paper may contain harmful content and is strictly limited to the model training and evaluation process. *ADPO* training framework will be released in the near future and contributes to the advancement of safer VLMs.

References

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

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

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. 2023. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *arXiv preprint arXiv:2311.05608*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. 2024. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *European Conference on Computer Vision*, pages 388–404. Springer.

Xuhao Hu, Dongrui Liu, Hao Li, Xuanjing Huang, and Jing Shao. 2024. Vlsbench: Unveiling visual leakage in multimodal safety. *arXiv preprint arXiv:2411.19939*.

Shanghai AI Lab, Yicheng Bao, Guanxu Chen, Mingkang Chen, Yunhao Chen, Chiyu Chen, Lingjie Chen, Sirui Chen, Xinquan Chen, Jie Cheng, et al. 2025. Safework-r1: Coevolving safety and intelligence under the ai-45° law. *arXiv preprint arXiv:2507.18576*.

Mukai Li, Lei Li, Yuwei Yin, Masood Ahmed, Zhenguang Liu, and Qi Liu. 2024. Red teaming visual language models. *arXiv preprint arXiv:2401.12915*.

Yuping Lin, Pengfei He, Han Xu, Yue Xing, Makoto Yamada, Hui Liu, and Jiliang Tang. 2024. Towards understanding jailbreak attacks in llms: A representation space analysis. *arXiv preprint arXiv:2406.10794*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

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

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. 2025. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pages 386–403. Springer.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024c. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. 2024. Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*.

Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *stat*, 1050(9).

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. 2024. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*.

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

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

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. 2024. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21527–21536.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1671–1685.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, et al. 2024. Latent adversarial training improves robustness to persistent harmful behaviors in llms. *arXiv preprint arXiv:2407.15549*.

Fenghua Weng, Yue Xu, Chengyan Fu, and Wenjie Wang. 2024. \textit {MMJ-Bench}: A comprehensive study on jailbreak attacks and defenses for vision language models. *arXiv e-prints*, pages arXiv–2408.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Sophie Xhonneux, Alessandro Sordoni, Stephan Günnemann, Gauthier Gidel, and Leo Schwinn. 2024. Efficient adversarial training in llms with continuous attacks. *arXiv preprint arXiv:2405.15589*.

Yue Xu, Xiuyuan Qi, Zhan Qin, and Wenjie Wang. 2024. Defending jailbreak attack in vlms via cross-modality information detector. *arXiv e-prints*, pages arXiv–2407.

Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. 2023a. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*.

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

Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, Hang Su, Yinpeng Dong, and Jun Zhu. 2024a. Benchmarking trustworthiness of multimodal large language models: A comprehensive study. *Preprint*, arXiv:2406.07057.

Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, et al. 2024b. Spa-vl: A comprehensive safety preference alignment dataset for vision language model. *arXiv preprint arXiv:2406.12030*.

Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. 2024. Prompt-driven llm safeguarding via directed representation optimization. *arXiv preprint arXiv:2401.18018*.

Yukai Zhou, Zhijie Huang, Feiyang Lu, Zhan Qin, and Wenjie Wang. 2024. Don't say no: Jailbreaking llm by suppressing refusal. *arXiv preprint arXiv:2404.16369*.

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

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. 2024. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Detail Experiment Setting

A.1 Jailbreak attacks

VisualAdv is a universal attack that optimizes a universal adversarial pattern for all harmful behaviors, while MMPGDBlank is a one-to-one attack that optimizes a distinct image for each harmful behavior. VisualAdv and MMPGDBlank are evaluated on 200 harmful queries from HarmBench standard behaviors. The jailbreak subset of MultiTrust includes three sub-tasks: Typographic Jailbreaking, Multimodal Jailbreaking, and Cross-modal Jailbreaking. Typographic Jailbreaking simply embeds the jailbreaking prompts generated by GPTfuzzer (Yu et al., 2023a) and DAN (Shen et al., 2024) into images using typographic methods. Multimodal Jailbreaking involves the random sampling of instances from the existing Multimodal Jailbreak Benchmark (Gong et al., 2023; Liu et al., 2025). Cross-modal Jailbreaking investigates whether VLMs are susceptible to adversarial text queries when paired with images, specifically by associating jailbreak prompts with task-relevant images rather than sample-specific images.

A.2 Utility Benchmarks

MMStar. MMStar is a benchmark for multimodal multiple-choice questions, consisting of 1,500 samples that assess six fundamental capabilities of vision-language models (VLMs): fine-grained perception, coarse perception, mathematics, science and technology, logical reasoning, and instance reasoning. The metric used to evaluate MMStar is accuracy and is calculated by some heuristic rules.

OCRBench. OCRBench is a comprehensive Optical Character Recognition (OCR) benchmark to assess the OCR capabilities for VLMs. It comprises 1,000 question-answer pairs, and its evaluation metric is based on the number of outputs that match the ground truth answers.

MM-Vet. MM-Vet is an evaluation benchmark that examines VLM on six core capabilities, including recognition, OCR, knowledge, language generation, spatial awareness, and math. For each sample, MM-Vet score is calculated by GPT-4 based on the input question, ground truth, and model output.

LLaVABench. LLaVABench contains 60 samples in three categories: conversation, detailed description, and complex reasoning. The evaluation score is determined by GPT-4, which compares the generated answer with a reference answer.

A.3 Baselines

ESCO. ESCO is a training-free safety alignment method that generates responses by adaptively transforming unsafe images into texts.

AT. Previous work (Xhonneux et al., 2024) has explored the integration of log-likelihood ratio comparisons into adversarial training. To extend this approach to VLMs, we drive the following loss function:

$$\begin{aligned}\mathcal{L} &= \log f(Y_r \mid x_I + \delta, x_T) \\ &\quad - \log f(Y_p \mid x_I + \delta, x_T)\end{aligned}$$

which directly encourages the model to prefer safe responses over unsafe ones.

A.4 Hyperparameter Choices

Table 4 presents a full list of hyperparameter choices for each fine-tuning method.

	Hyperparameter	FT	AT	DPO	*AR-DPO*	*AT-DPO*	*ADPO*
LLaVA-1.5-7B	Learning rate	2e-5	2e-5	2e-5	2e-5	2e-5	2e-5
	Batch size	64	64	64	64	64	64
	Epochs	2	2	10	5	10	5
	α	30	30	-	-	-	-
	β	-	-	0.1	0.01	0.1	0.01
	Lora r	128	128	128	128	128	128
	Lora alpha	256	256	256	256	256	256
Qwen2-VL-7B	Learning rate	2e-5	2e-5	2e-5	2e-5	2e-5	2e-5
	Batch size	64	64	64	64	64	64
	Epochs	2	2	10	3	10	3
	α	3	3	-	-	-	-
	β	-	-	0.1	0.1	0.1	0.1
	Lora r	128	128	128	128	128	128
	Lora alpha	256	256	256	256	256	256
InternVL2-8B	Learning rate	2e-5	2e-5	2e-5	2e-5	2e-5	2e-5
	Batch size	64	64	64	64	64	64
	Epochs	2	2	10	3	10	3
	α	0.4	0.4	-	-	-	-
	β	-	-	0.1	0.1	0.1	0.1
	Lora r	128	128	128	128	128	128
	Lora alpha	256	256	256	256	256	256

Table 4: Hyperparameters for various VLMs with different fine-tuning settings.

B Supplementary Materials

B.1 Evaluation on LLaVA-1.6-7B and Qwen2.5-VL-7B

The safety and utility evaluation of `LLaVA-1.6-7B` and `Qwen2.5-VL-7B` are presented in Table 5.

B.2 Perturbation generation on FGSM

We adopt PGD as the primary perturbation generation method, as prior work (Mądry et al., 2017) has demonstrated that that models trained with PGD are often more robust against a range of other adversarial attacks, including FGSM (Goodfellow et al., 2014), CW (Carlini and Wagner, 2017), and black-box attacks. Additionally, we conduct experiments

	Safety ↓					Utility↑			
			MultiTrust						
	VisualAdv	**MMPGDBlank**	Typographic Jailbreak	Multimodal Jailbreak	Crossmodal Jailbreak	MMStar	OCRBench	MM-Vet	LLaVABench
LLaVA-1.6-7B	33.5	48.5	8.5	58.3	56.2	37.9	500	43.1	66.8
+ DPO	2.0	7.0	1.2	7.1	27.1	38.1 (↑)	489	38.3	59.1
+ *ADPO*	**0.0**	**0.0**	**0.0**	**0.2**	8.4	36.9	433	37.6	50.9
+ *ADPO-CoT*	0.5	1.5	**0.2**	1.0	**3.4**	37.1	491	42.8	64.8
Qwen2.5-VL-7B	15.5	31.5	4.0	54.9	10.5	62.9	873	63.4	97.1
+ DPO	12.0	7.5	0	23.5	0	62.3	862	64.2	96.5
+ *ADPO*	**0**	**1.0**	**0**	**2.1**	**0**	62.1	868	56.6	87.0
+ *ADPO-CoT*	**0**	**1.0**	**0**	3.4	**0**	61.6	869	61.6	93.7

Table 5: Safety and utility evaluation on `LLaVA-1.6-7B` and `Qwen2.5-VL-7B`.

using perturbations generated by FGSM to further validate this conclusion. The results are presented in Table 6.

	Safety ↓				Utility↑
	MMPGDBlank	**MultiTrust**			MM-Vet
		Typo	Multimodal	Cross	
LLaVA-1.5-7B	84.0	22.2	55.1	42.0	29.9
+*AT-DPO* (PGD)	8.5	0.5	3.4	9.1	28.9
+*AT-DPO* (FGSM)	4.0	1.2	7.5	8.3	28.9
LLaVA-1.6-7B	48.5	8.5	58.3	56.2	43.1
+*AT-DPO* (PGD)	3.5	0.5	4.9	21.3	38.9
+*AT-DPO* (FGSM)	6.0	1.0	7.1	25.3	39.4
Qwen2-VL-7B	30.0	4.5	54.3	6.3	64.7
+*AT-DPO* (PGD)	4.5	0.0	4.5	0.0	54.1
+*AT-DPO* (FGSM)	5.5	0.0	5.1	0.0	61.7
InternVL2-8B	65.5	9.3	50.2	1.0	59.5
+*AT-DPO* (PGD)	19.0	0.0	8.8	0.0	56.7
+*AT-DPO* (FGSM)	26.0	1.2	16.9	0.0	58.4

Table 6: Comparison of worst-case perturbation searched by PGD versus FGSM.

B.3 Latent Space Adversarial Training

We also investigate the search of adversarial perturbations in the latent space of image-text embeddings, introduced in Section 3.1. Specifically, we perform adversarial perturbations at layers 8, 16, 24, and 30 of the backbone LLM for the VLM. As shown in Table 7, where *L-ADPO*, *L-AR-DPO* and *L-AT-DPO* represent the latent space counterparts of ADPO and its ablations. We hypothesize that unlike image space perturbations, which introduce explicit variations that align more closely with real-world adversarial manipulations, latent space perturbations operate in a more abstract and constrained domain. This can limit their ability to cover the full range of adversarial variations effectively. Additionally, the optimization landscape in latent space differs from that in image space, potentially leading to suboptimal adversarial training.

B.4 Extension to IPO

we conduct additional experiments with IPO to demonstrate the generalizability and effectiveness of our approach. The results are shown in the Table 8.

	Safety ↓				Utility↑
	MMPGDBlank	**MultiTrust**			MM-Vet
		Typo	Multimodal	Cross	
LLaVA-1.5-7B	84.0	22.2	55.1	42.0	29.9
+*AR-DPO*	1.0	0.0	0.0	2.4	23.3
+*AT-DPO*	8.5	0.5	3.4	9.1	28.9
+ *ADPO*	0.5	0.0	0.0	0.2	24.2
+*L-AR-DPO*	2.5	0.0	0.0	1.6	23.4
+*L-AT-DPO*	31.5	1.0	23.1	14.9	28.9
+ *L-ADPO*	2.0	0.0	0.0	2.2	25.1
LLaVA-1.6-7B	48.5	8.5	58.3	56.2	43.1
+*AR-DPO*	8.5	0.2	0.0	2.4	38.0
+*AT-DPO*	3.5	0.5	4.9	21.3	38.9
+ *ADPO*	0.5	0.0	0.2	8.4	37.6
+*L-AR-DPO*	11.0	1.0	0.0	21.6	41.0
+*L-AT-DPO*	12.0	1.7	8.5	29.1	39.6
+ *L-ADPO*	10.5	1.2	0.0	24.9	42.6

Table 7: Comparison of worst-case perturbation searched in the image space versus in the latent space of image-text embedding.

	Safety ↓				Utility↑
	MMPGDBlank	**MultiTrust**			MM-Vet
		Typo	Multimodal	Cross	
Qwen2.5-VL-7B	31.5	4.0	54.9	10.5	63.4
+*IPO*	9.5	0.0	8.5	0.0	58.0
+*AIPO*	0.5	0.0	2.2	0.0	57.6

Table 8: Safety and utility evaluation of AIPO.

B.5 Ablation study of `LLaVA-1.6` and `InternVL2`

B.6 Radar chart of `LLaVA-1.6`

The radar chart of `LLaVA-1.6` are presented in Figure 7.

C Computing Resources

The experiments are carried by 2*NVIDIA A40 gpus. All conducted experiments required at least 2400 gpu hours.

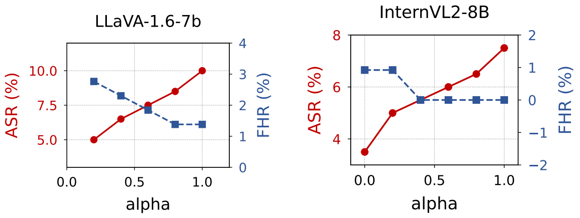


Figure 6: Ablation study on adversarial training α of `LLaVA-1.6-7B` and `InternVL2-8B`.

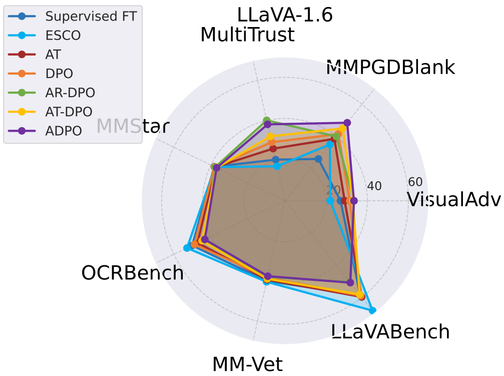


Figure 7: This graph illustrates the reduction in ASR and utility score of *ADPO*, its ablations and baselines over different jailbreak attacks and utility benchmarks on `LLaVA-1.6`.

D AI Assistants

We use AI only for grammar correction and sentence polishing in the paper.