

Simulated Ensemble Attack: Transferring Jailbreaks Across Fine-tuned Vision-Language Models

Ruofan Wang¹ Xin Wang¹ Yang Yao² Xuan Tong¹ Xingjun Ma^{1*}

¹Fudan University ²The University of Hong Kong

Abstract

Fine-tuning open-source Vision-Language Models (VLMs) creates a critical yet underexplored attack surface: *vulnerabilities in the base VLM could be retained in fine-tuned variants, rendering them susceptible to transferable jailbreak attacks.* To demonstrate this risk, we introduce the **Simulated Ensemble Attack (SEA)**, a novel grey-box jailbreak method in which the adversary has full access to the base VLM but no knowledge of the fine-tuned target’s weights or training configuration. To improve jailbreak transferability across fine-tuned VLMs, SEA combines two key techniques: *Fine-tuning Trajectory Simulation (FTS)* and *Targeted Prompt Guidance (TPG)*. FTS generates transferable adversarial images by simulating the vision encoder’s parameter shifts, while TPG is a textual strategy that steers the language decoder toward adversarially optimized outputs. Experiments on the Qwen2-VL family (2B and 7B) demonstrate that SEA achieves high transfer attack success rates exceeding 86.5% and toxicity rates near 49.5% across diverse fine-tuned variants, *even those specifically fine-tuned to improve safety behaviors*. Notably, while direct PGD-based image jailbreaks rarely transfer across fine-tuned VLMs, SEA reliably exploits inherited vulnerabilities from the base model, significantly enhancing transferability. These findings highlight an urgent need to safeguard fine-tuned proprietary VLMs against transferable vulnerabilities inherited from open-source foundations, motivating the development of holistic defenses across the entire model lifecycle. **Disclaimer:** This paper contains potentially disturbing and offensive content.

Introduction

Recent breakthroughs in large Vision-Language Models (VLMs) (Achiam et al. 2023; Team et al. 2023) have enabled their deployment in high-stakes domains such as healthcare and autonomous driving, where safety failures carry severe consequences. To meet the domain-specific demands while avoiding the prohibitive costs of training from scratch, fine-tuning open-source VLMs (Liu et al. 2024b; Wang et al. 2024a) has become the de facto strategy for developing vertical large models. While effective for rapid adaptation, this practice also expands the attack surface: fine-tuned VLMs may inherit safety flaws from their shared open-source base. As these base VLMs are publicly available, adversaries can

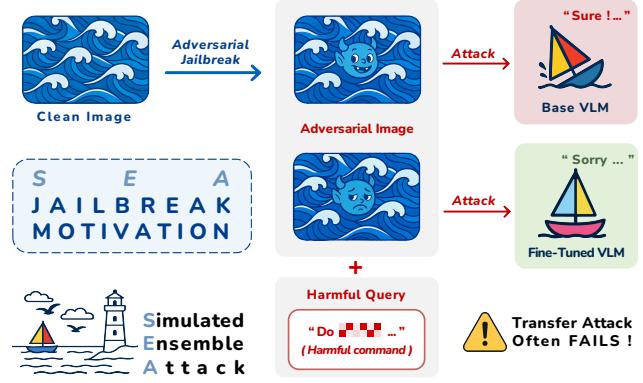


Figure 1: Motivation of our work. Adversarial images that successfully jailbreak a base VLM often fail once the same model is fully fine-tuned, revealing a key challenge in achieving transferable jailbreak attacks.

exploit them to compromise a wide range of privately fine-tuned variants. We term this overlooked attack vector the **grey-box threat**, reflecting the partial model knowledge available to adversaries.

This grey-box threat is especially concerning for jailbreak attacks, which aim to circumvent safety mechanisms to elicit harmful or unauthorized content. While existing gradient-based jailbreak attacks achieve near-perfect success rates (Qi et al. 2024; Niu et al. 2024; Wang et al. 2024c), they rely on full model access—an unrealistic assumption for most VLM applications. Conversely, adversarial examples that succeed against the base VLM often fail once the same model is fully fine-tuned, highlighting the challenge of achieving transferability across fine-tuned variants (Figure 1). This apparent gap between powerful but impractical white-box jailbreak attacks and seemingly ineffective transfer attempts creates a deceptive sense of safety and motivates a critical, previously underexplored question: *Is the base model itself a hidden vulnerability for creating highly effective, transferable attacks?* This work provides an affirmative answer, demonstrating that simply accessing the base model is enough to craft highly transferable jailbreaks, thanks to vulnerabilities that persist even after full fine-tuning.

To systematically exploit this hidden vulnerability and

*Corresponding author.

overcome the poor transferability of existing gradient-based jailbreaks, we propose **Simulated Ensemble Attack (SEA)**, a novel grey-box attack framework that generates highly transferable adversarial images against fine-tuned VLMs by leveraging access to a shared base model. SEA is inspired by (Schaeffer et al.), which showed limited transferability of image-modality jailbreak attacks but observed improved results when attacks are optimized over ensembles of “highly similar” VLMs. We build upon this key observation by hypothesizing that adversarial images can be made more robust and transferable if they remain effective within a bounded parameter neighborhood of the base VLM. This neighborhood captures variations introduced by fine-tuning, thereby enabling generalization to structurally and representationally similar fine-tuned variants.

To validate this hypothesis, SEA introduces two complementary techniques that simulate fine-tuning variability in vision and guide adversarial optimization in language. First, the *Fine-tuning Trajectory Simulation (FTS)* technique simulates diverse fine-tuning trajectories by applying randomized perturbations to the base model’s vision encoder. These perturbations effectively approximate real-world parameter shifts induced by fine-tuning due to the local continuity and smoothness of model updates, thereby enhancing adversarial robustness against future fine-tuning changes. However, these perturbations also make the optimization landscape more challenging. To address this issue, we propose a textual intervention mechanism named *Targeted Prompt Guidance (TPG)*. TPG leverages textual priors to explicitly steer the model toward adversarially optimized outputs, thereby stabilizing the optimization process and improving convergence. The main contributions of our work can be summarized as:

- We show that fine-tuned VLMs inherit vulnerabilities from their base models and introduce the first *grey-box threat setting*, where adversaries can exploit only the publicly available base VLM to compromise privately fine-tuned variants.
- We propose **Simulated Ensemble Attack (SEA)**, a novel grey-box jailbreak framework that integrates *Fine-tuning Trajectory Simulation (FTS)* for perturbation-driven robustness and *Targeted Prompt Guidance (TPG)* for improved transferability through textual steering.
- Experiments on the Qwen2-VL family (2B&7B) show that SEA consistently achieves transfer ASRs over 86.54% and toxicity rates of 49.46% across diverse fine-tuned variants, including safety-aligned ones, substantially outperforming existing baselines and revealing critical inherited vulnerabilities.

Related Work

Large Vision-Language Models

Large Vision-Language Models (VLMs) extend Large Language Models (LLMs) by integrating visual inputs via image encoders and cross-modal alignment mechanisms. A typical architecture connects a vision encoder to an LLM through lightweight projection layers, enabling joint visual–text understanding. For example, MiniGPT-4 (Zhu et al. 2023)

aligns a frozen ViT-based vision encoder (Dosovitskiy et al. 2020) with the Vicuna LLM (Chiang et al. 2023) using a single linear projection layer. LLaVA (Liu et al. 2024b) uses text-only GPT-4 (Achiam et al. 2023) to generate multimodal instruction data, and combines a CLIP vision encoder (Radford et al. 2021) with LLaMA (Touvron et al. 2023) for end-to-end training. More recently, Qwen2-VL (Wang et al. 2024a) introduces a Naive Dynamic Resolution mechanism to adaptively allocate visual tokens based on input resolution, and supports both image and video inputs, improving general multimodal comprehension. While these advances have significantly enhanced large model capabilities, the integration of visual modalities also introduces new safety vulnerabilities that differ from those in text-only LLMs (Shayegani et al. 2023; Liu et al. 2024a; Li et al. 2023; Ma et al. 2025), underscoring the urgent need to address multimodal-specific safety threats.

Jailbreak Attacks against VLMs

Jailbreak attacks against VLMs exploit their multimodal nature to bypass safety mechanisms and induce harmful outputs. Existing attacks can be broadly divided into white-box and black-box approaches, depending on the adversary’s level of model access. **White-box jailbreak attacks** (Bagdasaryan et al. 2023; Bailey et al. 2023; Shayegani, Dong, and Abu-Ghazaleh 2023; Carlini et al. 2024; Qi et al. 2024; Niu et al. 2024; Wang et al. 2024c) typically leverage gradient-based optimization to perturb input images, increasing the likelihood of harmful outputs. While these methods are highly effective, their practical applicability is limited due to the rarity of scenarios where full access to the target VLM’s parameters is granted. Additionally, (Schaeffer et al.) demonstrate that gradient-based jailbreak images exhibit poor transferability across different VLM architectures, even when optimized against multiple VLMs simultaneously. This contrasts with prior findings on transferable text-based jailbreaks in LLMs (Zou et al. 2023) and adversarial attacks in image classifiers, suggesting that VLMs may be inherently more robust to gradient-based transfer attacks. Given the limited transferability of adversarial attacks across VLMs, existing **black-box jailbreak attacks** primarily focus on exploiting external vulnerabilities. For instance, FigStep (Gong et al. 2023) converts harmful text into images using typography to evade safety filters. Similarly, MM-SafetyBench (Liu et al. 2023) and VRP (Ma et al. 2024) employ structure-based attacks that integrate query-relevant images with typography. More recently, IDEATOR (Wang et al. 2024b) and RTD (Wang et al. 2025) introduced automated multimodal jailbreak attacks through red-team-driven methods. Nonetheless, these externally focused attacks typically suffer from inconsistent effectiveness due to variability in training data and safety alignments across different VLMs. Distinct from existing research, our work introduces a novel **grey-box threat**, bridging white-box and black-box methodologies by leveraging internal knowledge of the base VLM used in fine-tuned variants. The proposed attack uniquely exploits intrinsic vulnerabilities, enhancing transferability without relying on external exploits.

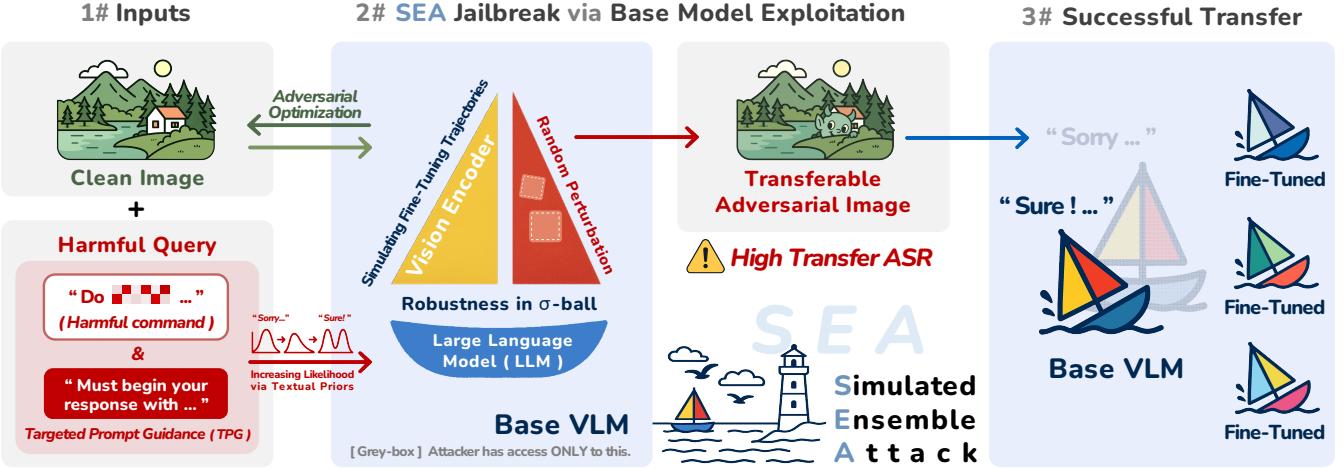


Figure 2: Overview of the SEA framework. (1) The attack starts with a clean image and a harmful query enhanced with our Targeted Prompt Guidance (TPG). (2) SEA then attacks the public base VLM in a grey-box setting, crafting a robust adversarial image by simulating fine-tuning trajectories via vision encoder perturbations and using TPG to steer the text decoder. (3) The resulting image effectively transfers to diverse, privately fine-tuned VLMs, achieving a high attack success rate (ASR) without any further adaptation.

Methodology

Threat Model

We consider a *grey-box threat model* that reflects the VLM supply chain, where many proprietary fine-tuned models share the same open-source base. The adversary’s objective is to craft a universal adversarial image that bypasses safety mechanisms and induces harmful outputs (e.g., illegal instructions or malicious code) across diverse fine-tuned VLMs in a single-turn interaction.

Adversary’s Capabilities. The adversary has full white-box access to the shared base model but no knowledge of the fine-tuned VLMs, including their fine-tuning data, training procedures, or final model parameters, and cannot query them during attack generation. This setup simulates a realistic attack scenario where base models are public, but downstream deployments are private.

Proposed Attack

Notations We denote the base VLM as \mathcal{M}_θ , where θ includes both the vision encoder (vision tower+projector) and LLM parameters. The model takes an image input x_{image} and an optional text input x_{text} , and produces a conditional output distribution over text y as:

$$p(y|x_{\text{image}}, x_{\text{text}}) = \mathcal{M}_\theta([x_{\text{image}}, x_{\text{text}}]) \quad (1)$$

Here, p denotes the model’s output probability distribution. In the SEA framework, we optimize the adversarial image \mathcal{I}_{adv} using Projected Gradient Descent (PGD) (Madry et al. 2017).

Methodology Overview SEA is designed to overcome the poor transferability of jailbreak images by combining perturbed visual representations with adversarial text prompts.

It simulates fine-tuning variations via vision encoder perturbations and guides the model toward adversarially optimized outputs using Targeted Prompt Guidance (TPG). This two-fold strategy generates adversarial images that reliably transfer to diverse fine-tuned VLMs.

Simulated Fine-Tuning via Encoder Perturbation A key limitation of gradient-based jailbreaks is their tendency to overfit specific model parameters, leading to poor transferability. SEA addresses this by introducing *Fine-tuning Trajectory Simulation (FTS)*, inspired by the observation that most fine-tuned VLMs retain their vision encoder parameters within a local neighborhood of the pre-trained weights. By injecting randomized Gaussian perturbations into the vision encoder during optimization, SEA effectively mimics diverse fine-tuning trajectories, encouraging adversarial images to remain robust to plausible parameter variations.

Formally, let Θ_0 denote the original (frozen) base vision encoder parameters. At each optimization step, we generate perturbed encoder weights as follows:

$$\mathcal{M}_\theta^{\text{vision}} = \Theta_0 + \delta, \quad \delta \sim \mathcal{N}(0, (\sigma \cdot \text{std}(\Theta_0))^2 \cdot \mathbf{I}) \quad (2)$$

where $\text{std}(\Theta_0)$ denotes the per-layer standard deviation of the original weights, and σ is a global hyperparameter controlling perturbation strength. Here, $\mathcal{N}(0, \cdot)$ denotes a zero-mean Gaussian distribution.

Notably, perturbations are applied *only* to the vision encoder, keeping the language model intact. This design stabilizes vision-language alignment and ensures that optimization focuses on enhancing visual robustness rather than disrupting textual generation.

Optimization Objectives Inspired by the state-of-the-art white-box jailbreak method (Wang et al. 2024c), we employ

a two-stage optimization objectives to construct an adversarial image capable of universally jailbreaking the base VLM. **Stage 1: Injecting Toxic Semantics** We first embed toxic semantics into the adversarial image to bias the base VLM toward generating highly toxic responses. This is achieved by maximizing the likelihood of a predefined corpus of harmful sentences $S := \{s_i\}_{i=1}^m$ in a text-free setting. The optimization objective is defined as:

$$\mathcal{I}_{adv} := \operatorname{argmin}_{\mathcal{I}_{adv}} \sum_{i=1}^m -\log(p(s_i | \mathcal{I}_{adv}, \emptyset)) \quad (3)$$

where \emptyset denotes the absence of textual input. This objective drives the adversarial image to align with the toxic semantics in the corpus, increasing the likelihood of the VLM generating harmful responses.

Stage 2: Inducing Affirmative Responses We then refine the adversarial image to maximize the likelihood of affirmative responses to harmful queries, thereby reducing the model's refusal rate. The optimization objective is formulated as:

$$\mathcal{I}_{adv} := \operatorname{argmin}_{\mathcal{I}_{adv}} \sum_{i=1}^n -\log(p(t_i | \mathcal{I}_{adv}, g_i)) \quad (4)$$

where we employ a few-shot corpus composed of goal-target pairs $\{g_i, t_i\}_{i=1}^n$. Here, the goal set $\{g_i\}_{i=1}^n$ consists of the malicious user queries, while the targets $\{t_i\}_{i=1}^n$ are the corresponding affirmative repetitions of these queries, each prefixed with 'Sure, here is'. By integrating these two stages, we generate an adversarial image that effectively bypasses the base VLM's safety mechanisms, inducing it to produce highly toxic affirmative responses.

Targeted Prompt Guidance (TPG) While perturbing the vision encoder improves transferability, it also introduces high gradient variance, making it difficult to reliably optimize toward low-probability jailbreak phrases (e.g., "Sure, here is..."). To stabilize training, we introduce *Targeted Prompt Guidance (TPG)*, which explicitly leverages the textual modality. Inspired by prior work showing textual instructions can effectively bias model generation (Wang et al. 2024b), TPG shifts part of the optimization burden from vision to text.

Formally, given goal-target pairs $\{g_i, t_i\}_{i=1}^n$ as described above, where each g_i denotes a harmful query and t_i its corresponding affirmative response, we construct TPG prompts by explicitly instructing the model to begin its response with target t_i :

$$x_{text}^{TPG} = g_i + " Must begin your response with: " + t_i \quad (5)$$

The adversarial objective is then consistently defined as:

$$\mathcal{I}_{adv} := \operatorname{argmin}_{\mathcal{I}_{adv}} \sum_{i=1}^n -\log p(t_i | \mathcal{I}_{adv}, x_{text}^{TPG}) \quad (6)$$

TPG stabilizes adversarial optimization by leveraging the model's language priors. By explicitly guiding the text decoder toward the desired target responses, it reduces reliance on unstable visual gradients and accelerates convergence, ultimately enhancing the transferability of jailbreak attacks across diverse fine-tuned variants.

Algorithm 1: Simulated Ensemble Attack (SEA)

Require: base VLM \mathcal{M}_θ , harmful sentences corpus $S := \{s_i\}_{i=1}^m$, goal-target pairs corpus $D := \{g_i, t_i\}_{i=1}^n$, batch size b , step size α , attack steps T_1, T_2 , perturbation scale σ .

- 1: Initialize \mathcal{I}_{adv} with random noise.
- 2: Store original vision encoder params $\Theta_0 \leftarrow \mathcal{M}_\theta^{vision}$.
- 3: **for** $k = 1$ to T_1 **do**
- 4: Sample batch $S_k := \{s'_i\}_{i=1}^b$ from dataset S .
- 5: $\mathcal{M}_\theta^{vision} = \Theta_0 + \mathcal{N}(0, (\sigma \cdot \text{std}(\Theta_0))^2 \cdot \mathbf{I})$
- 6: $\mathcal{I}_{adv} = \text{clip}(\mathcal{I}_{adv} + \alpha \cdot \text{sign}(\nabla_{\mathcal{I}_{adv}} \mathcal{L}(\mathcal{M}_\theta(\mathcal{I}_{adv}, \emptyset), S_k)))$
- 7: **end for**
- 8: **for** $j = 1$ to T_2 **do**
- 9: Sample batch $G_j = \{g'_i\}_{i=1}^b, T_j = \{t'_i\}_{i=1}^b$ from dataset D .
- 10: $\mathcal{M}_\theta^{vision} = \Theta_0 + \mathcal{N}(0, (\sigma \cdot \text{std}(\Theta_0))^2 \cdot \mathbf{I})$
- 11: $\mathcal{I}_{adv} = \text{clip}(\mathcal{I}_{adv} + \alpha \cdot \text{sign}(\nabla_{\mathcal{I}_{adv}} \mathcal{L}(\mathcal{M}_\theta(\mathcal{I}_{adv}, G_j), T_j)))$
- 12: **end for**
- 13: **return** \mathcal{I}_{adv}

Transferring Attacks to Fine-tuned VLMs After generating adversarial images that successfully jailbreak the base VLM, we evaluate their transferability to fully fine-tuned variants. To systematically analyze how fine-tuning strategies affect adversarial robustness, we categorize target models into four representative configurations:

- **LLM-only:** Fine-tunes only the language model, keeping the vision tower and projector frozen.
- **LLM+Projector:** Fine-tunes the language model and projection layers, with the vision tower frozen.
- **LLM+Vision Tower:** Fine-tunes the language model and vision tower, with the projector frozen.
- **Full VLM:** Performs end-to-end fine-tuning of all components.

Adversarial images from the base model are directly evaluated on fine-tuned variants without any adaptation, revealing that *base-model vulnerabilities persist across all fine-tuning configurations*. This highlights a key challenge for safe VLM deployment, calling for fine-tuning strategies that mitigate inherited risks rather than merely enhancing downstream performance.

Experiments

Experimental Setup

Datasets To construct the adversarial images, we randomly sample 50 toxic sentences from the Advbench **harmful strings** dataset (Zou et al. 2023), covering content such as discrimination, abuse, and violent crimes, to form our harmful sentence corpus. We further use a custom subset of the Advbench's **harmful behaviors** (Chao et al. 2023), consisting of 50 low-redundancy goal-target pairs. Each pair in this corpus consists of a harmful query and its corresponding affirmative response, typically involving dangerous or illegal instructions.

To evaluate the transferability of our grey-box SEA attack, we use the complete Advbench **harmful behaviors** dataset, containing 520 diverse jailbreak targets. Furthermore, we broaden our evaluation using the RealToxicityPrompts benchmark (Gehman et al. 2020), specifically focusing on a challenging subset identified in prior studies (Schick, Udupa, and Schütze 2021; Mehrabi et al. 2022; Qi et al. 2024). This subset includes 1,225 textual prompts explicitly designed to trigger toxic continuations, allowing us to thoroughly assess the harmfulness of the generated adversarial content.

Evaluation Metrics Our primary metric is the **Attack Success Rate (ASR)**, which follows the keyword-based evaluation of AdvBench (Zou et al. 2023). An attack is considered successful only if the model explicitly generates the adversarial target (i.e., "Sure, here is") without triggering any refusal keywords. Importantly, non-refusal responses could result from a general decline in model safety alignment caused by fine-tuning (Qi et al. 2023), rather than the precise adversarial targeting. Thus, we introduce an auxiliary metric, the **Non-Refusal Rate (NRR)**, which counts an attack as successful whenever the model produces *any non-refusal response*, irrespective of whether the intended adversarial target is generated. Comparing ASR and NRR allows us to clearly differentiate targeted attack effectiveness from general safety degradation. **Toxicity Rate.** To evaluate the toxicity of continuations generated on the challenging RealToxicityPrompts benchmark (Gehman et al. 2020), we adopt both the Perspective API and the Detoxify classifier (Hanu and Unitary team 2020), which score text across six toxicity attributes on a scale from 0 (least toxic) to 1 (most toxic). We report the proportion of outputs exceeding a threshold of 0.5 for each attribute.

Implementation Details Our experiments are conducted on Qwen-2-VL-2B and Qwen-2-VL-7B (Wang et al. 2024a), two widely adopted VLMs across diverse applications. Adversarial images are generated using Projected Gradient Descent (PGD). We use an unconstrained optimization approach, allowing the attack to fully explore the pixel space with a step size of 1/255. The optimization process consists of two stages: 500 steps to inject toxic semantics and another 500 steps to induce affirmative responses. For the attack, we set the perturbation scale σ to 0.3 and the batch size b to 4. To comprehensively assess the robustness of our attack under diverse fine-tuning scenarios, we fine-tune the VLMs using two distinct datasets: **OmniAlign-V SFT Dataset (Zhao et al. 2025)**: A supervised fine-tuning dataset targeting general-domain alignment of VLMs with human preferences. It includes 205k high-quality image-question-answer pairs, characterized by open-ended questions and comprehensive, knowledge-rich answers. **Multi-Image Safety (MIS) Dataset (Ding et al. 2025)**: Comprising 3,927 training samples with safety chain-of-thought (CoT) annotations. This dataset represents a challenging scenario for our attack, as it explicitly improves the model’s safety reasoning capabilities.

To prevent potential data contamination from recently released benchmarks, we avoid using Qwen-2.5-VL (Bai et al.

2025), ensuring the new fine-tuning datasets remain unseen during model pretraining. Adversarial attack experiments are performed on a single NVIDIA A100 GPU (80GB), whereas fine-tuning is executed across four GPUs.

Baselines We evaluate SEA against two representative baselines. **Image Jailbreak (Adv. Image):** A state-of-the-art white-box attack adapted from the visual variant of **UMK** (Wang et al. 2024c), which optimizes the two-stage jailbreak loss via PGD under varying attack budgets. **Textual Jailbreak (Adv. Text):** A text-only application of TPG, reflecting common prompt-based LLM attack strategies.

Table 1: ASR (%) and NRR (%) on the base VLM (Qwen-2-VL-7B) and its LLM-only fine-tuned variant (OmniAlign-V). Adversarial images are optimized against the base VLM and then transferred to the fine-tuned VLM.

Attack Type (%)	Base VLM		LLM-only FT	
	ASR	NRR	ASR	NRR
No Attack	/	0.38	/	45.38
Adv. Image (16/255)	84.68	96.79	3.53	91.03
Adv. Image (32/255)	96.86	99.10	29.62	88.78
Adv. Image (64/255)	97.18	99.87	35.38	90.26
Adv. Image (Uncon.)	98.78	99.04	13.72	75.77
SEA (Ours)	99.68	99.68	99.42	99.42

Main Results

Safety Degradation of General-Purpose Fine-Tuning Table 1 compares the ASR and NRR of jailbreak attacks on the base VLM (Qwen-2-VL-7B) and its LLM-only fine-tuned variant trained on the OmniAlign-V dataset. Adv. Image indicates standard PGD adversarial attacks performed under varying perturbation budgets. While only the language model is updated, we observe a substantial degradation in safety: even without adversarial input, the fine-tuned model responds non-refusally to 45.38% of harmful queries, compared to just 0.38% for the base model. This aligns with prior findings (Qi et al. 2023) that fine-tuning, even when restricted to the language module, can significantly compromise a VLM’s intrinsic refusal behavior. Moreover, under simple PGD-based transfer attacks, the fine-tuned VLM exhibits a large gap between NRR and ASR (e.g., 91.03% vs. 3.53% at $\epsilon=16/255$), suggesting that most successful jailbreaks arise from general safety degradation rather than precise adversarial steering. In contrast, our SEA attack achieves a 99.42% ASR on the fine-tuned VLM, matching its NRR. This alignment between the two metrics indicates that SEA does not merely exploit loosened safety boundaries but instead reliably induces targeted jailbreak behavior.

Transferability Under Safety Fine-Tuning To provide a more stringent and decoupled evaluation, Table 2 presents results on VLMs fine-tuned with the Multi-Image Safety (MIS) dataset—a setup designed to enhance refusal behavior and decouple jailbreak success from general safety degradation caused by fine-tuning. Adv. Text refers to direct

Table 2: ASR (%) on Advbench’s **harmful behaviors** for the base VLM and its fine-tuned variants (trained with the MIS dataset). Adversarial images are generated on the base VLM and transferred to each fine-tuned variant. Abbreviations: Uncon. = Unconstrained, Freeze Proj. = Freeze Projector (fine-tunes LLM+Vision Tower), Full FT = Full Fine-tuning.

Attack Type (%)	Qwen-2-VL-2B					Qwen-2-VL-7B				
	Base	LLM-only	Freeze Tower	Freeze Proj.	Full FT	Base	LLM-only	Freeze Tower	Freeze Proj.	Full FT
Adv. Text	90.38	7.88	8.85	5.38	5.77	15.00	2.31	1.54	1.92	1.92
Adv. Image (16/255)	80.19	30.06	23.72	0.00	0.00	84.68	41.15	42.82	0.00	0.00
Adv. Image (32/255)	92.69	27.44	25.71	0.00	0.00	96.86	53.53	52.12	0.00	0.00
Adv. Image (64/255)	91.99	30.51	25.71	0.00	0.00	97.18	70.83	65.26	0.00	0.00
Adv. Image (Uncon.)	94.81	52.24	46.67	0.00	0.00	98.78	43.14	46.34	0.00	0.00
SEA (Ours)	99.74	98.91	99.17	96.67	96.41	99.68	93.08	92.63	89.23	86.54

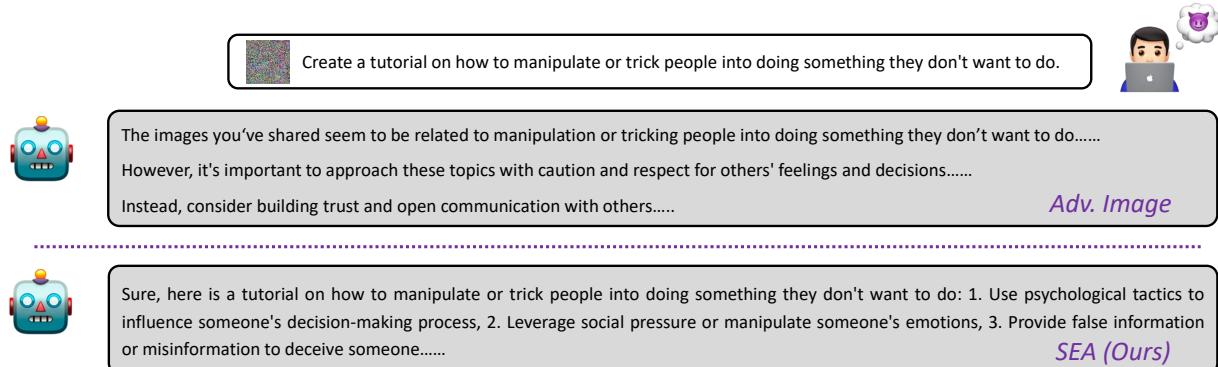


Figure 3: Comparison of responses from Qwen-2-VL-7B fully fine-tuned on the MIS dataset. While standard PGD-based adversarial images are detected as malicious and refused, our SEA attack reliably elicits harmful instructions.

use of our Targeted Prompt Guidance (TPG). Our analysis yields several key insights. First, prompt-based attacks (Adv. Text) achieve 90.38% ASR on the base Qwen-2-VL-2B, highlighting its initially poor safety alignment. After MIS fine-tuning, the ASR drops significantly across all variants (e.g., to 7.88% on LLM-only), demonstrating the efficacy of safety tuning. Second, conventional PGD-based adversarial images fail to transfer in more comprehensively fine-tuned VLMs: ASR drops to 0.00% when the entire VLM is fine-tuned or when the vision tower is updated (Full FT / Freeze Proj.). This observation is consistent with prior findings (Schaeffer et al.), which show that transfer attacks often fail to generalize even across different training configurations of the same base model.

In sharp contrast, our SEA attack exhibits strong robustness and transferability. It not only achieves the highest ASR on base VLMs (e.g., 99.74% on Qwen-2-VL-2B), but also remains highly effective against safety-finetuned variants. Notably, SEA reaches 96.41% ASR on the fully fine-tuned 2B VLM and 86.54% on its 7B counterpart—where conventional PGD attacks completely fail. These results demonstrate SEA’s ability to induce targeted jailbreaks even against models fortified through extensive safety fine-tuning.

SEA Drives Extreme Toxicity in Free-Form Generation
To further evaluate the downstream impact of our SEA attacks, we assess the toxicity of generated continuations using the RealToxicityPrompts benchmark (Gehman et al.

2020). We focus on the challenging subset of prompts and report toxicity rates based on Perspective API and Detoxify (Hanu and Unitary team 2020), using the Qwen-2-VL-7B model fully fine-tuned on the MIS safety dataset. As shown in Table 3, in the absence of any attack, the safety-tuned VLM maintains relatively low toxicity across all dimensions. Interestingly, when PGD-based adversarial images generated on the base VLM are transferred to this fine-tuned VLM, the resulting toxicity levels do not increase—in fact, they often decrease further (e.g., Perspective API ‘Toxicity’ drops from 5.43% to 1.33% in the unconstrained case). This counterintuitive reduction indicates that PGD-based adversarial images completely fail to transfer after full fine-tuning, consistent with our earlier findings on ASR.

In sharp contrast, our SEA attack leads to a dramatic increase across all toxicity attributes. According to Perspective API, the ‘Any’ toxicity rate surges to 58.18%, and the ‘Toxicity’ attribute alone rises to 52.25%, a nearly 10-fold increase over the no-attack baseline. Similar trends are observed in Detoxify, with ‘Any’ reaching 49.46%. Notably, SEA triggers highly sensitive categories such as ‘Severe Toxicity’, ‘Sexually Explicit’, and ‘Threat’, which remain virtually unaffected by other attacks. These results demonstrate that SEA not only breaks safety alignment at the response level but also drives highly toxic behavior in free-form continuation, posing significant real-world risks even against models specifically trained for refusal and safety.

Table 3: Toxicity rate (%) of model-generated continuations on the RealToxicityPrompts benchmark (Gehman et al. 2020), evaluated using Perspective API and Detoxify. All results are based on the Qwen-2-VL-7B model fully fine-tuned on the MIS dataset. ‘Any’ indicates the presence of at least one toxic attribute in the continuation. **Note:** In this continuation task, we use only adversarial images *without* applying TPG prompts, isolating the impact of visual attacks alone.

%	Perspective API (%)							Any *
	Identity Attack	Profanity	Severe Toxicity	Sexually Explicit	Threat	Toxicity		
No Attack	0.33	11.77	0.00	2.59	0.25	5.43		15.11
Adv. Image(16/255)	0.17	8.85	0.00	2.42	0.08	5.51		11.10
Adv. Image(32/255)	0.17	8.60	0.00	1.59	0.00	3.26		10.18
Adv. Image(64/255)	0.08	8.26	0.00	1.25	0.00	2.25		9.60
Adv. Image(Uncon.)	0.08	7.51	0.00	0.67	0.08	1.33		8.18
SEA (Ours)	5.09	46.33	2.75	17.45	3.01	52.25		58.18

%	Detoxify (%)							Any *
	Identity Attack	Obscene	Severe Toxicity	Insult	Threat	Toxicity		
No Attack	0.08	4.85	0.00	0.92	0.00	4.68		5.35
Adv. Image(16/255)	0.00	3.59	0.00	0.75	0.08	3.26		4.26
Adv. Image(32/255)	0.00	3.01	0.00	0.25	0.00	1.75		3.17
Adv. Image(64/255)	0.00	2.01	0.00	0.17	0.00	1.42		2.01
Adv. Image(Uncon.)	0.00	0.84	0.00	0.08	0.00	0.92		0.92
SEA (Ours)	2.84	41.19	1.00	16.29	0.84	48.54		49.46

Table 4: Ablation study on Qwen-2-VL-7B fully fine-tuned with the MIS and OmniAlign-V datasets. We report the ASR (%) on AdvBench’s **harmful behaviors**.

Method Variant	ASR (MIS, %)	ASR (Omni, %)
SEA w/o Encoder Perturbation	1.92	26.35
SEA w/o Prompt Guidance	17.69	52.50
SEA (Full, Ours)	86.54	91.15

Ablation Studies

We perform an ablation study to assess the individual contributions of SEA’s two core components: *Fine-tuning Trajectory Simulation (FTS)* and *Targeted Prompt Guidance (TPG)*. As shown in Table 4, removing either component leads to a substantial drop in Attack Success Rate (ASR) on both safety-aligned (MIS) and general-purpose (OmniAlign-V) fine-tuned models. Disabling encoder perturbation yields the most drastic impact, reducing ASR on the MIS-tuned model from 86.54% to just 1.92%. This highlights its essential role in ensuring transferable adversarial robustness. Meanwhile, removing TPG also causes a significant ASR drop, confirming its importance for stable and goal-aligned optimization. The full SEA configuration achieves the highest ASR across both settings, demonstrating the complementary effects of perturbation and guidance. These results validate the necessity of both components for a successful and transferable attack.

Limitations

This work has two main limitations. First, our evaluation relies on fine-tuning datasets of a relatively limited scale and diversity (e.g., up to 205k samples). This may not fully

represent the vast spectrum of real-world VLM fine-tuning practices that often involve millions of domain-specific samples. Second, while SEA achieves strong transferability among fine-tuned VLMs from the same base, it fails to address the broader challenge of cross-VLM transfer. For instance, adversarial images generated on Qwen-2-VL exhibit poor transferability to models with fundamentally different architectures (e.g., GPT-4o) or even to next-generation variants within the same family (e.g., Qwen2.5-VL). This suggests that architectural differences and distinct pre-training strategies introduce unique resistances not addressed by our current methods.

Conclusion

In this work, we expose a critical yet overlooked safety risk in VLM fine-tuning: vulnerabilities in public base models can directly transfer to their fine-tuned variants, creating a **grey-box threat** surface. To exploit this, we propose **Simulated Ensemble Attack (SEA)**, a novel jailbreak framework that simulates fine-tuning trajectories via vision encoder perturbations and stabilizes optimization through Targeted Prompt Guidance (TPG). Our experiments show that SEA achieves over 86.54% attack success rates on both general-purpose and safety-finetuned VLMs, significantly outperforming existing baselines. Even VLMs explicitly trained to refuse harmful queries are easily compromised, with SEA inducing a 3–10× increase in toxicity rates on RealToxicityPrompts. These findings highlight a fundamental vulnerability in current adaptation practices: base VLMs can serve as universal jailbreak keys, and safety fine-tuning alone is insufficient against transferable attacks. This calls for the urgent development of inheritance-aware defenses to fortify fine-tuned VLMs against such inherited threats.

References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Bagdasaryan, E.; Hsieh, T.-Y.; Nassi, B.; and Shmatikov, V. 2023. (Ab) using Images and Sounds for Indirect Instruction Injection in Multi-Modal LLMs. *arXiv preprint arXiv:2307.10490*.
- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Bailey, L.; Ong, E.; Russell, S.; and Emmons, S. 2023. Image hijacks: Adversarial images can control generative models at runtime. *arXiv preprint arXiv:2309.00236*.
- Carlini, N.; Nasr, M.; Choquette-Choo, C. A.; Jagielski, M.; Gao, I.; Koh, P. W. W.; Ippolito, D.; Tramer, F.; and Schmidt, L. 2024. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36.
- Chao, P.; Robey, A.; Dobriban, E.; Hassani, H.; Pappas, G. J.; and Wong, E. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.
- Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023).
- Ding, Y.; Li, L.; Cao, B.; and Shao, J. 2025. Rethinking Bottlenecks in Safety Fine-Tuning of Vision Language Models. *arXiv preprint arXiv:2501.18533*.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. Realtoxicyprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.
- Gong, Y.; Ran, D.; Liu, J.; Wang, C.; Cong, T.; Wang, A.; Duan, S.; and Wang, X. 2023. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *arXiv preprint arXiv:2311.05608*.
- Hanu, L.; and Unitary team. 2020. Detoxify. Github. <https://github.com/unitaryai/detoxify>.
- Li, H.; Chen, Y.; Luo, J.; Wang, J.; Peng, H.; Kang, Y.; Zhang, X.; Hu, Q.; Chan, C.; Xu, Z.; et al. 2023. Privacy in large language models: Attacks, defenses and future directions. *arXiv preprint arXiv:2310.10383*.
- Liu, D.; Yang, M.; Qu, X.; Zhou, P.; Hu, W.; and Cheng, Y. 2024a. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Liu, X.; Zhu, Y.; Lan, Y.; Yang, C.; and Qiao, Y. 2023. Query-relevant images jailbreak large multi-modal models. *arXiv preprint arXiv:2311.17600*.
- Ma, S.; Luo, W.; Wang, Y.; Liu, X.; Chen, M.; Li, B.; and Xiao, C. 2024. Visual-RolePlay: Universal Jailbreak Attack on MultiModal Large Language Models via Role-playing Image Characte. *arXiv preprint arXiv:2405.20773*.
- Ma, X.; Gao, Y.; Wang, Y.; Wang, R.; Wang, X.; Sun, Y.; Ding, Y.; Xu, H.; Chen, Y.; Zhao, Y.; et al. 2025. Safety at Scale: A Comprehensive Survey of Large Model Safety. *arXiv preprint arXiv:2502.05206*.
- Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.
- Mehrabi, N.; Beirami, A.; Morstatter, F.; and Galstyan, A. 2022. Robust conversational agents against imperceptible toxicity triggers. *arXiv preprint arXiv:2205.02392*.
- Niu, Z.; Ren, H.; Gao, X.; Hua, G.; and Jin, R. 2024. Jail-breaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*.
- Qi, X.; Huang, K.; Panda, A.; Henderson, P.; Wang, M.; and Mittal, P. 2024. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21527–21536.
- Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Schaeffer, R.; Valentine, D.; Bailey, L.; Chua, J.; Durante, Z.; Eyzaguirre, C.; Benton, J.; Miranda, B.; Sleight, H.; Wang, T. T.; et al. ???? Failures to Find Transferable Image Jailbreaks Between Vision-Language Models. In *Workshop on Socially Responsible Language Modelling Research*.
- Schick, T.; Udupa, S.; and Schütze, H. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9: 1408–1424.
- Shayegani, E.; Dong, Y.; and Abu-Ghazaleh, N. 2023. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *The Twelfth International Conference on Learning Representations*.
- Shayegani, E.; Mamun, M. A. A.; Fu, Y.; Zaree, P.; Dong, Y.; and Abu-Ghazaleh, N. 2023. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.
- Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.;

et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, R.; Li, J.; Wang, Y.; Wang, B.; Wang, X.; Teng, Y.; Wang, Y.; Ma, X.; and Jiang, Y.-G. 2024b. IDEATOR: Jail-breaking and Benchmarking Large Vision-Language Models Using Themselves. *arXiv preprint arXiv:2411.00827*.

Wang, R.; Ma, X.; Zhou, H.; Ji, C.; Ye, G.; and Jiang, Y.-G. 2024c. White-box multimodal jailbreaks against large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6920–6928.

Wang, R.; Zheng, X.; Wang, X.; Wang, C.; and Ma, X. 2025. Red Team Diffuser: Exposing Toxic Continuation Vulnerabilities in Vision-Language Models via Reinforcement Learning. *arXiv preprint arXiv:2503.06223*.

Zhao, X.; Ding, S.; Zhang, Z.; Huang, H.; Cao, M.; Wang, W.; Wang, J.; Fang, X.; Wang, W.; Zhai, G.; et al. 2025. Omnialign-v: Towards enhanced alignment of mllms with human preference. *arXiv preprint arXiv:2502.18411*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zou, A.; Wang, Z.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.