Cross-Model Transfer Attacks against Large Vision-Language Models via Model Diversity Enrichment and Stochastic Parameter Sampling

Zhenze Yang[1], **Xiaowen Cai**[2]*, **Junhao Dong**[3], **Guangke Chen**[1], **Guiyao Tie**[2], **Daizong Liu**[1]*
Hongyang He[4], **Zhiyuan Ma**[2], **Xiang Fang**[3], **Dengpan Ye**[1], **Jing Zhang**[1]
[1]Wuhan University [2]Huazhong University of Science and Technology
[3]Nanyang Technological University [4]University of Warwick
zhenzeyang0307@gmail.com,{xwcai,tgy,mzyth}@hust.edu.cn
{guangkechen,daizongliu,yedp,jingzhang.cv}@whu.edu.cn
junhao003@ntu.edu.sg,Hongyang.He@warwick.ac.uk,xfang9508@gmail.com

Abstract

Large Vision-Language Models (LVLMs) have achieved remarkable multimodal reasoning capabilities but remain critically vulnerable to adversarial perturbations. In practical black-box scenarios, severe architectural discrepancies, misaligned feature spaces, and divergent multimodal fusion strategies often undermine adversarial transferability, preventing perturbations crafted on surrogate models from generalizing to unseen target models. Existing ensemble-based LVLM attacks attempt to mitigate this by integrating multiple surrogates, yet they often fail to bridge the structural generalization gap across heterogeneous LVLM families. In this work, we propose a novel framework that enhances cross-model adversarial transferability via model diversity enrichment and stochastic parameter sampling. We show that transfer success is fundamentally governed by cross-model gradient variance. To mitigate this, we introduce (1) a diversity-induced continuous surrogate manifold that captures cross-model decision variability, and (2) stochastic parameter sampling during optimization to approximate the gradient distribution of unseen LVLM models. We provide a theoretical analysis bounding the expected adversarial transfer gain by gradient variance and prove that stochastic sampling reduces adversarial overfitting to surrogate decision boundaries. Extensive experiments across multiple state-of-the-art LVLMs demonstrate substantial improvements in black-box transferability under strict evaluation protocols.

1 Introduction

Large Vision-Language Models (LVLMs) have rapidly emerged as the dominant paradigm for multimodal reasoning [1, 2, 3, 4, 5, 6, 7] by combining pretrained visual encoders with powerful large language models [8, 9, 10]. Despite their strong performance, recent studies [11, 12, 13, 14, 15] reveal that LVLMs remain vulnerable to adversarial perturbations. In realistic threat scenarios, LVLM attackers often operate under a strict black-box setting [15, 16, 14], where the internal parameters of the target LVLM model are inaccessible. Consequently, adversarial examples against LVLMs must be crafted on surrogate models and rely on transferability.

Existing methodologies for crafting transferable adversarial examples against LVLMs primarily follow an ensemble-based paradigm [17, 18, 19, 20, 21], which attempts to mitigate surrogate-

*Corresponding Authors: Daizong Liu and Xiaowen Cai. This work was done when Zhenze Yang was an intern under the supervision of Daizong Liu (Project Leader, daizongliu@whu.edu.cn).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

specific overfitting by jointly optimizing across a discrete set of model instances. For instance, [21] employs a dynamic selection mechanism among multiple CLIP-based surrogates to improve generalization, while [20] introduces adaptive weighting to balance gradient contributions from different architectures. However, we argue that these ensemble-centric approaches are suboptimal for capturing the complex transfer dynamics of LVLM models. First, the computational scalability remains a prohibitive bottleneck, as maintaining an ensemble of multiple multi-billion-parameter LVLMs incurs extreme memory and FLOPs overhead. Second, and more critically, these methods treat transferability as an empirical by-product of discrete model averaging, failing to explicitly enrich the model diversity required to cover the vast parameter space of heterogeneous LVLMs. Consequently, they often settle for a sparse approximation of the gradient landscape, neglecting the distributional geometry of gradients that fundamentally governs cross-model propagation.

Based on the above observations, we conclude that adversarial perturbations optimized on a fixed surrogate often overfit to its specific gradient geometry, aligning with sharp, surrogate-specific loss directions that fail to generalize across heterogeneous architectures. We argue that this transferability failure fundamentally originates from *cross-model gradient variance*. If gradients differ significantly across models, optimizing against a single fixed parameter instance inevitably produces brittle, non-transferable perturbations. To bridge this gap, our work deviates from the conventional static ensemble paradigm and advocates for Model Diversity Enrichment through the construction of a continuous parameter manifold. Furthermore, rather than relying on a sparse set of discrete surrogates, we introduce Stochastic Parameter Sampling, which treats surrogate parameters as samples from a dynamic distribution. This approach forces the adversarial perturbation to stabilize across an enriched variety of decision boundaries, ensuring that the generated signals are not tied to isolated architectural artifacts but are instead aligned with universal multimodal vulnerabilities.

To this end, in this paper, we propose a principled stochastic optimization LVLM attack framework that explicitly accounts for parameter uncertainty during the adversarial generation process. Instead of treating surrogate parameters as a fixed point in the weight space, we model them as stochastic samples drawn from a dynamically constructed local parameter subspace. Specifically, our approach consists of two synergistic components: Model Subspace Diversification and Stochastic Parameter Sampling. First, we enrich the surrogate's diversity by generating a set of orthogonal low-rank weight perturbations based on singular value decomposition (SVD), effectively spanning a continuous manifold that captures the structural and alignment variability inherent in diverse LVLMs. Second, during each optimization iteration, we sample a fresh parameter configuration from this manifold, performing perturbation optimization over a stochastic ensemble induced by this weight-space diversification. This mechanism forces the adversarial perturbation to decouple from surrogate-specific "gradient spikes" and instead align with the principal, model-agnostic gradient directions. Theoretically, we demonstrate that this sampling process implicitly minimizes the cross-model gradient variance, leading to adversarial signals that are fundamentally more robust and transferable. Extensive experiments across multiple state-of-the-art LVLMs validate that our framework significantly outperforms existing ensemble-based methods in terms of both transfer success rate and computational efficiency.

In summary, our main contributions are as follows:

- We identify cross-model gradient variance as the fundamental cause of transferability failure in LVLMs, shifting the research focus from discrete model ensembling to continuous parameter distribution modeling.
- We propose a lightweight Model Subspace Diversification mechanism. By generating SVD-based orthogonal low-rank perturbations, we construct a continuous parameter manifold that captures the architectural and alignment variability of LVLMs without the prohibitive cost of multiple full-scale surrogates.
- We introduce a novel stochastic optimization framework that dynamically samples configurations from the enriched manifold. This forces adversarial perturbations to decouple from surrogate-specific artifacts and align with universal multimodal vulnerabilities, achieving state-of-the-art transfer success rates across diverse black-box LVLMs.

2 Related Work

Adversarial Attacks on LVLMs. LVLMs achieve remarkable multimodal reasoning by aligning pre-trained visual backbones with powerful large language models [22, 23, 24, 25, 26, 27]. However,

this complex integration inherently amplifies adversarial vulnerabilities, as the cross-modal projection layer introduces novel attack vectors not present in unimodal systems [28, 29, 30, 31]. Early research in this domain primarily adapts gradient-based optimization from CV tasks, crafting imperceptible perturbations through back-propagation to mislead model outputs [32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43]. While these methods demonstrate significant potency in white-box settings, their efficacy is often confined to a specific surrogate instance and a fixed prompt. We observe that such optimization leads the perturbation to overfit the specific *gradient geometry* of the surrogate, resulting in brittle adversarial examples that fail to generalize across heterogeneous LVLM architectures. Although recent efforts [13, 15] have explored cross-prompt transferability, they often rely on computationally expensive joint training, leaving cross-model architectural transfer largely unaddressed.

Adversarial Transferability. Improving the transferability of adversarial examples has been extensively studied in both 2D image and 3D point cloud domains [44, 45, 46, 47, 48, 49, 50, 51]. The consensus among these works is that adversarial perturbations tend to overfit the local decision boundaries of the surrogate model, failing to generalize to unseen architectures. To mitigate this, a prevalent line of research [47, 52, 53] introduces diverse input transformations (*e.g.*, scaling, translation) to promote data-level generalization. Complementarily, advanced momentum-based optimizers [54, 55, 56] have been proposed to stabilize the optimization trajectory and escape surrogate-specific local optima. Traditional ensemble attacks [57, 58] attempt to capture model-agnostic vulnerabilities by attacking multiple fixed models simultaneously. However, these methods rely on a discrete set of static surrogates, which is not only computationally expensive for LVLMs but also provides a sparse coverage of the underlying gradient manifold. While feature-disruption methods [59, 60] aim for model-agnostic designs, they often fail to account for the unique cross-modal alignment variability in LVLMs. Different from these data-centric or static ensemble approaches, our work investigates transferability from a parameter-distributional perspective.

3 Methodology

In this section, we present a detailed exposition of our proposed framework for enhancing the adversarial transferability across LVLMs. Our methodology is grounded in the observation that transfer failure is fundamentally a consequence of gradient distributional mismatch between surrogate and target models. We propose to bridge this gap through a stochastic optimization paradigm that operates on a diversified parameter manifold.

3.1 A Distribution-Aware Perspective on LVLMs' Adversarial Transferability

Let $\mathcal{F}(\mathbf{x}, \mathbf{p}; \theta)$ denote an LVLM, where $\mathbf{x} \in \mathcal{X}$ is the input image, $\mathbf{p} \in \mathcal{P}$ is the textual prompt, and $\theta \in \Theta \subseteq \mathbb{R}^d$ represents the high-dimensional parameters. In a transfer-based black-box attack setting, the adversary lacks the exact parameters θ_t of the target LVLM model and instead optimizes a perturbation δ on a surrogate LVLM model θ_s.

We define the adversarial transferability bottleneck of existing LVLM attacks through the lens of gradient geometry inconsistency. For any two model instances θ_i, θ_j, let $\mathbf{g}(\theta) = \nabla_{\mathbf{x}} \mathcal{L}(\mathcal{F}(\mathbf{x}, \mathbf{p}; \theta), y_{target})$ be the adversarial gradient, where y_{target} denotes the attacker-chosen target output. The inconsistency is characterized by the gradient variance across the model distribution $\mathcal{P}(\Theta)$:

$$\sigma^2_{grad}(\mathbf{x}) = \int_{\Theta} \|\mathbf{g}(\theta) - \bar{\mathbf{g}}(\mathbf{x})\|^2 d\mathcal{P}(\theta), \quad \text{where } \bar{\mathbf{g}}(\mathbf{x}) = \mathbb{E}_{\theta \sim \mathcal{P}(\Theta)}[\mathbf{g}(\theta)]. \tag{1}$$

If the optimizer only follows $\mathbf{g}(\theta_s)$, it inevitably aligns the perturbation $\boldsymbol{\delta}$ with surrogate-specific gradient noise (*i.e.*, high-frequency local artifacts) that are singular to θ_s but vanish or shift in θ_t. To mitigate this, we reformulate the attack as a Stochastic Robust Optimization problem:

$$\min_{\|\boldsymbol{\delta}\|_\infty \leq \epsilon} J(\boldsymbol{\delta}) = \mathbb{E}_{\theta \sim \mathcal{P}(\Theta|\theta_s)} \left[\mathcal{L}(\mathcal{F}(\mathbf{x} + \boldsymbol{\delta}, \mathbf{p}; \theta), y_{target}) \right], \tag{2}$$

where $\mathcal{P}(\Theta|\theta_s)$ is a local parameter distribution designed to simulate potential target LVLMs' architectural and alignment variability.

3.2 Overview

Our Motivation. The fundamental bottleneck in cross-model transferability lies in the gradient distribution gap between the surrogate model θ_s and the unseen target model θ_t. We observe

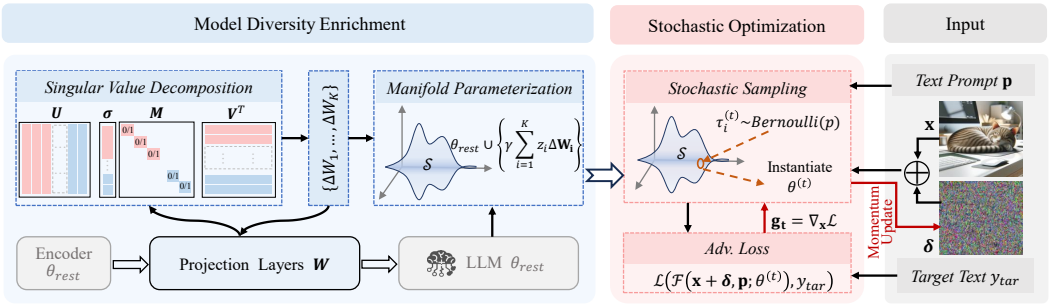


Figure 1: Overall pipeline of our proposed cross-model transfer attack.

that LVLMs, characterized by their massive parameter scales and complex cross-modal projection layers, possess highly non-smooth loss landscapes. Conventional transfer attacks, which optimize a perturbation $\boldsymbol{\delta}$ against a static θ_s or a static set $\{\theta_s\}$, inadvertently align the adversarial signal with surrogate-specific gradient spikes that are unique to the surrogate's specific architectural artifacts. Our motivation is to denoise these model-specific artifacts by shifting the optimization from such point-wise approach to a distributional approach optimizing over a continuous parameter manifold $\mathcal{S}$. By forcing the perturbation to remain effective across a diverse population of parameter configurations, we can extract the model-agnostic universal multimodal vulnerabilities.

Overall of Our Attack. We propose a systematic LVLM transfer-attack framework that transforms the adversarial generation process into a stochastic alignment problem. Our attack pipeline comprises two logically integrated stages in Figure 1: (1) Model Diversity Enrichment: We first expand the singular surrogate θ_s into an expressive, continuous parameter manifold $\mathcal{S}$. Instead of an expensive ensemble of multiple discrete models, we leverage Low-Rank Subspace Spanning via SVD to generate a set of orthogonal bases $\{\Delta\mathbf{W}_i\}$. These bases capture the latent structural variability of LVLMs, effectively creating a "dense virtual population" of surrogates that share the same high-level semantics but possess varying local decision boundaries. (2)Stochastic Parameter Sampling: Guided by the enriched manifold, we perform adversarial optimization via a dynamic sampling strategy. At each iteration t, we instantiate a fresh model configuration $\theta^{(t)} \sim \mathcal{S}$ and compute the gradient. This dynamic instantiation forces the perturbation δ to decouple from any single model's geometry and instead converge toward the principal, low-frequency gradient components that are invariant across the entire manifold. By integrating the two parts, our framework ensures that the resulting adversarial examples are not tied to architectural coincidences but are rooted in the fundamental cross-modal alignment flaws shared by the LVLM family, thus achieving superior transferability.

3.3 Model Subspace Diversification: Constructing the Continuous Surrogate Manifold

As established in our motivation, the primary obstacle to transferability is the singularity of the surrogate's gradient geometry. However, maintaining a diverse ensemble of multiple LVLMs—each typically exceeding 7B parameters—is computationally prohibitive for most adversarial scenarios. To circumvent this resource bottleneck while still capturing the necessary architectural variability, we propose to construct a lightweight yet expressive parameter subspace $\mathcal{S}$, hereafter referred to as the continuous surrogate manifold (here, "continuous" contrasts with discrete ensemble methods rather than implying strict topological continuity), extended from the original surrogate θ_s.

SVD-based Orthogonal Basis Generation. Our diversification strategy focuses on the vision-language projection layers and the cross-attention blocks, as these components are the primary drivers of multimodal decision boundaries and alignment behavior. To simulate the alignment gaps present in heterogeneous LVLMs, we perturb the latent semantic structure of these layers. Specifically, let $\mathbf{W} \in \mathbb{R}^{m\times n}$ be a critical weight matrix in the surrogate's projector. We first perform a Singular Value Decomposition to decouple its principal alignment directions: $\mathbf{W} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$, where $\mathbf{U}, \mathbf{V}$ are orthogonal matrices of singular vectors, and $\boldsymbol{\Sigma}$ is a diagonal matrix of singular values.

The key insight is that different models often emphasize different spectral components of the visual-linguistic mapping. To approximate this variability, we generate K structured basis matrices $\{\Delta\mathbf{W}_i\}_{i=1}^K$ by modulating the singular value spectrum:

$$\Delta\mathbf{W}_i = \mathbf{U} \cdot \mathrm{diag}(\boldsymbol{\sigma} \odot \mathbf{M}_i) \cdot \mathbf{V}^\top, \tag{3}$$

where $\boldsymbol{\sigma}$ is the vector of singular values from $\mathbf{\Sigma}$. $\mathbf{M}_i \in \{0,1\}^{\min(m,n)}$ is a disjoint low-rank binary mask. By ensuring that the masks $\{\mathbf{M}_i\}$ target non-overlapping segments of the singular value spectrum, we obtain a set of mutually orthogonal bases in the parameter space. This orthogonality is crucial as it ensures that each basis captures a unique "variation mode" of the model's decision logic.

Manifold Parameterization. The resulting diversified subspace is then defined as a linear span of these orthogonal bases:

$$\mathcal{S} = \left\{ \theta(\mathbf{z}) = \theta_{rest} \cup \{\gamma \sum_{i=1}^{K} z_i \Delta \mathbf{W}_i\} \mid z_i \in \{0,1\} \right\}, \tag{4}$$

where θ_s is split into projection layer parameters and all other parameters θ_{rest}. γ is a scaling factor that controls the diversity magnitude while maintaining the model's fundamental reasoning utility. Through this construction, we effectively span a high-dimensional volumetric variance within the parameter manifold. This manifold acts as a proxy for the distributional discrepancy between models, allowing us to simulate a vast population of potential victim LVLM models without the memory overhead of multi-model ensembling.

3.4 Stochastic Parameter Sampling Optimization

While the manifold $\mathcal{S}$ provides an expressive continuous subspace of victim LVLMs, performing a deterministic ensemble over the entire manifold is computationally prohibitive, as it would require simultaneous gradient backpropagation through multiple parameter configurations. To tractably harness the diversity of our manifold, we propose Stochastic Parameter Sampling (SPS). This approach transforms the static optimization problem into a stochastic optimization process, dynamically approximating the LVLM-agnostic gradient direction during the iterative attack.

Dynamic Instantiation via Stochastic Sampling. Rather than relying on fixed surrogates, the SPS mechanism instantiates a different identity from the manifold $\mathcal{S}$ at each optimization step. In each iteration t of the attack, we sample a stochastic configuration by drawing a random selection vector $\boldsymbol{\tau}^{(t)} \in \{0,1\}^K$, where each component $\tau_i^{(t)}$ follows a Bernoulli distribution with probability p. The sampled parameter configuration $\theta^{(t)}$ is then dynamically instantiated as:

$$\theta^{(t)} = \theta_{rest} \cup \left\{ \gamma \sum_{i=1}^{K} \tau_i^{(t)} \Delta \mathbf{W}_i \right\}. \tag{5}$$

Gradient Smoothing and Update. To ensure that the adversarial perturbation converges toward the most stable directions within the manifold, we incorporate a Momentum-based Gradient Smoothing mechanism. Let $\mathbf{g}(\theta^{(t)}) = \nabla_{\mathbf{x}} \mathcal{L}(\mathcal{F}(\mathbf{x}_t, \mathbf{p}; \theta^{(t)}), y_{target})$ be the gradient computed on the t-th sampled configuration. The update rule is formulated as:

$$\boldsymbol{m}_{t+1} = \mu \cdot \boldsymbol{m}_t + \frac{\mathbf{g}(\theta^{(t)})}{\|\mathbf{g}(\theta^{(t)})\|_1}, \quad \mathbf{x}_{t+1} = \text{Clip}_{\mathbf{x},\epsilon}\{\mathbf{x}_t - \alpha \cdot \text{sign}(\boldsymbol{m}_{t+1})\}. \tag{6}$$

where μ is momentum factor, $\boldsymbol{m}_t$ is the i-th gradient, and α is step size.

Why SPS Enhances Transferability? The causal logic behind SPS lies in the law of large numbers as applied to gradient geometry. By sampling a fresh $\theta^{(t)}$ at each iteration, the optimizer is effectively prevented from trapping the perturbation in surrogate-specific local optima. Essentially, SPS acts as a low-frequency filter: the noise components of the gradient that are unique to a specific $\theta^{(t)}$ cancel out over time, while the low-frequency gradient components that persist across the majority of the manifold remain. This stochastic blurring of the decision boundary ensures that the final perturbation δ is not an overfitted artifact of the surrogate, but rather a robust signal aligned with the intrinsic, model-agnostic vulnerabilities shared by the broader LVLM population. (Proof in Appendix A.2.)

3.5 Theoretical Analysis: Bounding the Transfer Success

To provide a rigorous foundation for our framework, we analyze the relationship between stochastic optimization and cross-model generalization. We demonstrate that our approach reduces the transfer gap by explicitly accounting for gradient inconsistency within $\mathcal{S}$. We treat the manifold $\mathcal{S}$ as a structured proxy for the model distribution $\mathcal{P}(\Theta)$ around the surrogate model θ_s.

Theorem 1 (Expected Gradient Alignment and Variance Penalty). *Let $\boldsymbol{\delta}_{det}$ be the perturbation optimized deterministically on a single surrogate, and $\boldsymbol{\delta}_{sps}$ be the perturbation optimized over the stochastic manifold $\mathcal{S}$ via SPS. For a target model θ_t drawn from the same underlying distribution $\mathcal{P}(\Theta)$, the ratio of their expected adversarial gains satisfies:*

$$\frac{\mathbb{E}_{\theta_s,\theta_t}[\langle \boldsymbol{g}_t, \boldsymbol{\delta}_{det} \rangle]}{\mathbb{E}_{\theta_t}[\langle \boldsymbol{g}_t, \boldsymbol{\delta}_{sps} \rangle]} \approx \underbrace{\left(1 + \frac{\sigma_{grad}^2}{||\bar{\mathbf{g}}||^2}\right)^{-1/2}}_{\text{Variance Penalty}}. \tag{7}$$

where $\bar{\mathbf{g}} = \mathbb{E}_{\theta \sim \mathcal{P}(\Theta)}[\nabla_x \mathcal{L}(x, \theta)]$ represents the universal model-agnostic gradient signal, and σ_{grad}^2 is the cross-model gradient variance.

Proof Sketch (Details are in Appendix A.1). We evaluate the expected inner product between the generated perturbation and the true target gradient $\boldsymbol{g}_t$. For a deterministic update $\boldsymbol{\delta}_{det}$ derived from a single surrogate, the expected alignment is severely degraded by the surrogate-specific high-frequency noise, resulting in the variance penalty term parameterized by σ_{grad}^2. Unlike deterministic attacks that overfit to isolated "gradient spikes" singular to θ_s—which strictly misaligns the perturbation on unseen targets—SPS forces $\boldsymbol{\delta}_{sps}$ to converge towards the universal gradient $\bar{\mathbf{g}}$ by averaging across the manifold $\mathcal{S}$. As shown in the bound, SPS completely circumvents this variance penalty, maximizing the transfer success.

3.6 Final Objectives for Generating Transferable Adversarial Examples against LVLMs

The ultimate objective of our framework is to synthesize a perturbation $\boldsymbol{\delta}$ that captures the universal multimodal failure modes across the LVLM family. By integrating Model Diversity Enrichment and Stochastic Parameter Sampling, we redefine the adversarial objective function as a multi-target optimization over the parameter manifold $\mathcal{S}$:

$$\boldsymbol{\delta}^* = \arg\min_{\|\boldsymbol{\delta}\|_\infty \leq \epsilon} \mathbb{E}_{\theta \in \mathcal{S}} \left[\mathcal{D}_{sem}\left(\mathcal{F}(\mathbf{x} + \boldsymbol{\delta}, \mathbf{p}; \theta), y_{target}\right)\right], \tag{8}$$

where $\mathcal{D}_{sem}$ denotes a semantic alignment loss. Our objective explicitly seeks a consensus direction within the LoRA-Manifold. This ensures that the generated $\boldsymbol{\delta}^*$ is not an accidental byproduct of a specific model's alignment artifacts, but a fundamental signal that disrupts the core vision-language projection mechanism shared by various LVLM architectures. The complete implementation of our proposed framework is summarized in Algorithm 1 of the Appendix.

4 Experiments

4.1 Experimental Setup

Evaluated LVLM Architectures. To verify cross-model transferability, we evaluate our attack on a diverse suite of open-source LVLMs following [13, 15, 16], including: (1) LLaVA-1.5 (Vicuna-7B backbone) [6], (2) MiniGPT-4 (Llama-2-7B-Chat backbone) [61], (3) BLIP-2 (OPT-2.7B backbone) [5], and (4) InstructBLIP (Vicuna-7B backbone) [62]. This selection encompasses various vision encoders and language model alignments, providing a rigorous testbed for architectural transferability.

LVLM Tasks and Datasets. Our evaluation spans three representative multimodal reasoning tasks: Image Captioning, Image Classification, and Visual Question Answering (VQA). We construct our evaluation set by sampling images from DALL-E [63], SVIT [64], and VQAv2 [65]. Task-specific instructions and prompts are adapted from the CroPA framework [13].

Evaluation Protocols. To evaluate the effectiveness of adversarial perturbations, we employ three metrics: (1) Semantic Similarity (SS): Measures the cosine similarity between the latent embeddings of the model's output and the target text, computed using SentenceTransformer [66]. (2) Exact Match (EM) and (3) Conditional Contain (CC): These metrics evaluate the word-level success rates by assessing precise overlaps or the presence of target semantic keywords in the generated response.

Hyper-parameter Settings. For the adversarial generation process, we set the maximum perturbation budget to $\epsilon = 16/255$ under the ℓ_∞ norm. We employ the momentum-stabilized optimizer with a momentum factor $\mu = 0.9$. The optimization spans $T = 500$ iterations with a step size $\alpha = \epsilon/T$. For

Table 1: Performance comparisons on the adversarial transferability across different LVLM models. The experimental results are calculated by the averaged semantic similarities (↑) and attack success rates (↑) on three tasks. Target text: "I am sorry".

Dataset	Source Model	LVLM Attack	LLaVA-1.5			MiniGPT-4			BLIP-2			InstructBLIP		
			SS	*EM*	*CC*	*SS*	*EM*	*CC*	*SS*	*EM*	*CC*	*SS*	*EM*	*CC*
DALL-E	LLaVA-1.5	PGD [67]	*0.964*	*96.1*	*96.1*	0.042	0.0	0.0	0.094	0.3	0.3	0.137	1.2	1.6
		CroPA [13]	*0.819*	*79.7*	*79.7*	0.043	0.0	0.0	0.093	0.0	0.0	0.139	3.1	3.4
		UniAtt [15]	*0.842*	*80.6*	*88.3*	0.186	12.5	17.9	0.267	16.2	23.8	0.303	22.8	29.5
		MPTA [16]	*0.813*	*80.4*	*80.4*	0.661	61.4	66.7	0.693	63.5	66.9	0.724	63.7	70.2
		Ours	*0.885*	*86.4*	*90.2*	**0.712**	**67.3**	**72.5**	**0.738**	**68.4**	**71.9**	**0.776**	**69.1**	**75.8**
	MiniGPT-4	PGD [67]	0.046	0.0	0.0	*0.823*	*79.7*	*79.7*	0.103	2.8	8.9	0.146	5.9	10.4
		CroPA [13]	0.051	0.0	0.0	*0.955*	*94.8*	*96.1*	0.125	9.0	11.1	0.166	3.4	3.4
		UniAtt [15]	0.298	15.4	22.7	*0.830*	*79.5*	*85.2*	0.338	23.1	27.7	0.316	16.2	16.2
		MPTA [16]	0.650	57.3	64.9	*0.860*	*84.2*	*84.5*	0.716	63.4	68.5	0.698	58.3	64.1
		Ours	**0.698**	**64.2**	**68.1**	*0.912*	*90.5*	*92.4*	**0.765**	**69.1**	**73.8**	**0.752**	**65.4**	**71.6**
	BLIP-2	PGD [67]	0.056	0.0	0.0	0.053	0.0	0.0	*0.608*	*58.2*	*64.7*	0.164	3.5	11.3
		CroPA [13]	0.059	0.0	0.0	0.057	0.7	2.0	*0.610*	*28.8*	*93.5*	0.199	8.1	17.2
		UniAtt [15]	0.397	24.8	31.2	0.359	22.6	29.4	*0.817*	*77.9*	*85.1*	0.275	13.4	15.8
		MPTA [16]	0.695	52.4	60.3	0.657	52.7	58.5	*0.755*	*75.1*	*81.4*	0.506	41.2	43.1
		Ours	**0.742**	**59.1**	**66.7**	**0.704**	**60.3**	**65.1**	*0.824*	*80.3*	*86.9*	**0.563**	**48.2**	**50.7**
	InstructBLIP	PGD [67]	0.048	0.0	0.0	0.047	0.0	0.0	0.128	1.0	1.9	*0.498*	*43.1*	*53.6*
		CroPA [13]	0.048	0.0	0.0	0.053	0.0	0.7	0.230	9.6	16.7	*0.842*	*81.7*	*82.3*
		UniAtt [15]	0.173	8.5	14.6	0.269	21.4	21.4	0.421	30.8	37.3	*0.854*	*79.6*	*83.7*
		MPTA [16]	0.448	40.2	45.9	0.473	43.5	46.8	0.539	46.9	54.6	*0.689*	*68.9*	*79.0*
		Ours	**0.513**	**48.4**	**52.3**	**0.542**	**50.1**	**56.7**	**0.607**	**55.8**	**62.1**	*0.887*	*85.4*	*88.2*
SVIT	LLaVA-1.5	PGD [67]	*0.968*	*96.7*	*96.7*	0.040	0.0	0.0	0.097	0.8	1.2	0.132	2.4	3.7
		CroPA [13]	*0.762*	*61.3*	*66.4*	0.054	0.0	0.0	0.091	0.3	0.9	0.125	1.5	1.9
		UniAtt [15]	*0.856*	*83.9*	*87.8*	0.211	13.2	21.0	0.285	16.6	27.4	0.324	25.8	25.8
		MPTA [16]	*0.825*	*82.4*	*82.4*	0.675	61.3	68.6	0.699	65.3	68.9	0.718	65.1	71.0
		Ours	*0.897*	*88.4*	*91.3*	**0.728**	**66.9**	**74.1**	**0.751**	**69.4**	**73.6**	**0.765**	**70.3**	**77.2**
	MiniGPT-4	PGD [67]	0.049	0.0	0.0	*0.882*	*86.9*	*87.6*	0.118	2.9	2.9	0.140	7.8	9.3
		CroPA [13]	0.047	0.0	0.0	*0.988*	*98.7*	*98.7*	0.191	13.1	13.1	0.162	4.6	4.6
		UniAtt [15]	0.315	16.8	24.2	*0.849*	*84.4*	*85.3*	0.352	24.7	30.3	0.428	30.0	32.3
		MPTA [16]	0.647	62.3	66.0	*0.904*	*89.2*	*89.5*	0.745	66.9	73.1	0.785	72.4	72.4
		Ours	**0.701**	**67.5**	**71.8**	*0.941*	*93.6*	*93.8*	**0.792**	**72.5**	**78.9**	**0.834**	**78.1**	**79.5**
	BLIP-2	PGD [67]	0.049	0.0	0.0	0.056	0.0	0.9	*0.642*	*61.4*	*68.0*	0.172	6.4	10.8
		CroPA [13]	0.051	0.0	0.0	0.064	0.7	3.5	*0.608*	*28.8*	*94.8*	0.194	8.3	16.1
		UniAtt [15]	0.375	26.1	33.7	0.392	23.8	31.4	*0.797*	*78.3*	*83.5*	0.257	12.5	18.6
		MPTA [16]	0.657	55.8	63.4	0.685	56.7	60.5	*0.754*	*75.0*	*82.5*	0.489	42.5	46.3
		Ours	**0.718**	**61.4**	**69.5**	**0.732**	**62.9**	**66.2**	*0.816*	*80.2*	*85.8*	**0.552**	**48.7**	**52.1**
	InstructBLIP	PGD [67]	0.046	0.0	0.0	0.037	0.0	0.0	0.133	4.0	6.5	*0.529*	*47.7*	*60.1*
		CroPA [13]	0.049	0.0	0.0	0.053	0.0	1.4	0.266	16.8	26.2	*0.859*	*83.7*	*85.0*
		UniAtt [15]	0.158	6.7	12.1	0.262	20.8	23.5	0.384	29.9	36.7	*0.836*	*77.9*	*84.5*
		MPTA [16]	0.482	44.6	47.4	0.521	47.1	53.0	0.599	49.0	55.5	*0.767*	*75.6*	*82.5*
		Ours	**0.551**	**51.9**	**55.8**	**0.593**	**54.4**	**59.1**	**0.665**	**57.2**	**63.8**	*0.912*	*88.1*	*90.4*

the Model Diversity Enrichment stage, we extract $K = 8$ orthogonal bases via SVD on projection layers. During the Stochastic Parameter Sampling phase, the sampling probability p is set to 0.5, and the diversity scale γ is set to 0.1 to maintain the surrogate's semantic integrity while maximizing gradient variance. All baselines are re-implemented under identical settings for a fair comparison.

4.2 Main Results

Cross-Model Transferability. Table 1 summarizes the performance of our proposed method against state-of-the-art LVLM attacks. The results demonstrate that our framework consistently achieves superior black-box transferability across all surrogate-target pairs. Notably, while baseline methods like PGD and CroPA exhibit high white-box performance, their efficacy collapses when transferred to heterogeneous architectures. In contrast, our method effectively bridges this alignment gap by achieving significant margins over the strongest baseline, MPTA, in all metrics. This substantial improvement validates: by diversifying the parameter subspace and employing stochastic sampling, we successfully decouple adversarial perturbations from surrogate-specific artifacts, thereby capturing universal multimodal vulnerabilities that are invariant across different vision-language alignments.

Table 2: The transfer-attack performance on complex target texts using adversarial examples generated on LLaVA-1.5 with DALL-E dataset.

Target Text	LVLM Attack	LLaVA-1.5			MiniGPT-4			BLIP-2			InstructBLIP		
		SS	*EM*	*CC*	*SS*	*EM*	*CC*	*SS*	*EM*	*CC*	*SS*	*EM*	*CC*
"A man holding a big doughnut at a festival."	PGD [67]	0.916	91.2	91.2	0.054	0.0	0.0	0.087	0.5	0.5	0.112	1.3	2.1
	CroPA [13]	0.782	76.9	76.9	0.040	0.0	0.0	0.087	0.0	0.0	0.108	2.8	3.0
	UniAtt [15]	0.821	80.1	85.5	0.204	17.8	21.7	0.291	18.9	28.3	0.312	23.4	26.8
	MPTA [16]	0.796	79.9	80.2	0.662	60.5	64.8	0.714	62.4	68.6	0.685	59.7	64.2
	Ours	**0.834**	**82.6**	**86.1**	**0.702**	**65.3**	**69.4**	**0.748**	**66.1**	**72.5**	**0.731**	**64.4**	**69.8**
"A photo of a teddy bear on a skateboard in Times Square."	PGD [67]	0.903	89.8	91.2	0.048	0.0	0.0	0.096	0.9	1.1	0.106	1.5	2.4
	CroPA [13]	0.786	78.2	78.2	0.056	0.0	0.0	0.099	0.0	0.4	0.115	3.2	3.2
	UniAtt [15]	0.834	82.7	85.9	0.216	18.6	21.5	0.285	19.6	25.4	0.298	20.1	24.6
	MPTA [16]	0.779	79.8	79.8	0.671	61.2	66.9	0.695	62.7	67.7	0.664	57.1	63.9
	Ours	**0.821**	**81.5**	**86.4**	**0.715**	**65.8**	**70.4**	**0.733**	**67.2**	**72.1**	**0.709**	**63.8**	**68.5**
"A beautiful bird with a black and white color in snow."	PGD [67]	0.907	90.1	90.1	0.065	0.0	0.0	0.113	0.0	0.4	0.124	0.8	1.2
	CroPA [13]	0.814	80.6	80.6	0.052	0.0	0.0	0.091	0.4	0.8	0.119	3.6	3.6
	UniAtt [15]	0.843	81.6	86.3	0.229	18.3	23.1	0.287	20.7	29.6	0.305	24.1	28.5
	MPTA [16]	0.815	80.9	82.2	0.676	61.5	68.2	0.729	60.9	68.4	0.692	61.3	65.5
	Ours	**0.862**	**84.4**	**88.7**	**0.724**	**66.1**	**71.9**	**0.768**	**65.5**	**73.2**	**0.744**	**67.3**	**71.0**

Table 3: Transfer-attack performance on the generated adversarial examples from MiniGPT-4 to GPT-4o and Claude-3.5.

LVLM Attack	GPT-4o			Claude-3.5		
	SS	*EM*	*CC*	*SS*	*EM*	*CC*
PGD [67]	0.036	0.0	0.0	0.048	0.0	0.0
CroPA [13]	0.057	0.0	0.0	0.052	0.0	0.0
UniAtt [15]	0.142	4.1	7.5	0.169	3.3	10.4
MPTA [16]	0.608	44.7	51.3	0.620	48.6	56.5
Ours	**0.653**	**52.6**	**58.9**	**0.667**	**56.3**	**61.7**

Table 4: Efficiency and performance comparison from LLaVA-1.5 to InstructBLIP.

Method	Config	VRAM	TFLOPS	SS	EM	CC	Speed
Single Surrogate	$N = 1$	14.2	1.0×	0.137	1.2	1.6	**4.2** it/s
Traditional Ensemble [20]	$N = 2$	28.5	2.1×	0.324	22.1	28.4	2.0 it/s
	$N = 4$	56.8	4.2×	0.512	45.3	51.2	1.1 it/s
	$N = 8$	OOM	8.5×	0.648	58.7	63.2	0.4 it/s
Ours	$K = 2$	14.4	1.0×	0.582	52.4	59.7	4.1 it/s
	$K = 4$	14.6	1.1×	0.704	61.3	68.9	3.9 it/s
	$K = 8$	14.8	1.1×	0.776	69.1	75.8	3.8 it/s
	$K = 16$	15.2	1.2×	**0.791**	**70.4**	**77.2**	3.5 it/s

Moreover, we evaluate the transferability of adversarial examples crafted on MiniGPT-4 to commercial LVLMs like GPT-4o [68] and Claude-3.5-Sonnet [69]. Table 3 indicates that our attack maintains superior performance across these applications.

Sensitivity to Diverse Target Semantics. To further evaluate the versatility of our framework, we extend the attack to complex, multi-entity target descriptions in Table 2. The results indicate that our method is highly insensitive to the complexity and content of the target text.

4.3 Attack Robustness and Efficiency

Robustness against Defenses. To further evaluate the robustness of our proposed attack, we employ JPEG Compression [70], Diffusion Restoration [71], FARE [72] and DPS [73] defenses to investigate whether our attack remains effective and transferable. As shown in Figure 2, our attack not only achieves the smallest performance degeneration compared to other methods, but also demonstrates the best transferability under different defenses, validating the effectiveness of our attack.

Analysis on Diversity and Efficiency. To evaluate our continuous manifold-based diversity against discrete model ensembling, we vary the ensemble size (N) of standard models. As shown in Table 4, traditional ensemble methods [20] suffer from diminishing returns as N increases, primarily because discrete models often share similar architectural biases. In contrast, by spanning a continuous spectral subspace, our method achieves higher transferability than an ensem-

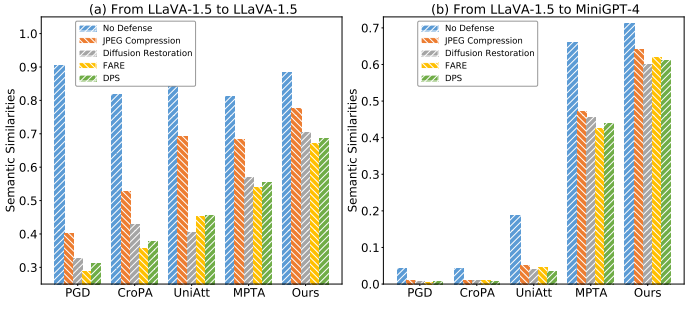


Figure 2: Investigation on the robustness against defenses.

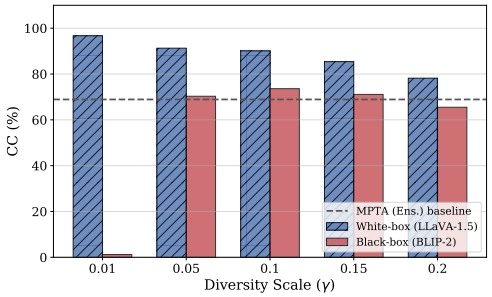


Figure 3: Impact of diversity scale γ on CC. Peak performance is observed around $\gamma = 0.1$.

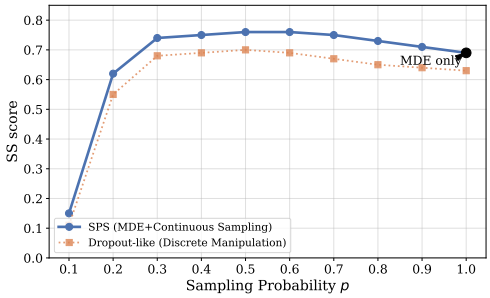


Figure 4: Comparison of SPS vs. Dropout-like scheme on SS score.

ble of 8 full-scale models. Moreover, traditional ensembles incur a linear increase in GPU memory and Times relative to the number of models. Conversely, our method maintains a near-constant memory footprint ($\sim$14.8GB for LLaVA-1.5), as it only requires storing low-rank basis matrices and performing additive weight modulation. Furthermore, the throughput (it/s) remains nearly $9\times$ faster than the $N = 8$ ensemble, making it a highly practical solution for large-scale adversarial red-teaming.

4.4 Ablation Studies

Effectiveness of Main Components. We first examine the necessity of our two core components: Model Diversity Enrichment (MDE) and Stochastic Parameter Sampling (SPS). We compare four configurations: (a) *Base*: A standard PGD attack on the single surrogate θ_s; (b) *MDE only*: Optimizing against a fixed ensemble of $K = 8$ orthogonal bases, equivalent to standard discrete ensembling; (c) *SPS only*: Applying stochastic sampling to the base model θ_s without manifold diversification; (d) *Full (Ours)*: The complete MDE-SPS framework. As shown in Table 5, while discrete ensembling (MDE only) provides a moderate boost over the *Base* attack, the improvement is limited by surrogate-specific biases. SPS only fails to gain significant transferability due to the lack of structural diversity. Notably, our *Full* method achieves a significant performance.

Sensitivity to Diversity Scale γ. The diversity scale γ controls the expansion magnitude of the parameter subspace. A small γ may fail to capture sufficient cross-model variance, while an excessively large γ could destroy the model's semantic functionality. We analyze this trade-off by varying γ from 0.01 to 0.2. Figure 3 illustrates that an optimal diversity magnitude exists that maximizes transferability without sacrificing white-box effectiveness.

Table 5: Ablation study of the proposed components (from LLaVA-1.5 to BLIP-2).

Component	LLaVA-1.5 (White-box)			BLIP-2 (Black-box)		
	SS	*EM*	*CC*	*SS*	*EM*	*CC*
(a) Base	0.968	96.7	96.7	0.097	0.8	1.2
(b) MDE only	0.849	82.6	85.7	0.642	62.2	65.5
(c) SPS only	0.951	94.2	95.8	0.112	1.5	2.4
(d) Full	0.897	88.4	91.3	**0.751**	**69.4**	**73.6**

Role of Bernoulli Sampling Probability p. In the SPS phase, the probability p determines the stochasticity of the dynamically instantiated surrogate. As shown in Figure 4, this confirms that explicitly constructing orthogonal bases (MDE) provides a more stable foundation for gradient denoising than random model manipulations.

5 Conclusion

In this paper, we investigate the fundamental bottleneck of adversarial transferability in LVLMs and identify surrogate-specific gradient variance as a primary cause of transfer failure. To mitigate this, we propose a novel framework that shifts the adversarial optimization paradigm from a deterministic point-wise search to a distributional manifold alignment. By synergizing Model Diversity Enrichment and Stochastic Parameter Sampling, our method effectively decouples adversarial perturbations from architectural artifacts and aligns them with universal multimodal vulnerabilities. Extensive experiments across diverse state-of-the-art LVLMs demonstrate that our framework consistently outperforms existing methods, achieving superior black-box transferability regardless of target architectural discrepancies or semantic complexity.

Acknowledgements

This work was supported by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123604). This work was also supported by the Department of Science and Technology of Hubei Province (No. 2026DJA002). This work was also supported by WHU-Kingsoft Joint Lab. The numerical calculations in this paper have been done in part on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[2] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

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

[4] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

[5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

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

[7] Daizong Liu, Tianyao Luo, Shuwei Huang, Junyi Yang, Siyuan Liang, Xiaojun Jia, Chun Pong Lau, Jianfeng Dong, Yan Liu, Yang Bai, et al. A survey on ai agent security: Reasoning, acting, and self-evolving. *AI Plus*, 1(1):100004–122, 2026.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

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

[11] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*, 2024.

[12] Daizong Liu, Tianyao Luo, Junhao Dong, Xiaowen Cai, Hao He, Weidong Wang, Runwei Guan, Jianfeng Dong, Pan Zhou, Bo Du, et al. Flexible targeted adversarial alignment with frequency-based visual focus for attacking large vision-language models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–18, 2026.

[13] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[14] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[15] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Xiang Fang, Keke Tang, Yao Wan, and Lichao Sun. Pandora's box: Towards building universal attackers against real-world large vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[16] Xiaowen Cai, Daizong Liu, Xiaoye Qu, Xiang Fang, Jianfeng Dong, Keke Tang, Pan Zhou, Lichao Sun, and Wei Hu. Towards building model/prompt-transferable attackers against large vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[17] Qi Guo, Shanmin Pang, Xiaojun Jia, Yang Liu, and Qing Guo. Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. *IEEE Transactions on Information Forensics and Security*, 2024.

[18] Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards large-scale self-supervised generation of targeted adversarial examples for vision-language models. *arXiv e-prints*, pages arXiv–2410, 2024.

[19] Zhaoyi Li, Xiaohan Zhao, Dong-Dong Wu, Jiacheng Cui, and Zhiqiang Shen. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of gpt-4.5/4o/o1. *arXiv preprint arXiv:2503.10635*, 2025.

[20] Xiaojun Jia, Sensen Gao, Simeng Qin, Tianyu Pang, Chao Du, Yihao Huang, Xinfeng Li, Yiming Li, Bo Li, and Yang Liu. Adversarial attacks against closed-source mllms via feature optimal alignment. *arXiv preprint arXiv:2505.21494*, 2025.

[21] Hanxun Huang, Sarah Erfani, Yige Li, Xingjun Ma, and James Bailey. X-transfer attacks: Towards super transferable adversarial attacks on clip. *arXiv preprint arXiv:2505.05528*, 2025.

[22] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. Benchmarking micro-action recognition: Dataset, method, and application. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6238–6252, 2024.

[23] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244, 2021.

[24] Daizong Liu, Wei Hu, and Xin Li. Point cloud attacks in graph spectral domain: When 3d geometry meets graph signal processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[25] Daizong Liu and Wei Hu. Seeing is not believing: Adversarial natural object optimization for hard-label 3d scene attacks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11886–11897, 2025.

[26] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078, 2020.

[27] Xianke Chen, Daizong Liu, Xun Yang, Xirong Li, Jianfeng Dong, Meng Wang, and Xu Wang. Prvr: Partially relevant video retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[28] Junhao Dong, Piotr Koniusz, Xinghua Qu, and Yew-Soon Ong. Stabilizing modality gap & lowering gradient norms improve zero-shot adversarial robustness of vlms. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 236–247, 2025.

[29] Fuyao Cai, Daizong Liu, Xiang Fang, Jixiang Yu, Keke Tang, and Pan Zhou. Imperceptible beam-sensitive adversarial attacks for lidar-based object detection in autonomous driving. In *IEEE International Conference on Multimedia & Expo 2025 (ICME 2025)*, 2025.

[30] Daizong Liu and Wei Hu. Explicitly perceiving and preserving the local geometric structures for 3d point cloud attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3576–3584, 2024.

[31] Xiaowen Cai, Daizong Liu, Runwei Guan, and Pan Zhou. Imperceptible transfer attack on large vision-language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[32] Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime. *arXiv preprint arXiv:2309.00236*, 2023.

[33] Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google's bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*, 2023.

[34] Xiaohan Fu, Zihan Wang, Shuheng Li, Rajesh K Gupta, Niloofar Mireshghallah, Taylor Berg-Kirkpatrick, and Earlence Fernandes. Misusing tools in large language models with visual adversarial examples. *arXiv preprint arXiv:2310.03185*, 2023.

[35] Kuofeng Gao, Yang Bai, Jiawang Bai, Yong Yang, and Shu-Tao Xia. Adversarial robustness for visual grounding of multimodal large language models. *arXiv preprint arXiv:2405.09981*, 2024.

[36] Daizong Liu, Xiaowen Cai, Pan Zhou, Xiaoye Qu, Lichao Sun, and Wei Hu. Are large vision-language models robust to adversarial visual transformations? *IEEE Transactions on Information Forensics and Security*, 2026.

[37] Hai Yan, Haijian Ma, Xiaowen Cai, Daizong Liu, Zenghui Yuan, Xiaoye Qu, Jianfeng Dong, Runwei Guan, Xiang Fang, Hongyang He, et al. Fit the distribution: Cross-image/prompt adversarial attacks on multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[38] Daizong Liu, Baoquan Chen, and Wei Hu. Spatial-spectral homogeneous attacks on physical-world large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 7114–7122, 2026.

[39] Daizong Liu, Xiaowen Cai, Junhao Dong, Zhongliang Guo, Xiaoye Qu, Runwei Guan, Xiang Fang, and Dengpan Ye. Attacking gray-box large vision-language models with adaptive svd-structured adversarial alignment. In *International Conference on Machine Learning*, 2026.

[40] Daizong Liu, Wangqin Liu, Xiaowen Cai, Pan Zhou, Runwei Guan, Xiaoye Qu, and Bo Du. Generating transferable attacks across large vision-language models using adversarial deformation learning. *Pattern Recognition*, page 113194, 2026.

[41] Nian Ai, Guangke Chen, Xiaowen Cai, Zhongliang Guo, Daizong Liu, Pan Zhou, and Ognjen Arandjelović. Attacking hard-label large vision-language models with model-sensitive adversarial patch designs. *Pattern Recognition*, page 114261, 2026.

[42] Daizong Liu, Junhao Dong, Zhiyuan Ma, Xiaoye Qu, Xiang Fang, Runwei Guan, Keke Tang, Jianfeng Dong, and Yew-Soon Ong. Forbid your attention: Fooling multimodal large language models by selectively removing intrinsic focus in spectral domain. *IEEE Transactions on Multimedia*, 2026.

[43] Daizong Liu, Junhao Dong, Xiang Fang, Hongyang He, Keyan Jin, Zhongliang Guo, Xiaoye Qu, and Keke Tang. Understanding and exploiting phase sensitivity for attacking large vision–language models. In *The 35th International Joint Conference on Artificial Intelligence*, 2026.

[44] Daizong Liu and Wei Hu. Imperceptible transfer attack and defense on 3d point cloud classification. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4727–4746, 2022.

[45] Weibin Wu, Yuxin Su, Michael R Lyu, and Irwin King. Improving the transferability of adversarial samples with adversarial transformations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9024–9033, 2021.

[46] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019.

[47] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4312–4321, 2019.

[48] Qianjiang Hu, Daizong Liu, and Wei Hu. Exploring the devil in graph spectral domain for 3d point cloud attacks. In *European Conference on Computer Vision*, pages 229–248. Springer, 2022.

[49] Yunbo Tao, Daizong Liu, Pan Zhou, Yulai Xie, Wei Du, and Wei Hu. 3dhacker: Spectrum-based decision boundary generation for hard-label 3d point cloud attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14340–14350, 2023.

[50] Daizong Liu, Wei Hu, and Xin Li. Robust geometry-dependent attack for 3d point clouds. *IEEE Transactions on Multimedia*, 26:2866–2877, 2023.

[51] Yang Liu, Daizong Liu, Zongming Guo, and Wei Hu. Cross-task knowledge transfer for semi-supervised joint 3d grounding and captioning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3818–3827, 2024.

[52] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019.

[53] Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14993–15002, 2022.

[54] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[55] Zhijin Ge, Hongying Liu, Wang Xiaosen, Fanhua Shang, and Yuanyuan Liu. Boosting adversarial transferability by achieving flat local maxima. *Advances in Neural Information Processing Systems*, 36:70141–70161, 2023.

[56] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1924–1933, 2021.

[57] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. Learning transferable adversarial examples via ghost networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11458–11465, 2020.

[58] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[59] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1161–1170, 2020.

[60] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018.

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

[62] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[63] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

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

[65] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[66] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[67] Xuanimng Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[68] OpenAI. Hello gpt-4o. *https://openai.com/index/hello-gpt4o/*, 2024.

[69] Anthropic. The claude 3 model family: Opus, sonnet, haiku. *https://claude.ai/*, 2024.

[70] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

[71] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

[72] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *ICML*, 2024.

[73] Qi Zhou, Tianlin Li, Qing Guo, Dongxia Wang, Yun Lin, Yang Liu, and Jin Song Dong. Defending lvlms against vision attacks through partial-perception supervision. *ICML*, 2025.

A Detailed Mathematical Derivations and Proofs

In this appendix, we provide the complete mathematical proofs for the theoretical claims made in Section 3. Specifically, we focus on proving Theorem 1, which establishes the relationship between cross-model gradient variance and the adversarial transferability gap.

A.1 Proof of Theorem 1: Expected Gradient Alignment and Variance Penalty

In transfer-based black-box attacks, the success of an adversarial perturbation $\boldsymbol{\delta}$ (bounded by $||\boldsymbol{\delta}||_2 \leq \eta$) is primarily determined by its alignment with the true gradient of the target model $\boldsymbol{g}_t = \nabla_x \mathcal{L}(x, \theta_t)$. (Note: While practical attacks often use L_∞ bounds with the sign operator, analyzing the L_2 bounded inner product is standard theoretical practice to maintain analytical tractability while preserving the geometric implications of gradient variance.)

Definitions and Assumptions: Let $\mathcal{P}(\Theta)$ be the distribution of LVLM parameters that share the same underlying modality gap. We define the model-agnostic expected gradient as:

$$\bar{\mathbf{g}} = \mathbb{E}_{\theta \sim \mathcal{P}(\Theta)}[\nabla_x \mathcal{L}(x, \theta)]. \tag{9}$$

For any specific model instance (either surrogate θ_s or target θ_t), its gradient $\mathbf{g}(\theta)$ can be decomposed into the universal gradient $\bar{\mathbf{g}}$ and a surrogate-specific noise component (high-frequency local artifacts) $\boldsymbol{\xi}(\theta)$:

$$\mathbf{g}(\theta) = \bar{\mathbf{g}} + \boldsymbol{\xi}(\theta). \tag{10}$$

where $\mathbb{E}_\theta[\boldsymbol{\xi}(\theta)] = \mathbf{0}$. The cross-model gradient variance is defined as the expected energy of this model-specific noise:

$$\sigma_{grad}^2 = \mathbb{E}_\theta[||\boldsymbol{\xi}(\theta)||^2]. \tag{11}$$

We assume that the low-frequency gradient component $\mathbf{g}_{LF}$ (as shown in Appendix A.2) captured within the continuous surrogate manifold $\mathcal{S}$ is semantically aligned with the expected gradient $\bar{\mathbf{g}}$ of the model distribution $\mathcal{P}(\Theta)$, i.e., $\mathbf{g}_{LF} \approx \bar{\mathbf{g}}$. This is guaranteed by our SVD-based construction, which spans the most significant alignment modes shared across the LVLM family.

Part I: The Expected Transfer Gain of Deterministic Attack (Single Surrogate)

In a standard attack, the perturbation is optimized against a single static surrogate θ_s. The optimal update direction is:

$$\boldsymbol{\delta}_{det} = \eta \frac{\boldsymbol{g}_s}{||\boldsymbol{g}_s||} = \eta \frac{\bar{\mathbf{g}} + \boldsymbol{\xi}_s}{||\bar{\mathbf{g}} + \boldsymbol{\xi}_s||}. \tag{12}$$

When this perturbation is applied to an unseen target model θ_t, the expected first-order adversarial gain is:

$$\mathbb{E}_{\theta_t}[\langle \boldsymbol{g}_t, \boldsymbol{\delta}_{det} \rangle] = \mathbb{E}_{\theta_t}\left[\left\langle \bar{\mathbf{g}} + \boldsymbol{\xi}_t, \eta \frac{\bar{\mathbf{g}} + \boldsymbol{\xi}_s}{||\bar{\mathbf{g}} + \boldsymbol{\xi}_s||} \right\rangle\right]. \tag{13}$$

Since the target model's noise $\boldsymbol{\xi}_t$ is independent of θ_s and has zero mean, the term involving $\boldsymbol{\xi}_t$ vanishes:

$$\mathbb{E}_{\theta_t}[\langle \boldsymbol{g}_t, \boldsymbol{\delta}_{det} \rangle] = \eta \frac{\langle \bar{\mathbf{g}}, \bar{\mathbf{g}} + \boldsymbol{\xi}_s \rangle}{||\bar{\mathbf{g}} + \boldsymbol{\xi}_s||} = \eta \frac{||\bar{\mathbf{g}}||^2 + \langle \bar{\mathbf{g}}, \boldsymbol{\xi}_s \rangle}{||\bar{\mathbf{g}} + \boldsymbol{\xi}_s||}. \tag{14}$$

To evaluate the expected performance across a general selection of surrogates, we take the expectation over θ_s. In high-dimensional spaces, we can approximate the expectation of the denominator using the first-order Taylor expansion (the Delta method, $\mathbb{E}[\sqrt{X}] \approx \sqrt{\mathbb{E}[X]}$):

$$\mathbb{E}_{\theta_s}[||\bar{\mathbf{g}} + \boldsymbol{\xi}_s||] \approx \sqrt{\mathbb{E}[||\bar{\mathbf{g}}||^2 + ||\boldsymbol{\xi}_s||^2 + 2\langle \bar{\mathbf{g}}, \boldsymbol{\xi}_s \rangle]} = \sqrt{||\bar{\mathbf{g}}||^2 + \sigma_{grad}^2}. \tag{15}$$

Given that $\mathbb{E}[\langle \bar{\mathbf{g}}, \boldsymbol{\xi}_s \rangle] = 0$, the expected transfer gain for a deterministic attack is bounded by:

$$\mathbb{E}_{\theta_s, \theta_t}[\langle \boldsymbol{g}_t, \boldsymbol{\delta}_{det} \rangle] \approx \eta \frac{||\bar{\mathbf{g}}||^2}{\sqrt{||\bar{\mathbf{g}}||^2 + \sigma_{grad}^2}}. \tag{16}$$

Part II: The Expected Transfer Gain of SPS (Manifold Sampling)

For our proposed SPS framework, the perturbation is generated by iteratively sampling from the orthogonal manifold $\mathcal{S}$, which effectively computes the expected gradient (as shown in Appendix A.2). Thus, the SPS update direction converges to:

$$\boldsymbol{\delta}_{sps} = \eta \frac{\bar{\mathbf{g}}}{||\bar{\mathbf{g}}||}. \tag{17}$$

When this perturbation is applied to the target model θ_t, the expected adversarial gain is:

$$\mathbb{E}_{\theta_t}[\langle \boldsymbol{g}_t, \boldsymbol{\delta}_{sps} \rangle] = \mathbb{E}_{\theta_t}\left[\left\langle \bar{\mathbf{g}} + \boldsymbol{\xi}_t, \eta \frac{\bar{\mathbf{g}}}{||\bar{\mathbf{g}}||} \right\rangle\right]. \tag{18}$$

Again, since $\mathbb{E}_{\theta_t}[\boldsymbol{\xi}_t] = \mathbf{0}$, the cross term vanishes entirely:

$$\mathbb{E}_{\theta_t}[\langle \boldsymbol{g}_t, \boldsymbol{\delta}_{sps} \rangle] = \eta \frac{\langle \bar{\mathbf{g}}, \bar{\mathbf{g}} \rangle}{||\bar{\mathbf{g}}||} = \eta ||\bar{\mathbf{g}}||. \tag{19}$$

Conclusion: By comparing the two expected gains, we can derive the transferability penalty ratio of the deterministic attack:

$$\frac{\text{Expected Gain}(\boldsymbol{\delta}_{det})}{\text{Expected Gain}(\boldsymbol{\delta}_{sps})} \approx \frac{||\bar{\mathbf{g}}||}{\sqrt{||\bar{\mathbf{g}}||^2 + \sigma_{grad}^2}} = \left(1 + \frac{\sigma_{grad}^2}{||\bar{\mathbf{g}}||^2}\right)^{-1/2}. \tag{20}$$

Proof completes. This theoretical bound rigorously demonstrates that the transferability of a conventional single-surrogate attack is heavily penalized by the cross-model gradient variance σ_{grad}^2. As the architectural differences (and thus the variance) increase, the deterministic alignment strictly degrades. In contrast, by operating on the expected manifold, SPS completely circumvents this variance penalty, extracting the pure universal vulnerability signal $\bar{\mathbf{g}}$.

A.2 Derivation of the Stochastic Smoothing Effect

We further analyze how SPS acts as a low-pass filter on the gradient manifold. Let $\mathbf{g}(\theta)$ be the gradient at parameter θ. We can decompose the gradient into a low-frequency component $\mathbf{g}_{LF}$ (shared across the subspace $\mathcal{S}$) and a high-frequency component $\mathbf{g}_{HF}$ (surrogate-specific noise):

$$\mathbf{g}(\theta) = \mathbf{g}_{LF} + \mathbf{g}_{HF}(\theta), \quad \text{where } \mathbb{E}_{\theta \in \mathcal{S}}[\mathbf{g}_{HF}(\theta)] = \mathbf{0}. \tag{21}$$

When we perform standard point-wise optimization on θ_s, the update is $\delta \propto \mathbf{g}_{LF} + \mathbf{g}_{HF}(\theta_s)$. The term $\mathbf{g}_{HF}(\theta_s)$ causes the perturbation to overfit to the local geometry of θ_s.

Under SPS, the effective update after T iterations is:

$$\delta_{SPS} \propto \frac{1}{T}\sum_{t=1}^{T}(\mathbf{g}_{LF} + \mathbf{g}_{HF}(\theta^{(t)})) = \mathbf{g}_{LF} + \frac{1}{T}\sum_{t=1}^{T}\mathbf{g}_{HF}(\theta^{(t)}). \tag{22}$$

By the Law of Large Numbers, as $T \to \infty$:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{g}_{HF}(\theta^{(t)}) \xrightarrow{a.s.} \mathbb{E}_{\theta}[\mathbf{g}_{HF}(\theta)] = \mathbf{0}. \tag{23}$$

Therefore, $\delta_{SPS} \to \mathbf{g}_{LF}$. This proves that SPS explicitly filters out architectural artifacts and converges to the semantic-level vulnerabilities that are invariant to parameter shifts, which explains the significant boost in transferability.

A.3 Geometric Orthogonality of SVD-based Diversification

The SVD-based initialization ensures that the generated model variants explore the parameter space efficiently. Let $\mathbf{W} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ be the projection matrix. We construct K basis matrices $\Delta\mathbf{W}_i = \mathbf{U} \cdot \text{diag}(\boldsymbol{\sigma} \odot \mathbf{M}_i) \cdot \mathbf{V}^\top$, where $\mathbf{M}_i$ are binary masking matrices.

Algorithm 1 Adversarial Transfer via Stochastic Parameter Sampling (SPS)

Input: Surrogate model θ_s, benign image $\mathbf{x}$, prompt $\mathbf{p}$, target response y_{target}, iterations T, budget ϵ, diversity scale γ.
Initialize: Adversarial image $\mathbf{x}_0 = \mathbf{x}$, perturbation $\boldsymbol{\delta}_0$, gradient $\boldsymbol{m}_0 = 0$.
Phase I: Model Diversity Enrichment
  Extract projection layer weights $\mathbf{W}$ from θ_s.
  Compute spectral decomposition (SVD): $\mathbf{W} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$.
  Construct K orthogonal manifold bases $\{\Delta\mathbf{W}_i\}_{i=1}^K$ via spectral masking.
Phase II: Stochastic Optimization
for $t = 0$ to $T-1$ **do**
  Sample random configuration: $\boldsymbol{\tau}^{(t)} \in \{0,1\}^K$ where $\tau_i^{(t)} \sim \text{Bernoulli}(p)$.
  Dynamic Instantiation: $\theta^{(t)} = \theta_{rest} \cup \left\{\gamma \sum_{i=1}^K \tau_i^{(t)} \Delta\mathbf{W}_i\right\}$.
  Compute forward pass: $y_{pred} = \mathcal{F}(\mathbf{x}_t, \mathbf{p}; \theta^{(t)})$.
  Calculate adversarial gradient: $\mathbf{g}_t = \nabla_{\mathbf{x}} \mathcal{L}(y_{pred}, y_{target})$.
  Gradient Smoothing: $\boldsymbol{m}_{t+1} = \mu \cdot \boldsymbol{m}_t + \mathbf{g}_t / \|\mathbf{g}_t\|_1$.
  Perturbation Update: $\mathbf{x}_{t+1} = \text{Clip}_{\mathbf{x},\epsilon}\{\mathbf{x}_t - \alpha \cdot \text{sign}(\boldsymbol{m}_{t+1})\}$.
end for
return $\mathbf{x}_T$.

Property 1 (Orthogonality): For any $i \neq j$, the Frobenius inner product satisfies $\langle \Delta\mathbf{W}_i, \Delta\mathbf{W}_j \rangle_F = 0$.

$$\langle \Delta\mathbf{W}_i, \Delta\mathbf{W}_j \rangle_F = \text{Tr}(\Delta\mathbf{W}_i^\top \Delta\mathbf{W}_j) = \text{Tr}(\mathbf{V}\text{diag}(\boldsymbol{\sigma} \odot \mathbf{M}_i)\mathbf{U}^\top \mathbf{U}\text{diag}(\boldsymbol{\sigma} \odot \mathbf{M}_j)\mathbf{V}^\top). \tag{24}$$

Utilizing the property $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ and the cyclic property of the trace:

$$\langle \Delta\mathbf{W}_i, \Delta\mathbf{W}_j \rangle_F = \text{Tr}(\text{diag}(\boldsymbol{\sigma} \odot \mathbf{M}_i)\text{diag}(\boldsymbol{\sigma} \odot \mathbf{M}_j)\mathbf{V}^\top\mathbf{V}) = \sum_k (\sigma_k M_{i,k})(\sigma_k M_{j,k}). \tag{25}$$

Since the masks $\mathbf{M}_i$ and $\mathbf{M}_j$ are designed to select disjoint sets of singular values, $(\sigma_k M_{i,k})(\sigma_k M_{j,k}) = 0$ for all k.

Geometric Implication: In the LoRA-Manifold, the basis matrices $\{\Delta\mathbf{W}_i\}$ form an orthogonal coordinate system. The stochastic instantiation $\theta^{(t)} = \theta_{rest} \cup \{\gamma \sum \tau_i^{(t)} \Delta W_i\}$ can be viewed as sampling from a K-dimensional hypercube centered at θ_s. This orthogonality maximizes the *sampling volume* within the parameter space, ensuring that the gradient expectation $\bar{\mathbf{g}}$ is estimated from the most diverse possible set of model configurations.

A.4 Details of Our Algorithm

We provide the details of our attack process in Algorithm 1. This algorithm consists of two stages: (1) Model Diversity Enhancement: We first extract the projection layer weights from a surrogate model θ_s. We then apply a low-rank subspace spanning technique based on Singular Value Decomposition to generate a set of orthogonal basis $\{\Delta\mathbf{W}_i\}_{i=1}^K$. (2) Stochastic Optimization: In each iteration t of the attack, we first sample a stochastic configuration by drawing a random selection vector $\boldsymbol{\tau}^{(t)} \in \{0,1\}^K$, where each component $\tau_i^{(t)}$ follows a Bernoulli distribution with probability p. Then, we dynamically instantiate the LVLM model $\theta^{(t)}$ based on $\tau_i^{(t)}$ and $\Delta\mathbf{W}_i$. Next, we feed the current perturbed image and the prompt into the model $\theta^{(t)}$ to compute the loss between the output and the target text, as well as the gradients with respect to the perturbed image. Finally, we introduce a momentum mechanism and utilize a sign function to control the step size during the perturbation update, thereby ultimately generating transferable adversarial examples.

B More Experimental Results

B.1 Experiments on More Datasets

We provide more performance comparisons on the adversarial transferability across different LVLMs on VQAv2 datasets as shown in Table 6. We can see that our attack still achieves the highest

Table 6: Performance comparisons on the adversarial transferability across different LVLM models (on VQAv2 dataset). The experimental results are calculated by the averaged semantic similarities (↑) and attack success rates (↑) on three tasks. Target text: "I am sorry".

Dataset	Source Model	LVLM Attack	LLaVA-1.5			MiniGPT-4			BLIP-2			InstructBLIP		
			SS	*EM*	*CC*	*SS*	*EM*	*CC*	*SS*	*EM*	*CC*	*SS*	*EM*	*CC*
VQAv2	LLaVA-1.5	PGD [67]	*0.968*	*96.1*	*96.1*	0.044	0.0	0.0	0.096	0.1	0.7	0.135	0.9	1.9
		CroPA [13]	*0.797*	*65.2*	*65.2*	0.023	0.0	0.0	0.084	0.0	0.8	0.120	1.1	1.4
		UniAtt [15]	*0.828*	*82.0*	*84.5*	0.172	10.9	18.1	0.253	22.1	22.1	0.298	23.7	30.6
		MPTA [16]	*0.833*	*83.7*	*83.7*	0.645	65.6	65.6	0.706	63.2	69.9	0.723	64.2	72.5
		Ours	*0.847*	*85.9*	*86.6*	**0.716**	**69.7**	**73.2**	**0.739**	**69.1**	**74.6**	**0.743**	**70.2**	**75.2**
	MiniGPT-4	PGD [67]	0.036	0.0	0.0	*0.866*	*85.0*	*85.0*	0.139	4.4	5.9	0.157	3.0	3.0
		CroPA [13]	0.037	0.0	0.0	*0.987*	*98.0*	*98.7*	0.187	10.5	11.9	0.166	3.4	3.7
		UniAtt [15]	0.373	24.9	35.8	*0.874*	*83.6*	*89.7*	0.380	31.1	38.5	0.309	22.4	27.6
		MPTA [16]	0.658	61.3	67.2	*0.910*	*89.9*	*90.0*	0.726	61.4	61.4	0.695	63.1	68.3
		Ours	**0.676**	**65.8**	**69.4**	*0.885*	*86.1*	*90.6*	**0.692**	**66.8**	**71.1**	**0.704**	**67.5**	**71.4**
	BLIP-2	PGD [67]	0.047	0.0	0.0	0.057	0.0	0.0	*0.649*	*61.4*	*70.6*	0.199	6.5	6.5
		CroPA [13]	0.046	0.0	0.0	0.060	0.0	1.3	*0.619*	*28.1*	*96.1*	0.209	7.4	14.9
		UniAtt [15]	0.406	27.9	35.3	0.385	23.8	32.0	*0.836*	*79.3*	*86.2*	0.284	15.7	20.1
		MPTA [16]	0.613	58.5	62.0	0.602	55.8	58.3	*0.744*	*73.6*	*82.9*	0.492	37.5	39.6
		Ours	**0.677**	**65.1**	**69.6**	**0.653**	**63.2**	**68.5**	*0.840*	*81.9*	*85.7*	**0.537**	**51.8**	**55.3**
	InstructBLIP	PGD [67]	0.037	0.0	0.0	0.049	0.0	0.0	0.132	0.8	4.1	*0.532*	*44.4*	*62.7*
		CroPA [13]	0.039	0.0	0.0	0.059	0.7	3.5	0.265	13.1	23.5	*0.869*	*85.0*	*87.6*
		UniAtt [15]	0.182	10.4	15.1	0.253	18.7	22.9	0.448	37.3	43.6	*0.858*	*81.5*	*83.9*
		MPTA [16]	0.464	43.9	48.7	0.562	54.3	58.6	0.577	55.9	64.1	*0.782*	*76.6*	*84.3*
		Ours	**0.544**	**52.4**	**57.7**	**0.638**	**61.9**	**65.6**	**0.651**	**63.2**	**67.0**	*0.892*	*87.3*	*91.2*

transferability. This substantial improvement validates: by diversifying the parameter subspace and employing stochastic sampling, we successfully decouple adversarial perturbations from surrogate-specific artifacts, thereby capturing universal multimodal vulnerabilities that are invariant across different vision-language alignments.

B.2 Stability across Iterations

We investigate how many sampling iterations T are required for SPS to converge to a stable gradient direction. As shown in Figure 5, the cosine similarity between $\delta_{SPS}^{(T)}$ and the "ground truth" ensemble gradient $\bar{\mathbf{g}}$ stabilizes rapidly.

With $T = 20$, the similarity already exceeds 0.92, whereas random sampling without our MDE-manifold requires $T > 100$ to reach similar stability. This efficiency is a direct result of our SVD-based basis construction, which spans the most significant directions of the parameter space, reducing the variance of the stochastic estimate.

B.3 Impact of Basis Number K on Performance and Memory

The number of basis matrices K determines the dimensionality of the sampled manifold. While a higher K provides finer-grained diversity, it also increases the storage requirement for the low-rank components. Table 7 shows that transferability saturates around $K = 8$. Specifically, increasing K from 8 to 16 only yields a marginal improvement (+0.015 SS) at the cost of doubling the storage for basis matrices. Thus, we select $K = 8$ as the default configuration for its optimal balance between red-teaming efficacy and resource efficiency.

Table 7: Sensitivity analysis of the number of basis matrices K from LLaVA-1.5 to InstructBLIP.

Basis Count K	2	4	8	16
SS	0.582	0.704	0.776	0.791
CC (%)	59.7	68.9	75.8	77.2
Storage (MB)	124	248	496	992

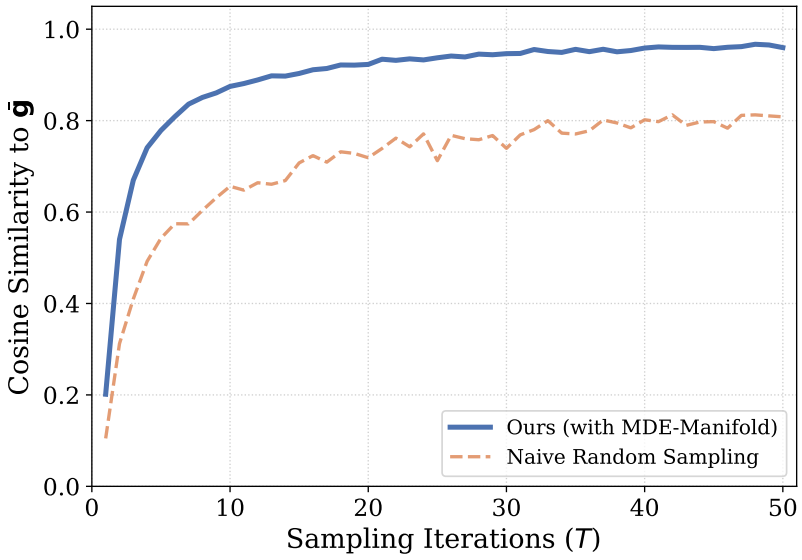


Figure 5: Cosine similarity between the SPS gradient and the manifold mean gradient as a function of iterations T. Our method achieves faster convergence due to structured diversity.

Table 8: Transferability across different vision tower architectures. SS scores are reported.

Target Model	Vision Tower	Baseline (MPTA)	Ours
LLaVA-1.5 (Surrogate)	CLIP-ViT-L/14	0.825 (White-box)	0.897
MiniGPT-v2	EVA-CLIP-G	0.452	**0.618**
InternLM-XC2	OpenAI-CLIP-L	0.512	**0.704**
InstructBLIP	ViT-G/14	0.488	**0.655**

C Further Quantitative Investigations

C.1 Generalization Across Vision Tower Architectures

A key challenge in LVLM transferability is the discrepancy between vision encoders (*e.g.*, CLIP-ViT-L vs. SigLIP). We evaluate Diverse-SPS by attacking LLaVA-1.5 (ViT-L/14) and transferring to models with heterogeneous vision towers.

As shown in Table 8, our method achieves significant transfer gains even when the target model uses a different pre-training paradigm (*e.g.*, InternLM-XComposer2 uses OpenAI-CLIP while MiniGPT-v2 uses EVA-CLIP). This suggests that our manifold sampling captures high-level multimodal alignment vulnerabilities that transcend the specific low-level visual features learned by different encoders.

C.2 Joint Sensitivity Analysis of γ and p

To understand the interaction between the diversity scale γ and the sampling probability p, we provide a landscape of the Transfer Success Rate (SS) in Figure 6. We observe a "stability ridge" where $p \in [0.4, 0.6]$ and $\gamma \in [0.08, 0.12]$. This plateau indicates that the framework is not overly sensitive to precise hyper-parameter tuning, provided that the manifold expansion and sampling stochasticity are properly balanced. The performance drop at high p and high γ confirms that excessive stochasticity without sufficient structure may introduce gradient directions that deviate too far from the semantic manifold.

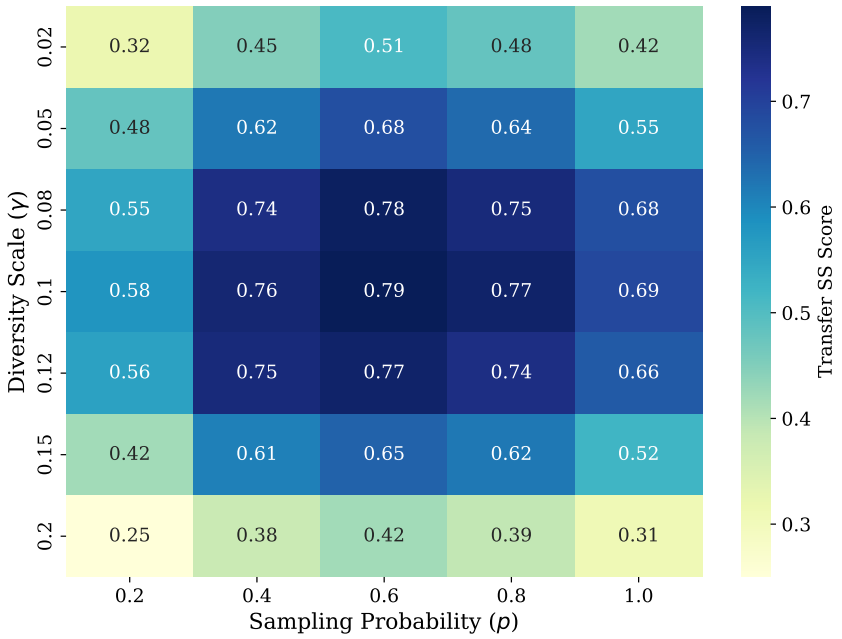


Figure 6: Joint sensitivity analysis of diversity scale γ and sampling probability p on the black-box transfer SS score (LLaVA-1.5 $\rightarrow$ BLIP-2). The "stability ridge" in the central region indicates that our Diverse-SPS framework is robust to hyper-parameter variations within a reasonable range.

D More Visualization Results

To further assess the effectiveness and transferability of the proposed attack method, we provide a multi-dimensional visualization analysis: (1) Adversarial examples generated against various target texts are illustrated in the Figure 7. The results demonstrate that the generated adversarial examples exhibit excellent visual imperceptibility. (2) Visualization results of transfer attacks targeting a variety of LVLMs are presented in the Figure 8. The results indicate that the generated adversarial examples can effectively utilize attacker-specified text labels to mislead unknown LVLMs into producing target outputs, thereby fully validating the cross-model transferability of our method. (3) Attacks targeting commercial closed-source LVLMs are illustrated in Figures 9, 10, 11 and 12. Specifically, these figures depict the results of transferring adversarial examples generated by MiniGPT-4 to attack GPT-4o and Claude-3.5. The results demonstrate that the generated adversarial examples can effectively attack advanced closed-source models, inducing them to generate descriptions consistent with the target text. This indicates that the synergistic interplay between parameter subspace diversification and random sampling strategies endows our method with superior transfer performance in cross-architecture scenarios.

E Limitations and Broader Impacts

Limitations. Future research directions for evaluating the security of LVLMs should prioritize physical-world adversarial attack scenarios, particularly in safety-critical deployments such as autonomous driving or robotic control systems. In such applications, adversarial examples must be crafted under real-world constraints, where input images are captured by physical sensors (*e.g.*, cameras or LiDAR) and subject to dynamic environmental interference (*e.g.*, lighting variations, motion blur, or sensor noise). Although our attack method is effective in theoretical settings, it still requires further improvement to maintain attack effectiveness under such complex real-world conditions.

Broader Impacts. This study presents a novel LVLM attack method for generating transferable adversarial examples against different LVLM models. Despite the rising interest in attacking LVLMs, our attack method shows that direct harm exists to LVLM applications. Our research aims to deepen

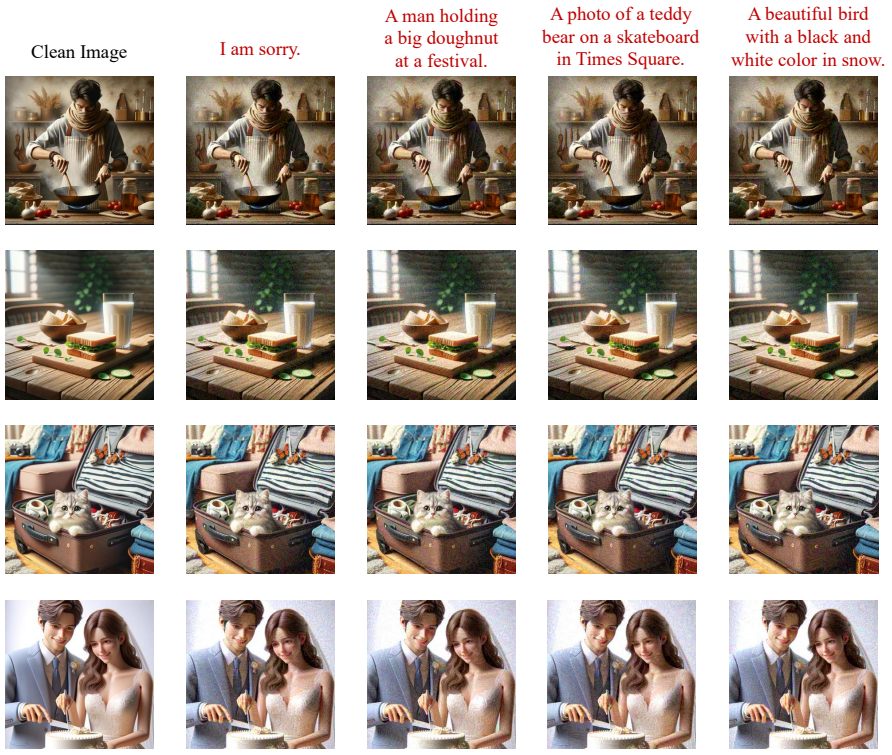


Figure 7: Adversarial examples generated against various target texts.

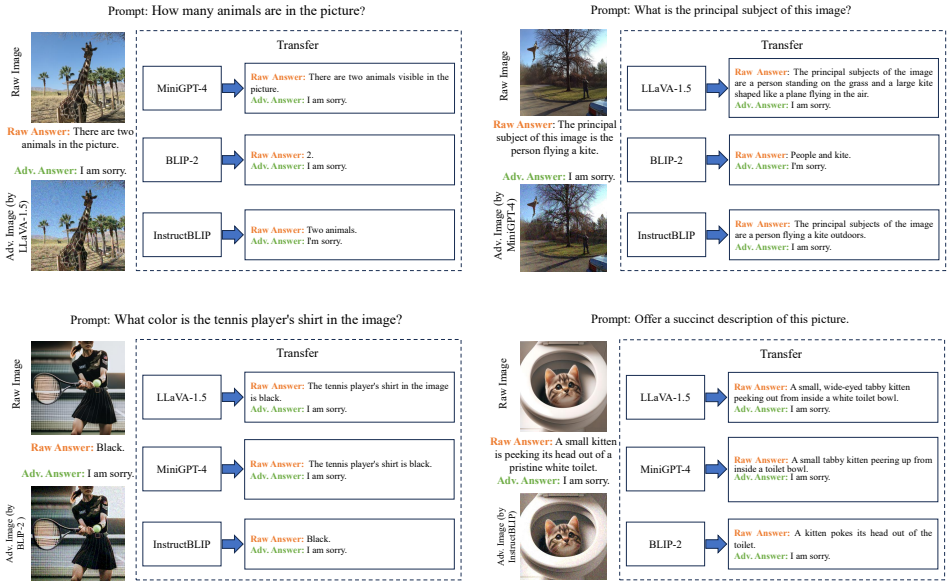


Figure 8: Visualizations of our transferable attack across different LVLM models.

the understanding of LVLM robustness and enhance their safety, promoting safer AI environments. However, we acknowledge the potential negative societal impact of our work and the presence of potentially offensive and harmful adversarial examples in our paper. It is possible that the developed attacking strategies could be misused to evade practically deployed systems and cause potential negative societal impacts. Specifically, our threat model assumes real-world access and targeted responses, which involve manipulating existing APIs such as GPT-4 (with visual inputs) and/or Midjourney on purpose, thereby increasing the risk if these vision-language APIs are implemented as

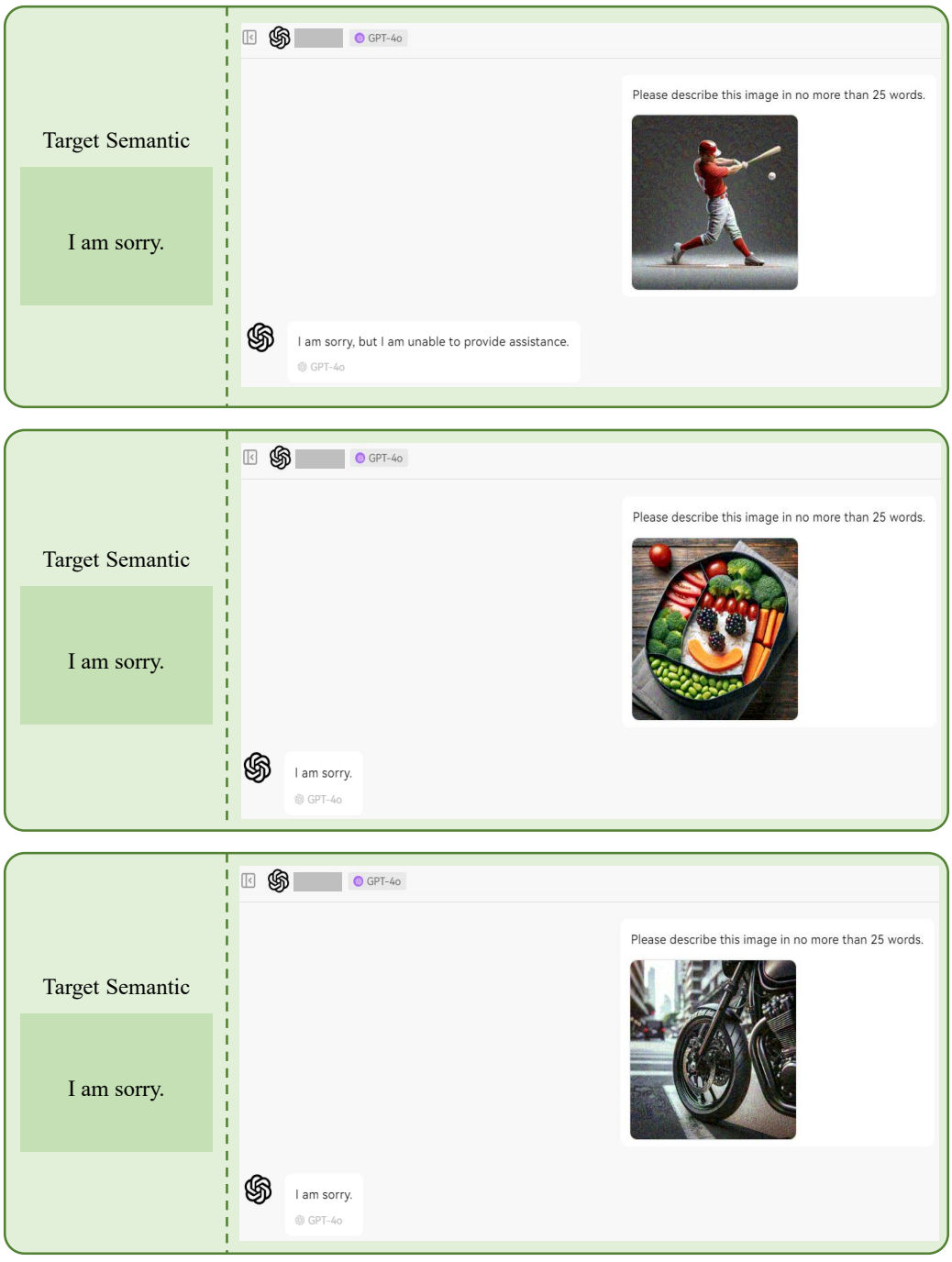


Figure 9: Visualization of transfer attacks where adversarial examples generated on MiniGPT-4 are transferred to the commercial closed-source model GPT-4o.

plugins in other products. We believe the contributions in our work point out new vulnerabilities in LVLMs, which could aid future research in making them more reliable and secure.

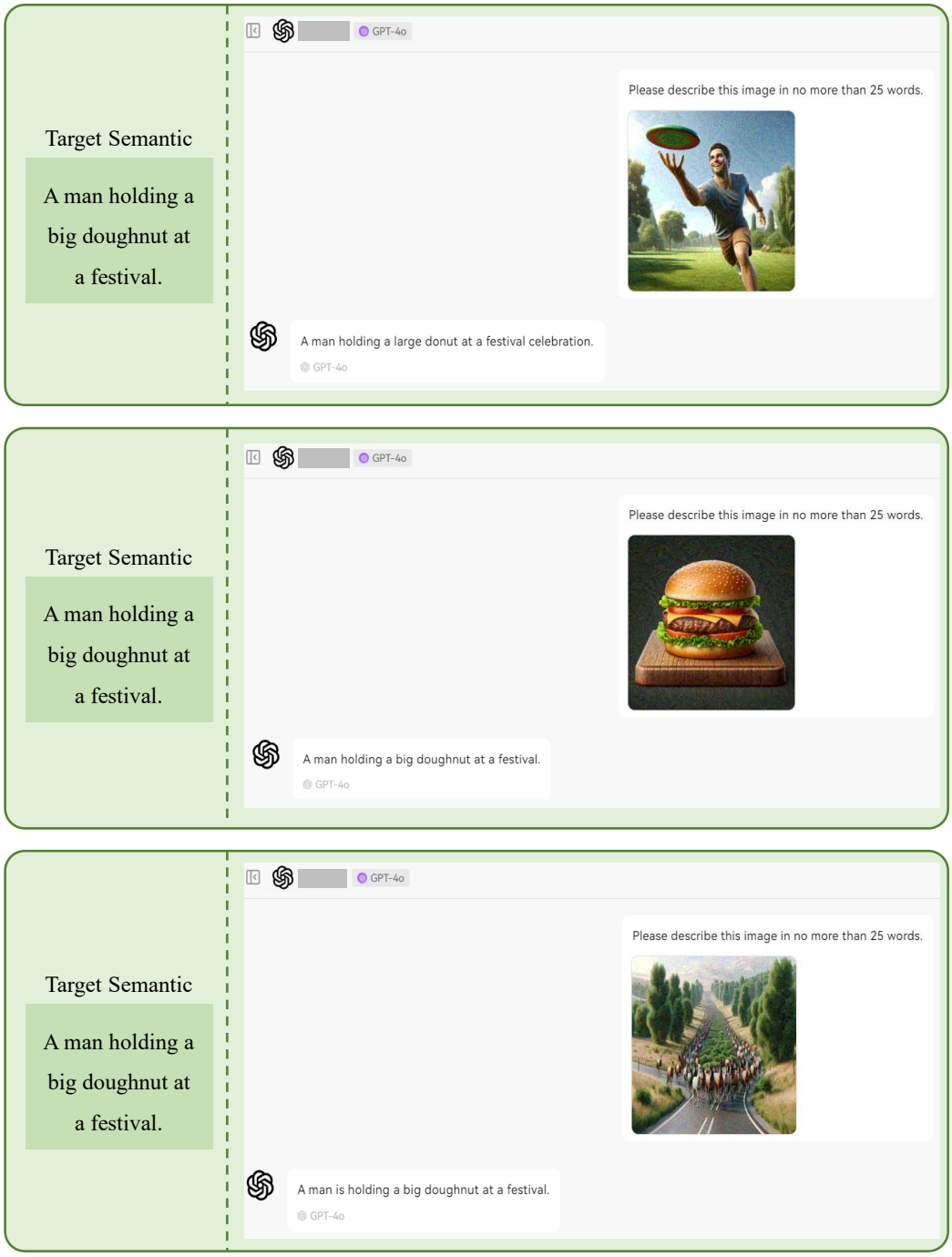


Figure 10: Visualization of transfer attacks where adversarial examples generated on MiniGPT-4 are transferred to the commercial closed-source model GPT-4o.

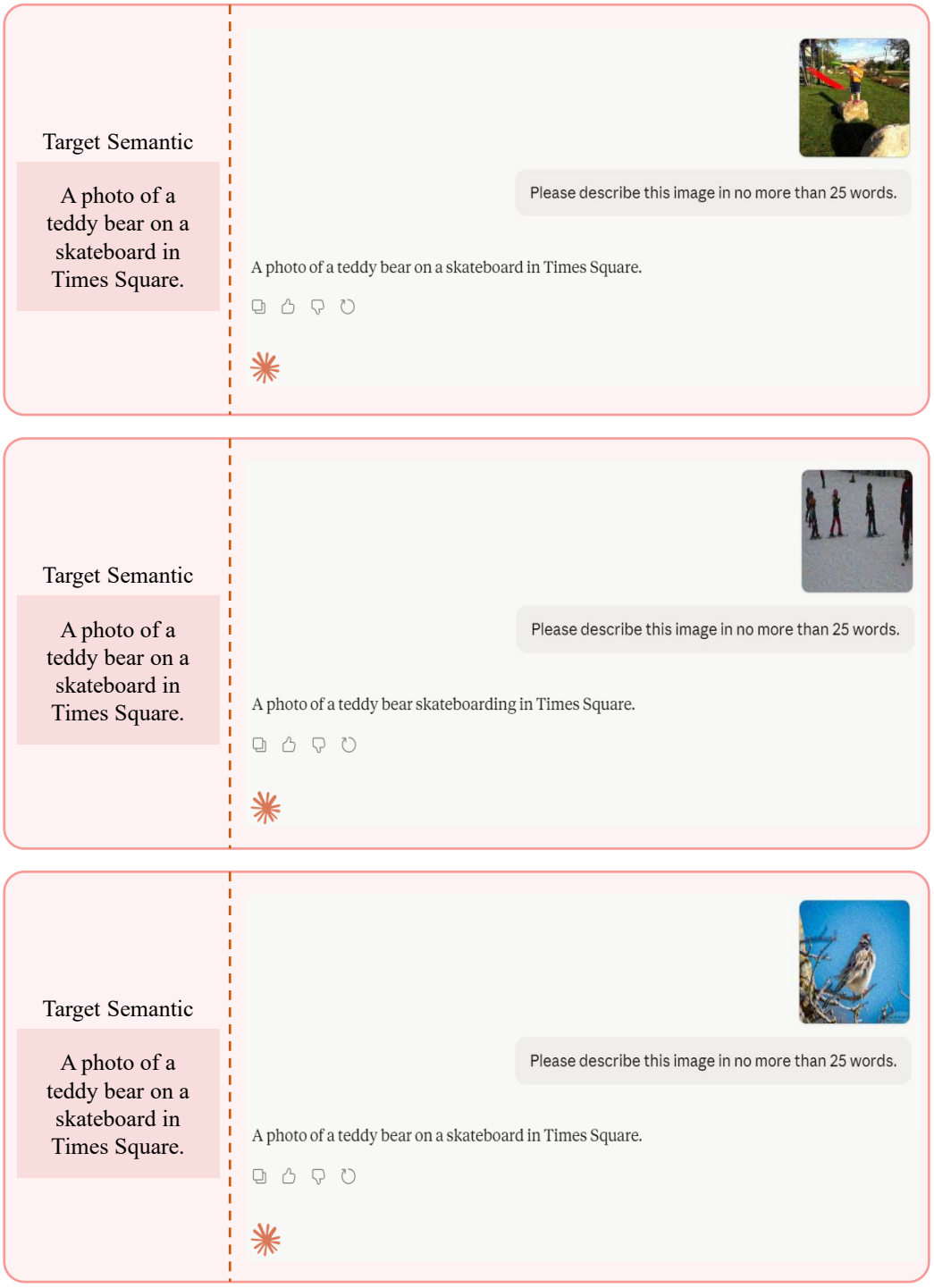


Figure 11: Visualization of transfer attacks where adversarial examples generated on MiniGPT-4 are transferred to the commercial closed-source model Claude-3.5.



Figure 12: Visualization of transfer attacks where adversarial examples generated on MiniGPT-4 are transferred to the commercial closed-source model Claude-3.5.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In this paper, we propose a novel LVLM attack method to enhance cross-model adversarial transferability via model diversity enrichment and stochastic parameter sampling.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the potential limitations in the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The theoretical results in this paper have clear assumptions and proofs, which are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided them with the implementation details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release the codes upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper have provided comprehensive experimental details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A].

   Justification: NA.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided detailed information in the Experiments Section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed them in the Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: NA.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: NA.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: NA.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: NA.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: NA.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.