Even Sailors Need Calm Seas: Taming the Geometry of VLMs for Fast Adversarial Fine-Tuning

Yuqing Wen[1,2*] **Junhao Dong**[1†] **Ting Peng**[1] **Xudong Zhang**[3] **Jiaming Zhang**[1]
Weiming Liu[4] **Zhengtao Yao**[5] **Xinghua Qu**[1] **Yew-Soon Ong**[1,6]

[1]Nanyang Technological University [2]National University of Singapore
[3]Tongji University [4]Zhejiang University
[5]University of Southern California [6]CFAR, IAIC, A*STAR

yqwen021@gmail.com, junhao003@ntu.edu.sg

Abstract

Vision-Language Models (VLMs) such as CLIP have demonstrated strong performance across tasks yet remain highly vulnerable to adversarial attacks. While adversarial fine-tuning has been demonstrated to be effective against such malicious inputs, its multi-step adversary generation scheme during fine-tuning further induces prohibitive computational costs for VLM backbones. Previous works in unimodal models have introduced single-step strategies to reduce this burden. However, we find naive single-step strategies fail on VLMs due to more severe catastrophic overfitting. In addition, we identify a novel failure mode termed *Perturbation Radius Overfitting*, where VLMs overfit to the specific training attack budget while paradoxically becoming fragile to weaker perturbations. We trace these failures to growing angular misalignment and magnitude surge of the local gradients, which compromise the local linearity essential for accurate single-step approximation. Guided by this analysis, we introduce a joint adversarial optimization scheme that actively rectifies the local geometry by suppressing both angular misalignment and magnitude surge along the perturbation path. Our approach consistently outperforms existing single-step methods across various datasets and VLM architectures, achieving state-of-the-art robustness while preserving efficiency. To our knowledge, this is the first work to systematically examine single-step adversarial fine-tuning on VLMs and establish the geometric foundations of their robustness.

1 Introduction

Vision-Language Models (VLMs) have dramatically evolved from research prototypes into foundational infrastructure for modern AI systems, powering open-vocabulary recognition [27], cross-modal retrieval [59], and serving as vision backbones for large vision-language assistants and embodied agents [4, 33, 63, 3]. Given their critical role across these domains, ensuring reliability is essential for societal trust and meeting emerging standards for AI robustness [10]. However, recent studies show that state-of-the-art VLMs and large multimodal models remain highly vulnerable to adversarial perturbations [61, 62], which can induce incorrect predictions with high confidence. A standard defense against such malicious inputs is adversarial fine-tuning, which optimizes a robust min–max objective, where the inner maximization requires iterative multi-step gradients (*e.g.*, PGD [36]) to generate perturbations. Though effective, such multi-step adversarial training incurs prohibitive computational costs [31, 60], particularly for large-scale VLMs [45].

*This work was done during Yuqing Wen's internship at Nanyang Technological University.
†Corresponding author.

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

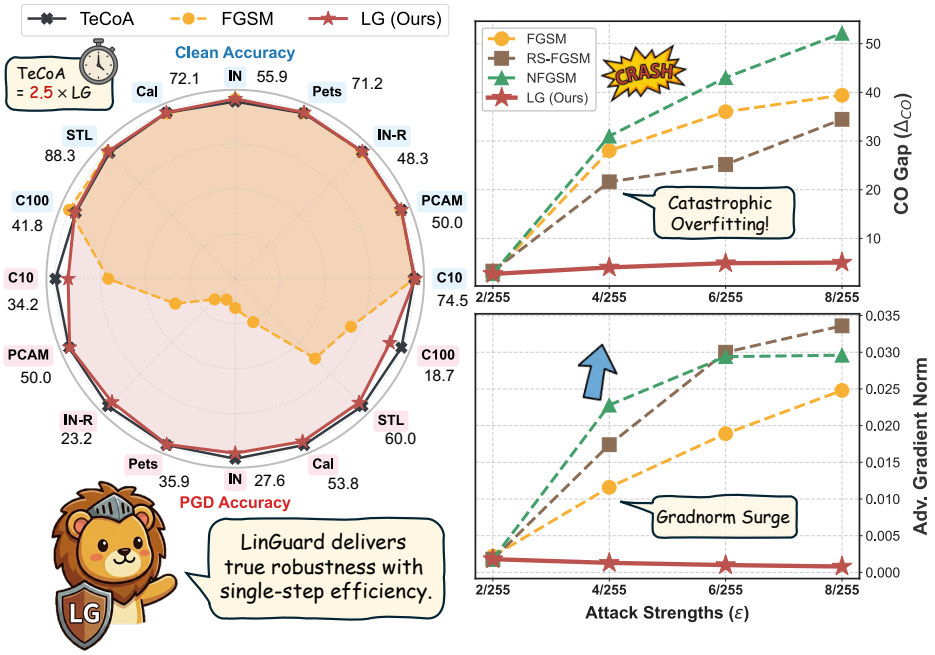


Figure 1: Illustration of CO in VLMs. LinGuard matches multi-step (TeCoA [38]) robustness while significantly reducing computational costs (Left). As the perturbation budget (ϵ) increases, existing single-step methods exhibit a widening CO Gap (top right) and a drastic surge in adversarial gradient norms (bottom right). Our method effectively mitigates CO and suppresses gradient norm growth.

To reduce the high cost induced by multi-step adversary generation, single-step alternatives [22] have been proposed. However, these methods typically suffer from Catastrophic Overfitting (CO) [57], a phenomenon where the model overfits its robustness to single-step attacks during training, leading to *a drastic collapse in defense against iterative multi-step attacks*. While mitigation strategies exist for unimodal models [51, 8], their efficacy in VLMs remains unexplored. We systematically investigate this transfer, revealing that direct application is insufficient. As shown in Figure 1, VLMs exhibit significantly more severe CO compared with single-modal (vision) models. Furthermore, we identify a novel failure mode of single-step training termed **Perturbation Radius Overfitting (PRO)** [1], where models appear robust to single-step attacks at the training perturbation budget yet remain vulnerable at smaller budgets, offering a principled perspective on the approximation failure underlying CO.

Prior studies on unimodal vision models have linked gradient properties to the generalization gap [52, 43]. To understand the root causes of vulnerability in VLMs, we present a systematic investigation into gradient implications for adversarial robustness in multimodal settings. By utilizing gradient norms to quantify local stability, we observe far more drastic gradient dynamics and severe directional divergence in VLMs. To formalize these observations, our theoretical analysis reveals an additional local curvature term intrinsic to VLMs that is absent in unimodal models, concluding that this geometric property explains why CO is significantly more severe and why the novel PRO phenomenon emerges in multimodal settings. Guided by this analysis, we introduce **Lin**earization **Guard**, termed **LinGuard**, a simple yet effective geometry-aware single-step adversarial fine-tuning framework that stabilizes adversarial gradients to enforce local linearity via a joint objective of *Angular Misalignment* and *Magnitude Surge* losses.

Extensive experiments conducted across 14 diverse zero-shot datasets demonstrate that our proposed LinGuard framework achieves state-of-the-art performance among single-step adversarial fine-tuning methods, effectively eradicating both CO and PRO while preserving competitive clean accuracy and training efficiency. Moreover, we show that the robustness induced by LinGuard can effectively transfer to a broad array of downstream vision-language tasks—including Visual Question Answering and image captioning—via a **plug-and-play** replacement of the original vision encoder with our

[1] Both theoretical and empirical analyses for PRO are provided in Section 4.3.

robustly fine-tuned version. By delivering these significant robustness gains without compromising utility, LinGuard offers a highly efficient path for securing foundational VLMs.

Our contributions are summarized as follows:

1. We present the first study of single-step adversarial fine-tuning in VLMs, revealing significantly more severe CO than unimodal models. We uncover Perturbation Radius Overfitting (PRO), where models overfit to the training budget yet remain fragile to weaker perturbations.
2. From the theoretical perspective, we trace CO and PRO to the breakdown of local linearity along the perturbation path, and attribute the VLM-specific severity to additional curvature induced by cosine-similarity classification in CLIP-style VLMs, which is absent in unimodal linear classifiers.
3. We propose LinGuard, a simple geometry-aware single-step framework combining Angular Misalignment and Magnitude Surge losses to restore local linearity throughout the ϵ-ball. Across 14 zero-shot transfer benchmarks, LinGuard achieves state-of-the-art single-step robustness, matching multi-step training at near–single-step cost while eradicating CO and PRO.

2 Related works

Vision-Language Models (VLMs) such as CLIP [45] have demonstrated remarkable performance across diverse multimodal tasks, yet they remain highly vulnerable to adversarial attacks [61, 62, 47, 10] in both unimodal and multimodal settings, raising concerns for downstream deployment. Broader robustness research spans knowledge distillation [16], few-shot learning [11], and evolutionary approaches to trustworthy AI [17]. Recent VLM studies explore weak-to-strong robustness generalization [15], multi-objective adversarial distillation [14], hierarchical robustness [18], and subspace alignment [12]. While Adversarial Training (AT) [36] stands as the primary defense paradigm, existing VLM methods predominantly rely on multi-step adversarial fine-tuning, such as TeCoA [38] and multimodal PGD-AT [48, 56]. These methods typically require K iterative attack steps, each involving a forward-backward pass, before every parameter update, with $K \geq 10$ in practice, incurring substantial computational overhead [60, 49, 57] when fine-tuning large VLM backbones on web-scale image-text corpora. In contrast, we investigate single-step adversarial fine-tuning to achieve comparable robustness at a fraction of the computational cost. Although the unimodal literature has long pursued efficient single-step alternatives that approximate the inner maximization with one gradient step [22], these methods are plagued by *Catastrophic Overfitting* (CO), where robustness against iterative attacks abruptly collapses [57] while single-step robustness remains misleadingly high. To stabilize training, prior works introduce learnable adversarial initialization [28] and random noise injection [8] to avoid overfitting to the single-step attack trajectory. Other approaches leverage prior knowledge [1] and study auxiliary regularizers, entropy gaps, and participation ratios to improve training stability [51, 39]. However, despite their effectiveness on unimodal backbones, we find that directly adapting these unimodal strategies to VLMs is insufficient. It not only exacerbates CO under the cosine-normalized cross-modal head but also triggers a novel failure mode, *Perturbation Radius Overfitting* (PRO), thereby highlighting the need for VLM-specific approaches. Geometrically, prior unimodal research links generalization to local linearity through gradient regularization [24, 19, 35, 46] and to flat minima through curvature regularization [21, 58, 40]. Violations of the linearity assumption undermine the stability of single-step approximations [1]. Building on these analyses, we investigate the geometric mechanisms underlying CO and PRO in VLMs by linking single-step linearization error to path-wise gradient variation and identifying normalization-induced excess curvature that amplifies gradient magnitude and angular drift. More comprehensive related works are in Appendix A.

3 Multimodal Single-Step Adversarial Fine-Tuning

Problem definition. We investigate efficient single-step adversarial fine-tuning for VLMs. Our work addresses two critical failure modes: Catastrophic Overfitting (CO), where robustness against iterative attacks collapses, and Perturbation Radius Overfitting (PRO), where models remain fragile to weaker attacks than the training budget. Our goal is to develop a geometry-aware optimization scheme that rectifies the local loss landscape, thereby achieving multi-step level robustness with near single-step efficiency while maintaining competitive clean accuracy.

VLM Architecture. We focus on CLIP-style Vision-Language Models [45] that align visual and textual representations in a shared embedding space. Let $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^d$ denote the image encoder parameterized by $\boldsymbol{\theta}$, and $h_{\boldsymbol{\phi}} : \mathcal{T} \to \mathbb{R}^d$ the text encoder parameterized by $\boldsymbol{\phi}$. For a C-class classification task, the prediction probability is given by the cosine similarity between the image embedding and fixed text embeddings $\mathbf{w}_c = h_{\boldsymbol{\phi}}(\mathbf{t}_c)$ derived from class prompts:

$$p(y = c|\mathbf{x}) = \frac{\exp(\langle f_{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{w}_c \rangle / \tau)}{\sum_{j=1}^{C} \exp(\langle f_{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{w}_j \rangle / \tau)}, \tag{1}$$

where τ is a learnable temperature parameter and $\langle \cdot, \cdot \rangle$ denotes the inner product.

Adversarial Fine-Tuning Objective. Following standard practice [38], we update only the image encoder and keep the text encoder *frozen*, yielding the minimax objective

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\max_{\|\boldsymbol{\delta}\|_\infty \leq \epsilon} \mathcal{L}_{ce}(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), \mathbf{W}, y) \right], \tag{2}$$

where $\mathcal{L}_{ce}$ is the cross-entropy loss, $\mathbf{W} = \{\mathbf{w}_c\}_{c=1}^{C}$ collects the fixed text embeddings, and $\boldsymbol{\delta}$ is constrained to the ℓ_∞-ball $\mathbb{B}_\infty(\mathbf{x}, \epsilon)$. The inner maximization is typically solved by multi-step Projected Gradient Descent (PGD) [36],

$$\boldsymbol{\delta}^{(k+1)} = \Pi_{\mathbb{B}(\mathbf{x},\epsilon)} \Big(\boldsymbol{\delta}^{(k)} + \alpha \cdot \operatorname{sign} \Big(\nabla_{\mathbf{x}} \mathcal{L}_{ce}(\mathbf{x} + \boldsymbol{\delta}^{(k)}, y; \boldsymbol{\theta}) \Big) \Big), \tag{3}$$

which requires $K \geq 10$ forward-backward passes per parameter update, becoming prohibitive when fine-tuning large-scale VLMs.

Single-Step Approximation. To mitigate this overhead, single-step adversarial training approximates the inner maximization with a single gradient step, typically via the Fast Gradient Sign Method (FGSM) [22], $\boldsymbol{\delta}_{\text{fgsm}} = \epsilon \cdot \operatorname{sign}(\nabla_{\mathbf{x}} \mathcal{L}_{ce}(\mathbf{x}, y; \boldsymbol{\theta}))$. This reduces the cost to nearly that of standard training, but its efficacy hinges on the local linearity of the loss landscape, whose violation produces *Catastrophic Overfitting* (CO) [57]. However, our experiments reveal that directly adapting these single-step recipes to VLMs aggravates CO and triggers the previously undocumented PRO.

Insight. Unlike unimodal classifiers, VLMs' cross-modal alignment induces a more complex loss landscape where local linearity is harder to preserve, heuristically explaining the emergence of CO and PRO. Section 4 formalizes the geometric source, and Section 5 translates it into a remedy.

4 Theoretical Analysis

4.1 From the Empirical CO Gap to the Optimization Gap of FGSM

Catastrophic overfitting is the empirical discrepancy between robustness under the single-step training attack and that under a stronger multi-step adversary. At the dataset level, $\Delta_{\text{CO}} := \mathcal{A}_{\text{FGSM}} - \mathcal{A}_{\text{multi}}$, where $\mathcal{A}_{\text{FGSM}}$ and $\mathcal{A}_{\text{multi}}$ are the robust accuracies under FGSM and under a strong iterative attack (*e.g.*, PGD or AutoAttack), respectively.

As an accuracy-level quantity, Δ_{CO} is not directly amenable to analysis. We pass to its loss-level counterpart. With $\boldsymbol{\delta}_{\text{fgsm}}$ defined in Section 3 and the true inner maximizer $\boldsymbol{\delta}^\star(\mathbf{x}) \in \arg\max_{\|\boldsymbol{\delta}\|_\infty \leq \epsilon} \ell(\mathbf{x} + \boldsymbol{\delta}, y)$, we define the per-example **adversarial objective gap**

$$\Delta_{\text{adv}}(\mathbf{x}, \epsilon) := \ell(\mathbf{x} + \boldsymbol{\delta}^\star, y) - \ell(\mathbf{x} + \boldsymbol{\delta}_{\text{fgsm}}, y) \;\geq\; 0, \tag{4}$$

which measures the suboptimality of FGSM in solving the inner maximization and is the loss-level source of Δ_{CO}. However, Δ_{adv} remains intractable as it involves the true maximizer $\boldsymbol{\delta}^\star$. We therefore replace it by a quantity depending only on the loss surface, the local **linearization error**

$$E_{\text{lin}}(\mathbf{x}, \epsilon) := \sup_{\|\boldsymbol{\delta}\|_\infty \leq \epsilon} \left| \ell(\mathbf{x} + \boldsymbol{\delta}, y) - \ell(\mathbf{x}, y) - \nabla_{\mathbf{x}} \ell(\mathbf{x}, y)^\top \boldsymbol{\delta} \right|, \tag{5}$$

the worst-case deviation of ℓ from its first-order Taylor expansion on $\mathbb{B}_\infty(\mathbf{x}, \epsilon)$. Unlike Δ_{adv}, E_{lin} is intrinsic to the loss surface and independent of any specific attack.

Theorem 4.1 (Linearization error controls FGSM)**.** *If $\ell(\cdot, y)$ is differentiable on $\mathbb{B}_\infty(\mathbf{x}, \epsilon)$, then*

$$0 \le \Delta_{\text{adv}}(\mathbf{x}, \epsilon) \le 2\, E_{\text{lin}}(\mathbf{x}, \epsilon). \tag{6}$$

(Proof in Appendix J.)

Remark. Up to a factor of two, the optimization-level suboptimality of FGSM is fully captured by a purely geometric property of the loss surface: how much ℓ deviates from its first-order Taylor expansion inside the perturbation ball.

4.2 Magnitude Surge and Angular Misalignment as Hessian Projections

By Theorem 4.1, FGSM suboptimality reduces to controlling E_{lin}. We bound E_{lin} by geometric quantities of the loss surface, beginning with the Hessian.

Theorem 4.2 (Curvature-induced bound)**.** *Let $\nu(\mathbf{x}, \epsilon) := \sup_{\|\boldsymbol{\eta}\|_\infty \le \epsilon} \|H_{\mathbf{x}} \ell(\mathbf{x} + \boldsymbol{\eta}, y)\|_2$. If $\ell(\cdot, y)$ is twice differentiable on $\mathbb{B}_\infty(\mathbf{x}, \epsilon)$, then*

$$E_{\text{lin}}(\mathbf{x}, \epsilon) \;\le\; \tfrac{1}{2}\, d\, \epsilon^2\, \nu(\mathbf{x}, \epsilon), \tag{7}$$

where d is the input dimension. (Proof in Appendix J.)

The bound is tight in derivation but the Hessian itself is hard to optimize directly. We therefore replace it by an easily computable proxy, the path-wise gradient variation $V(\mathbf{x}, \boldsymbol{\eta}) := \|\nabla_{\mathbf{x}} \ell(\mathbf{x} + \boldsymbol{\eta}, y) - \nabla_{\mathbf{x}} \ell(\mathbf{x}, y)\|_2$.

Theorem 4.3 (Path-wise gradient-variation bound)**.** *For any $\boldsymbol{\delta}$,*

$$\ell(\mathbf{x} + \boldsymbol{\delta}, y) - \ell(\mathbf{x}, y) - \nabla_{\mathbf{x}} \ell(\mathbf{x}, y)^\top \boldsymbol{\delta} \;=\; \int_0^1 \big(\nabla_{\mathbf{x}} \ell(\mathbf{x} + t\boldsymbol{\delta}, y) - \nabla_{\mathbf{x}} \ell(\mathbf{x}, y)\big)^\top \boldsymbol{\delta}\, dt, \tag{8}$$

and consequently $E_{\text{lin}}(\mathbf{x}, \epsilon) \;\le\; \sqrt{d}\, \epsilon \cdot \sup_{\|\boldsymbol{\eta}\|_\infty \le \epsilon} V(\mathbf{x}, \boldsymbol{\eta})$. (Proof in Appendix J.)

The integral form shows E_{lin} is governed by the entire interior path of $\mathbb{B}_\infty(\mathbf{x}, \epsilon)$, not just a boundary point. We next decompose V to identify what makes it large and how curvature generates it.

Proposition 4.4 (Geometric structure of V)**.** *Let $\mathbf{g} = \nabla_{\mathbf{x}} \ell(\mathbf{x}, y)$, $\mathbf{g}' = \nabla_{\mathbf{x}} \ell(\mathbf{x} + \boldsymbol{\eta}, y)$, and define the path $\gamma(t) = \mathbf{x} + t\boldsymbol{\delta}$ with $\mathbf{g}(t) = \nabla_{\mathbf{x}} \ell(\gamma(t), y)$ and $\mathbf{u}(t) = \mathbf{g}(t)/\|\mathbf{g}(t)\|_2$ (assuming $\mathbf{g}(t) \neq 0$). Then*

$$V(\mathbf{x}, \boldsymbol{\eta})^2 \;=\; \underbrace{\big(\|\mathbf{g}'\|_2 - \|\mathbf{g}\|_2\big)^2}_{\textit{magnitude surge}} \;+\; \underbrace{2\|\mathbf{g}'\|_2 \|\mathbf{g}\|_2 \big(1 - \cos(\mathbf{g}, \mathbf{g}')\big)}_{\textit{angular misalignment}}, \tag{9}$$

$$\tfrac{d}{dt}\|\mathbf{g}(t)\|_2 = \mathbf{u}(t)^\top H_{\mathbf{x}} \ell(\gamma(t), y)\, \boldsymbol{\delta}, \qquad \tfrac{d}{dt}\mathbf{u}(t) = \frac{\big(I - \mathbf{u}(t)\mathbf{u}(t)^\top\big)\, H_{\mathbf{x}} \ell(\gamma(t), y)\, \boldsymbol{\delta}}{\|\mathbf{g}(t)\|_2}. \tag{10}$$

(Proof in Appendix J.)

Equation 9 decomposes V into magnitude and angular components. Equation 10 shows that both are generated by the same Hessian along the perturbation path: the parallel projection $\mathbf{u}^\top H \boldsymbol{\delta}$ drives magnitude growth, while the orthogonal projection $(I - \mathbf{u}\mathbf{u}^\top) H \boldsymbol{\delta}$ drives directional drift.

Empirically, the magnitude surge $|\|\mathbf{g}'\|_2 - \|\mathbf{g}\|_2|$ manifests as the inflation of the adversarial input-gradient norm $\|\mathbf{g}'\|_2$ that accompanies CO. Through the chain $\Delta_{\text{CO}} \to \Delta_{\text{adv}} \to E_{\text{lin}} \to V$, our analysis pinpoints this gradient-norm explosion as a geometric driver of CO, directly visualized in Figure 1 and reported as Adv. GN in Section 6.2.

Remark. V is the path-wise control quantity for E_{lin} (Theorem 4.3). Its magnitude and angular components are the parallel and orthogonal projections of the same Hessian along the adv. path (Proposition 4.4). Suppressing V requires controlling both, and the same curvature governs both.

VLM-specific excess curvature. The cosine-normalized head induces a structurally non-zero curvature term $R_{\text{norm}} \sim 1/(\tau\|\mathbf{z}\|_2^2)$ that is identically zero for linear heads. This term inflates V above its unimodal counterpart, mechanistically explaining why CO is more severe on VLMs. Detailed Hessian-level derivations are deferred to Appendix H.

4.3 PRO as an Interior-Path Failure

Beyond CO, we identify a failure mode previously undocumented in single-step studies on unimodal classifiers [57, 1, 8]. As shown in Figure 2, single-step VLMs paradoxically lose robustness at interior radii $\epsilon' < \epsilon_{\text{train}}$, violating the natural expectation that defense at ϵ_{train} implies defense at any smaller ϵ'. We term this anomaly **Perturbation Radius Overfitting (PRO)**: there exists $\epsilon' < \epsilon_{\text{train}}$ such that

$$\mathbb{E}[\ell(\mathbf{x}+\boldsymbol{\delta}_{\text{ss}}(\epsilon'), y)] > \mathbb{E}[\ell(\mathbf{x}+\boldsymbol{\delta}_{\text{ss}}(\epsilon_{\text{train}}), y)] + \gamma, \gamma > 0. \quad (11)$$

PRO admits a geometric interpretation: along any perturbation ray, the loss is non-decreasing whenever directional alignment is positive and gradient magnitude is bounded away from zero (Proposition H.3 in Appendix H). When either condition fails inside the ball, the loss can rise at an intermediate radius. Linear heads have $R \equiv 0$, whereas the cosine-head excess curvature $R_{\text{norm}} \sim 1/(\tau\|\mathbf{z}\|_2^2)$ inflates both magnitude growth and rotation rate, promoting violations of the monotonicity condition.

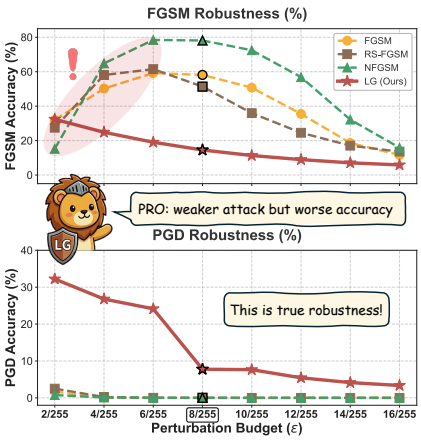


Figure 2: **Illustration of PRO.** Robust accuracy of single-step VLMs fine-tuned at $\epsilon_{\text{train}} = 8/255$ across diverse perturbation radii during evaluation.

A feasible interior perturbation with higher loss certifies the suboptimality of the single-step training attack. This exposes a failure of the single-step approximation used during training.

Remark. PRO is the failure of an interior-path monotonicity condition. The excess curvature R_{norm} induced by cosine normalization promotes this failure by inflating both magnitude growth and directional rotation. This also offers a principled perspective on how single-step approximation errors can undermine adversarial training and contribute to CO.

5 LinGuard: Simple yet Effective Geometry-Aware Fine-Tuning

Section 4 reduces FGSM suboptimality to controlling the path-wise gradient variation V on $\mathbb{B}_\infty(\mathbf{x}, \epsilon)$, with V^2 admitting an angular/magnitude decomposition driven by the input-space Hessian along the perturbation path (Theorem 4.3, Proposition 4.4). Guided by this decomposition, our **LinGuard** framework adds two lightweight regularizers to single-step adversarial fine-tuning, one per component, on a shared interior sample $\boldsymbol{\eta} \sim \mathcal{U}(\mathbb{B}_\infty(\epsilon))$—a Monte-Carlo proxy for $\sup_\eta V$—at minimal cost beyond FGSM. Throughout this section, write $\mathbf{g} = \nabla_{\mathbf{x}}\ell(\mathbf{x}, y; \boldsymbol{\theta})$ and $\mathbf{g}_{\boldsymbol{\eta}} = \nabla_{\mathbf{x}}\ell(\mathbf{x}+\boldsymbol{\eta}, y; \boldsymbol{\theta})$.

5.1 Angular Regularizer

The angular component of V^2 in Proposition 4.4 is $2\|\mathbf{g}_{\boldsymbol{\eta}}\|_2\|\mathbf{g}\|_2\big(1 - \cos(\mathbf{g}, \mathbf{g}_{\boldsymbol{\eta}})\big)$. We penalize this term directly through the **Angular Misalignment (AM) loss**

$$\mathcal{L}_{AM}(\mathbf{x}, y; \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{U}(\mathbb{B}_\infty(\epsilon))}\big[1 - \cos(\mathbf{g}, \mathbf{g}_{\boldsymbol{\eta}})\big]. \quad (12)$$

$\mathcal{L}_{AM}$ adopts the gradient-alignment idea introduced for unimodal fast adversarial training by GradAlign [1], instantiated here as the angular surrogate of the path-wise variation V on VLMs. Acting on an interior sample is essential: Theorem 4.3 shows that E_{lin} is governed by gradient variation along the entire interior path, and Proposition H.3 characterizes PRO as a violation of an interior directional condition. By enforcing this condition along the perturbation path, $\mathcal{L}_{AM}$ addresses CO and PRO simultaneously.

5.2 Magnitude Regularizer

The magnitude component $(\|\mathbf{g}_{\boldsymbol{\eta}}\|_2 - \|\mathbf{g}\|_2)^2$ is generated by the path integral $\int_0^1 \mathbf{u}(t)^\top H_{\mathbf{x}}\ell(\gamma(t), y)\,\boldsymbol{\delta}\,dt$ from Equation 10, where the Hessian acts along the current ascent direction $\mathbf{u}(t)$. We attack this term at its local source: setting the probing direction $\mathbf{v} = \mathbf{g}_{\boldsymbol{\eta}}/\|\mathbf{g}_{\boldsymbol{\eta}}\|_2$, the **Magnitude Surge (MS) loss** is

$$\mathcal{L}_{MS}(\mathbf{x}, y; \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\eta}}\Big[\big\|\nabla_{\mathbf{x}}\ell(\mathbf{x}+\boldsymbol{\eta}+h\mathbf{v}, y; \boldsymbol{\theta}) - \mathbf{g}_{\boldsymbol{\eta}}\big\|_2^2\Big], \quad (13)$$

Table 1: **Zero-shot Performance.** We report the detailed Robust accuracy under PGD-20 attacks ($\epsilon = 4/255$) and the average Clean accuracy across 14 datasets. Our method achieves robustness comparable to multi-step TeCoA while maintaining competitive clean average accuracy. Notably, the multi-step TeCoA is approximately **2.5 × slower** than our method in terms of training time.

	Robust Accuracy (PGD-20, $\epsilon = 4/255$)															Clean Acc
Method	ImageNet	STL10	CIFAR-10	CIFAR-100	ImageNet-R	ImageNet-S	Stanf.Cars	OxfordPet	Flower102	DTD	EuroSAT	FGVC	PCAM	Caltech101	**Avg.**	**Avg.**
TeCoA (Multi-step)	27.56	60.03	34.24	18.73	23.23	14.80	3.97	35.92	9.16	12.87	11.67	0.84	49.98	53.79	25.48	44.58
FGSM	4.48	37.65	24.17	13.05	3.77	1.44	0.27	4.50	2.49	4.41	0.02	0.00	18.04	14.00	9.16	44.90
RS-FGSM	9.34	45.80	23.02	12.20	4.74	1.38	1.18	12.40	4.08	9.10	0.13	0.00	22.41	20.23	11.86	**47.72**
NFGSM	4.16	37.85	19.99	9.76	2.97	1.19	0.19	3.11	2.11	4.79	0.00	0.00	9.49	12.77	7.74	45.82
LinGuard	**26.72**	**58.45**	**31.82**	**17.44**	**22.52**	**14.79**	**3.95**	**35.79**	**9.84**	**12.61**	**7.24**	**0.87**	**49.81**	**52.65**	**24.61**	45.35

with finite-difference step $h > 0$. Taylor expansion yields $\mathcal{L}_{MS}/h^2 \to \|H_{\mathbf{x}}\ell(\mathbf{x}+\boldsymbol{\eta}, y)\,\mathbf{v}\|_2^2$, so $\mathcal{L}_{MS}$ is a finite-difference surrogate for the Hessian action along the locally attack-relevant direction [40]. Probing at $\mathbf{x}+\boldsymbol{\eta}$ rather than $\mathbf{x}$ is necessary because $\mathbf{u}(t)$ has already rotated inside the perturbation ball, so the clean-input direction $\mathbf{g}/\|\mathbf{g}\|_2$ is no longer the relevant axis.

5.3 Objective Function

LinGuard combines the single-step adversarial classification loss with the two regularizers:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{ce}(\mathbf{x}+\boldsymbol{\delta}_{\text{ss}}, y; \boldsymbol{\theta}) + \lambda_{AM}\mathcal{L}_{AM} + \lambda_{MS}\mathcal{L}_{MS}, \tag{14}$$

where $\boldsymbol{\delta}_{\text{ss}}$ is a generic single-step perturbation (*e.g.*, FGSM or its variants) and $\lambda_{AM}, \lambda_{MS}$ are regularization coefficients.

The two regularizers are not interchangeable. By Proposition 4.4, alignment alone does not bound the magnitude surge: setting $\mathbf{g}_{\boldsymbol{\eta}} = \alpha\mathbf{g}$ yields $V = |\alpha - 1|\|\mathbf{g}\|_2$ even when $\cos(\mathbf{g}, \mathbf{g}_{\boldsymbol{\eta}}) = 1$. Conversely, bounding the Hessian at a single point does not control endpoint alignment when gradients drift along the path. The two terms of V^2 are linked by the integrated magnitude identity

$$\big|\|\mathbf{g}_{\boldsymbol{\eta}}\|_2 - \|\mathbf{g}\|_2\big| \;\le\; \int_0^1 \big\|H_{\mathbf{x}}\ell(\mathbf{x}+t\boldsymbol{\eta}, y)\,\boldsymbol{\eta}\big\|_2\, dt, \tag{15}$$

which makes the division of labour explicit: $\mathcal{L}_{MS}$ suppresses Hessian action at its local source, while $\mathcal{L}_{AM}$ suppresses the directional drift accumulated along the interior path. Sharing $\boldsymbol{\eta}$ couples both regularizers on a common local witness of V and reuses $\mathbf{g}_{\boldsymbol{\eta}}$ as $\mathbf{v}$, avoiding an extra backward pass.

For cosine-normalized VLMs, Proposition H.2 further shows that the orthogonal action $(I - \mathbf{u}\mathbf{u}^\top)R_{\text{norm}}\boldsymbol{\delta}$ inflates the rotation rate of the ascent direction with operator norm $\sim 1/(\tau\|\mathbf{z}\|_2^2)$, so the angular term dominates V^2 in the regimes where single-step training fails, namely small temperature τ, small feature norm, and hard examples.

6 Experiments

6.1 Experimental Setup

Datasets. Following [38, 56], we fine-tune CLIP on the ImageNet training set [9] and evaluate on 14 zero-shot classification datasets across diverse scenarios, with extensions to vision-language understanding tasks including VQA and Image Captioning. More details are in Appendix C.

Implementation details. We adopt CLIP-ViT-B/32 [45] as the main backbone and ViT-L/14 for understanding tasks. Adversarial fine-tuning uses the ℓ_∞-norm threat model with $\epsilon = 4/255$ and step size $\alpha = \epsilon$, optimized by AdamW with cosine annealing for 2 epochs and regularization weights $\lambda_{AM} = 0.4$, $\lambda_{MS} = 1.2$. We compare **LinGuard** against the multi-step TeCoA [38] and three single-step baselines: FGSM [22], RS-FGSM [57], and NFGSM [8]. We report robust accuracy under PGD-20 ($\alpha = \epsilon/4$) and AutoAttack (AA) [7]. More details are in Appendix C.

6.2 Main Results

Evaluations across 14 datasets. Table 1 reports zero-shot performance across 14 datasets. Standard single-step methods suffer from severe Catastrophic Overfitting (CO) in the VLM setting, yielding

near-random robust accuracy ($< 10\%$). In contrast, LinGuard recovers robustness (24.61%) comparable to the multi-step TeCoA (25.48%) while achieving higher average clean accuracy (**45.35%** vs. 44.58%) at significantly lower training cost, demonstrating that our geometric constraints close the robustness gap between single-step and multi-step optimization.

Evaluation on Diverse Radii and Geometric Analysis. Table 2 presents the performance breakdown under varying perturbation radii ϵ, detailing accuracy against multi-step attacks. Beyond standard metrics, we explicitly calculate the **CO Gap**, defined as the difference between single-step and multi-step robust accuracy ($\Delta_{\text{CO}} = \mathcal{A}_{\text{FGSM}} - \mathcal{A}_{\text{AA}}$). To further verify our theoretical analysis, we also report the **Adversarial Gradient Norm** (Adv. GN), calculated as the ℓ_2-norm of input gradients at PGD-20 adversarial points, serving as an empirical proxy for the magnitude surge identified in Section 4.2. As evidenced in Table 2, at a low perturbation radius ($\epsilon = 2/255$), the CO phenomenon is not pronounced, with all methods exhibiting minimal CO Gaps and comparable performance. However, as ϵ increases, traditional single-step adversarial fine-tuning methods suffer from increasingly severe CO. For instance, the CO Gap for NFGSM explodes to 43.02% at $\epsilon = 6/255$. Crucially, this degradation corresponds to a drastic surge in gradient norms (rising to 0.0294), confirming the magnitude surge predicted by our analysis and indicating a violation of the linear approximation. In contrast, our LinGuard effectively suppresses the growth of gradient norms, thereby preserving local linearity and sustaining high resistance to multi-step attacks. Consequently, LinGuard achieves state-of-the-art performance, maintaining a high multi-step robust accuracy even under strong perturbations (*e.g.*, **18.17%** at $\epsilon = 6/255$) where baseline methods collapse.

Table 2: Comparison of our method with other adv. fine-tuning approaches under **diverse ϵ**. Robust accuracies are averaged across the 14 zero-shot classification datasets.

Radius	Method	PGD-20	AA	CO Gap	Adv. GN
$\epsilon = 2/255$	FGSM	33.53	32.11	2.74	0.0022
	RS-FGSM	31.50	30.21	3.27	0.0018
	NFGSM	33.55	32.13	2.74	0.0018
	LinGuard	**33.65**	**32.23**	**2.69**	**0.0018**
$\epsilon = 4/255$	FGSM	9.16	4.80	28.00	0.0116
	RS-FGSM	11.86	8.55	21.61	0.0174
	NFGSM	7.74	4.52	31.00	0.0228
	LinGuard	**24.61**	**22.94**	**4.03**	**0.0013**
$\epsilon = 6/255$	FGSM	1.62	0.79	36.02	0.0189
	RS-FGSM	3.32	1.54	25.17	0.0300
	NFGSM	0.59	0.21	43.02	0.0294
	LinGuard	**18.17**	**16.24**	**4.90**	**0.0010**
$\epsilon = 8/255$	FGSM	0.10	0.02	39.48	0.0248
	RS-FGSM	0.13	0.04	34.55	0.0336
	NFGSM	0.04	0.01	52.17	0.0296
	LinGuard	**13.96**	**12.18**	**5.02**	**0.0008**

Transferability across diverse CLIP architectures. We further evaluate the generalizability of LinGuard across three CLIP backbones spanning transformer and convolutional families. Table 3 reports ImageNet robustness at $\epsilon = 4/255$ for ViT-B/32, ViT-L/14, and ResNet-50 under PGD-20 [36] and AA [7]. Across all three architectures, **LinGuard** achieves the strongest PGD-20 and AA robustness while maintaining competitive clean accuracy. On ViT-B/32, LinGuard reaches **26.72%** PGD-20 and **23.90%** AA, substantially outperforming the strongest single-step baseline, RS-FGSM, which obtains 9.34% PGD-20 and 6.00% AA, while FGSM and NFGSM collapse below 5%. On ViT-L/14, LinGuard achieves **47.82%** PGD-20 and **45.70%** AA, whereas FGSM and NFGSM exhibit near-zero robustness. On ResNet-50, LinGuard still leads with **8.08%** PGD-20 and **4.02%** AA, while all baselines collapse below 2%. Note that *the smaller margin is likely due to the stronger influence of the convolutional encoder's input-space curvature term* $\nabla^2_{\mathbf{x}} f_k$ in Eq. (16), which lies outside the cosine-head normalization effect targeted by R_{norm}. The consistent improvements across transformer- and convolution-based encoders further confirm that the geometric

Table 3: **Transferability across CLIP architectures.** Accuracy (%) on the ImageNet test set under PGD-20 and AutoAttack (AA) at $\epsilon = 4/255$.

Architecture	Method	Clean	PGD-20	AA
ViT-B/32	FGSM	55.82	4.48	1.38
	RS-FGSM	**60.16**	9.34	6.00
	NFGSM	56.18	4.16	1.70
	LinGuard	55.88	**26.72**	**23.90**
ViT-L/14	FGSM	74.32	0.32	0.00
	RS-FGSM	**79.18**	42.44	30.44
	NFGSM	76.62	14.82	4.20
	LinGuard	75.44	**47.82**	**45.70**
ResNet-50	FGSM	53.82	1.86	0.98
	RS-FGSM	53.32	1.94	0.80
	NFGSM	52.30	0.78	0.80
	LinGuard	**56.06**	**8.08**	**4.02**



Figure 3: **Image captioning examples with LLaVA 1.5.** We compare vision encoders from different methods on clean and adversarial inputs. Each example shows the target image (left), captions generated from clean inputs (middle), and captions generated under PGD-50 attacks (right).

mechanism formalized in Section 4—suppressing magnitude surge and angular misalignment along the perturbation path—is architecture-agnostic.

Extension to Generative Tasks (VQA & Captioning). To demonstrate the versatility and scalability of our proposed method, we extend our evaluation to Large Vision-Language Models (LVLMs) across four standard benchmarks: COCO [32] and Flickr30k [44] for image captioning (evaluated by CIDEr), and TextVQA [50] and VQAv2 [23] for visual question answering (evaluated by Accuracy). We compare LinGuard against the standard CLIP model and three representative single-step adversarial fine-tuning baselines (FGSM, RS-FGSM, and NFGSM), training with a perturbation budget of $\epsilon = 4/255$. We report performance on both clean data and under strong attacks at $\epsilon = 4/255$. As shown in Table 4, the standard CLIP model achieves high clean performance but lacks any robustness. Among the baselines, FGSM suffers from severe CO, collapsing to near-zero robustness similar to standard CLIP. While RS-FGSM and NFGSM offer some improvement, they still struggle to maintain stability under multi-step attacks. In contrast, LinGuard demonstrates superior stability across all datasets, achieving the highest robust performance (Average 19.47), outperforming the baselines. This indicates that our geometric constraints effectively generalize to the multimodal generation paradigm, preserving the quality of generated text even under intense adversarial perturbations.

Table 4: **Robustness Evaluation on Generative Tasks.** Comparison of **LinGuard** with single-step baselines on VQA & Captioning tasks under Clean and PGD-50 protocols. **Bold**/underline denote best/second-best results.

Method	COCO		Flickr30k		TextVQA		VQAv2		Average	
	Cln.	Rob.	Cln.	Rob.	Cln.	Rob.	Cln.	Rob.	Cln.	Rob.
CLIP	117.19	6.45	78.91	0.96	37.44	0.84	72.78	12.92	76.58	5.29
FGSM	**91.90**	5.56	**53.21**	0.99	17.32	1.84	**63.64**	13.76	**56.52**	5.54
RS-FGSM	87.29	21.25	52.02	9.67	17.88	6.42	61.40	23.88	54.65	15.31
NFGSM	87.77	17.19	49.56	5.17	15.76	4.34	62.00	23.14	53.77	12.46
LinGuard	88.80	**28.41**	51.19	**14.40**	**18.66**	**7.00**	62.78	**28.08**	55.36	**19.47**

Qualitative Visualization. Figure 3 presents image captioning examples under PGD-50 attacks. Single-step baselines exhibit semantic drift, describing a running dog as "a dead bird" or a skiing

scene as "people on a beach". LinGuard better preserves the main scene content under attack. Additional qualitative examples are provided in Appendix D.

6.3 Ablation Study

Hyperparameter sensitivity. We further analyze LinGuard's components via independent ablation studies, isolating the Angular Misalignment (λ_{AM}) and Magnitude Surge (λ_{MS}) losses as shown in Figure 4.

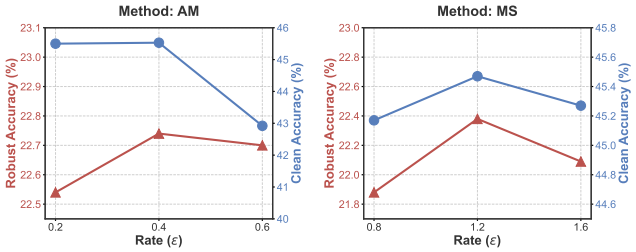


Figure 4: **Ablation Study on Hyperparameters. Left:** Impact of the Angular Misalignment loss (λ_{AM}) with fixed $\lambda_{MS} = 0$. **Right:** Impact of the Magnitude Surge loss (λ_{MS}) with fixed $\lambda_{AM} = 0$. The plots utilize dual y-axes to show the trade-off between Robust Accuracy (Red, Left Axis) and Clean Accuracy (Blue, Right Axis).

As shown in Figure 4 (Left), increasing λ_{AM} initially leads to an improvement in Robust Accuracy (AA) while maintaining Clean Accuracy. However, a larger λ_{AM} (*e.g.*, 0.6) introduces a slight trade-off in clean accuracy, suggesting that excessive alignment constraints may hinder benign feature learning. Thus, a moderate strength is sufficient to reduce the CO Gap while preserving clean performance. Similarly, Figure 4 (Right) shows robustness peaks at $\lambda_{MS} = 1.2$. Consequently, we adopt the combination of $\lambda_{AM} = 0.4$ and $\lambda_{MS} = 1.2$ to maximize robustness while minimizing the trade-off in clean generalization.

Optimization collapse under extreme regularization. While the visualization focuses on the effective range, extreme regularization strength (*e.g.*, $\lambda_{MS} = 2.0$) can trigger an optimization collapse. As shown in Figure 5, the normal regime ($\lambda = 1.2$) yields healthy convergence with both training clean and robust accuracies improving, whereas the collapse regime ($\lambda = 2.0$) exhibits a pathological failure: a clear crossover occurs around step 2500 where training clean accuracy drops to random chance ($\sim$10%) while training robust accuracy continues to rise. This divergence reveals that excessive regularization dominates the loss landscape, with the optimizer satisfying the geometric constraints by sacrificing semantic information. The apparent surge in robust accuracy further stems from gradient obfuscation, as the over-suppressed gradients prevent the inner maximization from locating valid adversarial directions. A balanced regularization strength is therefore critical for the optimal trade-off between geometric flatness and semantic classification. Full ablation tables are provided in Appendix E.

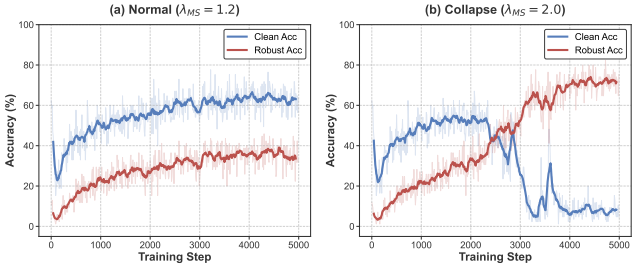


Figure 5: **Training dynamics under different regularization strengths.** The plots display smoothed **Training Clean** and **Training Robust** accuracy (100-step moving average) alongside raw logs. **(a)** Normal setting ($\lambda = 1.2$). **(b)** Extreme setting ($\lambda = 2.0$).

7 Conclusion

We uncover that severe *Catastrophic Overfitting* and *Perturbation Radius Overfitting* in single-step VLM fine-tuning stem from a breakdown of local linearity in high-dimensional landscapes. Specifically, sharp gradient surges and directional misalignment cause single-step updates to diverge from the worst-case adversaries. To address this, we propose a geometry-aware framework that rectifies the loss surface by suppressing both angular misalignment and magnitude surge along the perturbation path. Our approach effectively eradicates these failures, achieving robustness comparable to expensive multi-step methods while maintaining the efficiency critical for foundation models. Future work will explore advanced gradient calculation acceleration strategies to improve efficiency.

Acknowledgements

This research is supported by the MTI under its AI Centre of Excellence for Manufacturing (AIMfg) (Award W25MCMF014), the A*STAR-NTU Research Joint Lab for Smart and Sustainable Advanced Manufacturing, partly by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative, the Centre for Frontier Artificial Intelligence Research, Institute of Advanced Intelligence and Computing, A*Star, and the College of Computing and Data Science at Nanyang Technological University. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, and Infocomm Media Development Authority.

References

[1] Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33:16048–16059, 2020.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[6] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[7] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[8] Pau de Jorge Aranda, Adel Bibi, Riccardo Volpi, Amartya Sanyal, Philip Torr, Grégory Rogez, and Puneet Dokania. Make some noise: Reliable and efficient single-step adversarial training. *Advances in Neural Information Processing Systems*, 35:12881–12893, 2022.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, and Francisco Herrera. Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation. *Information Fusion*, 99:101896, 2023.

[11] Junhao Dong, Yuan Wang, Xiaohua Xie, Jianhuang Lai, and Yew-Soon Ong. Generalizable and discriminative representations for adversarially robust few-shot learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):5480–5493, 2024.

[12] Junhao Dong, Piotr Koniusz, Liaoyuan Feng, Yifei Zhang, Hao Zhu, Weiming Liu, Xinghua Qu, and Yew-Soon Ong. Robustifying zero-shot vision language models by subspaces alignment. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21037–21047. IEEE, 2025.

[13] 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.

[14] Junhao Dong, Jiao Liu, Xinghua Qu, and Yew-Soon Ong. Confound from all sides, distill with resilience: Multi-objective adversarial paths to zero-shot robustness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 624–634, 2025.

[15] Junhao Dong, Cong Zhang, Xinghua Qu, Zejun Ma, Piotr Koniusz, and Yew Soon Ong. Robust superalignment: Weak-to-strong robustness generalization for vision-language models. *Advances in Neural Information Processing Systems*, 38:18345–18377, 2025.

[16] Junhao Dong, Raoof Zare Moayedi, Yew-Soon Ong, and Seyed-Mohsen Moosavi-Dezfooli. Allies teach better than enemies: Inverse adversaries for robust knowledge distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[17] Junhao Dong, Chenkai Wang, Xuanhui Lin, Mingrong Gong, Siyu Wang, Yuqing Wen, Jiao Liu, Catherine Huang, Gary G Yen, Xin Yao, et al. Evolutionary computation for trustworthy ai: From attacks and defenses to self-evolving era. *arXiv preprint arXiv:2610.02996*, 2026.

[18] Junhao Dong, Yifei Zhang, Hao Zhu, Yew-Soon Ong, and Piotr Koniusz. Hierarchically robust zero-shot vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 37642–37652, June 2026.

[19] Harris Drucker and Yann Le Cun. Improving generalization performance using double back-propagation. *IEEE transactions on neural networks*, 3(6):991–997, 1992.

[20] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[21] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

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

[23] 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.

[24] Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

[25] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[26] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

[27] Xiaohu Huang, Hao Zhou, Kun Yao, and Kai Han. Froster: Frozen clip is a strong teacher for open-vocabulary action recognition. *arXiv preprint arXiv:2402.03241*, 2024.

[28] Xiaojun Jia, Yong Zhang, Baoyuan Wu, Jue Wang, and Xiaochun Cao. Boosting fast adversarial training with learnable adversarial initialization. *IEEE Transactions on Image Processing*, 31: 4417–4430, 2022.

[29] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[30] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[31] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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

[35] Chunchuan Lyu, Kaizhu Huang, and Hai-Ning Liang. A unified gradient regularization family for adversarial examples. In *2015 IEEE international conference on data mining*, pages 301–309. IEEE, 2015.

[36] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[37] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[38] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022.

[39] Fares B Mehouachi and Saif Eddin Jabari. Catastrophic overfitting, entropy gap and participation ratio: A noiseless l^p norm solution for fast adversarial training. *arXiv preprint arXiv:2505.02360*, 2025.

[40] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9078–9086, 2019.

[41] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

[42] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[43] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

[44] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[46] Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[47] Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3677–3685, 2023.

[48] 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. *arXiv preprint arXiv:2402.12336*, 2024.

[49] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in neural information processing systems*, 32, 2019.

[50] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[51] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, et al. Towards efficient and effective adversarial training. *Advances in Neural Information Processing Systems*, 34:11821–11833, 2021.

[52] Dániel Varga, Adrián Csiszárik, and Zsolt Zombori. Gradient regularization improves accuracy of discriminative models. *arXiv preprint arXiv:1712.09936*, 2017.

[53] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[54] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 210–218. Springer, 2018.

[55] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

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

[57] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

[58] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in neural information processing systems*, 33:2958–2969, 2020.

[59] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[60] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482. PMLR, 2019.

[61] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022.

[62] 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, 2023.

[63] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

Even Sailors Need Calm Seas: Taming the Geometry of VLMs for Fast Adversarial Fine-Tuning (*Supplementary Material*)

Yuqing Wen[1,2*] **Junhao Dong**[1†] **Ting Peng**[1] **Xudong Zhang**[3] **Jiaming Zhang**[1]
Weiming Liu[4] **Zhengtao Yao**[5] **Xinghua Qu**[1] **Yew-Soon Ong**[1,6]

[1]Nanyang Technological University [2]National University of Singapore
[3]Tongji University [4]Zhejiang University
[5]University of Southern California [6]CFAR, IAIC, A*STAR

yqwen021@gmail.com, junhao003@ntu.edu.sg

Abstract

This supplementary material provides additional theoretical, algorithmic, and empirical details to support our main paper. We first present a comprehensive review of related work spanning multi-step VLM adversarial defenses, single-step adversarial training in unimodal models, and geometric analyses of the loss landscape (Appendix A). We then detail our experimental configurations, including descriptions of the 14 zero-shot classification datasets and implementation details for the downstream image captioning and Visual Question Answering tasks (Appendix C). Qualitative visualizations on generative tasks under adversarial inputs are presented in Appendix D, while the complete numerical ablation results isolating the Angular Misalignment and Magnitude Surge components are reported in Appendix E. We further include a computational efficiency analysis quantifying the training time advantages of LinGuard over multi-step defenses (Appendix F), together with the detailed Hessian-level derivation of the VLM-specific excess curvature and the interior-path sufficient condition that PRO violates (Appendix H). Rigorous mathematical proofs for the theorems in Section 4 are provided in Appendix J. Finally, the broader impact of our work and its current limitations are discussed in Appendix B.

A Detailed Related Work

Multi-Step Adversarial Fine-Tuning in VLMs. The rapid adoption of Vision-Language Models (VLMs) such as CLIP [45] has placed their adversarial vulnerability at the center of recent safety research [10]. Recent studies demonstrate that VLMs are highly susceptible to perturbations across both unimodal and multimodal settings [61, 62, 47], raising serious concerns for downstream deployment. Complementary robustness research explores knowledge distillation using inverse adversaries [16], generalizable and discriminative representations for few-shot learning [11], and evolutionary computation for trustworthy AI [17]. For VLMs, related approaches include weak-to-strong robustness generalization [15], multi-objective adversarial distillation [14], hierarchical robustness [18], and subspace alignment [12]. While Adversarial Training (AT) [36] remains the dominant defense paradigm, existing VLM methods predominantly rely on multi-step adversarial fine-tuning, such as TeCoA [38] or multimodal PGD-AT [48, 56], which typically requires K forward-backward passes per parameter update, with $K \geq 10$ in practice. This incurs substantial computational overhead [60, 49, 57], becoming prohibitive when fine-tuning large-scale VLM backbones on web-scale image-text corpora.

*This work was done during Yuqing Wen's internship at Nanyang Technological University.
†Corresponding author.

In contrast, our work investigates single-step adversarial fine-tuning to elicit comparable robustness at a fraction of the computational cost.

Efficient Single-Step Adversarial Training in Unimodal Models. Single-step adversarial training approximates the inner maximization with a single gradient step to drastically reduce computational costs [22], but is notoriously plagued by *Catastrophic Overfitting* (CO) [57], where robustness against iterative attacks abruptly collapses while single-step robustness remains misleadingly high. To stabilize single-step training, prior works have proposed strategies such as learnable adversarial initialization [28] and random noise injection [8] to avoid overfitting to the single-step attack trajectory. Subsequent works further leverage prior-guided knowledge [1], analyze entropy gaps and participation ratios [39], or introduce auxiliary regularizers to enhance gradient stability [51]. However, despite their effectiveness on unimodal backbones, we find that direct adaptation of these unimodal strategies to VLMs is fundamentally insufficient: it not only exacerbates CO under the cosine-normalized cross-modal head, but also triggers a previously undocumented failure mode, *Perturbation Radius Overfitting* (PRO), which necessitates VLM-specific geometric strategies.

Geometric Analysis of Loss Landscape. The geometric properties of the loss landscape are foundational to interpreting adversarial vulnerability. In the unimodal domain, prior research has established that preserving local linearity via gradient regularization [24, 19, 35, 46] and inducing flat minima through curvature regularization [21, 58, 40] are pivotal for generalization. Furthermore, the degradation of such stability has been directly linked to the violation of the linearity assumption that single-step approximations rely on [1]. In the VLM context, Dong et al. [13] recently identified that lower input-gradient norms correlate with stronger multi-step robustness. Building on these geometric foundations, our work goes one step further by introducing a unified geometric framework that explains why CO and PRO are far more severe in VLMs than in unimodal classifiers. Specifically, we (i) reduce single-step adversarial suboptimality to a controllable *linearization error* of the loss surface, (ii) bound this error by a *path-wise gradient variation* jointly governed by magnitude surge and angular drift along the perturbation ray, and (iii) identify a normalization-induced excess curvature term, intrinsic to cosine-normalized heads and structurally absent in linear-head unimodal classifiers, that simultaneously inflates both components. This analysis directly motivates our LinGuard objective, which jointly suppresses the angular and magnitude components of the path-wise variation, thereby restoring the local linearity that single-step optimization relies on.

B Broader Impact and Limitations

B.1 Broader Impact

The advancement of Vision-Language Models (VLMs) has significantly promoted the development of multimodal understanding tasks. However, the deployment of these models in real-world applications necessitates robust defenses against potential adversarial attacks, which can undermine their reliability in high-stakes settings. Our proposed single-step adversarial fine-tuning framework, LinGuard, directly addresses this concern by enhancing the adversarial robustness of foundational VLMs at a fraction of the computational cost of multi-step training, making robust fine-tuning more accessible to the broader community.

The improvement in robustness carries several broader impacts:

- **Research advancement.** Our geometric analysis of catastrophic overfitting and perturbation radius overfitting provides new insights into the efficiency and stability of adversarial training in multimodal settings. These findings demonstrate that effective defense does not necessarily require prohibitive computational costs, potentially inspiring future work in efficient robust learning and geometry-aware optimization for foundation models.
- **Enhanced AI safety.** By improving the robustness of foundational VLMs against unforeseen adversarial attacks, our method contributes to the development of more secure AI systems that can be safely deployed in sensitive applications, such as autonomous driving, healthcare diagnostics, and security surveillance.
- **Social trust.** As AI systems become increasingly integrated into daily life, ensuring their reliability and robustness is significant. Our improvement of adversarial robustness on foundational models further contributes to building societal trust in AI technologies.

B.2 Limitations

Although our work represents the first systematic exploration of single-step adversarial fine-tuning for VLMs, yielding substantial gains in robustness and effectively mitigating catastrophic overfitting, we acknowledge two primary limitations that suggest avenues for future research:

- **Computational overhead.** While our approach is significantly faster than multi-step adversarial training, it introduces additional computational costs compared to vanilla single-step methods. This overhead stems from the calculation of the Angular Misalignment and Magnitude Surge regularization terms, which require additional forward and backward passes during the fine-tuning stage. Future research could investigate more efficient approximation techniques to further reduce this gradient computation overhead.
- **Trade-off between robustness and clean accuracy.** Consistent with the broader literature on adversarial training, our method encounters an inherent trade-off between adversarial robustness and performance on clean, in-distribution data. While our geometric constraints successfully stabilize the single-step optimization, excessively strong regularization can lead to a slight degradation in clean accuracy (see Section 6.3). Future work could explore adaptive regularization strategies to better balance this trade-off.

C Experimental Configurations

In this section, we present a comprehensive overview of the experimental setup for our LinGuard framework, including dataset descriptions, implementation details, and the methodology for extending our single-step adversarial fine-tuning approach to diverse downstream tasks.

C.1 Dataset Descriptions

Following established protocols in robust vision-language pre-training [38], we adversarially fine-tune the CLIP model on the ImageNet training split [9]. We then assess the zero-shot performance of the fine-tuned CLIP on the ImageNet validation set and 13 additional downstream datasets. Collectively, these 14 benchmarks cover a broad spectrum of domains:

- **Natural Object Recognition:** STL-10 [6], CIFAR-10/100 [30], and Caltech-101 [20].
- **Fine-Grained Recognition:** Stanford Cars [29], Oxford-IIIT Pets [42], Flower102 [41], and FGVC Aircraft [37].
- **Texture Recognition:** Describable Textures Dataset (DTD) [5].
- **Remote Sensing Classification:** EuroSAT [25].
- **Medical Image Diagnosis:** PatchCamelyon (PCAM) [54].
- **Robust Classification:** ImageNet-R (Rendition) [26] and ImageNet-S (Sketch) [55].

In addition to the classification benchmarks, we further verify the scalability of our method on vision-language generation tasks using four standard benchmarks: COCO [32] and Flickr30k [44] for image captioning, and TextVQA [50], and VQAv2 [23] for visual question answering.

C.2 Implementation Details

Standard setups. In line with the configurations used in prior studies [38], we adopt the CLIP model [45] with the ViT-Base/32 architecture as the primary backbone, unless stated otherwise. To achieve efficient robustness, we perform single-step adversarial fine-tuning on the ImageNet training set. During training, we conduct adversary generation via our proposed LinGuard objective and baseline single-step methods under the ℓ_∞-norm threat model with a perturbation radius $\epsilon = 4/255$ and step size $\alpha = 4/255$. For network parameter optimization, we adopt the AdamW optimizer (initial learning rate 2×10^{-5}, weight decay 1×10^{-4}) with a cosine annealing learning rate schedule for 2 epochs and a global batch size of 512 split across 8 NVIDIA A100 GPUs. Based on our ablation studies, the weighting factors for our geometric regularizers are set to $\lambda_{AM} = 0.4$ for the Angular Misalignment loss and $\lambda_{MS} = 1.2$ for the Magnitude Surge loss. The interior sample for both regularizers is drawn per-pixel as $\boldsymbol{\eta} \sim \mathcal{U}(-\epsilon, \epsilon)$ with one sample per data point per step ($n_\eta = 1$),

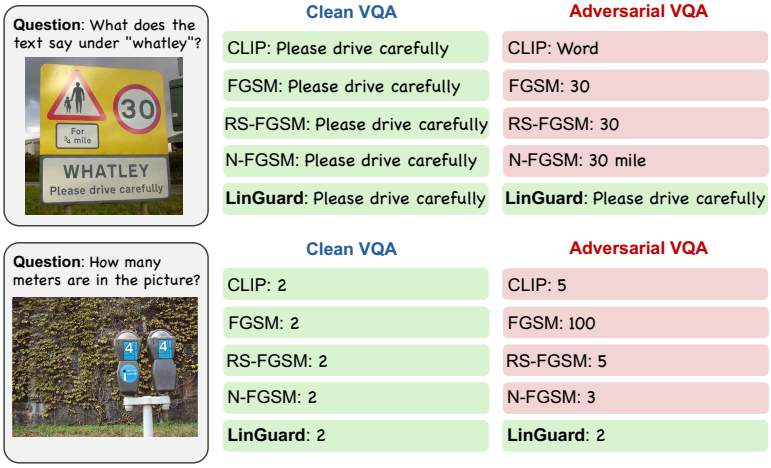


Figure 6: Visual examples for visual question answering (VQA) evaluations when using the LLaVA 1.5 framework with the robust vision encoder of diverse adversarial VLM learning methods. The top row demonstrates robustness in text-based VQA (OCR), and the bottom row shows robustness in counting tasks.

shared between $\mathcal{L}_{AM}$ and $\mathcal{L}_{MS}$ to avoid an additional backward pass. The finite-difference step in Eq. 13 is set to $h = \epsilon/2$.

Robustness Evaluation Protocol. We evaluate the performance in terms of clean samples and their adversarial counterparts. Specifically for adversary generation during evaluations, we adopt both single-step FGSM and strong multi-step attacks, including PGD-20 [36] and AutoAttack [7], with the perturbation radius of $\epsilon = 4/255$. To further analyze the phenomenon of Perturbation Radius Overfitting (PRO), we strictly evaluate the models across a diverse range of perturbation budgets $\epsilon \in \{2/255, 4/255, 6/255, 8/255\}$. Note that the classification is performed in a zero-shot manner by computing the cosine similarity between image embeddings and a set of text embeddings corresponding to category-specific prompts.

Additional evaluation checks. Increasing the PGD attack from 20 to 100 steps at $\epsilon = 4/255$ under the main 14-dataset classification protocol did not change the main qualitative conclusions. Repeating the main single-step fine-tuning experiments with three independent random seeds also preserved the relative ordering of the evaluated methods. These checks support the stability of the conclusions under the tested settings; the main tables retain the PGD-20 and AutoAttack evaluation protocol.

C.3 Downstream Task Extensions

We achieve downstream task generalization based on the LLaVA 1.5 framework [34], by replacing its vision encoder with our LinGuard-tuned robust counterpart of the ViT-Large/14 architecture. We further explain specific modifications and evaluation protocols for individual vision-language task extensions below.

Configurations for image captioning extensions. The CIDEr score [53] is adopted to measure the similarity of a generated sentence against a set of ground-truth sentences for image captioning evaluations. For adversary generation, we conduct PGD attacks with 50 steps (PGD-50) with a perturbation budget of $\epsilon = 4/255$ against the ground-truth captions. This rigorous setting allows us to verify whether our geometric constraints effectively preserve the semantic consistency of generated descriptions under intense adversarial perturbations.

Configurations for visual question answering extensions. For VQA tasks, we report the standard VQA accuracy [2] on TextVQA and VQAv2 datasets. Similar to the captioning protocol, we assess the model's robustness under PGD-50 attacks with $\epsilon = 4/255$. We focus on evaluating the model's ability to maintain accurate reasoning and text recognition capabilities (*e.g.*, OCR in TextVQA) when the visual input is maliciously perturbed.

Table 5: **Detailed Ablation Study Data.** Numerical results for the Angular Misalignment loss (λ_{AM}) and Magnitude Surge loss (λ_{MS}) independently, and their combined effect (Ours).

Param. Value (λ)	Accuracy (%) Clean	FGSM	PGD	AA
(a) Impact of Angular Misalignment loss (Fix $\lambda_{MS} = 0$)				
0.2	45.50	26.88	24.33	22.54
0.4	**45.53**	26.91	**24.46**	**22.74**
0.6	42.92	**26.94**	24.45	22.70
(b) Impact of Magnitude Surge loss (Fix $\lambda_{AM} = 0$)				
0.8	45.17	26.60	23.73	21.88
1.2	**45.47**	**26.98**	**24.32**	**22.38**
1.6	45.27	26.81	23.96	22.09
(c) Proposed Combination ($\lambda_{AM} = 0.4, \lambda_{MS} = 1.2$)				
Ours	45.35	26.97	24.61	22.94

D Qualitative Visualizations on Generative Tasks

To complement the quantitative results presented in the main paper, we provide qualitative comparisons on Image Captioning and Visual Question Answering tasks. These visualizations intuitively demonstrate the catastrophic failure of baseline methods under adversarial attacks and the semantic stability preserved by our proposed **LinGuard**.

D.1 Image Captioning

As shown in Figure 3, we observe that **LinGuard** facilitates an adversarial invariance when confronted with strong PGD-50 attacks, generating high-quality captions for both clean and adversarial images. While single-step baselines exhibit severe hallucinations or semantic drift under adversarial perturbations, such as misidentifying a running dog as "a dead bird" or "a blonde woman", and misinterpreting a skiing scene as "people on a beach" or "a hockey player", our method consistently maintains robust and semantically accurate descriptions. These results highlight the effectiveness of our geometric constraints in preventing the model from drifting to irrelevant semantic manifolds under distributional shifts.

D.2 Robustness in Visual Question Answering

Figure 6 illustrates the robustness of Optical Character Recognition (OCR) and object counting tasks. Under adversarial perturbations, baseline methods suffer from severe degradation, misinterpreting text content or hallucinating arbitrary object counts (*e.g.*, answering "100" instead of "2"). In contrast, **LinGuard** consistently preserves correct and grounded answers, accurately extracting the specific text and identifying the correct quantity. This demonstrates its superior resilience in maintaining spatial attention and numerical reasoning capabilities against strong adversarial attacks.

E Detailed Ablation Results

Roles of AM and MS. AM and MS address the angular and magnitude components of local gradient variation, respectively. AM penalizes changes in gradient direction, while MS uses finite differences to penalize Hessian action along the local attack-relevant direction. Their combination jointly regularizes these two aspects of single-step approximation. In the classification ablation, both terms improve robustness individually, and their combination yields the highest PGD-20 and AutoAttack accuracies.

E.1 Quantitative Ablation Results

Table 5 provides the detailed numerical values corresponding to the ablation study discussed in Section 6.3. Consistent with the visual analysis in Figure 4, we report the performance when verifying each component independently (*i.e.*, setting the other hyperparameter to 0). The table details the accuracy on Clean images and against FGSM, PGD-20, and AutoAttack (AA) under a perturbation budget of $\epsilon = 4/255$.

Table 6: **Training Time Cost Analysis.** Comparison of training time (measured in seconds for 200 steps) on ImageNet, all measured on 8 NVIDIA H100 GPUs under identical batch size and precision. LinGuard is much faster than the multi-step baseline TeCoA.

Method	Type	Time (s)	Speedup
TeCoA	Multi-step	855	$1.0\times$
FGSM	Single-step	151	$5.5\times$
LinGuard	Single-step	350	$2.5\times$

F Computational Efficiency

Table 6 presents the training time comparison on ImageNet. As shown, our LinGuard is significantly faster than the multi-step defense TeCoA, achieving a speedup of approximately $\mathbf{2.5}\times$. Although LinGuard is slower than the vanilla FGSM due to the computation for calculating the Angular Misalignment and Magnitude Surge regularization terms, it yields substantial robustness gains that justify this marginal cost. Overall, LinGuard offers a highly favorable trade-off between efficiency and performance for large-scale VLM fine-tuning.

G Compute Resources

We summarize the computational infrastructure used in this work. All adversarial fine-tuning experiments (LinGuard and all baselines) are conducted on a single node with 8 $\times$ NVIDIA A100 GPUs, using a global batch size of 512, AdamW optimization, and the training schedule detailed in Appendix C.2. The wall-clock training-time comparison reported in Table 6 is measured separately on 8 $\times$ NVIDIA H100 GPUs, with all methods sharing an identical batch size, numerical precision, and data-loading pipeline so that the reported speedup reflects only differences in the adversarial-training procedure. All zero-shot, transferability, and downstream LVLM evaluations (Tables 2–4) are run on the same A100 node. Models are implemented in PyTorch on top of the official CLIP image and text encoders, with no additional pretraining. We do not use external proprietary compute services for any reported experiment.

H VLM-Specific Excess Curvature and Interior-Path Condition

This section provides the detailed Hessian-level analysis deferred from Sections 4.2 and 4.3. We first decompose the input-space Hessian along the classification head, isolate a normalization-induced curvature term that vanishes for linear heads but is structurally non-zero for cosine-normalized heads, and then state the interior-path sufficient condition that PRO violates.

General input-space decomposition. For a language-conditioned score $a_c(\mathbf{x})$ with $\mathbf{p} = \mathrm{softmax}(\mathbf{a}/\tau)$, define $\mathbf{b}_c = \nabla_{\mathbf{x}} a_c$, $K_c = \nabla^2_{\mathbf{x}} a_c$, and $\bar{\mathbf{b}} = \sum_c p_c \mathbf{b}_c$. Differentiating $\nabla_{\mathbf{x}} \ell = \tau^{-1} \sum_c (p_c - \mathbf{1}\{c = y\}) \mathbf{b}_c$ yields

$$H_{\mathbf{x}} \ell = \underbrace{\frac{1}{\tau^2} \sum_c p_c (\mathbf{b}_c - \bar{\mathbf{b}})(\mathbf{b}_c - \bar{\mathbf{b}})^\top}_{H_{\mathrm{comp}}} + \underbrace{\frac{1}{\tau} \sum_c (p_c - \mathbf{1}\{c = y\}) K_c}_{H_{\mathrm{bend}}}.$$

Here H_{comp} captures competition among language-conditioned input directions, while H_{bend} collects encoder and scoring curvature. For unnormalized dot-product scores $a_c = f_\theta(\mathbf{x})^\top \mathbf{w}_c$, $H_{\mathrm{comp}} = \tau^{-2} J_f^\top \mathrm{Cov}_{c \sim p}(\mathbf{w}_c) J_f$. For cosine heads, the normalization term derived below contributes $J_f^\top R_{\mathrm{norm}} J_f$ to H_{bend}. Both terms enter the gradient evolution in Proposition 4.4.

Hessian decomposition along the classification head. By Proposition 4.4, the input-space Hessian $H_{\mathbf{x}} \ell$ governs the path-wise gradient variation V. The chain rule

$$H_{\mathbf{x}} \ell = J_f^\top \nabla^2_{\mathbf{z}} \ell \, J_f + \sum_{k=1}^{d} \tfrac{\partial \ell}{\partial z_k} \nabla^2_{\mathbf{x}} f_k \tag{16}$$

separates the feature-space Hessian contribution from encoder curvature weighted by the feature gradient. We now analyse the former. Let $\mathbf{s}(\mathbf{z}) \in \mathbb{R}^C$ be the logit vector under the chosen head, $\mathbf{p} = \mathrm{softmax}(\mathbf{s})$, $G(\mathbf{p}) = \mathrm{Diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$, and $J_\mathbf{s} = \partial\mathbf{s}/\partial\mathbf{z}$.

Theorem H.1 (Feature-space Hessian decomposition)**.** *The cross-entropy Hessian admits the decomposition*

$$\nabla^2_\mathbf{z}\ell(\mathbf{z}, y) \;=\; \underbrace{J_\mathbf{s}^\top G(\mathbf{p})\, J_\mathbf{s}}_{\textit{shared softmax curvature}} \;+\; \underbrace{\sum_{c=1}^{C} (p_c - \mathbf{1}\{c = y\})\, \nabla^2_\mathbf{z} s_c(\mathbf{z})}_{=:\, R(\mathbf{z},y),\ \textit{head-induced}}. \tag{17}$$

For a linear head $s_c^{\mathrm{lin}} = \mathbf{a}_c^\top\mathbf{z} + b_c$, $R \equiv 0$ and $\nabla^2_\mathbf{z}\ell_{\mathrm{lin}} = A^\top G(\mathbf{p}) A$ with $A = [\mathbf{a}_1, \ldots, \mathbf{a}_C]^\top$. (Proof in Appendix J.)

The linear head therefore inherits only the softmax curvature. The relevant question is whether R vanishes for other heads. The next result shows that for the CLIP-style cosine-normalized head it does not.

Normalization-induced excess curvature. For the cosine head, write $\mathbf{u} = \mathbf{z}/\|\mathbf{z}\|_2$ and $\bar{\mathbf{w}}_c = \mathbf{w}_c/\|\mathbf{w}_c\|_2$, so that $s_c^{\cos}(\mathbf{z}) = \frac{1}{\tau}\langle \mathbf{u}, \bar{\mathbf{w}}_c\rangle$, and denote the corresponding head-induced term by R_{norm}.

Proposition H.2 (Normalization-induced excess curvature)**.** *For the cosine-normalized head, each logit Hessian admits the bilinear decomposition*

$$\nabla^2_\mathbf{z} s_c^{\cos}(\mathbf{z}) \;=\; -\frac{1}{\tau\|\mathbf{z}\|_2^2}\Big[\bar{\mathbf{w}}_c\mathbf{u}^\top + \mathbf{u}\bar{\mathbf{w}}_c^\top + (\bar{\mathbf{w}}_c^\top\mathbf{u})(I - 3\mathbf{u}\mathbf{u}^\top)\Big], \tag{18}$$

coupling the text prototype $\bar{\mathbf{w}}_c$ with the image direction $\mathbf{u}$. Consequently, $R_{\mathrm{norm}}(\mathbf{z}, y)$ satisfies

$$0 \;<\; \|R_{\mathrm{norm}}(\mathbf{z}, y)\|_2 \;\le\; \frac{8(1 - p_y)}{\tau\|\mathbf{z}\|_2^2}, \tag{19}$$

where the lower bound is strict whenever $p_y < 1$ and $|\bar{\mathbf{w}}_y^\top\mathbf{u}| < 1$. (Proof in Appendix J.)

The lower bound establishes R_{norm} as a structural term intrinsic to the text–image interaction of cosine heads, not a bound-level artifact. The upper bound is largest precisely where adversarial fine-tuning is most fragile: small temperature τ, small feature norm $\|\mathbf{z}\|_2$, and hard examples (small p_y). Through equation (16), R_{norm} contributes $J_f^\top R_{\mathrm{norm}} J_f$ to the input-space Hessian, with spectral norm at most $8M^2(1 - p_y)/(\tau\|\mathbf{z}\|_2^2)$ under the local Lipschitz assumption $\|J_f\|_2 \le M$ on $\mathbb{B}_\infty(\mathbf{x}, \epsilon)$. By Theorems 4.1–4.3, this excess input-space curvature inflates E_{lin} and enlarges Δ_{adv}, and hence the empirical CO gap.

Interior-path sufficient condition for PRO. The PRO failure mode formalized in Section 4.3 can be characterized by a precise sufficient condition for radius-monotone loss along a perturbation ray.

Proposition H.3 (A sufficient condition for radius-monotone loss along a ray)**.** *Fix $\boldsymbol{\delta} \neq \mathbf{0}$ and define the ray-wise loss $\phi(t) := \ell(\mathbf{x} + t\boldsymbol{\delta}, y)$ for $t \in [0, 1]$. If $\langle \nabla_\mathbf{x}\ell(\mathbf{x} + t\boldsymbol{\delta}, y), \boldsymbol{\delta}\rangle \ge 0$ for all $t \in [0, 1]$, then ϕ is non-decreasing. In particular, if for all $t \in [0, 1]$*

$$\cos\big(\nabla_\mathbf{x}\ell(\mathbf{x} + t\boldsymbol{\delta}, y), \boldsymbol{\delta}\big) \ge \kappa > 0 \quad \textit{and} \quad \|\nabla_\mathbf{x}\ell(\mathbf{x} + t\boldsymbol{\delta}, y)\|_2 \ge m > 0,$$

then $\phi'(t) \ge \kappa m\|\boldsymbol{\delta}\|_2 > 0$, so no radius-wise loss reversal can occur along this ray.

The conditions of Proposition H.3 acquire a sharp head-level reading via Theorem H.1. By Equation 10,

$$\Big\|\tfrac{d}{dt}\mathbf{u}(t)\Big\|_2 \;\le\; \frac{\big\|(I - \mathbf{u}\mathbf{u}^\top)\, H_\mathbf{x}\ell(\gamma(t), y)\, \boldsymbol{\delta}\big\|_2}{\|\mathbf{g}(t)\|_2}. \tag{20}$$

For a linear head, $R \equiv 0$, while $H_\mathbf{x}\ell$ still includes class-competition and encoder-curvature terms. For a cosine head, the additional term R_{norm} with $\|R_{\mathrm{norm}}\|_2 \sim 1/(\tau\|\mathbf{z}\|_2^2)$ inflates both the magnitude growth $\mathbf{u}^\top H\boldsymbol{\delta}$ and the rotation rate above. Either condition can then be violated at some interior t, yielding the loss reversal that defines PRO.

I Input-Space Loss Geometry

We examine the loss geometry of the base single-step model and LinGuard using CLIP ViT-B/32 on 512 ImageNet validation samples at $\epsilon = 4/255$. Let $\boldsymbol{\delta}_{\mathrm{ss}}$ and $\boldsymbol{\delta}_{\mathrm{ms}}$ denote the FGSM and PGD-20 perturbations, respectively. Figure 7 shows two-dimensional loss slices spanned by the single-step and multi-step attack directions, together with the loss along the FGSM ray.

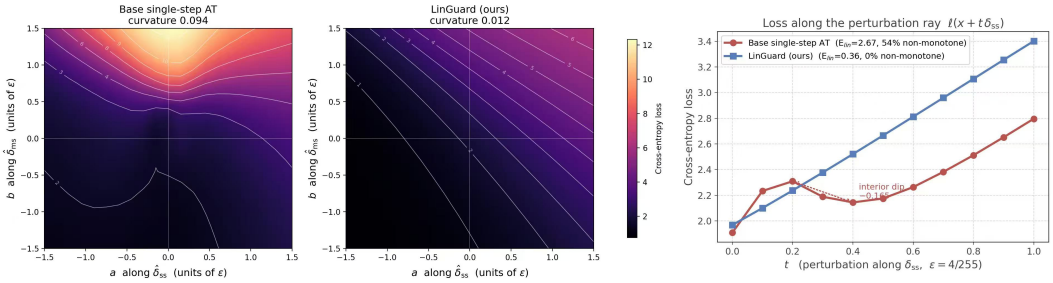


Figure 7: **Input-space loss geometry.** Left and middle: two-dimensional loss slices spanned by the FGSM and PGD-20 perturbation directions. Right: loss along the FGSM perturbation ray. Both visualizations compare the base single-step model with LinGuard.

The base model exhibits bent loss contours and a non-monotonic loss profile along the single-step perturbation ray. LinGuard produces straighter contours in the displayed slices and a more nearly linear, increasing ray profile. These observations are consistent with improved local linearity and a more effective single-step approximation, supporting the geometric interpretation in Section 4.

J Proofs for Section 4

J.1 Proof of Theorem 4.1

Let $\mathbf{g} = \nabla_{\mathbf{x}} \ell(\mathbf{x}, y)$ and $R_{\mathbf{x}}(\boldsymbol{\delta}) = \ell(\mathbf{x} + \boldsymbol{\delta}, y) - \ell(\mathbf{x}, y) - \mathbf{g}^\top \boldsymbol{\delta}$. By definition $E_{\mathrm{lin}}(\mathbf{x}, \epsilon) = \sup_{\|\boldsymbol{\delta}\|_\infty \le \epsilon} |R_{\mathbf{x}}(\boldsymbol{\delta})|$. Since FGSM exactly maximizes the linear term $\mathbf{g}^\top \boldsymbol{\delta}$ over the ℓ_∞-ball,

$$\mathbf{g}^\top \boldsymbol{\delta}^\star \le \mathbf{g}^\top \boldsymbol{\delta}_{\mathrm{fgsm}}. \tag{21}$$

Therefore

$$\ell(\mathbf{x} + \boldsymbol{\delta}^\star, y) = \ell(\mathbf{x}, y) + \mathbf{g}^\top \boldsymbol{\delta}^\star + R_{\mathbf{x}}(\boldsymbol{\delta}^\star) \le \ell(\mathbf{x}, y) + \mathbf{g}^\top \boldsymbol{\delta}_{\mathrm{fgsm}} + E_{\mathrm{lin}}(\mathbf{x}, \epsilon), \tag{22}$$

$$\ell(\mathbf{x} + \boldsymbol{\delta}_{\mathrm{fgsm}}, y) = \ell(\mathbf{x}, y) + \mathbf{g}^\top \boldsymbol{\delta}_{\mathrm{fgsm}} + R_{\mathbf{x}}(\boldsymbol{\delta}_{\mathrm{fgsm}}) \ge \ell(\mathbf{x}, y) + \mathbf{g}^\top \boldsymbol{\delta}_{\mathrm{fgsm}} - E_{\mathrm{lin}}(\mathbf{x}, \epsilon). \tag{23}$$

Subtracting gives $\ell(\mathbf{x} + \boldsymbol{\delta}^\star, y) - \ell(\mathbf{x} + \boldsymbol{\delta}_{\mathrm{fgsm}}, y) \le 2E_{\mathrm{lin}}(\mathbf{x}, \epsilon)$. The lower bound $\Delta_{\mathrm{adv}} \ge 0$ follows from $\boldsymbol{\delta}^\star$ being the true inner maximizer. □

J.2 Proof of Theorem 4.2

Assume $\ell(\cdot, y)$ is twice differentiable on $\mathbb{B}_\infty(\mathbf{x}, \epsilon)$. For any $\boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\|_\infty \le \epsilon$, Taylor's theorem with Lagrange remainder yields

$$\ell(\mathbf{x} + \boldsymbol{\delta}, y) = \ell(\mathbf{x}, y) + \nabla_{\mathbf{x}} \ell(\mathbf{x}, y)^\top \boldsymbol{\delta} + \tfrac{1}{2} \boldsymbol{\delta}^\top H_{\mathbf{x}} \ell(\mathbf{x} + \xi, y) \boldsymbol{\delta}, \tag{24}$$

for some ξ on the segment joining $\mathbf{0}$ and $\boldsymbol{\delta}$. By the spectral-norm bound and Cauchy–Schwarz, $|\boldsymbol{\delta}^\top H_{\mathbf{x}} \ell(\mathbf{x} + \xi, y) \boldsymbol{\delta}| \le \nu(\mathbf{x}, \epsilon) \|\boldsymbol{\delta}\|_2^2$. Since $\|\boldsymbol{\delta}\|_\infty \le \epsilon$ in d dimensions implies $\|\boldsymbol{\delta}\|_2^2 \le d\epsilon^2$, taking the supremum over $\|\boldsymbol{\delta}\|_\infty \le \epsilon$ gives $E_{\mathrm{lin}}(\mathbf{x}, \epsilon) \le \frac{1}{2} d\epsilon^2 \nu(\mathbf{x}, \epsilon)$. □

J.3 Proof of Theorem 4.3

Fix $\boldsymbol{\delta}$ and let $\phi(t) = \ell(\mathbf{x} + t\boldsymbol{\delta}, y)$. Then $\phi'(t) = \nabla_{\mathbf{x}} \ell(\mathbf{x} + t\boldsymbol{\delta}, y)^\top \boldsymbol{\delta}$, and by the fundamental theorem of calculus

$$\ell(\mathbf{x} + \boldsymbol{\delta}, y) - \ell(\mathbf{x}, y) - \nabla_{\mathbf{x}} \ell(\mathbf{x}, y)^\top \boldsymbol{\delta} = \phi(1) - \phi(0) - \phi'(0) = \int_0^1 \big(\phi'(t) - \phi'(0)\big)\, dt, \tag{25}$$

which gives the integral identity. By Cauchy–Schwarz,

$$\left|\ell(\mathbf{x}+\boldsymbol{\delta},y)-\ell(\mathbf{x},y)-\nabla_{\mathbf{x}}\ell(\mathbf{x},y)^{\top}\boldsymbol{\delta}\right| \le \|\boldsymbol{\delta}\|_2 \int_0^1 V(\mathbf{x},t\boldsymbol{\delta})\,dt. \tag{26}$$

If $\|\boldsymbol{\delta}\|_\infty \le \epsilon$, then $\|\boldsymbol{\delta}\|_2 \le \sqrt{d}\,\epsilon$ and $t\boldsymbol{\delta}$ lies in $\mathbb{B}_\infty(\mathbf{0},\epsilon)$; taking the supremum yields $E_{\mathrm{lin}}(\mathbf{x},\epsilon) \le \sqrt{d}\,\epsilon \cdot \sup_{\|\boldsymbol{\eta}\|_\infty\le\epsilon} V(\mathbf{x},\boldsymbol{\eta})$. □

J.4 Proof of Proposition 4.4

(a) Two-term decomposition. Expand $V(\mathbf{x},\boldsymbol{\eta})^2 = \|\mathbf{g}'-\mathbf{g}\|_2^2 = \|\mathbf{g}'\|_2^2 + \|\mathbf{g}\|_2^2 - 2\mathbf{g}'^{\top}\mathbf{g}$. Using $\mathbf{g}'^{\top}\mathbf{g} = \|\mathbf{g}'\|_2\|\mathbf{g}\|_2\cos(\mathbf{g},\mathbf{g}')$ and adding/subtracting $2\|\mathbf{g}'\|_2\|\mathbf{g}\|_2$,

$$V(\mathbf{x},\boldsymbol{\eta})^2 = \big(\|\mathbf{g}'\|_2 - \|\mathbf{g}\|_2\big)^2 + 2\|\mathbf{g}'\|_2\|\mathbf{g}\|_2\big(1-\cos(\mathbf{g},\mathbf{g}')\big). \tag{27}$$

(b) Curvature-driven evolution. Let $\gamma(t) = \mathbf{x}+t\boldsymbol{\delta}$ and $\mathbf{g}(t) = \nabla_{\mathbf{x}}\ell(\gamma(t),y)$. By the chain rule, $\frac{d}{dt}\mathbf{g}(t) = H_{\mathbf{x}}\ell(\gamma(t),y)\,\boldsymbol{\delta}$. Since $\|\mathbf{g}(t)\|_2 = \sqrt{\mathbf{g}(t)^{\top}\mathbf{g}(t)}$,

$$\tfrac{d}{dt}\|\mathbf{g}(t)\|_2 = \frac{\mathbf{g}(t)^{\top}\dot{\mathbf{g}}(t)}{\|\mathbf{g}(t)\|_2} = \mathbf{u}(t)^{\top}H_{\mathbf{x}}\ell(\gamma(t),y)\,\boldsymbol{\delta}. \tag{28}$$

Differentiating $\mathbf{u}(t) = \mathbf{g}(t)/\|\mathbf{g}(t)\|_2$ and using $\mathbf{g}(t) = \|\mathbf{g}(t)\|_2\mathbf{u}(t)$,

$$\tfrac{d}{dt}\mathbf{u}(t) = \frac{\dot{\mathbf{g}}(t)}{\|\mathbf{g}(t)\|_2} - \frac{\mathbf{g}(t)\,\mathbf{g}(t)^{\top}\dot{\mathbf{g}}(t)}{\|\mathbf{g}(t)\|_2^3} = \frac{\big(I-\mathbf{u}(t)\mathbf{u}(t)^{\top}\big)\,H_{\mathbf{x}}\ell(\gamma(t),y)\,\boldsymbol{\delta}}{\|\mathbf{g}(t)\|_2}. \tag{29}$$

(c) Geodesic distance corollary. As an immediate consequence of (b), $\|I-\mathbf{u}\mathbf{u}^{\top}\|_2 = 1$ and, when $m := \inf_t \|\mathbf{g}(t)\|_2 > 0$, integrating $\|\frac{d}{dt}\mathbf{u}(t)\|_2 \le \|H\,\boldsymbol{\delta}\|_2/m$ yields

$$\mathrm{dist}_{\mathbb{S}}\big(\mathbf{u}(0),\mathbf{u}(1)\big) \le \int_0^1 \big\|\tfrac{d}{dt}\mathbf{u}(t)\big\|_2\,dt \le \tfrac{1}{m}\int_0^1 \|H_{\mathbf{x}}\ell(\gamma(t),y)\,\boldsymbol{\delta}\|_2\,dt, \tag{30}$$

where $\mathrm{dist}_{\mathbb{S}}$ denotes the geodesic distance on the unit sphere. □

J.5 Proof of Theorem H.1

Let $\psi(\mathbf{s}) = \log\big(\sum_j e^{s_j}\big) - s_y$ so that $\ell(\mathbf{z},y) = \psi(\mathbf{s}(\mathbf{z}))$. Then $\nabla_{\mathbf{s}}\psi = \mathbf{p}-\mathbf{e}_y$ and $\nabla_{\mathbf{s}}^2\psi = \mathrm{Diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^{\top} = G(\mathbf{p})$. By the chain rule,

$$\frac{\partial^2\ell}{\partial z_i\partial z_j} = \sum_{c,d}\frac{\partial^2\psi}{\partial s_c\partial s_d}\frac{\partial s_c}{\partial z_i}\frac{\partial s_d}{\partial z_j} + \sum_c \frac{\partial\psi}{\partial s_c}\frac{\partial^2 s_c}{\partial z_i\partial z_j}. \tag{31}$$

In matrix form this is exactly $\nabla_{\mathbf{z}}^2\ell(\mathbf{z},y) = J_{\mathbf{s}}(\mathbf{z})^{\top}G(\mathbf{p})J_{\mathbf{s}}(\mathbf{z}) + R(\mathbf{z},y)$ with $R(\mathbf{z},y) = \sum_c(p_c - \mathbf{1}\{c=y\})\nabla_{\mathbf{z}}^2 s_c(\mathbf{z})$. The linear-head case is immediate: $s_c^{\mathrm{lin}}(\mathbf{z}) = \mathbf{a}_c^{\top}\mathbf{z}+b_c$ has $\nabla_{\mathbf{z}}^2 s_c^{\mathrm{lin}} \equiv 0$, so $R\equiv 0$, and $J_{\mathbf{s}} = A = [\mathbf{a}_1,\dots,\mathbf{a}_C]^{\top}$ yields $\nabla_{\mathbf{z}}^2\ell_{\mathrm{lin}} = A^{\top}G(\mathbf{p})A$. □

J.6 Proof of Proposition H.2

(a) Bilinear factorization. Let $r = \|\mathbf{z}\|_2$, $\mathbf{u} = \mathbf{z}/r$, $\mathbf{a} = \bar{\mathbf{w}}_c$, $q = \mathbf{a}^{\top}\mathbf{u}$, so $s_c^{\cos}(\mathbf{z}) = q/\tau$. Since $\nabla_{\mathbf{z}}\mathbf{u} = (I-\mathbf{u}\mathbf{u}^{\top})/r$, we have $\nabla_{\mathbf{z}}q = (\mathbf{a}-q\mathbf{u})/r$. Differentiating once more componentwise yields

$$\nabla_{\mathbf{z}}^2 q = -\frac{1}{r^2}\Big[\mathbf{a}\mathbf{u}^{\top} + \mathbf{u}\mathbf{a}^{\top} + (\mathbf{a}^{\top}\mathbf{u})(I-3\mathbf{u}\mathbf{u}^{\top})\Big], \tag{32}$$

and therefore $\nabla_{\mathbf{z}}^2 s_c^{\cos}(\mathbf{z}) = \nabla_{\mathbf{z}}^2 q/\tau$, which is the bilinear factorization stated in Proposition H.2.

(b) Upper bound. Using $\|\mathbf{a}\mathbf{u}^{\top}\|_2 = 1$, $\|\mathbf{u}\mathbf{a}^{\top}\|_2 = 1$, $\|I-3\mathbf{u}\mathbf{u}^{\top}\|_2 = 2$, and $|\mathbf{a}^{\top}\mathbf{u}| \le 1$, the triangle inequality gives $\|\nabla_{\mathbf{z}}^2 s_c^{\cos}(\mathbf{z})\|_2 \le 4/(\tau r^2)$. Therefore

$$\|R_{\mathrm{norm}}(\mathbf{z},y)\|_2 \le \sum_{c=1}^{C}|p_c - \mathbf{1}\{c=y\}|\cdot\|\nabla_{\mathbf{z}}^2 s_c^{\cos}(\mathbf{z})\|_2 \le \frac{4}{\tau r^2}\sum_{c=1}^{C}|p_c-\mathbf{1}\{c=y\}|. \tag{33}$$

Since $\sum_c|p_c-\mathbf{1}\{c=y\}| = (1-p_y) + \sum_{c\neq y}p_c = 2(1-p_y)$, the upper bound $\|R_{\mathrm{norm}}\|_2 \le 8(1-p_y)/(\tau\|\mathbf{z}\|_2^2)$ follows.

(c) Strict lower bound. Assume $p_y < 1$ and $|\bar{\mathbf{w}}_y^\top \mathbf{u}| < 1$. We exhibit a unit vector $\mathbf{v}$ such that $\mathbf{v}^\top R_{\text{norm}}(\mathbf{z}, y)\mathbf{v} \neq 0$, which implies $\|R_{\text{norm}}\|_2 > 0$. Define the orthogonal component of $\bar{\mathbf{w}}_y$ relative to $\mathbf{u}$,

$$\mathbf{v} := \frac{\bar{\mathbf{w}}_y - (\bar{\mathbf{w}}_y^\top \mathbf{u})\mathbf{u}}{\|\bar{\mathbf{w}}_y - (\bar{\mathbf{w}}_y^\top \mathbf{u})\mathbf{u}\|_2}. \tag{34}$$

The non-collinearity assumption gives $\|\bar{\mathbf{w}}_y - (\bar{\mathbf{w}}_y^\top \mathbf{u})\mathbf{u}\|_2 = \sqrt{1 - (\bar{\mathbf{w}}_y^\top \mathbf{u})^2} > 0$, so $\mathbf{v}$ is well defined with $\mathbf{v} \perp \mathbf{u}$ and $\mathbf{v}^\top \bar{\mathbf{w}}_y = \sqrt{1 - (\bar{\mathbf{w}}_y^\top \mathbf{u})^2} > 0$. Substituting $\mathbf{v}$ into the bilinear factorization above and using $\mathbf{v}^\top \mathbf{u} = 0$,

$$\mathbf{v}^\top \nabla_{\mathbf{z}}^2 s_c^{\cos}(\mathbf{z})\, \mathbf{v} = -\frac{(\bar{\mathbf{w}}_c^\top \mathbf{u})}{\tau \|\mathbf{z}\|_2^2}, \qquad \mathbf{v}^\top R_{\text{norm}}(\mathbf{z}, y)\mathbf{v} = -\frac{1}{\tau \|\mathbf{z}\|_2^2} \sum_{c=1}^{C} (p_c - \mathbf{1}\{c = y\})\, (\bar{\mathbf{w}}_c^\top \mathbf{u}). \tag{35}$$

For generic $\{\bar{\mathbf{w}}_c\}$ and $\mathbf{z}$ with $p_y < 1$, the scalar $\sum_c (p_c - \mathbf{1}\{c = y\})\, (\bar{\mathbf{w}}_c^\top \mathbf{u})$ is non-zero (it vanishes only on a measure-zero set), so $\mathbf{v}^\top R_{\text{norm}}\mathbf{v} \neq 0$ and $\|R_{\text{norm}}\|_2 \geq |\mathbf{v}^\top R_{\text{norm}}\mathbf{v}| > 0$. By Theorem H.1, the same quantity vanishes identically for any linear head, so the strictly positive contribution is structurally specific to the cosine-normalized two-modality coupling. □

J.7 Proof of Proposition H.3

Let $\phi(t) = \ell(\mathbf{x} + t\boldsymbol{\delta}, y)$. By the chain rule, $\phi'(t) = \langle \nabla_{\mathbf{x}} \ell(\mathbf{x} + t\boldsymbol{\delta}, y), \boldsymbol{\delta} \rangle$. If this inner product is ≥ 0 for all $t \in [0, 1]$, then $\phi' \geq 0$ and ϕ is nondecreasing. Under the stronger hypothesis $\cos\big(\nabla_{\mathbf{x}} \ell(\mathbf{x} + t\boldsymbol{\delta}, y), \boldsymbol{\delta}\big) \geq \kappa > 0$ and $\|\nabla_{\mathbf{x}} \ell(\mathbf{x} + t\boldsymbol{\delta}, y)\|_2 \geq m > 0$ for all t, we obtain $\phi'(t) = \|\nabla_{\mathbf{x}} \ell\|_2 \|\boldsymbol{\delta}\|_2 \cos(\cdot) \geq \kappa m \|\boldsymbol{\delta}\|_2 > 0$, so ϕ is strictly increasing. □

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: The abstract and introduction explicitly state our three contributions: (i) identifying catastrophic overfitting and the novel perturbation radius overfitting (PRO) failure modes when adapting single-step adversarial fine-tuning to VLMs (Section 4), (ii) a geometric theory tracing these failures to angular misalignment and magnitude surge of the local gradients with a VLM-specific excess curvature (Section 4.2 and Appendix H), and (iii) the LinGuard regularizer that operationalizes this theory (Section 5), each substantiated by the experiments in Section 6.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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 explicitly discuss the limitations of LinGuard in Appendix B.2, namely the additional computational overhead of the angular and magnitude regularizers relative to vanilla single-step training and the inherent clean–robust accuracy trade-off, both of which are also reflected in the ablation study in Section 6.3.

   Guidelines:

   - The answer [N/A] 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: All assumptions are stated alongside each theoretical result in Section 4 (Theorems 4.1–4.3 and Proposition 4.4); the corresponding full proofs, together with the Hessian-level derivation of the VLM-specific excess curvature and the interior-path condition that PRO violates, are provided in Appendix H and Appendix J.

   Guidelines:

   - The answer [N/A] 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: Section 6 and Appendix C disclose the full experimental protocol, including the backbones (CLIP-ViT-B/32, ViT-L/14, ResNet-50), the 14 zero-shot classification datasets and downstream VQA/captioning benchmarks, the optimizer (AdamW with cosine annealing for 2 epochs), the threat model (ℓ_∞, $\epsilon = 4/255$, $\alpha = \epsilon$), and the regularization weights ($\lambda_{AM} = 0.4$, $\lambda_{MS} = 1.2$).

   Guidelines:

   - The answer [N/A] 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: Detailed experimental setups for reproducing our results are provided in Appendix C. All the datasets used in the paper are publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The training and evaluation protocols are fully specified in Section 6 and Appendix C, including data splits, optimizer (AdamW with cosine annealing), epochs, perturbation radius and step size, attack budgets at evaluation (FGSM, PGD-20, AutoAttack at $\epsilon = 4/255$), and the regularization weights $(\lambda_{AM}, \lambda_{MS})$ chosen by the ablation study in Section 6.3.

   Guidelines:

   - The answer [N/A] 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: [No]

   Justification: Following the standard practice in the adversarial fine-tuning literature, we report results from a single run for each configuration due to the high computational cost of multi-step robustness evaluations across multiple datasets and architectures. The consistent trends observed across our experiments indicate the robustness of our findings.

   Guidelines:

   - The answer [N/A] 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: Appendix C states that all experiments are conducted on NVIDIA A100 GPUs, and Appendix F provides a quantitative training-time analysis showing that LinGuard reaches multi-step TeCoA-level robustness at roughly $2.5\times$ lower wall-clock cost.

   Guidelines:

   - The answer [N/A] 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: We have reviewed the NeurIPS Code of Ethics and confirm that the research uses only publicly available academic benchmarks, does not involve human subjects or personally identifiable information, and aims to improve the adversarial robustness rather than the offensive capability of foundation models.

Guidelines:

- The answer [N/A] 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: Appendix B.1 discusses the broader impact of LinGuard, covering positive impacts such as more efficient and accessible robust fine-tuning of foundational VLMs, enhanced AI safety in high-stakes applications, and increased social trust in deployed VLM systems, alongside the implicit risk that improved single-step training could be misused if applied outside defensive contexts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We do not release any new pre-trained generative model, image generator, or scraped dataset; LinGuard is a fine-tuning method applied to existing publicly released CLIP backbones, and the produced robust encoders are intended as a defensive contribution rather than a high-risk asset.

Guidelines:

- The answer [N/A] 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: All existing assets used in this paper, including the CLIP backbones, the LLaVA 1.5 framework, the FGSM/RS-FGSM/NFGSM/TeCoA baselines, the AutoAttack evaluation suite, and the 14 zero-shot classification datasets together with the COCO/Flickr30k/TextVQA/VQAv2 generative benchmarks, are publicly released academic resources and are credited via citations in Section 6 and Appendix C; we use them strictly for non-commercial research in accordance with their terms of use.

    Guidelines:

    - The answer [N/A] 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: This paper does not release any new dataset or model at submission time.

    Guidelines:

    - The answer [N/A] 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: This paper does not involve crowdsourcing experiments or any research with human subjects.

Guidelines:

- The answer [N/A] 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: This paper does not involve any human subjects, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as a component of the core methodology of this research. The LLaVA 1.5 framework appears only as an off-the-shelf evaluation backbone for downstream image captioning and VQA in Section 6 and Appendix C, with our LinGuard-tuned vision encoder substituted in a plug-and-play manner.

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.