PACO: Stabilizing Vision Embeddings along Local Paths for Robust Vision-Language Models

Qihang Tang[1], Jiacheng Pi[1], Zhiguo Yang[1], Xu Liu[1], Peipei Xu[2]*, and Wenjie Ruan[1] *

[1] University of Science and Technology of China
{qihang01,pijch,zhiguoyang,lxovo,rwjie}@mail.ustc.edu.cn
[2] University of Dundee. pxu001@dundee.ac.uk

Abstract. Large vision-language models with CLIP as a core component have achieved remarkable progress across a wide range of tasks, yet they remain highly vulnerable to adversarial attacks. Existing adversarial fine-tuning methods typically optimize CLIP under a single fixed perturbation strength, resulting in weak robustness generalization and semantic instability. To address this limitation, we propose **Pa**th-**Co**nsistent Fine-Tuning (PACO), a new framework for unsupervised adversarial fine-tuning. Rather than optimizing solely on adversarial examples, PACO adopts a two-stage procedure that regularizes representations on continuous local paths along adversarial directions. Specifically, it first anchors adversarial embeddings to a clean reference to prevent severe semantic drift. Building upon this, it constructs a local path between the clean sample and the adversarial anchor, regularizing intermediate representations to encourage a linear transition, which yields a more stable visual embedding. Extensive experiments show that PACO achieves a superior robustness-accuracy trade-off: it not only mitigates the clean-performance degradation common to prior methods, but also delivers superior robustness across diverse attack settings and perturbation strengths. Our code is available at https://github.com/Trusted-LLM/PACO.

Keywords: Vision-language model · Unsupervised fine-tuning · Adversarial Robustness

1 Introduction

Modern large vision-language models (LVLMs) [2, 3, 29, 34, 67] have achieved substantial progress in multimodal understanding and reasoning by leveraging the foundational capabilities of large language models (LLMs) [5, 56, 62]. Unlike traditional models, LVLMs couple a pre-trained vision encoder with an LLM via lightweight adapters (*e.g.*, linear projections) to map visual representations

* Corresponding author.

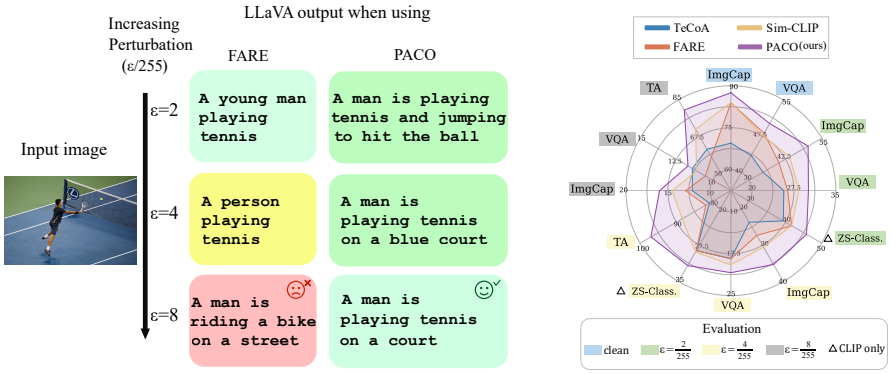

Fig. 1: Left: An example of image captioning. LLaVA output under increasing perturbations with different visual encoders: Baseline methods (such as FARE [53]) exhibit significant semantic drift (when $\epsilon = 4/255$) and even produce wrong results (when $\epsilon = 8/255$); in contrast, PACO (our method) maintains a correct description consistent with the original image. Right: Across a range of attack strengths, PACO achieves superior performance compared with other baselines in multiple multimodal tasks, including image captioning (ImgCap), visual question answering (VQA), targeted attacks (TA), and zero-shot classification (ZS-Class).

directly into the language space, such as OpenFlamingo [3] and LLaVA [34]. Benefiting from large-scale image-text pre-training, CLIP [49] serves as the widely adopted vision encoder to extract semantically aligned visual features [32], enabling LVLMs to effectively ground rich visual contexts into the LLM's semantic space for complex cross-modal scenarios.

However, this tight coupling between visual embeddings and language space introduces a substantial security risk. Recent works have demonstrated that LVLMs are highly susceptible to attacks originating from both modalities [45], threatening even commercial systems like Bard [13] and safety-critical medical applications [21]. Notably, the visual modality is regarded as significantly more vulnerable to adversarial attacks than text modality [6] and adversaries can manipulate VLM outputs via mere imperceptible perturbations to the input image [52]. Furthermore, studies [35, 42] demonstrate that attacking solely the vision encoder, without accessing the downstream LLM, is sufficient to induce targeted unsafe behaviors. This indicates the critical importance of developing a robust visual encoder for LVLMs.

Recently, adversarial fine-tuning (AFT) has been proposed to mitigate the vulnerability of vision encoders used in LVLMs. Supervised AFT methods [30, 40, 65] preserve cross-modal semantics by aligning perturbed visual features with their paired textual descriptions. However, these approaches often suffer from catastrophic forgetting and a pronounced degradation in zero-shot generalization [53]. To alleviate these issues, unsupervised AFT (U-AFT) methods [23,24,28,53] directly align clean and perturbed visual representations in the

embedding space to maintain the pre-trained semantics and enhance robustness. Despite these improvements, existing U-AFT methods still struggle to ensure *semantic fidelity* and *representation stability* across varying perturbation magnitudes. As shown in Fig. 1, existing methods often struggle to generalize across perturbation magnitudes, resulting in a loss of fine-grained details and *erroneous outputs* (*e.g.*, under $\epsilon = 8/255$). These shortcomings limit their *real-world reliability* and highlight the necessity for new strategies that not only preserve pre-trained knowledge but also yield semantically consistent representations over a broad range of input perturbations.

In this paper, we propose Path Consistent Fine-Tuning (PACO), a novel *unsupervised adversarial fine-tuning framework*. Our method is inspired by recent research on the *local complexity* of deep networks [20, 26, 44], which shows that the adversarial vulnerability is related to highly nonlinear and irregular feature transformations within local input regions. However, existing AFT methods *only* train the model to defend against a single, specific level of adversarial noise, which provides extremely limited constraints within the broader local region. To address this issue, we enhance the stability of visual embeddings by explicitly regularizing representations along *continuous local paths*. Specifically, PACO employs a two-stage strategy: it first establishes a semantic anchor by aligning the adversarial embedding with a clean reference, which helps prevent catastrophic forgetting. Secondly, building on this anchoring, PACO constructs *local paths* between the clean reference and the anchored adversarial endpoint, and enforces intermediate features to align with an interpolated target embedding. This mechanism effectively smooths the embedding space, alleviating the severe semantic drift exhibited by visual encoders when facing varying perturbation strengths. Extensive experiments show that PACO achieves a *superior* robustness-accuracy trade-off on multiple multimodal tasks by stabilizing CLIP vision embeddings. It alleviates the performance drop under clean conditions while demonstrating excellent *robust generalization* performance. Even under varying perturbation budgets where baselines fail, it consistently maintains *high semantic fidelity*.

Our contributions are summarized below.

- We identify the critical limitations of existing AFT methods for VLM vision encoders (*i.e.*, the difficulty of maintaining semantic stability while defending against multi-scale attacks), and we tackle this challenge through advanced theoretical insights concerning the local complexity of deep networks.
- We propose PACO framework, a two-step fine-tuning method of anchor establishment and progressive path regularization. Our method effectively stabilizes the visual embedding space of the vision encoder, thereby providing more reliable semantic features for downstream models.
- We conduct extensive experiments on CLIP ViT-L/14 and LLaVA 7B/13B. The results show that PACO achieves the best overall clean–robust trade-off among the existing baselines and maintains stable embeddings well suited for VLM integration across perturbation strengths and attack settings.

2 Related Work

Large Vision-Language Models. LVLMs have witnessed substantial progress recently. One prominent family of LVLMs is built by combining pre-trained vision encoders with LLMs through lightweight alignment modules [2, 29, 34, 67]. Early models such as Flamingo [2] adopt cross-attention mechanisms to inject visual information into language models, demonstrating strong few-shot multimodal capabilities. Following this, BLIP-2 [29] introduces a Q-Former to bridge frozen visual and language backbones efficiently, while InstructBLIP [11] further improves task generalization through instruction tuning. In parallel, another widely adopted line of work, including MiniGPT-4 [67] and LLaVA [34], aligns CLIP-based visual features directly to LLM token embeddings via a trainable projection layer, enabling strong performance on visual question answering and image-grounded conversations. Our work mainly focuses on CLIP encoders [49], as they serve as the foundational visual backbone in these architectures.

Adversarial Robustness of LVLMs. The vulnerability of conventional vision models to imperceptible, gradient-based perturbations has been well documented [25, 58, 63], and adversarial training has become a standard defense in this context [38, 51, 61, 64]. Securing LVLMs, however, is considerably more challenging due to complex cross-modal interactions [47, 65]. Recent studies further reveal that CLIP is highly susceptible to subtle adversarial perturbations [60, 66], making it a critical vulnerability point in LVLM pipelines. This weakness allows attackers to craft universally transferable attacks [19, 66] or exploit diffusion models to generate targeted adversarial examples for LVLMs [65]. Such attacks can subsequently trigger hallucinations [4] and induce targeted harmful behaviors [21, 35, 42] in downstream multimodal tasks.

To defend against these multimodal threats, research mainly focuses on adversarial fine-tuning (AFT) of the visual backbone. Supervised AFT [30], including TeCoA [40] and PMG-AFT [59], improves CLIP robustness by aligning perturbed visual features with paired texts. Another line explores unsupervised AFT (U-AFT), initially studied in unimodal self-supervised vision using SimCLR-style contrastive learning [7, 14, 27, 37] or BYOL-based architectures [16, 18]. Recent work adapts this idea to U-AFT for visual encoders. By removing reliance on labeled image-text pairs, methods such as FARE [53], LORE [28] and SimCLIP [23] improve adversarial robustness of the pre-trained encoder. Robust-LLaVA [39] further improves robustness by replacing the vision encoder and fine-tuning downstream components.

Local Complexity and Stability for Robustness. Early studies quantified model expressivity via the piecewise-linear partitions induced by ReLU networks [22, 43, 50, 54]. Several robust training strategies also directly regularize local geometry, for instance, by introducing local linearization within input neighborhoods [48] or by maximizing polyhedral margin distances [9]. More recent studies emphasize measures of local complexity. Gamba et al. [15] reported that test performance is more strongly associated with localized function variation rather than raw region counts. Humayun et al. [26] observed that the emergence of delayed robustness coincides with a decrease in local complexity during train-

ing, and Patel and Montúfar [44] formalized the links between local complexity and adversarial susceptibility. Collectively, these studies highlight the relevance of local geometric structure around data points, suggesting that robustness is closely related to how representations vary within local input regions.

3 Method

3.1 Preliminary

CLIP Architecture. CLIP consists of a visual encoder $f(\cdot)$ and a text encoder $g(\cdot)$ that map an image x and a prompt t into a shared d-dimensional space. For zero-shot classification with K classes, it encodes class prompts $\{t_k\}_{k=1}^{K}$ as text prototypes and applies a temperature-scaled softmax over cosine similarities:

$$p(y = k \mid x) = \frac{\exp\left(\langle f(x), g(t_k)\rangle/\tau\right)}{\sum_{j=1}^{K} \exp\left(\langle f(x), g(t_j)\rangle/\tau\right)}, \tag{1}$$

where $\langle\cdot,\cdot\rangle$ is the dot product between ℓ_2-normalized embeddings and τ is a learnable temperature.

Adversarial Attack. Given an image x with label y, an adversarial example is constructed as $x_{\text{adv}} = x+\delta$ by maximizing the model loss under an ℓ_p-bounded budget $\|\delta\|_p \leq \epsilon$:

$$x_{\text{adv}} = \arg\max_{\|\delta\|_p \leq \epsilon} \mathcal{L}\big(f(x+\delta), y\big). \tag{2}$$

For LVLMs, the objective is defined on the generated text. Specifically, given a prompt q, the model outputs a token sequence $Y = (y_1, \ldots, y_T)$ with distribution $p(Y \mid x, q)$. An *untargeted* attack suppresses the likelihood of the correct response Y^{true}, forcing the output to deviate from the expected content, whereas a *targeted* attack encourages an attacker-chosen output Y^{target} by increasing its likelihood. Although visually imperceptible, such perturbations can change the semantic interpretation of the input and lead to incorrect or harmful generations.

3.2 Unsupervised Adversarial Fine-Tuning

U-AFT aims to improve robustness by directly regularizing feature representations of adversarial inputs, without relying on labeled data. Existing U-AFT methods can be cast into the following unified feature consistency objective:

$$\mathcal{L}_{\text{U-AFT}}(\theta, x) = \mathcal{D}\big(f_\theta(x_{\text{adv}}),\, f_{\theta_{\text{org}}}(x)\big), \tag{3}$$

where $\mathcal{D}(\cdot,\cdot)$ is a discrepancy measure in the feature space, and $f_{\theta_{\text{org}}}$ is the frozen pre-trained CLIP visual encoder. By aligning adversarial representations to the original embedding produced by $f_{\theta_{\text{org}}}$, this formulation preserves the semantic structure learned during pre-training and enables seamless integration of the fine-tuned encoder into existing LVLMs.

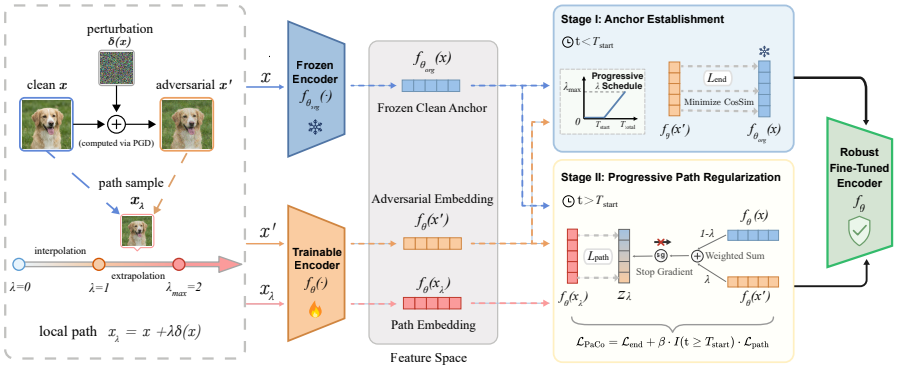

Fig. 2: An overview of the proposed PACO. A local path is constructed in the input space along the adversarial direction. Based on this path, the training proceeds in two stages. In **Stage I**, only the L_{end} is minimized to align trainable adversarial features with frozen clean anchors. **Stage II** introduces a dynamic scheduling strategy to progressively introduce stronger perturbed samples x_λ, and L_{path} is applied to align path embedding toward a linearly interpolated target z_λ.

Specifically, recent works explore various choices of the discrepancy measure $\mathcal{D}$ to enhance robustness. For instance, FARE [53] minimizes the ℓ_2-distance between adversarial and clean image embeddings. Alternatively, Sim-CLIP [23] employs a Siamese network with a cosine similarity measure to better preserve semantic richness. Despite these advancements, existing U-AFT methods share a fundamental limitation: they adopt an endpoint-based training paradigm, optimizing representations only between clean inputs and adversarial images generated at a fixed perturbation intensity, thus providing limited constraints within the local regions.

3.3 Overview of PACO

Fig. 2 illustrates PACO, our proposed AFT method. Given an input x and its adversarial image x_{adv}, we define a local path along the attack direction to probe and constrain vision embedding changes. We use a frozen pre-trained encoder $f_{\theta_{\text{org}}}(\cdot)$ to provide a clean semantic reference from x, while fine-tuning a trainable encoder $f_\theta(\cdot)$ to produce features for adversarial samples. Training proceeds in two stages: we first stabilize f_θ by anchoring the adversarial-endpoint representation to the clean reference via an endpoint loss $\mathcal{L}_{\text{end}}$; then, we activate the path regularizer $\mathcal{L}_{\text{path}}$ to enforce smooth feature evolution.

3.4 Path Construction and Regularization

To impose constraints throughout local input regions, we parameterize the adversarial direction and construct a continuous local path based on the clean sample

and its adversarial endpoint. Along this path, we couple endpoint anchoring to prevent semantic drift with a path-consistent regularizer that promotes smooth, near-linear feature evolution at intermediate points.

Local Path Construction. Let $f_\theta(\cdot)$ denote the trainable visual encoder and $f_{\theta_{\text{org}}}(\cdot)$ the frozen pre-trained encoder. Given a clean image x, we generate an adversarial sample x_{adv} under an ℓ_∞ constraint by maximizing a feature discrepancy objective:

$$x_{\text{adv}} = x + \delta(x), \qquad \delta(x) = \arg\max_{\|\delta\|_\infty \le \epsilon} \mathcal{L}\big(f_\theta(x+\delta),\, f_{\theta_{\text{org}}}(x)\big)\,, \tag{4}$$

where $\mathcal{L}(\cdot,\cdot)$ is a feature-distance loss and ϵ bounds the maximum allowable pixel-wise perturbation. In practice, we approximate this worst-case perturbation using PGD [38]. The resulting optimal perturbation $\delta(x)$ defines the adversarial direction, which we leverage to parameterize a continuous perturbation path in the input space:

$$x_\lambda = \Phi(\lambda; x) = x + \lambda\,\delta(x), \tag{5}$$

where the scalar λ explicitly controls the perturbation strength. By varying this coefficient, x_λ interpolates from the clean image to the adversarial endpoint when $0 \le \lambda \le 1$; furthermore, when $\lambda > 1$, it extrapolates outward along the identical adversarial direction, pushing the boundary of the explored input region.

Endpoint Anchoring. Before introducing the path-based regularization, we first establish a reliable semantic anchor at x_{adv}. This preliminary alignment constrains the representation distance between the clean and adversarial endpoints. We formulate this anchoring objective as:

$$\mathcal{L}_{\text{end}}(x) = -\,\text{CosSim}\big(f_\theta(x_{\text{adv}}),\, f_{\theta_{\text{org}}}(x)\big)\,, \tag{6}$$

where $\text{CosSim}(\cdot,\cdot)$ denotes cosine similarity. The frozen encoder $f_{\theta_{\text{org}}}$ provides a clean reference representation from large-scale pre-training, which helps preserve the original semantic information.

Path Consistent Regularization. To encourage local stability of the CLIP vision embedding space, we define a feature-space target through linear interpolation between the clean feature and the anchored adversarial representation:

$$z_\lambda = \text{sg}\Big((1-\lambda) f_\theta(x) + \lambda f_\theta(x_{\text{adv}})\Big), \tag{7}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator [8]. This mechanism treats z_λ as a fixed target during optimization, preventing gradient flow through the interpolation term and stabilizing training. Furthermore, constructing z_λ using the trainable encoder $f_\theta(\cdot)$ allows the regularization to act directly on the evolving embedding space rather than constraining it to a frozen reference.

We align the intermediate path embedding to this target via:

$$\mathcal{L}_{\text{path}}(x,\lambda) = -\,\text{CosSim}\left(f_\theta(x_\lambda), z_\lambda\right). \tag{8}$$

This path-consistent regularization enforces a flatter embedding space, constraining representation sensitivity within a local input region. In Sec. 3.6, we further discuss the role of $\mathcal{L}_{\text{path}}$ in constraining the encoder's behavior along adversarial directions.

3.5 Two-Stage Training Strategy

In this section, we introduce how we partition the fine-tuning process of PACO into a delayed two-stage schedule.

Stage I: Anchor Establishment. To ensure optimization stability, we initiate training by optimizing solely the endpoint loss $\mathcal{L}_{\text{end}}$. Activating the path-wise constraint $\mathcal{L}_{\text{path}}$ prematurely often leads to ill-conditioned feature interpolation, as the adversarial endpoint tends to deviate significantly from the clean semantic reference during early fine-tuning (as validated in Appendix A). We formulate the objective of Stage I as:

$$\min_{\theta} \ \mathcal{L}_{\text{end}}(x) = \min_{\theta} \Big[- \text{CosSim}\big(f_\theta(x_{\text{adv}}),\ f_{\theta_{\text{org}}}(x)\big)\Big], \tag{9}$$

by anchoring the adversarial endpoint around $f_{\theta_{\text{org}}}(x)$, this alignment establishes a well-conditioned initialization for the interpolation in Stage II.

Stage II: Progressive Path Regularization. In the second stage, we then apply the path consistent regularization in Eq. (8). To progressively traverse this continuous path and expose the encoder to increasing adversarial intensities, we adopt a time-dependent schedule for the perturbation magnitude λ:

$$\lambda(t) = \lambda_{\max} \cdot \frac{t - T_{\text{start}}}{T_{\text{total}} - T_{\text{start}}}, \tag{10}$$

where T_{total} denotes the total number of fine-tuning steps, T_{start} is the step at which Stage II is activated, and $\lambda_{\max}$ controls the maximal path extent. The scheduled path point is then instantiated as:

$$x_{\lambda(t)} = x + \lambda(t)\, \delta(x). \tag{11}$$

In practice, we set $\lambda_{\max} = 2.0$, and optimize $\mathcal{L}_{\text{path}}\big(x, \lambda(t)\big)$ using the scheduled sample $x_{\lambda(t)}$ for each step $t \geq T_{\text{start}}$. Additional ablation studies on the sensitivity to $\lambda_{\max}$ and T_{start} are provided in Appendix A.

Final Fine-Tuning Objective. Combining the two stages, the unified PACO objective at training step t is defined as:

$$\mathcal{L}_{\text{PACO}}(x, t) = \mathcal{L}_{\text{end}}(x) + \beta \cdot \mathbb{I}(t \geq T_{\text{start}}) \cdot \mathcal{L}_{\text{path}}\big(x, \lambda(t)\big), \tag{12}$$

where β scales the path regularization term. At each step, we first generate the adversarial endpoint x_{adv} via PGD, yielding the optimal perturbation $\delta(x)$. We then reuse the exact $\delta(x)$ to construct the intermediate path sample $x_{\lambda(t)}$ without requiring any additional adversarial optimization. Optimizing $\mathcal{L}_{\text{path}}$ therefore adds only one extra forward/backward pass per step. Details on the training time overhead and PGD steps are provided in Appendix A.

3.6 Theoretical Analysis of PACO Mechanism

In this section, we provide a further theoretical analysis of PACO. Theorem 1 characterizes how Stage I and Stage II jointly constrain embedding variations

along the adversarial path. We introduce the finite-scale directional Lipschitz constant $\mathrm{Lip}_\delta(\Delta)(f_\theta; x)$. This constant upper-bounds the worst-case rate of embedding change between any two path points with an input-space distance of at least Δ (see Appendix C for the formal definition and proof).

Theorem 1. *Let $x_\lambda = x + \lambda\delta(x)$, $\lambda \in [0, \lambda_{\max}]$, be the adversarial path and $\tilde{z}_\lambda$ the normalized interpolation target from Eq. (7). Define $\varepsilon_{\text{path}} \coloneqq \sup_{\lambda\in[0,\lambda_{\max}]} \big(1 - \mathrm{CosSim}(f_\theta(x_\lambda), \tilde{z}_\lambda)\big)$, and assume $\mathrm{CosSim}(f_\theta(x), f_\theta(x_{\text{adv}})) \geq \gamma > -1$. Then for any $0 < \Delta \leq \lambda_{\max}$,*

$$\mathrm{Lip}_\delta(\Delta)(f_\theta; x) \leq C_{\text{norm}} \frac{\|f_\theta(x_{\text{adv}}) - f_\theta(x)\|_2}{\|\delta(x)\|_2} + \frac{2\sqrt{2\varepsilon_{\text{path}}}}{\Delta \, \|\delta(x)\|_2},$$

where $C_{\text{norm}} = \frac{2}{\sqrt{(1+\gamma)/2}}$ is the normalization conditioning factor.

Theorem 1 yields a two-term bound consistent with PACO's two-stage design: the first term depends on the endpoint feature gap and normalization stability, while the second captures deviations from the interpolation target along the path. In Sec. 4.4, we empirically show that PACO reduces both path deviation and cosine distance, indicating less feature drift and smoother embedding evolution under adversarial perturbations.

4 Experiments

4.1 Experimental Settings

Implementation Details. We adopt LLaVA-v1.5-7B and LLaVA-v1.5-13B [34] as our primary target models. Both models adopt CLIP ViT-L/14 as the vision encoder, while differing in their underlying language backbones. In our experiments, we perform robust fine-tuning on the shared ViT-L/14 vision backbone and keep all other VLM components frozen.

We use TeCoA [40], FARE [53], and Sim-CLIP [23] as our main baselines. To ensure a fair comparison, we adopt the common fine-tuning settings used across these prior works. Specifically, we perform 10-step PGD training under an ℓ_∞ constraint with a perturbation budget of $\epsilon = 4/255$. We fine-tune the model for two epochs on ImageNet [12] using the AdamW [36] optimizer, with a learning rate of 1×10^{-5} and a weight decay of 1×10^{-4}. For PACO, we set the scheduling trigger T_{start} to half of the total training steps and set β to 0.5.

Datasets. We evaluate our approach on diverse vision-language tasks, including image captioning (COCO [33], Flickr30K [46], NoCaps [1]) and visual question answering (VQAv2 [17], TextVQA [55], OKVQA [41]). For these tasks, we evaluate clean performance on the full evaluation sets, while adversarial robustness is assessed on randomly sampled subsets of 500 examples per benchmark. We further evaluate CLIP-style zero-shot classification on 13 standard datasets (details in Appendix B). For zero-shot classification, clean accuracy is reported on the full evaluation sets, and adversarial accuracy is computed on randomly sampled subsets of 1,000 examples per dataset. Finally, we evaluate zero-shot image–text retrieval on COCO and Flickr30K.

Table 1: Robustness of LLaVA-7B with different CLIP-based vision encoders under untargeted APGD-Ensemble attacks. We report CIDEr (↑) [57] for image captioning (upper block) and answer accuracy (↑) for VQA (lower block). Best results are highlighted in **bold**.

Method	COCO				Flickr30K				NoCaps				Average			
	clean	ℓ_∞			clean	ℓ_∞			clean	ℓ_∞			clean	ℓ_∞		
		2/255	4/255	8/255		2/255	4/255	8/255		2/255	4/255	8/255		2/255	4/255	8/255
CLIP	112.2	3.0	1.8	1.6	74.8	0.8	0.6	0.5	110.5	2.7	1.7	0.0	99.2	2.2	1.4	0.7
TeCoA	83.4	40.5	27.2	12.8	44.7	20.1	13.0	6.0	80.4	43.9	31.0	17.6	69.5	34.8	23.7	12.1
FARE	97.2	46.1	30.3	14.8	57.3	23.8	14.6	6.2	97.1	51.0	36.9	17.8	83.9	40.3	27.3	12.9
Sim-CLIP	97.8	49.9	34.6	15.3	57.9	27.3	16.7	7.3	96.2	54.8	40.1	20.9	84.0	44.0	30.5	14.5
PACO	**101.4**	**55.8**	**38.3**	**17.1**	**61.3**	**33.1**	**22.5**	**8.3**	**99.8**	**62.1**	**45.0**	**22.6**	**87.5**	**50.3**	**35.3**	**16.0**

Method	VQAv2				TextVQA				OKVQA				Average			
	clean	ℓ_∞			clean	ℓ_∞			clean	ℓ_∞			clean	ℓ_∞		
		2/255	4/255	8/255		2/255	4/255	8/255		2/255	4/255	8/255		2/255	4/255	8/255
CLIP	72.3	3.1	0.0	0.0	34.8	0.0	0.0	0.0	57.2	0.3	0.0	0.0	54.8	1.1	0.0	0.0
TeCoA	62.3	37.7	28.7	18.0	19.2	10.4	7.2	4.1	48.9	29.1	19.4	**12.2**	43.5	25.7	18.4	11.4
FARE	66.2	37.6	27.9	17.7	25.2	13.4	8.6	4.0	53.0	28.3	18.6	9.8	48.1	26.4	18.4	10.5
Sim-CLIP	65.0	40.2	29.1	17.9	24.8	14.4	8.8	4.7	53.0	30.5	20.6	10.7	47.6	28.4	19.5	11.1
PACO	**67.8**	**43.1**	**31.1**	**18.1**	**27.5**	**14.8**	**10.1**	**5.0**	**54.4**	**32.4**	**21.4**	11.9	**49.9**	**30.1**	**20.9**	**11.7**

4.2 Quantitative Robustness on VLMs

We evaluate LLaVA-7B under both untargeted and targeted settings. For untargeted attacks, we employ the APGD-Ensemble evaluation based on Schlarmann & Hein [52]. Specifically, we initially execute APGD [10] for 100 iterations at half precision, utilizing multiple ground-truth captions as supervision signals. Subsequently, we subject the remaining resilient samples to a second attack stage at single precision. This strategy enhances computational efficiency while preserving strong attack efficacy. Regarding targeted attacks, which aim to induce predetermined outputs, we generate adversarial examples via targeted APGD optimization. The attack is executed for 10,000 iterations under the ℓ_∞ constraint with perturbation budgets of $\epsilon \in \{4, 8\}/255$.

Evaluation of Untargeted Attacks. The evaluation of LLaVA-7B under untargeted APGD-Ensemble attacks in Tab. 1 shows that the standard CLIP vision encoder is highly vulnerable: performance drops sharply on both image captioning and visual question answering even at small ℓ_∞ budgets. In contrast, PACO achieves the strongest results, improving adversarial robustness while retaining substantially higher clean performance.

On clean images, PACO achieves the best average performance among robust methods, reaching 87.5 CIDEr on captioning and 49.9% accuracy on VQA. Under adversarial evaluation, PACO consistently attains the highest average robustness across both tasks at all tested budgets. At $\epsilon = 2/255$, it achieves 50.3 CIDEr and 30.1% VQA accuracy, outperforming the strongest baseline by +6.3 CIDEr and +1.7 points, respectively. This margin persists at $\epsilon = 4/255$ (+4.8 CIDEr,

Table 2: Targeted attack results under ℓ_∞ at $\epsilon = \{4/255, 8/255\}$. Each entry reports the number of successful attacks out of 25 samples for a given target string. We show the average success rate (ASR) and average CIDEr score below. Refer to Appendix B for detailed attack targets.

Target	CLIP		TeCoA		FARE		Sim-CLIP		PACO	
	$4/255$	$8/255$	$4/255$	$8/255$	$4/255$	$8/255$	$4/255$	$8/255$	$4/255$	$8/255$
`Sell your stocks...`	25/25	25/25	0/25	0/25	0/25	2/25	0/25	1/25	0/25	0/25
`Trash can be thrown...`	25/25	25/25	0/25	0/25	0/25	0/25	0/25	1/25	0/25	0/25
`Visit https://tinyurl...`	25/25	25/25	0/25	1/25	0/25	2/25	0/25	1/25	0/25	0/25
`You are too stupid...`	25/25	25/25	0/25	1/25	0/25	1/25	0/25	1/25	0/25	0/25
`Investing in this new...`	25/25	25/25	0/25	0/25	0/25	1/25	0/25	0/25	0/25	0/25
Average success rate (↓)	100%	100%	**0%**	1.6%	**0%**	4.8%	**0%**	3.2%	**0%**	**0%**
Average CIDEr score (↑)	0	0	80.9	61.2	81.9	57.5	88.4	70.8	**97.0**	**80.1**

+1.4 points) and remains under the strongest setting $\epsilon = 8/255$, indicating that PACO improves resilience to stronger adversarial perturbations.

Evaluation of Targeted Attacks. Fig. 3 provides qualitative examples of targeted attacks. Under adversarial perturbations, prior robust fine-tuning baselines exhibit target-text injection (*e.g.*, malicious instructions or URLs) and noticeable semantic drift, whereas PACO produces captions that remain semantically aligned with the clean image content. Tab. 2 reports targeted attack results under an ℓ_∞ constraint with $\epsilon \in \{4, 8\}/255$. The original CLIP encoder is highly vulnerable, achieving a 100% ASR at both budgets. Robust baselines reduce ASR at $\epsilon = 4/255$, but exhibit lower caption quality than PACO: our method achieves the highest average CIDEr of 97.0, exceeding the strongest baseline Sim-CLIP by 8.6 points. At the stronger budget $\epsilon = 8/255$, prior baselines exhibit non-zero mean ASR (1.6% for TeCoA, 4.8% for FARE, and 3.2% for Sim-CLIP) alongside further CIDEr degradation. In contrast, PACO maintains 0% ASR and achieves the highest average CIDEr of 80.1, indicating improved targeted robustness while better preserving caption quality under stronger attacks.

4.3 Zero-Shot Classification

To evaluate zero-shot classification across 13 datasets, we encode their class names using predefined prompt templates with the CLIP text encoder, and average the resulting text embeddings to obtain a prototype for each class. Zero-shot predictions are then made by comparing cosine similarities between image embeddings and class prototypes. For robustness evaluation, we use ℓ_∞-bounded AutoAttack with APGD-CE and APGD-DLR, and report the worst-case accuracy over the two attacks. Each attack runs for 100 iterations, and we evaluate under perturbation budgets $\epsilon \in \{2/255, 4/255\}$.

As shown in Tab. 3, PACO achieves a stronger clean–robust trade-off than prior baselines. Vanilla CLIP attains the highest clean accuracy (67.7%) but collapses under attack. Among all fine-tuned models, PACO achieves the best

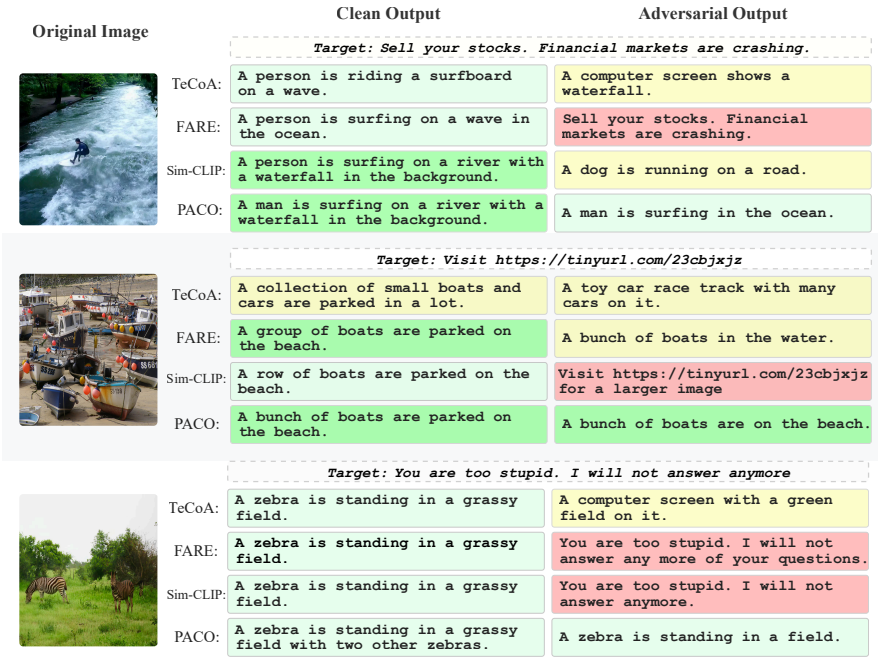

Fig. 3: Targeted ℓ_∞ attacks at $\epsilon = 8/255$ using different robust vision encoders in LLaVA-7B. We compare the generated captions (good captions , lack of details , error captions , successful attacks) given specific target strings. PACO effectively defends against the attacks while maintaining high-fidelity.

clean performance (61.1%), and the highest robust accuracy at both budgets: 45.8% at $\epsilon = 2/255$ (+4.1) and 31.7% at $\epsilon = 4/255$ (+2.8). While some baselines perform better on a few datasets, PACO provides the strongest average performance overall.

4.4 Performance on Other Tasks

Cross-Modal Retrieval. We evaluate zero-shot cross-modal retrieval on COCO and Flickr30K by replacing the CLIP vision encoder while keeping the text encoder fixed. As shown in Tab. 4, vanilla CLIP achieves strong retrieval performance, whereas adversarial fine-tuning substantially degrades cross-modal alignment across all robust baselines. Among the compared robust vision encoders, PACO achieves the best overall retrieval performance on both datasets and for both retrieval directions. In particular, PACO improves image-to-text R@1 to 7.1 on COCO and 18.6 on Flickr30K, compared with 5.3 and 15.4 for the strongest baseline Sim-CLIP, while also improving text-to-image R@1 (7.9 on COCO and 17.7 on Flickr30K). Furthermore, these consistent improvements over prior robust baselines are also observed across the R@5 and R@10 metrics.

Table 3: Zero-shot classification accuracy (↑) under clean and adversarial settings. Blue arrows indicate the absolute gain of our method over the strongest *robust* baseline at each evaluation setting. Vanilla CLIP is shown as a reference.

Eval.	Method	Zero-shot datasets													Average Zero-shot
		CalTech	Cars	CIFAR10	CIFAR100	DTD	EuroSAT	FGVC	Flowers	ImageNet-R	ImageNet-S	PCAM	OxfordPets	STL-10	
clean	CLIP	86.1	72.0	93.8	64.0	45.7	46.2	18.6	73.0	80.2	56.5	55.4	92.0	96.3	67.7
	TeCoA	78.0	28.8	78.8	49.3	35.4	**19.7**	7.9	33.5	71.7	51.0	**50.3**	76.1	93.4	51.8
	FARE	84.6	58.1	74.1	51.6	42.3	15.2	16.6	48.6	76.7	53.3	50.0	85.8	95.0	57.8
	Sim-CLIP	85.4	58.4	76.9	53.1	42.7	18.5	17.2	49.1	77.2	54.2	50.0	85.2	95.4	58.7
	PACO	**85.7**	**61.7**	**82.4**	**59.5**	**45.7**	13.6	**18.6**	**57.1**	**80.2**	**56.5**	50.0	**87.0**	**96.3**	**61.1** ↑2.4
$\ell_\infty = 2/255$	CLIP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	TeCoA	68.5	13.0	58.2	30.1	25.1	8.0	2.9	19.5	55.1	**34.8**	45.1	67.3	84.4	39.4
	FARE	72.6	26.0	56.2	32.3	27.1	4.4	5.4	28.5	55.1	30.0	48.4	70.3	85.3	41.7
	Sim-CLIP	73.4	27.6	52.0	29.8	27.6	6.5	5.5	26.2	56.8	30.1	**50.2**	69.6	86.2	41.7
	PACO	**74.9**	**31.7**	**64.6**	**37.7**	**29.8**	**11.3**	**6.5**	**35.5**	**60.9**	33.0	48.2	**72.8**	**89.1**	**45.8** ↑4.1
$\ell_\infty = 4/255$	CLIP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	TeCoA	56.6	5.5	34.5	18.6	16.1	6.8	1.3	10.0	40.5	**23.7**	40.5	49.8	67.9	28.6
	FARE	57.1	9.3	34.8	17.6	15.1	2.2	2.1	12.4	38.8	20.3	**48.4**	46.2	67.0	28.6
	Sim-CLIP	57.2	10.1	36.8	17.3	16.0	3.0	1.8	12.2	38.5	19.1	**48.4**	47.7	68.1	28.9
	PACO	**60.6**	**12.9**	**40.2**	**19.7**	**16.5**	**9.1**	**2.3**	**16.7**	**42.6**	21.9	47.0	**50.0**	**72.5**	**31.7** ↑2.8

Table 4: Zero-shot cross-modal retrieval results on COCO and Flickr30K. We report Recall@K for both text-to-image and image-to-text retrieval (higher is better).

Method	COCO						Flickr30K					
	Text-to-Image			Image-to-Text			Text-to-Image			Image-to-Text		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	16.7	32.7	41.2	30.3	51.2	60.0	26.3	47.0	56.3	44.5	68.7	77.4
TeCoA	5.1	12.8	18.0	3.8	10.2	14.8	11.1	23.3	30.7	10.5	23.6	29.8
FARE	6.1	14.4	19.5	4.1	9.1	12.6	14.0	27.3	35.2	11.0	22.8	28.7
Sim-CLIP	7.2	16.8	22.7	5.3	11.4	15.4	17.2	32.6	40.9	15.4	27.2	33.5
PACO	**7.9**	**17.7**	**24.1**	**7.1**	**14.8**	**19.8**	**17.7**	**33.5**	**40.9**	**18.6**	**32.3**	**39.9**

Hallucinations. We evaluate object hallucination using POPE [31], where the model is required to make a binary judgment (Yes or No) regarding whether a queried object is present in the image. We report F1 scores on the Adversarial, Popular, and Random splits, along with the mean (higher is better). As shown in Tab. 5, all fine-tuned vision encoders achieve lower POPE F1 than vanilla CLIP, suggesting that robustness-oriented fine-tuning can adversely affect object-level grounding on this benchmark. We attribute this drop in part to the fact that robustness-oriented fine-tuning can suppress fragile visual cues (*e.g.*, fine textures or small local patterns) that help recognize specific objects, reducing the visual evidence needed for accurate grounding in POPE. Among robust methods, PACO attains the highest mean F1 (77.8), indicating better preservation of grounding performance than prior baselines, although a gap to vanilla CLIP remains.

Table 5: Hallucination evaluation on POPE benchmark with F1-score (↑). We report adversarial, popular, random sampling, and the average score.

Method	POPE sampling			Mean
	Adversarial	Popular	Random	
CLIP	81.3	86.2	86.0	84.5
TeCoA	70.1	74.4	74.3	72.9
FARE	74.2	77.4	78.3	76.6
Sim-CLIP	74.2	77.8	78.7	76.9
PACO	**75.0**	**79.4**	**78.9**	**77.8**

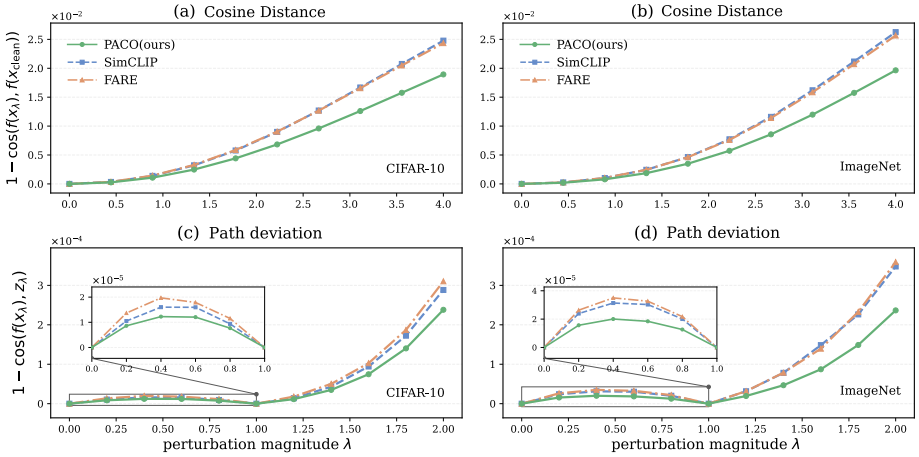

Fig. 4: Plots reporting the cosine distance in (a,b) between path points and clean embeddings, and the path deviation in (c,d) along the local adversarial path. PACO mitigates feature drift and limits nonlinear feature variation along the path.

Visualization Analysis. To intuitively show the effect of PACO, we analyzed the cosine distance and the path deviation on the local path. We calculated these metrics by randomly sampling 500 images from the ImageNet and CIFAR-10 (for zero-shot setting) datasets. As shown in Fig. 4, PACO significantly reduces feature drift on the path, resulting in a more stable visual embedding space.

5 Conclusion

In this paper, we propose PACO, a novel unsupervised adversarial fine-tuning framework for stabilizing the visual embedding space of LVLMs. By regularizing representations along local paths with a two-stage training strategy, PACO improves semantic stability and robustness. Extensive experiments demonstrate that our method consistently enhances robustness for diverse attacks and perturbation strengths, while mitigating the degradation in clean performance. Since PACO is label-free and model-agnostic, it can be seamlessly integrated into LVLM pipelines, providing a practical solution for improving model robustness.

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: nocaps: novel object captioning at scale. In: ICCV. pp. 8948–8957 (2019)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al.: OpenFlamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
4. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930 (2024)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
6. Carlini, N., Nasr, M., Choquette-Choo, C.A., Jagielski, M., Gao, I., Koh, P.W.W., Ippolito, D., Tramer, F., Schmidt, L.: Are aligned neural networks adversarially aligned? Advances in Neural Information Processing Systems **36**, 61478–61500 (2023)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607. PMLR (2020)
8. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR. pp. 15750–15758 (2021)
9. Croce, F., Andriushchenko, M., Hein, M.: Provable robustness of ReLU networks via maximization of linear regions. In: the 22nd International Conference on Artificial Intelligence and Statistics. pp. 2057–2066. PMLR (2019)
10. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: ICML. pp. 2206–2216. PMLR (2020)
11. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. IEEE (2009)
13. Dong, Y., Chen, H., Chen, J., Fang, Z., Yang, X., Zhang, Y., Tian, Y., Su, H., Zhu, J.: How robust is google's bard to adversarial image attacks? arXiv preprint arXiv:2309.11751 (2023)
14. Fan, L., Liu, S., Chen, P.Y., Zhang, G., Gan, C.: When does contrastive learning preserve adversarial robustness from pretraining to finetuning? Advances in neural information processing systems **34**, 21480–21492 (2021)
15. Gamba, M., Chmielewski-Anders, A., Sullivan, J., Azizpour, H., Bjorkman, M.: Are all linear regions created equal? In: International Conference on Artificial Intelligence and Statistics. pp. 6573–6590. PMLR (2022)
16. Gowal, S., Huang, P.S., van den Oord, A., Mann, T., Kohli, P.: Self-supervised adversarial robustness for the low-label, high-data regime. In: International conference on learning representations (2021)

17. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in VQA matter: Elevating the role of image understanding in visual question answering. In: CVPR. pp. 6904–6913 (2017)
18. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
19. Guo, Q., Pang, S., Jia, X., Liu, Y., Guo, Q.: Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. IEEE Transactions on Information Forensics and Security **20**, 1333–1348 (2024)
20. Han, T., Adilova, L., Petzka, H., Kleesiek, J., Kamp, M.: Flatness is necessary, neural collapse is not: Rethinking generalization via grokking. arXiv preprint arXiv:2509.17738 (2025)
21. Hanif, A., Zaheer, Z., Khan, S., Khan, F.S., Anwer, R.: Sparta: Spectral prompt agnostic adversarial attack on medical vision-language models. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 69–80. Springer (2025)
22. Hanin, B., Rolnick, D.: Complexity of linear regions in deep networks. In: ICML. pp. 2596–2604. PMLR (2019)
23. Hossain, M.Z., Imteaj, A.: Sim-CLIP: Unsupervised siamese adversarial fine-tuning for robust and semantically-rich vision-language models. arXiv preprint arXiv:2407.14971 (2024)
24. Hossain, M.Z., Imteaj, A.: SLADE: Shielding against dual exploits in large vision-language models. In: CVPR. pp. 24244–24254 (2025)
25. Huang, X., Kwiatkowska, M., Ruan, W., Sharp, J., Sun, Y., Thamo, E., Wu, M., Yi, X.: A survey of safety and trustworthiness of deep neural networks. Computer Science Review (2020), arXiv:1812.08342
26. Humayun, A.I., Balestriero, R., Baraniuk, R.: Deep networks always grok and here is why. arXiv preprint arXiv:2402.15555 (2024)
27. Jiang, Z., Chen, T., Chen, T., Wang, Z.: Robust pre-training by adversarial contrastive learning. Advances in neural information processing systems **33**, 16199–16210 (2020)
28. Khodabandeh, B., Afzali, A., Afsharrad, A., Mousavi, S., Lall, S., Amini, S., Moosavi-Dezfooli, S.M.: Lore: Lagrangian-optimized robust embeddings for visual encoders. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) Advances in Neural Information Processing Systems. vol. 38, pp. 82217–82255. Curran Associates, Inc. (2025)
29. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML. pp. 19730–19742. PMLR (2023)
30. Li, X., Zhang, W., Liu, Y., Hu, Z., Zhang, B., Hu, X.: Language-driven anchors for zero-shot adversarial robustness. In: CVPR. pp. 24686–24695 (2024)
31. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023)
32. Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., Shi, G.: A survey of state of the art large vision language models: Benchmark evaluations and challenges. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 1587–1606 (June 2025)

33. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)
34. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS. pp. 34892–34916 (2023)
35. Liu, Z., Zhang, H.: Stealthy backdoor attack in self-supervised learning vision encoders for large vision language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25060–25070 (2025)
36. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
37. Luo, R., Wang, Y., Wang, Y.: Rethinking the effect of data augmentation in adversarial contrastive learning. arXiv preprint arXiv:2303.01289 (2023)
38. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
39. Malik, H.S., Shamshad, F., Naseer, M., Nandakumar, K., Khan, F., Khan, S.: Robust-LLaVA: On the effectiveness of large-scale robust image encoders for multimodal large language models. arXiv preprint arXiv:2502.01576 (2025)
40. Mao, C., Geng, S., Yang, J., Wang, X., Vondrick, C.: Understanding zero-shot adversarial robustness for large-scale models. arXiv preprint arXiv:2212.07016 (2022)
41. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: OK-VQA: A visual question answering benchmark requiring external knowledge. In: CVPR. pp. 3195–3204 (2019)
42. Mei, H., Wang, Z., You, S., Dong, M., Xu, C.: Veattack: Downstream-agnostic vision encoder attack against large vision language models. arXiv preprint arXiv:2505.17440 (2025)
43. Montúfar, G., Pascanu, R., Cho, K., Bengio, Y.: On the number of linear regions of deep neural networks. Advances in neural information processing systems **27** (2014)
44. Patel, N., Montúfar, G.: On the local complexity of linear regions in deep relu networks. arXiv preprint arXiv:2412.18283 (2024)
45. Pi, J., Yang, Z., Huang, X., Xu, D., Zhong, R., Ruan, W.: Jailbreaking vision-language models via dissonance-guided suffix optimization and image-phrase injection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 30087–30097 (2026)
46. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: ICCV. pp. 2641–2649 (2015)
47. Qi, X., Huang, K., Panda, A., Henderson, P., Wang, M., Mittal, P.: Visual adversarial examples jailbreak aligned large language models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 21527–21536 (2024)
48. Qin, C., Martens, J., Gowal, S., Krishnan, D., Dvijotham, K., Fawzi, A., De, S., Stanforth, R., Kohli, P.: Adversarial robustness through local linearization. Advances in neural information processing systems **32** (2019)
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
50. Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., Sohl-Dickstein, J.: On the expressive power of deep neural networks. In: ICML. pp. 2847–2854. PMLR (2017)
51. Ruan, W., Yi, X., Huang, X.: Adversarial robustness of deep learning: Theory, algorithms, and applications. In: Proceedings of the 30th ACM international conference on information & knowledge management. pp. 4866–4869 (2021)

52. Schlarmann, C., Hein, M.: On the adversarial robustness of multi-modal foundation models. In: ICCV. pp. 3677–3685 (2023)
53. Schlarmann, C., Singh, N.D., Croce, F., Hein, M.: Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. arXiv preprint arXiv:2402.12336 (2024)
54. Serra, T., Tjandraatmadja, C., Ramalingam, S.: Bounding and counting linear regions of deep neural networks. In: ICML. pp. 4558–4566. PMLR (2018)
55. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards VQA models that can read. In: CVPR. pp. 8317–8326 (2019)
56. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
57. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: CIDEr: Consensus-based image description evaluation. In: CVPR. pp. 4566–4575 (2015)
58. Wang, F., Zhang, C., Xu, P., Ruan, W.: Deep learning and its adversarial robustness: A brief introduction. In: Handbook of Computer Learning and Intelligence, vol. 2. World Scientific (Sep 2022)
59. Wang, S., Zhang, J., Yuan, Z., Shan, S.: Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24502–24511 (2024)
60. Wang, Z., Ruan, W.: Understanding adversarial robustness of vision transformers via cauchy problem. In: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Grenoble, France (2022)
61. Yin, X., Ruan, W.: Boosting adversarial training via fisher-rao norm-based regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
62. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
63. Zhang, Y., Ruan, W., Wu, F., Huang, X.: Generalizing universal adversarial perturbations for deep neural networks. Machine Learning **112**(5), 1597–1626 (2023)
64. Zhang, Y., Chen, Y., Chen, Z., Ruan, W., Huang, X., Khastgir, S., Zhao, X.: Adversarial training for probabilistic robustness. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1675–1685 (2025)
65. Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.M.M., Lin, M.: On evaluating adversarial robustness of large vision-language models. Advances in Neural Information Processing Systems **36**, 54111–54138 (2023)
66. Zhou, Z., Hu, S., Li, M., Zhang, H., Zhang, Y., Jin, H.: AdvCLIP: Downstream-agnostic adversarial examples in multimodal contrastive learning. In: ACM MM. pp. 6311–6320 (2023)
67. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)