BOOSTING ADVERSARIAL ROBUSTNESS OF VISION-LANGUAGE PRE-TRAINING MODELS AGAINST MULTIMODAL ADVERSARIAL ATTACKS

Youze Wang, Wenbo Hu, Qin Li, Richang Hong
Hefei University of Technology
{wangyouze,liqin}@mail.hfut.edu.cn
{wenbohu,hongrc}@hfut.edu.cn

ABSTRACT

Vision-language pre-training (VLP) models, known for their generalization across multimodal tasks, are increasingly deployed in perturbation-sensitive environments, highlighting the need for improved adversarial robustness. Recent studies have revealed VLP models' vulnerability to multimodal adversarial attacks, which exploit interactions across multiple modalities to uncover deeper weaknesses than single-modal attacks. Methods like Co-attack, SGA, and VLP-attack leverage cross-modal interactions to more effectively challenge models' robustness. To counter these threats, adversarial fine-tuning has emerged as a key strategy. Our approach refines vision encoders using Multi-granularity Aligned Visual Adversarial Fine-tuning, which enhances robustness by expanding the vision semantic space and aligning features across perturbed and clean models. Extensive experiments demonstrate that our method offers superior robustness to multimodal adversarial attacks while preserving clean performance on downstream V+L tasks.

1 INTRODUCTION

Vision-language pre-training (VLP) modelsRadford et al. (2021); Li et al. (2021); Yang et al. (2022); Li et al. (2023) have become pivotal in various vision-language tasks such as image-text retrieval Cao et al. (2022), visual entailment Xie et al. (2019) and visual question answering Antol et al. (2015), demonstrating broad generalization capabilities across different domains. Given their increasing deployment in perturbation-sensitive environments, the need to improve their adversarial robustness of VLP models is pressing Zhang et al. (2022); Lu et al. (2023); Wang et al. (2025).

Adversarial robustness introduces vulnerabilities in deep learning (DL) models through subtle, often imperceptible perturbations to input data, including text and images Goodfellow et al. (2014); Wang et al. (2023); Li et al. (2020). Based on it, multimodal adversarial attacks have been developed, posing a more realistic threat in scenarios involving complex multimodal interactions. Co-attack Zhang et al. (2022) explores these vulnerabilities in VLP models under a white-box setting by perturbing both textual and visual modalities with a sequential text-then-image approach in single image-text pairs. SGA Lu et al. (2023) enhances the transferability of multimodal adversarial examples by exploiting cross-modal interactions with set-level guidance, thereby increasing attack efficacy. Furthermore, VLP-attack Wang et al. (2025) employs cross-modal contrastive learning to disrupt the alignment of image-text pairs from various perspectives, thereby misleading the multimodal understanding of VLP models. Compared to single-modal attacks Madry et al. (2017); Gao et al. (2020); Wang & He (2021), multimodal adversarial attacks exploit interactions across multiple modalities, revealing additional vulnerabilities and increasing the likelihood of identifying effective attack vectors against defenses designed primarily for image attacks.

To counteract the growing threat of adversarial attacks, numerous defense mechanisms have been developed to enhance the robustness of DL models. Adversarial training has proven particularly effective, as demonstrated by recent studies Sankaranarayanan et al. (2018); Jin et al. (2023). Subsequent research Mao et al. (2023); Schlarmann et al. (2024) has further shown that fine-tuning the visual embeddings of vision-language models can improve robustness to adversarial image per-

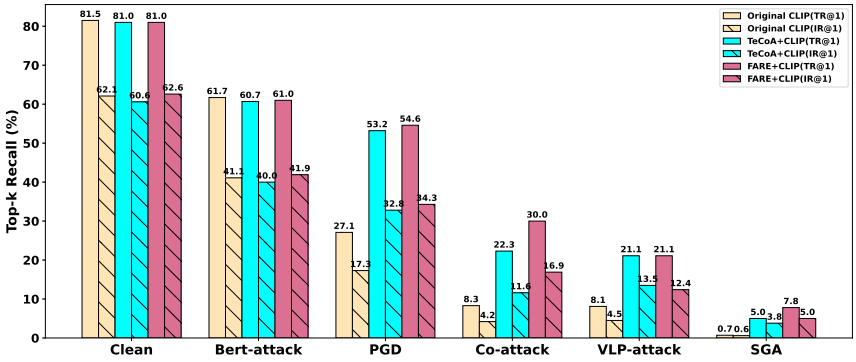


Figure 1: Comparison of CLIP with different visual encoders under various adversarial attacks. Despite original CLIP's high performance in image-text retrieval tasks on Flickr30k dataset, it exhibits vulnerabilities to adversarially constructed inputs. While adversarial fine-tuning strategies (e.g. TeCoA) tailored for image attacks enhance robustness against such threats (e.g. PGD), they prove less effective against multimodal attacks (e.g. SGA).

turbations (e.g. PGD Goodfellow et al. (2014) and APGD Croce & Hein (2020)) while largely preserving the original model features. However, as shown in Figure 1, a comparison of PGD, Co-attack, VLP-attack and SGA attacks on both original and adversarially fine-tuned CLIP (TeCoA and FARE) demonstrates that these methods offer only limited improvements in robustness to multimodal adversarial perturbations. We attribute this to the fact that unlike single-modal attacks, multimodal adversarial examples generated through image-text interactions can disrupt the contextual integrity of image-text pairs and introduce additional perturbation information, posing a greater threat to VLP models. These findings highlight the need to incorporate image-text interactions into adversarial fine-tuning and mitigate overfitting on adversarial images within a multimodal context.

This paper addresses the critical yet underexplored challenge of fine-tuning vision encoders in VLP models to enhance robustness against multimodal adversarial attacks. We propose **Multi-granularity Aligned Visual Adversarial fine-tuning (MAVA)**, a novel method to fortify VLP models against multimodal adversarial threats. Our approach fine-tunes vision encoders at three granularities to improve VLP model performance under image-text perturbations. First, we employ **Cross-Modal Supervision** (CMS) to leverage cross-modal interactions, aligning adversarial image features with text descriptions from multiple perspectives and mitigating the impact of perturbations based on image-text interactions. Second, we introduce **Vision Semantic Space Expansion** (VSSE) to incorporate neighboring data points around adversarial images to tune the visual features, expanding the vision semantic space and reducing overfitting on adversarial data. Third, we apply **Semantic Consistency Alignment** (SCA) to minimize divergence between feature representations of clean images in fine-tuned and non-fine-tuned models, preserving pre-training benefits. These strategies achieve an optimal balance between clean data accuracy and robustness to multimodal adversarial examples.

We conduct a comprehensive evaluation of the adversarial robustness of VLP models under multimodal attack scenarios, assessing the effectiveness of adversarial fine-tuning against multimodal threats. The experimental results demonstrate that MAVA maintains performance closer to clean data scenarios, even under multimodal adversarial attacks. The main contributions are summarized as follows:

- We demonstrate that existing adversarial fine-tuning methods are less effective against multimodal adversarial attacks;
- We propose MAVA, a novel adversarial fine-tuning method that effectively utilizes cross-modal supervision, vision semantic space expansion, and semantic consistency alignment to enhance multimodal adversarial robustness;
- Extensive experiments show that MAVA have better robustness to l_∞ bounded attacks, while perserving much closer the clean performance of VLP models on downstream V+L tasks.

2 RELATED WORKS

2.1 VISION-LANGUAGE PRE-TRAINING MODELS

Vision-Language Pre-training (VLP) models aim to improve multimodal task performance by leveraging large-scale pre-training on image-text pairs. A prominent example is CLIP Radford et al. (2021), which excels in image-text matching and zero-shot multimodal tasks by effectively integrating visual and textual information. ALBEF Li et al. (2021) builds on this with image-text contrastive learning to align and fuse representations through a multimodal encoder. TCL Yang et al. (2022) introduces triple contrastive learning, utilizing complementary information from multiple views to enhance representation learning and modality alignment. BLIP Li et al. (2022) and BLIP2 Li et al. (2023) further advance vision-language alignment. The adversarial robustness of these VLP models is crucial for their practical deployment and effectiveness in real-world applications.

2.2 MULTIMODAL ADVERSARIAL ATTACKS

The vulnerability of VLP mdoels to adversarial attacks remains a critical concern. Zhang's study investigates these vulnerabilities in VLP models within a white-box setting, providing valuable insights into multimodal adversarial attack construction and robustness enhancement strategies Zhang et al. (2022). Building on this, Lu et al. (2023) improve the transferability of multimodal adversarial samples by leveraging cross-modal interactions and data augmentation. Wang et al. Wang et al. (2025) explore the use of contrastive learning to disrupt image-text feature alignment from different perspectives. These multimodal adversarial attacks pose significant threats to the practical deployment of VLP models, particularly in downstream applications. In response, our study introduces a novel adversarial fine-tuning method designed to enhance the adversarial robustness of VLP systems, addressing these challenges.

2.3 ADVERSARIAL FINE-TUNING

Deep neural networks are vulnerable to adversarial attacks, prompting the development of various defense strategies, including adversarial purification Nie et al. (2022), data transformation Bhagoji et al. (2018); Dziugaite et al. (2016); Guo et al. (2017), and adversarial training Jin et al. (2023); Sankaranarayanan et al. (2018). Among these, adversarial training has proven most effective by incorporating adversarial examples into the training process. Recent studies Mao et al. (2023); Schlarmann et al. (2024) have adapted adversarial fine-tuning for vision-language models, primarily targeting image-based attacks by incorporating adversarial images during fine-tuning. However, these methods often fail against multimodal adversarial attacks, as shown in Figure 1. In contrast, our approach focuses on fine-tuning the vision encoder of VLP models to address multimodal adversarial threats, improving robustness and aligning the embeddings of adversarial images and text with those of clean images, thereby preserving modality coherence under adversarial conditions.

3 METHODS

VLP models effectively bridge the gap between visual and language understanding, enhancing V+L tasks. This study investigates adversarial fine-tuning to improve the robustness of VLP models against multimodal adversarial scenarios. While current methods boost performance through adversarial fine-tuning against images attacks, they often fall short against multimodal adversarial challenges. To overcome this, we introduce *Multi-granularity Aligned Visual Adversarial fine-tuning* (MAVA), specifically designed to enhance the adversarial robustness of VLP models to multimodal attacks.

3.1 PRELIMINARIES

Let $\mathcal{F}_\theta(\cdot)$ represent a vision-language pre-training (VLP) model with parameters θ. Given an image-text pair (v, t), the model generates representations $(\mathcal{E}_v, \mathcal{E}_t) = \mathcal{F}_\theta(v, t)$. For alignment tasks between image and text modalities, such as image-text retrieval, a standard VLP model seeks to minimize a contrastive loss, promoting cross-modal alignment and intra-modal consistency between $\mathcal{E}_v$ and $\mathcal{E}_t$.

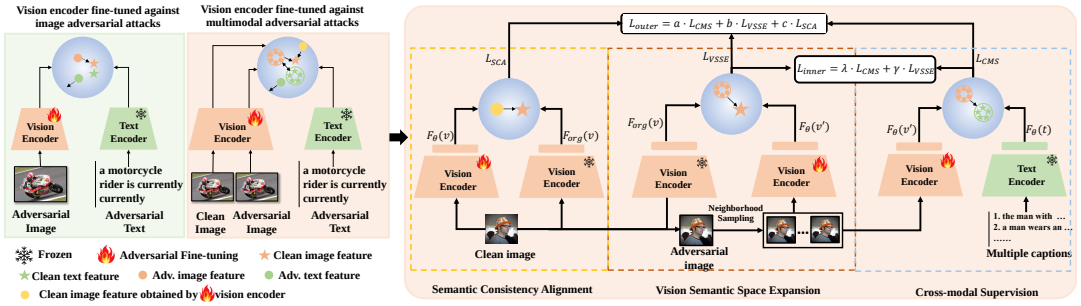


Figure 2: **Illustration of MAVA.** We propose an adversarial fine-tuning method that utilize cross-modal supervision, vision semantic space expansion, and semantic consistency alignment to enhance multimodal adversarial robustness.

Multimodal Adversarial attacks. An adversary typically optimizes for an additive pixel-level perturbation δ_v to the image to generate adversarial image $v' = v + \delta_v$, and word (character)-level perturbation δ_t to the text to generate adversarial text $t' = t + \delta_t$, which can mislead $\mathcal{F}_\theta$ to make wrong predictions:

$$\begin{cases} (v', t') = \arg\max_{(v', t')} \mathcal{L}((v', t'), y) \\ \text{s.t.} \quad ||v' - v||_\infty \leq \epsilon \\ \text{s.t.} \quad \text{similarity}(t', t) \leq \beta, \end{cases} \tag{1}$$

where ϵ and β constrain perturbations on images and texts, respectively, ensuring the attacks remain perceptually invisible. y indicates whether the images and texts are aligned. The defender's role is to correct the model's predictions against such attacks and enhance the robustness of VLP models to multimodal adversarial perturbations.

3.2 Multi-granularity Aligned Visual Adversarial Fine-tuning

We propose the Multi-granularity Aligned Visual Adversarial Fine-tuning method for VLP models as an effective defense against multimodal adversarial attacks. Unlike prior methods, such as TeCoA Mao et al. (2023) and FARE Schlarmann et al. (2024), which primarily target robustness against image-based attacks, our MAVA addresses a comprehensive multimodal attack paradigm. To balance accuracy on clean data with robustness to multimodal adversarial examples, MAVA integrates three novel strategies within the standard adversarial training framework, as shown in Figure 2.

Vision Semantic Space Expansion. Previous adversarial fine-tuning methods have primarily focused on single adversarial data points during each epoch's optimization, leading to overfitting on adversarial images. Instead, we ensure that features of neighboring data points around the current adversarial images remain close to the unperturbed features $\mathcal{F}_{org}(v)$ of the original VLP model. This strategy aims to expand the vision encoder's semantic space and mitigate overfitting. Specifically, we sample N examples from the neighborhood of v' to compute the loss $\mathcal{L}_{\text{VSSE}}$:

$$\mathcal{L}_{\text{VSSE}}(\theta) = \underset{v \sim \mathcal{D}}{E} \left[\frac{1}{N} \sum_{i=1}^{N} \| \mathcal{F}(v' + r_i) - \mathcal{F}_{org}(v) \|_2^2 \right], \tag{2}$$

where $r_i \sim U[-(\beta \cdot \epsilon)^d, (\beta \cdot \epsilon)^d]$, and $U[a^d, b^d]$ stands for the uniform distribution $\mathcal{D}$ in d dimensions. N is the number of sampling examples. As $\mathcal{L}_{\text{VSSE}}$ goes to zeors, the visual features given by the fine-tuned model for clean images is the same as the one by the original model, which implies that when presented with adversarial images, the vision encoder can generate more normal and stable features.

Cross-modal Supervision. Cross-modal interactions are essential for VLP models in vision-language tasks such as image-text retrieval, where text provides unique supervisory signals for each

image from various perspectives. Here, we leverage these interactions to expand the vision semantic space. Specifically, paired text information guides the optimization of the vision encoder, aligning neighboring datapoints around adversarial image features more closely with text-based supervision. This approach allows image features to integrate gradients from multiple supervisory sources, as follows:

$$\mathcal{L}_{\text{CMS}}(\theta) = \underset{v\sim\mathcal{D}}{E}\left[\frac{1}{M}\sum_{i=1}^{M}\frac{\mathcal{F}(t_i)\cdot\frac{1}{N}\sum_{j=1}^{N}\mathcal{F}(v'+r_j)}{\parallel\mathcal{F}(t_i)\parallel\parallel\frac{1}{N}\sum_{j=1}^{N}\mathcal{F}(v'+r_j)\parallel}\right], \tag{3}$$

where t is a captions set related to the image v, with set size is M. We optimize the vision semantic space to align adversarial features from the vision encoder closely with their corresponding captions.

Semantic Consistency Alignment. Adversarial fine-tuning can significantly alter image features, potentially diminishing the benefits derived from pre-training. To address this issue, we propose imposing a constraint on the divergence between the feature representations of clean images from fine-tuned and non-fine-tuned VLP models. This constraint is designed to ensure that the image features from fine-tuned models retain task-agnostic general knowledge during adversarial fine-tuning. We measure this divergence using the l_2 distance between the feature representations, as follows:

$$\mathcal{L}_{\text{SCA}}(\theta) = \underset{v\sim\mathcal{D}}{E}\left[\frac{1}{2}\parallel\mathcal{F}(v)-\mathcal{F}_{org}(v)\parallel_2^2\right], \tag{4}$$

By integrating $\mathcal{L}_{\text{CMS}}$ and $\mathcal{L}_{\text{SCA}}$ with the expansion of the base adversarial fine-tuning loss $\mathcal{L}_{\text{VSSE}}$, we direct the adversarial fine-tuning process to not only enhance adversarial robustness but also maintain clean performance. The overall training objective is formulated as follows:

$$\mathcal{L}_{\text{inner}} = \mathcal{L}_{\text{CMS}} + \mathcal{L}_{\text{VSSE}}, \tag{5}$$

$$\mathcal{L}_{\text{outer}} = a\cdot\mathcal{L}_{\text{CMS}} + b\cdot\mathcal{L}_{\text{VSSE}} + c\cdot\mathcal{L}_{\text{SCA}}, \tag{6}$$

where a, b, c are the hyper-parameters that can control the strength of different constraint terms.

Based on the above method, we iteratively optimize the parameters θ of the vision encoder in $\mathcal{F}_\theta$ to minimize the aforementioned objective (Eq. 6) on the generated adversarial examples(Eq. 5):

$$\min\mathcal{L}_{\text{outer}}\left[\max_{\delta\in B(v,\epsilon)}\mathcal{L}_{\text{inner}}(\theta, v+\delta)\right]. \tag{7}$$

4 Experiments and Results

We conduct experiments to evaluate the adversarial robustness of VLP models where the vision encoder is replaced with an adversarially fine-tuned version across various downstream V+L tasks.

4.1 Experiments Setting

Examized VLP models. For image-text retrieval, we assessed four vision-language pretraining (VLP) models: CLIP Radford et al. (2021), ALBEF Li et al. (2021), TCL Yang et al. (2022), and BLIP Li et al. (2022), each employing ViT-B/16 as the vision encoder. For visual entailment tasks, we focused on ALBEF and TCL, while for visual grounding, we exclusively evaluated TCL.

Datasets and Downstream Tasks. Following Co-attack Zhang et al. (2022), SGA Lu et al. (2023), and VLP-attack Wang et al. (2025), we evaluate the effectiveness of our method using test sets from MSCOCO Lin et al. (2014), Flickr30K Plummer et al. (2015), SNLI-VE Xie et al. (2019), and RefCOCO+ Yu et al. (2016). Specifically, we use MSCOCO's test set to fine-tune the vision encoders, Flickr30K's test set for image-text retrieval, SNLI-VE for Visual Entailment (VE), and RefCOCO+ for Visual Grounding (VG). For cross-task transferability evaluation, following SGA, we fine-tune the vision encoders on Flickr30K's test set and assess performance on MSCOCO's test set.

Setting. The main drawback of adversarial training/fine-tuning is the degradation of clean performance. For consistency with FARE Schlarmann et al. (2024), we use $\epsilon = \frac{2}{255}$ and $\epsilon = \frac{4}{255}$ for adversarial fine-tuning, denoting the VLP models as MAVA^2 and MAVA^4 (resp. TeCoA^2 and

Table 1: Clean and adversarial evaluation on Flickr30k dataset for image-text retrieval.

VLP models	Adversarial fine-tuning	clean		Co-attack				SGA			
				2/255		4/255		2/255		4/255	
		TR↑	IR↑	TR↑	IR↑	TR↑	IR↑	TR↑	IR↑	TR↑	IR↑
CLIP_{ViT}	Origin	81.5	62.1	8.3	4.2	1.5	0.5	0.7	0.6	0.3	0.1
	TeCoA^2	81.0	60.6	22.3	11.6	9.7	18.3	5.0	3.8	1.7	1.8
	FARE^2	81.0	62.6	30.0	16.9	12.8	7.7	7.8	5.0	2.8	1.8
	MAVA^2	**81.0**	**66.2**	**43.6**	**29.8**	**31.0**	**20.7**	**20.3**	**36.0**	**10.2**	**7.2**
	TeCoA^4	76.7	59.3	29.8	17.4	16.6	10.2	10.3	6.3	5.2	3.3
	FARE^4	**81.0**	64.7	26.0	13.4	10.4	4.5	5.1	3.1	1.8	1.2
	MAVA^4	80.9	**65.9**	**32.9**	**25.2**	**20.6**	**16.6**	**11.5**	**9.8**	**5.0**	**5.1**
ALBEF	Origin	94.9	84.5	27.3	23.5	13.0	11.9	2.3	2.5	1.1	1.0
	TeCoA^2	93.0	**85.7**	34.8	27.9	16.1	14.7	9.5	9.3	5.7	4.5
	FARE^2	94.6	84.5	74.3	53.7	62.8	46.2	19.1	14.6	7.9	6.3
	MAVA^2	**95.2**	85.0	**74.7**	**54.2**	**63.6**	**46.4**	**20.3**	**15.8**	**8.3**	**6.4**
	TeCoA^4	94.0	**84.5**	32.9	23.4	16.9	12.6	10.6	9.0	4.2	3.0
	FARE^4	90.7	80.2	75.7	53.9	70.8	50.3	11.4	9.9	5.2	3.8
	MAVA^4	**94.1**	84.3	**77.8**	**55.6**	**72.6**	**52.5**	**19.2**	**13.3**	**9.9**	**6.6**
TCL	Origin	94.9	84.0	13.5	13.0	2.8	3.1	1.7	1.1	0.5	0.3
	TeCoA^2	90.9	**84.7**	28.1	26.4	8.2	9.3	5.0	3.9	1.0	0.9
	FARE^2	93.1	82.5	57.4	25.2	32.9	33.5	12.8	9.7	3.3	2.6
	MAVA^2	**94.6**	83.7	**67.0**	**48.9**	**41.5**	**31.8**	**14.1**	**11.6**	**4.0**	**3.2**
	TeCoA^4	94.3	**83.5**	17.1	14.4	5.9	7.2	4.6	3.7	1.3	0.8
	FARE^4	94.0	82.9	56.6	39.7	34.6	25.9	13.0	9.7	3.3	2.7
	MAVA^4	**94.3**	83.4	**63.0**	**45.7**	**46.9**	**33.6**	**18.9**	**14.7**	**5.5**	**4.9**
BLIP	Origin	97.2	87.3	22.4	17.1	6.8	6.1	6.2	5.3	2.1	2.2
	TeCoA^2	96.8	86.6	36.6	24.7	14.4	10.9	13.2	**18.6**	7.4	9.8
	FARE^2	96.5	85.9	60.5	40.1	49.0	29.3	16.5	13.6	2.7	2.6
	MAVA^2	**97.5**	**87.4**	**60.6**	**41.3**	**53.3**	**35.6**	**22.6**	17.7	**11.2**	**8.8**
	TeCoA^4	95.8	84.6	33.1	23.2	12.3	10.0	13.7	15.0	6.2	7.5
	FARE^4	95.5	84.1	**61.4**	**42.5**	44.2	29.7	19.4	15.7	5.0	5.1
	MAVA^4	**97.3**	**87.3**	60.4	42.0	**54.1**	**36.8**	**24.2**	**18.6**	**11.6**	**9.4**

Note: the reported results are TR@1 and IR@1.

TeCoA^4, FARE^2 and FARE^4). Notably, the text encoders of all VLP models are fixed in these experiments.

Multimodal Adversarial Attacks and Defences. We use two adversarial fine-tuning methods (TeCoA Mao et al. (2023), FARE Schlarmann et al. (2024)) as baselines to evaluate three multi-modal adversarial attacks (Co-attack Zhang et al. (2022), SGA Lu et al. (2023), VLP-attack Wang et al. (2025)) against VLP models. The detailed description is referred App. A.1.

4.2 Quantitative Robustness Evaluation of VLP Models

First, we evaluate the clean and robust performance on several tasks native to the VLP models for l_∞ -perturbation strengths of $\epsilon = \frac{2}{255}$ and $\epsilon = \frac{4}{255}$.

Implementations. We employ PGD Goodfellow et al. (2014) -based adversarial attack with Eq. 5 as the loss degrades the VLP models' performance. We conduct PGD with 10 epochs at each iteration. The number of sampling neighbors is set to 3. The hyperparameters a, b, c in Eq. 6 are set to 0.002, 0.4, 0.6 separately. We perform adversarial fine-tuning on the vision encoder of VLP models for only 10 epochs. Leveraging the multiple text descriptions for each image in the MSCOCO and Flickr30k test sets, we fine-tune on the MSCOCO test set and evaluate on the Flickr30k test set for image-text retrieval, the SNLI-VE test set for visual entailment, and the RefCOCO+ test set for visual grounding to assess the generalization of the fine-tuned vision encoder in VLP models. The whole method is implemented by Pytorch Paszke et al. (2019). The implementation details about all tasks are referred to App. A.1.

Table 2: Clean and adversarial evaluation on SNLI-VE dataset for visual entailment.

VLP models	Adversarial fine-tuning	Clean	Co-attack (Acc)↑		SGA (Acc)↑	
			2/255	4/255	2/255	4/255
ALBEF	Origin	83.3	21.9	17.5	54.2	50.7
	TeCoA^2	83.6	25.2	19.1	56.8	52.6
	FARE^2	81.8	35.1	28.2	59.1	57.1
	MAVA^2	**83.7**	**44.0**	**42.9**	**67.9**	**68.4**
	TeCoA^4	83.5	24.1	18.6	55.9	52.3
	FARE^4	82.7	31.4	24.5	58.9	56.1
	MAVA^4	**83.6**	**41.7**	**40.8**	**61.1**	**61.3**
TCL	Origin	79.3	21.0	18.9	54.1	49.5
	TeCoA^2	**79.8**	23.9	20.1	57.3	53.1
	FARE^2	77.2	**30.3**	**27.2**	57.9	56.2
	MAVA^2	79.2	27.8	22.9	**63.0**	**61.0**
	TeCoA^4	**79.8**	23.8	20.0	57.4	52.9
	FARE^4	79.0	26.2	21.6	57.2	53.4
	MAVA^4	79.1	**34.4**	**30.8**	**63.0**	**62.5**

Main Results. We adversarially fine-tune vision encoder of VLP models for various V+L tasks using different adversarial fine-tuning methods and examine their performance on the select 4 VLP models. The results are shown in Table 1 and Table 2. Several observations are summarized in the following context.

First, the original VLP models exhibit the best clean performance but lack robustness against multimodal adversarial attacks, showing significant performance degradation. Among the adversarially fine-tuned models, those tuned using our proposed MAVA method demonstrate superior robustness to such attacks, while maintaining comparable clean performance to other baseline methods.

Additionally, all examined methods focus solely on adversarially fine-tuning the vision encoder of VLP models to defend against adversarial examples. Among baseline methods, FARE consistently outperforms TeCoA under different perturbation budgets ($\epsilon = \frac{2}{255}$ and $\frac{4}{255}$), except for TeCoA^4 and FARE^4. However, TeCoA generally aligns more closely with the original clean performance, particularly in visual entailment tasks. Our MAVA surpasses these baselines in both adversarial robustness and clean performance across three V+L downstream tasks, effectively meeting our objective of enhancing the robustness of VLP models against multimodal adversarial attacks. Furthermore, SGA continues to demonstrate a more effective attack strategy compared to Co-attack across various adversarial fine-tuning methods. Detailed results of adversarial fine-tuning, including PGD and VLP-attack, and the performance in visual grounding tasks are provided in App. A.2.

4.3 Transfer-based Attacks

Attack Setup. Following the settings of SGA Lu et al. (2023), we evaluate the performance of adversarial fine-tuning methods on two transfer-based settings: cross-modal transferability and cross-task transferability, where ALBEF is used as the surrogate model to craft adversarial examples.

4.3.1 Cross-Modal Transferability.

Image-Text Retrieval. We evaluate the effectiveness of adversarial fine-tuning methods in a transfer-based adversarial attack setting, with results summarized in Table 3. Transfer attacks are useful when adversaries lack white-box access to target models but have access to a surrogate model. Despite architectural and parameter differences, adversarial examples transfer effectively across ALBEF, TCL, and BLIP. However, when target VLP models employ robust vision encoders, their performance improves significantly. Compared to baselines, our MAVA demonstrates superior robustness, with MAVA^2 achieving the best results when paired with either TCL or BLIP.

Visual Entailment. The transferability of adversarial examples against robust vision encoders in the visual entailment task is similar to that observed in the image-text retrieval, as detailed in App. A.3.

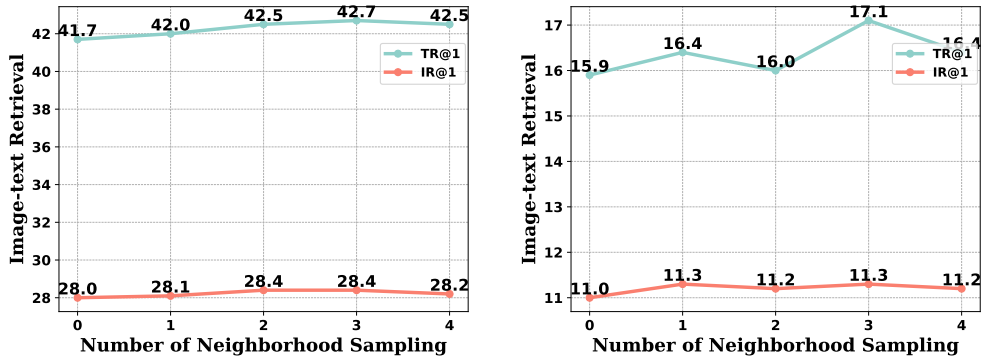
Table 3: The transfer-based attack results on the Flickr30k dataset for image-text retrieval. The surrogate model is ALBEF, we select TCL and BLIP as the target models to evaluate adversarial robustness in a transfer-based setting.

VLP model	Vision encoder	clean		SGA							
				2/255				4/255			
		TR@1	IR@1	TR@1	TR@5	IR@1	IR@5	TR@1	TR@5	IR@1	IR@5
TCL	Origin	94.9	84.0	51.1	75.0	38.2	62.0	33.7	56.8	26.5	47.3
	TeCoA^2	90.9	**84.7**	48.4	67.3	42.7	66.2	35.0	52.5	31.5	54.3
	FARE^2	93.1	82.5	61.4	82.9	45.2	77.8	46.8	70.3	35.2	58.9
	MAVA^2	**94.6**	83.7	**64.9**	**85.0**	**46.5**	**70.8**	**50.2**	**74.5**	**37.2**	**60.8**
	TeCoA^4	88.8	**83.5**	44.9	62.9	41.9	65.9	31.5	48.6	30.9	53.2
	FARE^4	94.0	82.9	59.9	82.1	45.0	69.0	46.4	71.0	35.4	58.9
	MAVA^4	**94.3**	83.4	**64.6**	**85.7**	**47.1**	**71.5**	**51.9**	**75.6**	**38.5**	**62.9**
BLIP	Origin	97.2	87.3	70.8	88.3	50.6	74.3	55.8	75.3	39.8	63.2
	TeCoA^2	**96.8**	**86.6**	72.0	89.4	52.3	75.5	59.2	78.3	42.7	66.5
	FARE^2	96.5	85.9	70.9	89.2	52.2	75.5	58.6	80.6	43.1	66.5
	MAVA^2	96.2	86.4	**75.5**	**91.4**	**55.1**	**77.7**	**65.9**	**85.3**	**47.7**	**71.6**
	TeCoA^4	**95.8**	**84.6**	71.5	88.8	**52.3**	75.3	58.0	78.1	42.6	66.2
	FARE^4	95.5	84.1	71.4	89.2	53.0	76.3	62.4	82.0	46.0	69.8
	MAVA^4	95.5	84.4	**71.5**	**89.6**	**53.0**	**76.4**	**62.7**	**82.6**	**46.1**	**69.8**

4.3.2 Cross-Task Transferability

Cross-modal interactions and alignments are fundamental to multimodal learning across various V+L tasks. Here we evaluate the effectiveness of MAVA in a cross-task transferability setting.

Image Captioning. In our experiment, we generate adversarial images using ALBEF with an image-text retrieval objective, then directly attack the target model, BLIP Li et al. (2022), on the image captioning task. Table 8 in App. A.4 shows the performance of BLIP with a robust vision encoder on image captioning. The results indicate notable improvements in adversarial robustness for BLIP in the cross-task transferability scenario. The details can be found in App. A.4.



(a) Co-attack (b) SGA

Figure 3: The impact of neighborhood sampling number on defense against multimodal adversarial attacks.

4.4 Ablation Studies

We conduct ablation studies on MAVA to further verify our design. In this context, we only report the performance of CLIP for analysis on image-text retrieval tasks.

Strategies of CMS, VSSE, and SCA. We first evaluate the impact of the proposed strategies—CMS, VSSE, and SCA—on image-text retrieval performance. As shown in Table 4, each strategy independently improves performance to some extent. When combined, they maximize adversarial robustness against diverse attacks. This suggests that CMS's unique supervisory signals from paired

Table 4: Evaluation of the proposed method on image-text retrieval tasks with CLIP model.

Methods	Vision encoder	Normal		Co-attack				SGA			
				2/255		4/255		2/255		4/255	
		TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
MAVA $_{w/CMS}$	MAVA^2	81.5	66.8	42.4	29.6	29.6	18.9	19.3	12.7	10.2	6.4
	MAVA^4	81.5	62.1	15.9	13.2	11.6	10.7	8.8	4.8	2.4	3.0
MAVA $_{w/VSSE}$	MAVA^2	80.5	65.9	42.9	29.6	31.0	20.6	19.3	12.9	10.2	6.4
	MAVA^4	79.3	65.8	30.2	22.0	19.5	14.5	9.4	4.2	3.6	3.2
MAVA$_{w/SCA}$	MAVA^2	48.1	45.5	25.6	26.6	21.8	22.4	16.0	16.9	12.0	12.1
	MAVA^4	46.4	40.3	22.6	20.4	18.4	16.6	10.6	9.2	6.8	7.0
MAVA	MAVA^2	81.0	66.2	43.6	29.8	31.0	20.7	20.3	36.0	10.2	7.2
	MAVA^4	80.9	65.9	32.9	25.2	20.6	16.6	11.5	9.8	5.0	5.1

w/X means a variant of with X being removed; The reported results are TR@1 and IR@1.

image-text data, VSSE's enhancement of the vision semantic space, and SCA's mitigation of performance degradation collectively contribute to the adversarial robustness of VLP models. Notably, removing SCA significantly reduces the performance of MAVA, highlighting its critical role in MAVA by constraining the divergence between feature representations of clean images in fine-tuned and non-fine-tuned VLP models. GradCAM visualizations in APP. A.6 demonstrate that MAVA effectively enhances the robustness of VLP models against multimodal adversarial attacks.

Number of neighborhood samples. To mitigate overfitting on adversarial images v' and enhance model robustness, we incorporate neighboring data points around the current v' to refine the vision semantic space during each iteration. Specifically, we sample N examples from the neighborhood of v' to compute the loss $\mathcal{L}_{\text{VSSE}}$. Figure 3 shows the effect of varying the number of neighborhood samples in MAVA^4 against Co-attack and SGA with $\epsilon = \frac{2}{255}$. As the number of neighborhood samples increases from 0 to 3, CLIP with MAVA^4 becomes more robust. However, when the number of samples reaches 4, robustness against Co-attack and SGA decreases. Based on these findings, we use 3 neighborhood samples to mitigate overfitting on adversarial images. The effect of the number of neighborhood samples in MAVA^4 against VLP-attack is referred to App. A.5.

4.5 In-depth Analyses

We present several in-depth analyses demonstrating MVAV's effectiveness under varying conditions.

Number of Iterations. To further analyze the impact of iterations on defense against multimodal adversarial attacks, we examine the effect of varying iterations during adversarial fine-tuning of CLIP with $\epsilon = \frac{2}{255}$, which are provided in App. B.1.

Multi Captions for An Image. We also investigate the impact of number of captions on adversarial robustness of vision encoder. The number of captions varies from 1 to M, where M represents the total number of caption pairs for each image, with each caption offering a distinct perspective, which is referred to App. B.2.

Theoretical Analysis of MAVA. To further understand MAVA, we provide a theoretical analysis of its core mechanisms, which are provided in App. B.3.

Scalability and Computational Cost of MAVA. We present a detailed analysis of the scalability and computational cost of MAVA, which is refreed to App. B.4 and B.5.

5 Conclusion

This paper presents a comprehensive study of multimodal adversarial robustness in VLP models. Our findings reveal that existing adversarial fine-tuning methods are inadequate against multimodal attacks. To address this, we propose a novel Multi-granularity Aligned Visual Adversarial Fine-tuning approach, which integrates cross-modal supervision, vision semantic space expansion, and semantic consistency alignment to improve model robustness on downstream V+L tasks. Future work will focus on optimizing our method and extending the framework to incorporate perturbations from other modalities, such as text.

REFERENCES

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pp. 382–398. Springer, 2016.

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*, pp. 2425–2433, 2015.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Arjun Nitin Bhagoji, Daniel Cullina, Chawin Sitawarin, and Prateek Mittal. Enhancing robustness of machine learning systems via data transformations. In *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–5. IEEE, 2018.

Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*, 2022.

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

Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.

Lianli Gao, Qilong Zhang, Jingkuan Song, and Heng Tao Shen. Patch-wise++ perturbation for adversarial targeted attacks. *arXiv preprint arXiv:2012.15503*, 2020.

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

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

Gaojie Jin, Xinping Yi, Dengyu Wu, Ronghui Mu, and Xiaowei Huang. Randomized adversarial training via taylor expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16447–16457, 2023.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*, 2020.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 102–111, 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *ICLR*, 2023.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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*, pp. 2641–2649, 2015.

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*, pp. 8748–8763. PMLR, 2021.

Swami Sankaranarayanan, Arpit Jain, Rama Chellappa, and Ser Nam Lim. Regularizing deep networks using efficient layerwise adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

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*, pp. 4566–4575, 2015.

Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1924–1933, 2021.

Youze Wang, Wenbo Hu, and Richang Hong. Iterative adversarial attack on image-guided story ending generation. *IEEE Transactions on Multimedia*, 2023.

Youze Wang, Wenbo Hu, Yinpeng Dong, Hanwang Zhang, Hang Su, and Richang Hong. Exploring transferability of multimodal adversarial samples for vision-language pre-training models with contrastive learning. *IEEE Transactions on Multimedia*, 2025.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15671–15680, 2022.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 69–85. Springer, 2016.

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*, pp. 5005–5013, 2022.

Monty-Maximilian Zühlke and Daniel Kudenko. Adversarial robustness of neural networks from the perspective of lipschitz calculus: A survey. *ACM Computing Surveys*, 2024.

SUPPLEMENTARY MATERIAL : BOOSTING ADVERSARIAL ROBUSTNESS OF VISION-LANGUAGE PRE-TRAINING MODELS AGAINST MULTIMODAL ADVERSARIAL ATTACKS

A ADDITIONAL EXPERIMENTS

A.1 EXPERIMENTS SETTING

Implementation details. We apply PGD Goodfellow et al. (2014) to generate adversarial images over 10 epochs, with a step size of 1.0 across all tasks. In our MAVA approach, the learning rate is set to 1e-3 for image-text retrieval (1e-4 for CLIP), visual entailment, and image captioning, and 1e-4 for visual grounding. The weight decay for all VLP models is 1e-4. For the FARE Schlarmann et al. (2024) method, to maintain clean performance close to that of standard VLP models, the learning rate is 1e-3 for image-text retrieval, except for ALBEF with $\epsilon = \frac{2}{255}$ and TCL with $\epsilon = \frac{2}{255}$ and $\epsilon = \frac{4}{255}$, where it is set to 1e-4. For visual entailment and visual grounding tasks, the learning rate is also 1e-4. In TeCoA Schlarmann et al. (2024), a learning rate of 1e-3 is used across all tasks. For all baseline methods, we choose the learning rate that yields the best adversarial robustness while maintaining clean performance comparable to standard VLP models. We use 3 neighborhood samples around adversarial images and 5 image captions for per image. The vision encoders in the VLP models are fine-tuned over 10 epochs using four A100 GPUs, while the text encoders remain fixed. The entire implementation is conducted using PyTorch Paszke et al. (2019).

Multimodal Adversarial Attacks. To demonstrate the effectiveness of our proposed method, in the work, we select three multimodal adversarial attacks as the baseline methods.

- **Co-attack** Zhang et al. (2022) is a multimodal adversarial attack against vision-language pre-training models that adopts a step-wise mechanism that first perturbs the discrete inputs (text) and then perturbs the continuous inputs (image) given the text perturbation, which is designed for white-box attack manner.
- **SGA** Lu et al. (2023) is a multimodal transfer-based adversarial attack that investigates the adversarial transferability of recent VLP models through modalities interactions and data augmentation.
- **VLP-attack** Wang et al. (2025) is a multimodal transfer-based adversarial attack that provides contrastive learning with sufficient image-text variations to perturb the inherent structures in the benign samples and the contextual integrity of image-text pairs from different views.

All the multimodal adversarial attacks have shown superior performance than single-modal adversarial attacks, such as PGD and Bert-attack.

Adversarial Fine-Tuning Methods. We use two baseline adversarial fine-tuning approaches to evaluate the effectiveness of our MAVA method:

- **TeCoA** Mao et al. (2023) employs a text-guided contrastive adversarial training loss to align text embeddings with adversarial image features, which are then used for model and visual prompt tuning.
- **FARE** Schlarmann et al. (2024) constrains the features of adversarial images to remain close to those of the unperturbed original CLIP model, while preserving the original model's feature representations.

A.2 QUANTITATIVE ROBUSTNESS EVALUATION OF VLP MODELS

We present additional experimental results on adversarial fine-tuning in visual grounding tasks, evaluating the performance of PGD and VLP-attack against adversarial fine-tuning methods in image-text retrieval.

Table 5: Clean and adversarial evaluation on Refcoco+ dataset for visual grounding.

Vision encoder	Clean			Co-attack						SGA					
				2/255			4/255			2/255			4/255		
	Val	TestA	TestB	Val	TestA	TestB	Val	TestA	TestB	Val	TestA	TestB	Val	TestA	TestB
Origin	58.4	65.9	46.2	32.8	36.3	29.3	31.5	34.0	28.2	39.9	45.2	33.9	36.6	40.2	32.0
TeCoA^2	57.4	**64.9**	45.8	45.1	50.4	36.1	43.2	48.5	35.5	46.2	51.0	37.8	43.9	48.8	36.3
FARE^2	56.7	63.3	46.0	46.1	51.8	37.4	44.8	50.2	36.6	47.2	52.8	38.5	45.7	51.4	37.9
MAVA^2	**57.9**	63.8	**46.2**	**47.0**	**52.7**	**39.0**	**46.4**	**51.0**	**38.4**	**48.8**	**53.2**	**39.8**	**47.5**	**53.3**	**39.0**
TeCoA^4	**57.6**	**65.2**	45.9	43.1	48.3	35.1	40.4	45.5	33.9	44.7	49.9	36.8	41.8	46.8	35.8
FARE^4	56.7	63.1	**46.0**	46.3	51.9	37.4	45.3	50.9	36.9	47.0	52.3	38.4	45.5	50.9	38.0
MAVA^4	56.9	63.1	45.9	**46.8**	**51.9**	**38.4**	**46.7**	**51.2**	**38.1**	**48.0**	**53.0**	**39.6**	**46.8**	**51.9**	**39.4**

Table 6: Clean and adversarial evaluation on Flickr30k dataset for image-text retrieval.

VLP models	Vision encoder	clean		PGD				VLP-attack			
				2/255		4/255		2/255		4/255	
		TR↑	IR↑	TR↑	IR↑	TR↑	IR↑	TR↑	IR↑	TR↑	IR↑
CLIP_{ViT}	Origin	81.5	62.1	27.5	16.4	9.4	7.9	9.1	5.4	3.1	2.5
	TeCoA^2	81.0	60.6	53.7	32.2	29.5	20.2	21.1	13.6	11.0	7.1
	FARE^2	80.5	62.6	52.9	31.3	24.6	15.1	21.1	12.5	8.0	5.0
	MAVA^2	**81.0**	**66.2**	**76.0**	**59.4**	**66.3**	**49.2**	**44.1**	**31.1**	**29.1**	**19.5**
	TeCoA^4	76.7	59.3	60.2	41.2	42.9	28.6	29.2	20.4	17.3	14.3
	FARE^4	**81.0**	64.7	61.7	40.2	33.6	21.8	24.1	14.7	8.9	5.9
	MAVA^4	80.9	**65.9**	**66.7**	**52.3**	**51.6**	**42.9**	**34.4**	**29.1**	**20.0**	**19.2**
ALBEF	Origin	94.9	84.5	31.9	23.6	12.0	9.3	14.1	9.7	5.2	4.1
	TeCoA^2	93.0	85.7	50.1	39.8	23.7	19.9	35.1	28.2	24.1	16.2
	FARE^2	94.6	84.5	80.0	66.5	64.4	50.8	58.2	51.6	27.0	19.8
	MAVA^2	**95.2**	**85.0**	**86.0**	**73.3**	**69.7**	**55.6**	**63.0**	**45.1**	**31.6**	**23.6**
	TeCoA^4	94.0	84.5	47.7	39.0	23.2	19.3	38.0	30.9	30.2	23.1
	FARE^4	90.7	80.2	82.7	71.9	81.4	70.6	66.4	45.8	42.0	28.6
	MAVA^4	**94.1**	**84.3**	**88.8**	**77.6**	**84.7**	**70.5**	**70.7**	**49.9**	**51.1**	**36.7**
TCL	Origin	94.9	84.0	42.2	30.2	18.8	11.8	12.6	8.8	4.9	4.0
	TeCoA^2	90.9	84.7	56.7	43.1	40.1	29.3	36.8	24.6	8.0	6.6
	FARE^2	93.1	82.5	78.4	65.8	64.0	51.3	35.8	23.7	7.6	5.5
	MAVA^2	**94.6**	**83.7**	**81.0**	**66.4**	**65.0**	**51.6**	**39.4**	**28.3**	**12.8**	**9.0**
	TeCoA^4	94.3	**83.5**	57.1	44.2	38.6	30.1	40.8	30.6	18.9	15.1
	FARE^4	94.0	82.9	76.5	62.4	65.0	51.6	60.8	42.2	49.8	36.2
	MAVA^4	**94.3**	83.4	**88.1**	**75.8**	**87.0**	**73.6**	**64.0**	**47.0**	**54.3**	**39.3**
BLIP	Origin	97.2	87.3	43.0	31.2	20.6	10.8	10.8	6.6	3.4	2.8
	TeCoA^2	96.8	86.6	60.2	46.6	45.4	36.1	40.2	34.8	28.6	22.5
	FARE^2	96.5	85.9	76.8	62.8	66.7	58.2	44.2	39.4	34.4	27.8
	MAVA^2	**97.5**	**87.4**	**80.6**	**72.5**	**78.4**	**69.7**	50.8	43.6	42.9	37.2
	TeCoA^4	95.8	84.6	61.6	48.1	46.2	36.2	51.0	39.8	32.4	30.0
	FARE^4	95.5	84.1	80.8	69.6	78.7	67.1	57.3	51.1	40.7	38.1
	MAVA^4	**97.3**	**87.3**	**83.8**	**72.1**	**81.4**	**71.4**	**61.0**	**56.8**	**46.2**	**43.8**

Note: the reported results are TR@1 and IR@1.

For the visual grounding task, we assess the clean and robust performance of VLP models under l_∞ -perturbation strengths of $\epsilon = \frac{2}{255}$ and $\epsilon = \frac{4}{255}$. The results, presented in Table 5, show that the performance of various adversarial fine-tuning methods on the VG task aligns closely with results from image-text retrieval and visual entailment tasks. Notably, SGA underperforms compared to Co-attack, likely because it requires multiple descriptive texts, and simple augmentations fail to provide alignment-preserving enhancements for image-text pairs in the Refcoco+ dataset. Overall, our MAVA method outperforms all baselines.

Table 7: The transfer-based attack results on the SNLI-VE dataset for visual entailment tasks. The surrogate model is ALBEF, we select TCL as the target models to evaluate adversarial robustness in a cross-modal transfer-based setting.

VLP models	Vision encoder	Clean (Acc↑)	SGA (Acc↑) 2/255	SGA (Acc↑) 4/255
TCL	Origin	83.3	58.5	46.4
	TeCoA^2	78.9	64.8	65.5
	FARE^2	79.5	64.4	65.0
	MAVA^2	**79.5**	**65.4**	**66.3**
	TeCoA^4	78.8	64.7	65.7
	FARE^4	79.2	64.8	65.6
	MAVA^4	**79.2**	**65.3**	**66.6**

To further demonstrate the effectiveness of our MAVA, we present the performance of VLP-attack and PGD against adversarial fine-tuning methods in Table 6. For the image-based attack PGD, all adversarial fine-tuning methods show improved defense. However, for the multimodal VLP-attack, the results are consistent with those in Table 1 and Table 2. Overall, the findings indicate that MAVA provides the best defense against both image and multimodal adversarial attacks.

A.3 Cross-Modal Transferability

We evaluate the effectiveness of adversarial fine-tuning methods under transfer-based adversarial attacks, with results summarized in Table 7. Transfer attacks are relevant when adversaries lack white-box access but have access to a surrogate model. Using ALBEF as the surrogate, we transfer adversarial examples to evaluate TCL on visual entailment tasks. As shown in Table 7, the results align with those in Table 3, demonstrating that MAVA consistently outperforms other methods in resisting multimodal adversarial attacks under a cross-modal transfer-based setting.

A.4 Details of Cross-Task Transferability

In image captioning, an image is encoded into a feature vector and decoded into a language description. In our experiment, we generate adversarial images using ALBEF with an image-text retrieval objective, then directly attack the target model, BLIP Li et al. (2022). Due to lack of text encoder in BLIP, we only use $\mathcal{L}_{\text{VSSE}}$ and $\mathcal{L}_{\text{SCA}}$ in $\mathcal{L}_{\text{outer}}$ to adversarially fine-tune the vision encoder of BLIP with test set in the Flickr30k dataset:

$$\mathcal{L}_{\text{outer}} = b \cdot \mathcal{L}_{\text{VSSE}} + c \cdot \mathcal{L}_{\text{SCA}}, \tag{8}$$

We evaluate performance on the MSCOCO dataset using several metrics, including BLEU Papineni et al. (2002) (B), METEOR Banerjee & Lavie (2005) (M), ROUGE Lin (2004) (R), CIDEr Vedantam et al. (2015) (C), and SPICE Anderson et al. (2016) (S) with pycocoevalcap tool[1]. The results in Table 8 indicate notable improvements in adversarial robustness for BLIP in the cross-task transferability scenario.

A.5 Number of neighborhood samples

To mitigate overfitting on adversarial images v' and enhance model robustness, we incorporate neighboring data points around v' to refine the vision semantic space during each iteration. Specifically, we sample N examples from the neighborhood of v' to compute the loss $\mathcal{L}_{\text{VSSE}}$. As illustrated in Figure 4, increasing the number of neighborhood samples in MAVA^4 improves robustness against VLP-attack with $\epsilon = \frac{2}{255}$. Unlike the results in Figure 3, robustness consistently improves as the number of samples increases from 0 to 4. Notably, when the number is 3 or 4, MAVA demonstrates superior defense against multimodal adversarial attacks. Based on these findings, we adopt 3 as the optimal configuration to mitigate overfitting on adversarial images.

[1]https://github.com/salaniz/pycocoevalcap

Table 8: **Cross-Task Transferability:ITR⟶IC**. Adversarial examples crafted from imae-text retrieval (ITR) to attack image captioning (IC) on MSCOCO. The baseline means the original performance of IC on clean data. The ϵ of adversarial attacks is $\frac{2}{255}$.

Vision Encoder	Attacks	B@4	M	R	C	S
Origin	Baseline	39.7	31.0	60.0	133.3	23.8
	Co-attack	37.4	29.8	58.4	125.5	22.8
	SGA	34.8	28.4	56.3	116.0	21.4
MAVA[4]	Co-attack	38.2	30.4	59.0	127.4	23.4
	SGA	36.4	29.6	57.4	118.8	22.8

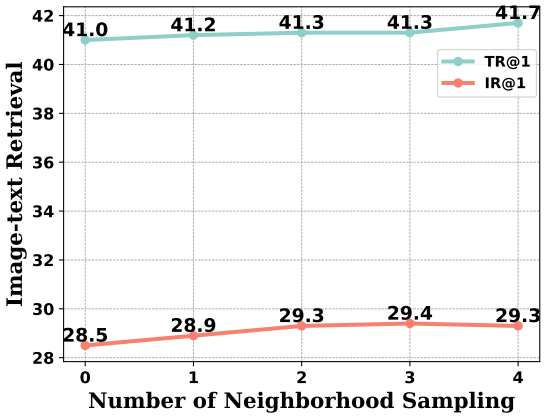


Figure 4: The impact of neighborhood sampling number on defense against multimodal adversarial attacks.

A.6 VISUALIZATION

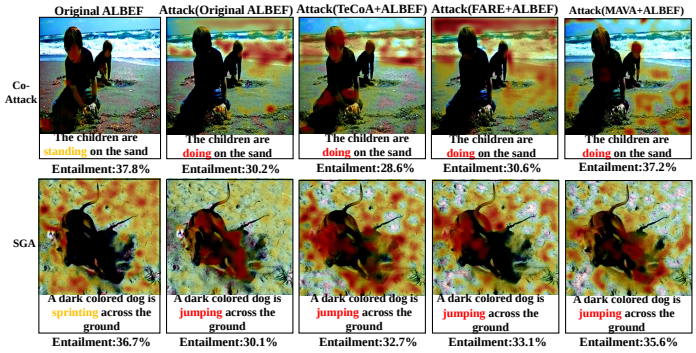
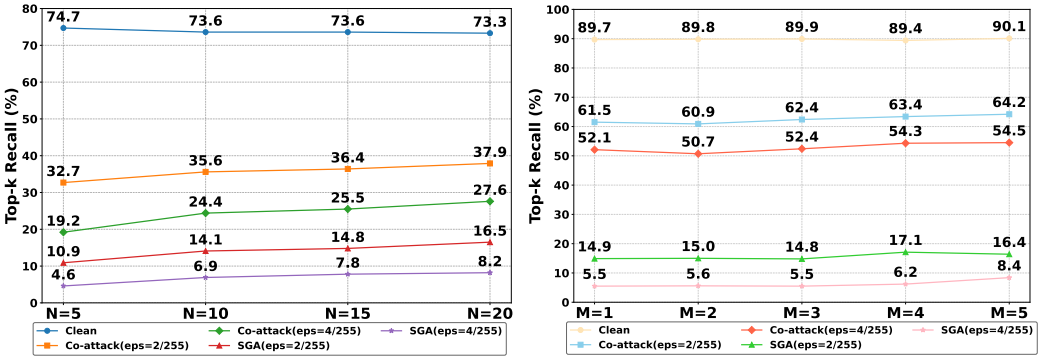


Figure 5: GradCAM visualizations for three adversarial fine-tuning methods on the VE task under Co-attack and SGA.

We present visualizations of adversarial examples from various multimodal attacks against adversarial fine-tuning methods, as shown in Figure 5. The heatmaps reveal that the attention of ALBEF shifts notably for adversarial examples when the vision encoder is replaced with different adversarially fine-tuned versions. Among the methods, MAVA demonstrates the best improvement in adversarial robustness against multimodal adversarial attack compared to other baselines.



(a) The impact of number of iterations. (b) The impact of number of captions.

Figure 6: The impact of iteration count and caption number on defense against multimodal adversarial attacks. We report the average of IR@1 and TR@1 for every adversarial attack.

B In-depth analyses

B.1 Number of Iterations

To further analyze the impact of iterations on defense against multimodal adversarial attacks, we examine the effect of varying iterations during adversarial fine-tuning of CLIP with $\epsilon = \frac{2}{255}$, as shown in Figure 6 (a). Increasing the number of iterations enhances CLIP's robustness but slightly reduces the clean performance of the fine-tuned vision encoder. For consistency with TeCoA, all methods are fine-tuned for 10 epochs.

B.2 Multi Captions for An Image

We also investigate the impact of number of captions on adversarial robustness of vision encoder. The number of captions varies from 1 to M, where M represents the total number of caption pairs for each image, with each caption offering a distinct perspective. As shown in Figure 6 (b), increasing the number of captions enhances the adversarial robustness of ALBEF (fine-tuning vision encoder with $\epsilon = \frac{2}{255}$) against multimodal attacks. These results highlight the effectiveness of cross-modal supervision in our approach.

B.3 Theoretical Analysis of the MAVA

The proposed Multi-granularity Aligned Visual Adversarial Fine-tuning (MAVA) enhances the adversarial robustness of VLP models against multimodal attacks through three key strategies. Below, we provide a theoretical explanation of its core mechanisms.

B.3.1 Optimization Framework for Adversarial Fine-tuning

MAVA adopts a min-max optimization framework (Eq. 7):

$$\min_{\theta} \underbrace{E_{(v,t)\sim D}\left[\max_{\delta\in B(v,\epsilon)} \mathcal{L}_{\text{inner}}(\theta, v+\delta)\right]}_{\text{Adversarial Attack Generation}} + a\cdot\mathcal{L}_{\text{CMS}} + b\cdot\mathcal{L}_{\text{VSSE}} + c\cdot\mathcal{L}_{\text{SCA}} \tag{9}$$

where the **inner maximization** generates adversarial perturbations to maximize the loss $\mathcal{L}_{\text{inner}}$, while the outer minimization optimizes model parameters θ under multi-granularity constraints. This framework theoretically improves robustness by fine-tune the model on worst-case adversarial examples, ensuring parameter updates account for perturbed inputs.

B.3.2 SMOOTHNESS CONSTRAINT VIA VISION SEMANTIC SPACE EXPANSION (VSSE)

$$\mathcal{L}_{\text{VSSE}}(\theta) = \underset{v\sim\mathcal{D}}{E}\left[\frac{1}{N}\sum_{i=1}^{N} \parallel \mathcal{F}(v' + r_i) - \mathcal{F}_{org}(v) \parallel_2^2\right], \tag{10}$$

where r_i denotes uniformly sampled perturbations. The theoretical implications are two fold:

Local Smoothness: By penalizing abrupt feature changes in the neighborhood of adversarial examples $(v' + r_i)$, $\mathcal{L}_{\text{VSSE}}$ prevents overfitting to specific adversarial perturbations.

Feature Stability: Minimizing the L_2-distance between perturbed and original features $\mathcal{F}_{org}(v)$, aligns with Lipschitz continuity assumptions Zühlke & Kudenko (2024), limiting the sensitivity of visual representations to input perturbations.

B.3.3 CROSS-MODAL SUPERVISION (CMS) FOR SEMANTIC ALIGNMENT

$\mathcal{L}_{\text{CMS}}$ leverages textual supervision to guide adversarial feature learning.

$$\mathcal{L}_{\text{CMS}}(\theta) = \underset{v\sim\mathcal{D}}{E}\left[\frac{1}{M}\sum_{i=1}^{M} \frac{\mathcal{F}(t_i)\cdot\hat{\mathcal{F}}(v')}{\parallel \mathcal{F}(t_i) \parallel\parallel \hat{\mathcal{F}}(v') \parallel}\right], \tag{11}$$

where $\hat{\mathcal{F}}(v') = \frac{1}{N}\sum_{j=1}^{N}\mathcal{F}(v' + r_j)$. This strategy ensures:

Multimodal Consistency: Maximizing cosine similarity between perturbed image features $(\hat{\mathcal{F}}(v'))$ and paired text embeddings $\mathcal{F}(t_i)$ preserves semantic alignment under attacks.

Gradient Diversity: Textual supervision introduces additional gradient directions during optimization, preventing the vision encoder from converging to suboptimal local minima and improving cross-modal generalization.

B.3.4 LOSS TRADE-OFFS AND GENERALIZATION GUARANTEES

The composite loss

$$\mathcal{L}_{\text{outer}} = a\cdot\mathcal{L}_{\text{CMS}} + b\cdot\mathcal{L}_{\text{VSSE}} + c\cdot\mathcal{L}_{\text{SCA}}, \tag{12}$$

requires careful tuning of hyperparameters (a, b, c) to balance clean-data accuracy and adversarial robustness. Theoretically:

Hyperparameter Sensitivity: Overweighting b may suppress semantic learning, while a small a weakens cross-modal alignment.

Generalization Bounds: Following PAC-Bayes theory, the generalization error of adversarially fine-tuned models is bounded by perturbation radius ϵ and model complexity. MAVA reduces the effective complexity through feature smoothness and cross-modal alignment, tightening the generalization bound.

B.4 SCALABILITY

Scalability is a critical consideration for adversarial fine-tuning methods deployed in real-world applications. Here, we analyze scalability from two key perspective.

Model Architecture Compatibility. Our proposed MAVA is evaluated on standard VLP models (e.g., CLIP, ALBEF, BLIP) with ViT-B/16 vision encoders. While scaling to larger architectures (e.g., ViT-L/H) would increase computational demands due to higher parameter counts, MAVA's design—including neighborhood sampling and cross-modal alignment—is architecture-agnostic, suggesting compatibility with larger models given sufficient resources.

Data Sampling Efficiency. Vision Semantic Space Expansion (VSSE) involves sampling N perturbed images per adversarial example. While increasing N linearly raises computational cost, it may enhance robustness by providing more comprehensive feature alignment. Similarly, Cross-modal Supervision (CMS) leverages M captions per image, with larger M improving supervision quality at the expense of proportional computational overhead. The scalability of the method depends on an effective trade-off between N and M, where smaller values can be adopted to improve efficiency without significantly degrading performance.

B.5 COMPUTATIONAL COST

The time complexity of MAVA is primarily determined by the following components:

1. Adversarial Example Generation.

- **Inner Maximization:** For each training iteration, MAVA generates adversarial examples by solving the inner maximization problem (e.g., using PGD). The time complexity depends on the number of PGD steps (K), the dimensionality of the input space (d) and the cost of computing gradients for the VLP model $\mathcal{F}_\theta$.
- **Complexity:** $O(K \cdot d \cdot T_{\text{gradient}})$ where T_{gradient} is time to compute gradients for $\mathcal{F}_\theta$.

2.Vision Semantic Space Expansion (VSSE).

- **Neighborhood Sampling:** For each adversarial example, MAVA samples N points in its neighborhood and computes their features.
- **Complexity:** $O(N \cdot d \cdot T_{\text{forward}})$ where T_{forward} is the time for a forward pass through $\mathcal{F}_\theta$.

3. Cross-Modal Supervision (CMS).

- **Text-Image Alignment:** For each image, MAVA computes the similarity between its features and M paired text embeddings.
- **Complexity:**$O(M \cdot d \cdot T_{\text{forward}})$

4. Semantic Consistency Alignment (SCA).

Measures divergence between fine-tuned and original features $(O(d))$, but requires storing original model features, increasing memory consumption.

5. Overall Time Complexity.

The total time complexity per iteration is: $O(K \cdot d \cdot T_{\text{gradient}} + N \cdot d \cdot T_{\text{forward}} + M \cdot d \cdot T_{\text{forward}} + d)$