An Inverse-Adversarial and Difficulty-Adaptive Robust Vision–Language Model

Ruobing Xu[1,2], Junhao Dong[3], and Xiaohua Xie[1,2]*

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, 510006, China
[2] Guangdong Province Key Laboratory of Information Security Technology, Guangzhou, 510006, China
[3] Nanyang Technological University, 639798, Singapore
xurb5@mail2.sysu.edu.cn
junhao003@ntu.edu.sg
xiexiaoh6@mail.sysu.edu.cn

Abstract. Vision-language models (VLMs) pre-trained on large-scale datasets demonstrate strong generalization in few-shot scenarios, often surpassing conventional visual models on natural samples. However, they remain vulnerable to adversarial perturbations, and directly applying adversarial training may disrupt the cross-modal alignment, degrading generalization to novel classes. To mitigate these issues, we propose an Inverse-Adversarial and Difficulty-Adaptive Robust Vision-Language Model (IADA-RVLM) for robust few-shot classification. Specifically, to constrain the optimization trajectory, our method jointly leverages adversarial and inverse-adversarial samples, where inverse-adversarial samples serve as semantic anchors to guide representations toward high-confidence regions. To improve robustness, we incorporate visual prompt tuning and text adapters to inject task-specific knowledge into both modalities. To balance robustness and generalization under varying scenarios, we quantify adversarial transfer difficulty via semantic distances and dynamically adjust the fusion of general and specialized knowledge. Extensive experiments across multiple datasets and attack settings demonstrate that our method achieves a superior overall trade-off between robustness and generalization.

Keywords: Robust Vision-Language Models · Few-Shot Learning · Inverse-Adversarial Learning · Adversarial Transfer Difficulty

1 Introduction

Pretrained vision-language models (VLMs) trained on large-scale image-text pairs have achieved strong performance across a wide range of downstream tasks [18,19]. By aligning visual representations with natural language semantics in a shared embedding space, such models enable strong generalization across

* Corresponding author: Xiaohua Xie.

tasks and domains. A representative example is CLIP [24], which learns contrastive image–text alignment and achieves remarkable few-shot recognition performance. Besides, its variants exhibit strong representation transferability under prompt-based adaptation [36]. These capabilities make CLIP-style VLMs particularly attractive in scenarios where data are scarce, such as medical imaging, remote sensing, and scenarios where acquiring labeled data is difficult or costly.

Despite their impressive generalization on natural images, recent studies reveal that VLMs remain highly vulnerable under adversarial perturbations [17]. Ensuring their safety is particularly crucial in data-scarce, high-stakes applications where model predictions can substantially affect real-world decisions, such as medical image analysis and traffic sign recognition. Adversarial examples crafted for conventional vision models transfer effectively to multimodal architectures, and CLIP-style models can even exhibit weaker robustness than conventional vision models [13,16,25,31]. More critically, directly applying standard adversarial training or adversarial fine-tuning to VLMs often disrupts the pre-trained cross-modal alignment, leading to a severe degradation of natural accuracy and generalization ability, especially in few-shot regimes [6,26,28,30]. This robustness–generalization conflict indicates that improving robustness in VLMs cannot simply rely on classical adversarial learning paradigms. Recent studies show that adversarial perturbations in multimodal models not only distort visual representations but also propagate misalignment into the shared embedding space, thereby degrading cross-modal semantic consistency [34, 37]. Such structural vulnerability distinguishes VLMs from single-modal vision models and calls for multimodal-aware designs under limited-data adaptation.

To reduce the cost of full-parameter adversarial fine-tuning on large VLMs, recent robust adaptation methods mainly adopt parameter-efficient tuning strategies based on prompts or lightweight adapters, e.g., visual prompt tuning [2, 21], text prompt tuning [15,31], multimodal prompt frameworks [14,38], and adapter-based robust tuning such as AdvCLIP-LoRA [9]. Although these approaches improve adversarial robustness to some extent, they still have limitations: existing methods often perform adversarial fine-tuning on base classes with fixed adaptation policies, which introduces domain-specific priors and may harm generalization to unseen novel classes. Moreover, most approaches treat adversarial samples uniformly across datasets and categories [31], ignoring that adversarial perturbations may induce drastically different levels of semantic destruction and robust transfer difficulty. As a consequence, a single fixed robustness-enhancement strategy is suboptimal: for low-difficulty adversarial cases, excessive task-specific adaptation may cause overfitting and reduce generalization, whereas for high-difficulty cases, relying solely on frozen general knowledge is insufficient to establish stable decision boundaries.

To address these issues, we propose **IADA-RVLM**, an Inverse-Adversarial and Difficulty-Adaptive Robust Vision-Language Model for robust few-shot CLIP adaptation. IADA-RVLM is designed as an integrated framework to improve adversarial robustness while preserving generalization. To stabilize optimization, we construct inverse-adversarial samples [4, 5] as semantic anchors that guide

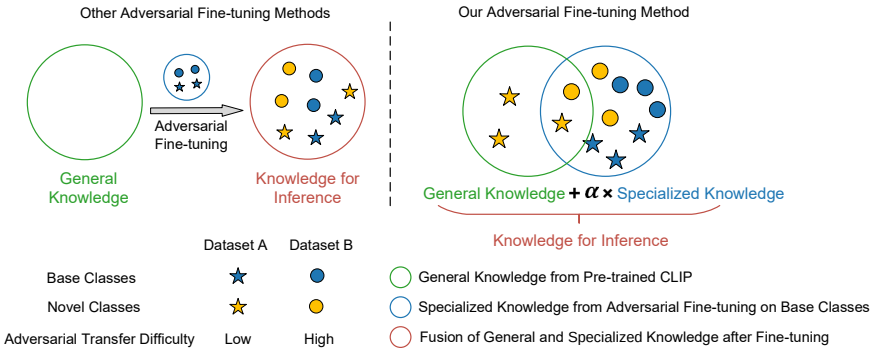


Fig. 1: Illustration of the difference between existing methods and our approach. Existing methods (left) adversarially fine-tune on base classes, but the resulting representation shift can hurt generalization on natural samples. Our method (right) adaptively adjusts the fusion ratio between general and specialized knowledge based on robust transfer difficulty, achieving better robustness and generalization across scenarios.

representations toward high-confidence regions, mitigating erroneous alignment caused by misclassified natural samples in standard adversarial training. Moreover, we propose a parameter-efficient expertise module that combines visual prompt learning (VPT) with a lightweight text adapter (TA) to inject specialized knowledge from few-shot data. VPT improves adversarial feature separability and TA refines decision boundaries. To balance performance on natural and adversarial samples, we design a difficulty-adaptive knowledge fusion mechanism (see Figure 1). It quantifies class-level adversarial transfer difficulty via semantic distances in the embedding space and dynamically adjusts the usage ratio of general and specialized knowledge during inference. This design preserves general semantics for low-difficulty adversarial samples while strengthening task-specific knowledge for high-difficulty ones, achieving stable performance under heterogeneous adversarial scenarios.

Extensive experiments are conducted on datasets with diverse adversarial transfer difficulties, including Caltech101 [7], OxfordPets [23], Flowers102 [22], DTD [3], and EuroSAT [12]. We follow a base-to-novel evaluation protocol by training on base classes and testing on novel classes, where natural samples evaluate generalization and adversarial samples evaluate robustness. These results demonstrate that IADA-RVLM consistently achieves superior robustness–generalization trade-offs under strong attacks, outperforming state-of-the-art robust prompt-based VLMs.

Our main contributions are summarized as follows:

– We propose **IADA-RVLM**, an inverse-adversarial and difficulty-adaptive robust vision-language model for few-shot robust classification. By jointly tuning with adversarial and inverse-adversarial samples, our method stabi-

lizes the optimization trajectory and boosts both robustness and generalization without large-scale re-pretraining.
– We introduce a parameter-efficient expertise adaptation module combining visual prompt tuning and a lightweight text adapter to inject specialized knowledge. Moreover, a difficulty-adaptive mechanism based on semantic distances dynamically balances general and specialized knowledge across adversarial scenarios of varying difficulty.
– Extensive experiments and visualizations across multiple datasets and attack settings demonstrate that the proposed method achieves a superior trade-off between robustness and generalization.

2 Related Work

2.1 Adversarial Learning

Adversarial training is a primary defense strategy for improving robustness. Its core idea is to incorporate adversarial samples during training so that the model learns perturbation-stable features and becomes more resilient at inference time. Goodfellow *et al.* [11] first advocated training with adversarial samples, and Madry *et al.* [20] formalized adversarial training as a robust min–max optimization problem that minimizes the classification loss under worst-case bounded perturbations:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\max_{\|\delta\|\le\epsilon} \mathcal{L}\big(f_\theta(x+\delta), y\big)\Big]. \tag{1}$$

Here, θ denotes model parameters, δ is a bounded perturbation, and $\mathcal{L}(\cdot)$ is the classification loss. Despite its effectiveness, standard adversarial training is computationally expensive and may suffer from limited robust generalization and degraded natural accuracy. Recent work addresses these issues by improving robust feature quality and transferring robust knowledge. For example, Mao *et al.* [21] designed text-guided contrastive adversarial losses to enhance semantic consistency. Robust distillation methods transfer robustness from teachers to students via robust soft labels [10, 39], while multi-teacher distillation further improves stability on both natural and adversarial inputs [33]. Andonian *et al.* [1] proposed progressive self-distillation to reduce the reliance on external teachers.

2.2 Vision-Language Models and Efficient Adaptation

Vision-language models (VLMs) learn joint representations from paired image–text data. CLIP [24] is a representative framework that aligns visual and textual embeddings via contrastive learning. Given a batch of N image–text pairs, CLIP optimizes a symmetric InfoNCE objective:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}\Big(\mathcal{L}_{\text{i2t}} + \mathcal{L}_{\text{t2i}}\Big), \qquad \mathcal{L}_{\text{i2t}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\langle v_i, t_i\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle v_i, t_j\rangle/\tau)}, \tag{2}$$

where $v_i = f_{\text{img}}(x_i)$ and $t_i = f_{\text{text}}(s_i)$ denote normalized embeddings, and τ is a temperature parameter. To adapt large VLMs efficiently, parameter-efficient tuning has been widely explored, mainly falling into two categories: prompt learning and adapter-based tuning. Prompt learning optimizes learnable contexts or tokens to adapt VLMs while keeping most parameters frozen [14, 35, 36]. Adapter-based tuning inserts lightweight modules (e.g., linear or bottleneck adapters) while freezing the backbone, providing additional capacity with limited trainable parameters [8, 27, 32]. These methods are attractive for few-shot adaptation because they reduce overfitting risk and computational cost compared with full fine-tuning.

2.3 Robust Vision-Language Models

Most robust vision–language models improve adversarial robustness by incorporating adversarial fine-tuning. Since full-parameter adversarial fine-tuning is expensive for large VLMs, existing methods typically adopt parameter-efficient prompt-based adaptation. On the vision side, C-AVP [2] and TeCoA-VPT [21] introduce learnable visual prompts to enhance the stability of visual representations under perturbations. C-AVP optimizes class-wise adversarial visual prompts, while TeCoA-VPT uses text-guided contrastive adversarial objectives to reinforce correct image–text alignment. On the text side, AdvPT [31] aligns learnable text prompts with adversarial image features via an adversarial embedding bank, whereas APT [15] focuses on directly learning robust textual prompts. Beyond single-modality prompts, multimodal frameworks such as AdvMaPLe [14] and FAP [38] jointly optimize visual and textual prompts to further improve robustness in few-shot settings. Despite these advances, most methods adversarially fine-tune on specific domains, which may introduce domain-biased priors and degrade generalization to unseen classes. Moreover, they often treat adversarial samples uniformly across datasets and classes, overlooking heterogeneous robust transfer difficulty.

3 Method

We first introduce inverse-adversarial fine-tuning in Section 3.1 and describe how to jointly train with adversarial and inverse-adversarial samples as semantic anchors. We then present our parameter-efficient expertise adaptation module in Section 3.2, which combines visual prompt tuning and a lightweight text adapter. Finally, we detail the difficulty-adaptive mechanism in Section 3.3, which estimates robust transfer difficulty and dynamically balances general versus specialized knowledge at inference time.

3.1 Inverse-Adversarial Fine-tuning

In conventional adversarial training, misclassified or low-confidence natural samples can provide unreliable supervision and harm robustness. To address this,

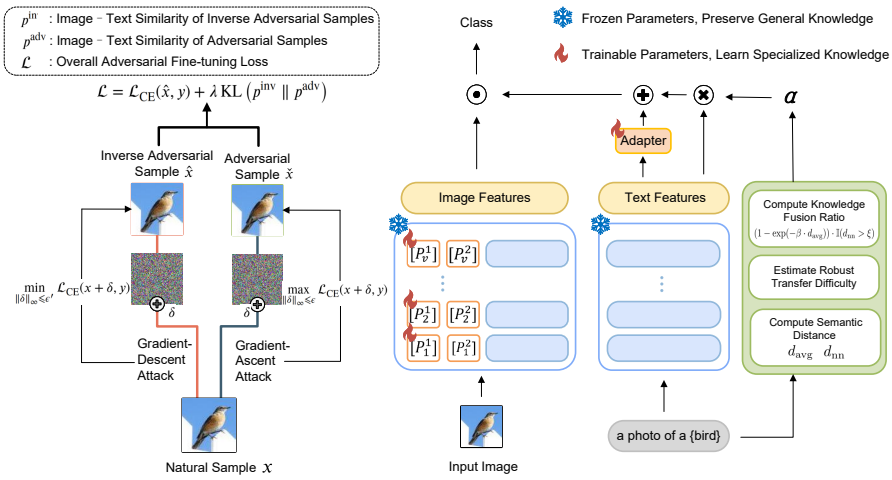


Fig. 2: Overall framework of the proposed method. During adversarial fine-tuning, the model is jointly trained with adversarial and inverse-adversarial samples, and incorporates visual prompt learning and a text adapter to inject task-relevant knowledge. Meanwhile, we design a difficulty-adaptive mechanism to estimate the robust transfer difficulty of adversarial samples and dynamically adjust the fusion ratio between general and specialized knowledge.

we introduce inverse-adversarial learning [5] and jointly train on adversarial and inverse-adversarial samples to constrain the optimization trajectory, improving robustness while mitigating erroneous supervision.

Generation of Inverse Adversarial Samples Standard adversarial training implicitly assumes that each natural sample provides reliable supervision. However, when a sample is low-confidence or already misclassified, its adversarial counterpart may further drift toward incorrect regions, which can amplify erroneous alignment and lead to robust overfitting in few-shot CLIP fine-tuning.

Inverse-adversarial learning addresses this issue by moving in the opposite direction of attacks. It minimizes the classification loss within a bounded neighborhood, pushing the input toward a higher-confidence region of the correct class. In CLIP's joint embedding space, this encourages the image feature to align with the corresponding text semantics, providing a more reliable semantic anchor.

Formally, given an input image x and its label y, the inverse-adversarial sample is obtained by solving:

$$\check{x} = \arg \min_{\|\delta\|_\infty \le \epsilon'} \mathcal{L}_{\text{CE}}(x+\delta, y), \tag{3}$$

where ϵ' denotes the maximum magnitude of inverse perturbation and $\mathcal{L}_{\mathrm{CE}}$ is the cross-entropy loss. Under CLIP classification, image–text similarities are transformed by softmax into a predictive distribution, based on which the cross-entropy loss is computed. We generate $\check{x}$ via the following projected gradient descent updates:

$$\check{x}^{(k+1)} = \Pi_{\mathcal{B}(x,\epsilon')} \left(\check{x}^{(k)} - \alpha' \cdot \mathrm{sign} \left(\nabla_{\check{x}^{(k)}} \mathcal{L}_{\mathrm{CE}}(\check{x}^{(k)}, y) \right) \right). \tag{4}$$

Here, α' is the step size and $\Pi_{\mathcal{B}}$ denotes the projection operator enforcing the perturbation constraint.

Joint Training with Adversarial and Inverse-Adversarial Samples Although inverse-adversarial samples provide reliable semantic anchors, they do not directly enhance defensive capability. Therefore, adversarial samples are still necessary to improve robustness under attacks. Our model adopts joint training with both adversarial samples and inverse-adversarial samples, combining their complementary roles. For an adversarial sample $\hat{x}$, its generation objective maximizes the classification loss:

$$\hat{x} = \arg \max_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}_{\mathrm{CE}}(x + \delta, y). \tag{5}$$

During fine-tuning, model parameters are updated by minimizing the cross-entropy loss on adversarial samples:

$$\theta \leftarrow \arg \min_{\theta} \mathcal{L}_{\mathrm{CE}}(\hat{x}, y). \tag{6}$$

Furthermore, we introduce inverse-adversarial guidance to constrain the optimization direction of adversarial samples, avoiding erroneous semantic alignment caused by misclassified natural samples. Specifically, we compute the image–text similarity distributions of adversarial and inverse-adversarial samples over textual prototypes:

$$p_j^{\mathrm{adv}} = \frac{\exp\left(\langle f_{\mathrm{img}}(\hat{x}), f_{\mathrm{text}}(t_j) \rangle / \tau\right)}{\sum_k \exp\left(\langle f_{\mathrm{img}}(\hat{x}), f_{\mathrm{text}}(t_k) \rangle / \tau\right)}, \tag{7}$$

$$p_j^{\mathrm{inv}} = \frac{\exp\left(\langle f_{\mathrm{img}}(\check{x}), f_{\mathrm{text}}(t_j) \rangle / \tau\right)}{\sum_k \exp\left(\langle f_{\mathrm{img}}(\check{x}), f_{\mathrm{text}}(t_k) \rangle / \tau\right)}. \tag{8}$$

We then impose a distribution alignment constraint by minimizing the KL divergence between them:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}}(\hat{x}, y) + \lambda \, \mathrm{KL}\left(p^{\mathrm{inv}} \parallel p^{\mathrm{adv}}\right), \tag{9}$$

where λ is a weighting coefficient balancing robustness improvement and semantic consistency. This joint optimization enables the model to maintain stable image–text alignment under adversarial perturbations while preventing incorrect semantic drift induced by low-confidence samples, resulting in a more robust and stable fine-tuning process.

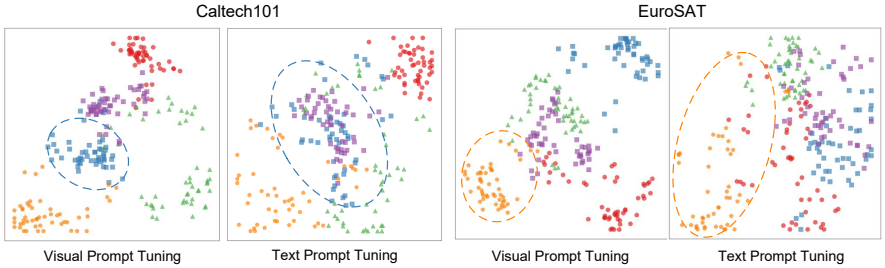


Fig. 3: Feature distribution comparison between adversarial fine-tuning with visual prompt tuning and text prompt tuning.

3.2 Expertise Adaptation Module

In few-shot learning, fine-tuning on base classes can inject task-specific expertise, but full-parameter updates easily overfit under limited supervision, especially under attack. Therefore, we propose a parameter-efficient expertise adaptation module that combines prompt learning and adapter-based tuning on the visual and textual encoders, improving robustness and generalization.

Visual Prompt Tuning In VLMs, prompts can be introduced into either the visual encoder or the text encoder. When applied to the visual encoder, this strategy is referred to as Visual Prompt Tuning (VPT). Prompts on the text encoder correspond to Text Prompt Tuning (TPT). VPT inserts a small set of learnable prompt tokens into intermediate Transformer layers so they can interact with patch features via self-attention and steer visual representations. Formally, assume that the CLIP visual encoder consists of L Transformer layers. At layer ℓ, the input is composed of the class token $\mathbf{c}^{(\ell)}$ and patch embeddings $\mathbf{E}^{(\ell)}$. VPT introduces K learnable prompt vectors $\mathbf{P}^{(\ell)} = \{\mathbf{p}_1^{(\ell)}, \ldots, \mathbf{p}_K^{(\ell)}\}$, which are concatenated with the original features and fed into the Transformer block:

$$\left[\mathbf{c}^{(\ell+1)}, \mathbf{E}^{(\ell+1)}\right] = \text{Transformer}^{(\ell)}\Big(\left[\mathbf{c}^{(\ell)}, \mathbf{E}^{(\ell)}, \mathbf{P}^{(\ell)}\right]\Big), \tag{10}$$

where $\mathbf{P}^{(\ell)}$ denotes learnable prompt parameters, while the remaining backbone parameters are kept frozen. This injects task-specific guidance without updating the backbone, while TPT adjusts class text embeddings via learnable contexts.

In adversarial fine-tuning, perturbations can reduce inter-class separability and mix features in the embedding space. VPT acts directly on the visual encoder to reshape representations during feature extraction. VPT reduces intra-class variance and enlarges inter-class margins, making feature clusters more compact under adversarial fine-tuning. Together with inverse-adversarial training, it provides stable visual constraints that support semantic guidance (Figure 3, left). In contrast, TPT modifies text-side class prototypes while keeping visual features

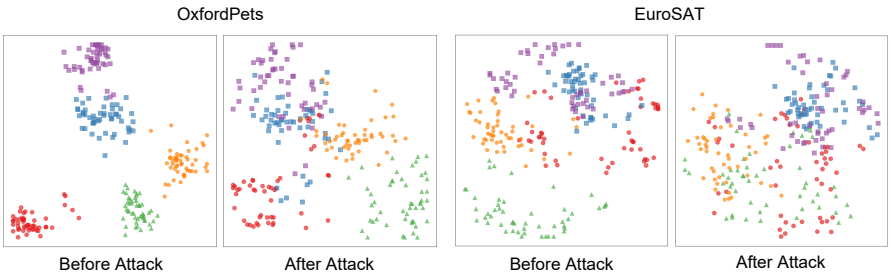


Fig. 4: Feature distribution changes before and after adversarial attacks on OxfordPets and EuroSAT. Different datasets exhibit significantly different degrees of feature destruction, indicating heterogeneous robust transfer difficulty.

fixed, which can overfit base classes when adversarial perturbations entangle visual features (Figure 3, right). Therefore, our expertise adaptation module adopts VPT as the prompt learning strategy.

Text Adapter Adaptation Although VPT optimizes the visual feature space, it does not explicitly adjust the model's semantic decision structure. In high-difficulty adversarial scenarios, visual separability alone may be insufficient. Thus, we introduce an adapter-based design. In vision-language models, adapters can be inserted on either modality side (text or vision). We adopt a lightweight linear text adapter. Let the frozen text encoder output be $\tilde{\mathbf{t}} \in \mathbb{R}^d$. The text adapter performs a linear mapping after the frozen encoder output:

$$\mathbf{t} = \mathrm{TxtAdapt}(\tilde{\mathbf{t}}) = \mathbf{A}^\top \tilde{\mathbf{t}} + \mathbf{b}, \tag{11}$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\mathbf{b} \in \mathbb{R}^d$ are learnable parameters. This design preserves CLIP's general knowledge while injecting task-specific expertise. Compared with visual adapters, text adapters are often safer for adversarial fine-tuning in CLIP, since text embeddings serve as class prototypes and can be adjusted with minimal impact on the frozen visual encoder. VPT and TA are complementary: VPT structures visual features, while TA refines semantic prototypes.

3.3 Difficulty-Adaptive Mechanism

In adversarially robust fine-tuning, adversarial samples from different datasets exhibit varying transfer difficulty. We define a metric to quantify robust adversarial transfer difficulty, inspired by Relative Transfer Difficulty (RTD) [29]. Specifically, let $f(\cdot)$ denote a random classifier with accuracy $1/C$, and let $g(\cdot)$ denote a 5-shot baseline model trained on frozen CLIP features with a simple classifier head. Let Prec_f and $\mathrm{Prec}_g^{\mathrm{adv}}$ denote the accuracy of the random classifier and the adversarial accuracy of the baseline model, respectively. We define

Table 1: Natural accuracy (Nat, %), robust accuracy (Rob, %) under 100-step PGD, and their harmonic mean (HM, unitless) on novel classes across datasets with different robust transfer difficulty levels.

Model	Caltech101			OxfordPets			Flowers102			DTD			EuroSAT		
	Nat	Rob	HM	Nat	Rob	HM	Nat	Rob	HM	Nat	Rob	HM	Nat	Rob	HM
CLIP	95.34	2.80	5.44	91.41	0.08	0.16	74.47	0.00	0.00	54.69	0.00	0.00	47.59	0.00	0.00
C-AVP	74.81	46.38	57.24	52.30	15.37	23.73	45.23	15.77	23.39	13.23	7.17	9.30	25.50	10.68	15.05
TeCoA-VPT	72.92	45.92	56.36	48.72	17.51	25.77	42.72	16.24	23.54	22.05	7.88	11.61	20.16	11.97	15.02
AdvPT	77.32	30.62	43.87	70.55	14.84	24.52	50.40	18.99	27.57	35.38	14.13	20.20	25.39	9.07	13.37
APT	78.57	48.98	60.34	68.27	22.98	34.38	51.81	19.26	28.09	34.49	16.36	22.21	25.61	11.22	15.60
AdvMaPLe	74.47	52.12	61.35	70.56	28.16	40.25	49.90	21.07	29.65	21.27	9.97	13.57	26.78	7.83	12.12
FAP	76.11	50.30	60.56	72.13	26.07	38.30	46.67	21.19	29.16	35.57	17.07	23.06	28.72	12.01	16.94
IADA-RVLM	76.54	52.52	**62.27**	72.16	28.43	**40.63**	50.64	24.36	**32.91**	35.02	20.62	**26.02**	28.33	13.76	**19.02**

robust transfer difficulty as:

$$\mathrm{RTD}^{\mathrm{adv}} = \frac{\mathrm{Prec}_f}{\mathrm{Prec}_g^{\mathrm{adv}}} = \frac{1/C}{\mathrm{Prec}_g^{\mathrm{adv}}}. \tag{12}$$

A larger $\mathrm{RTD}^{\mathrm{adv}}$ indicates that stable decision boundaries are harder to establish in few-shot adversarial learning, implying higher robust transfer difficulty. We observe difficulty differences across datasets (see Figure 4), which motivates dynamically balancing general versus specialized knowledge at inference time. While $\mathrm{RTD}^{\mathrm{adv}}$ provides a dataset-level characterization, it cannot capture class-level variations within the same dataset. Thus, we further model difficulty at the class level. In vision-language models, text embeddings serve as class-discriminative vectors, and their geometry reflects semantic relations. If a target class is close to learned classes in text space, transfer is easier, whereas distant classes lead to higher transfer difficulty. Accordingly, we quantify sample difficulty via semantic distances in the text embedding space. Let $\mathbf{t}_{\mathrm{eval}}$ denote the text feature of the evaluation class, and $\{\mathbf{t}_j\}_{j=1}^{C}$ denote the text features of base classes. We define the average semantic distance and nearest-neighbor distance as:

$$d_{\mathrm{avg}} = 1 - \frac{1}{C}\sum_{j=1}^{C} \mathrm{sim}(\mathbf{t}_{\mathrm{eval}}, \mathbf{t}_j)\,, \tag{13}$$

$$d_{\mathrm{nn}} = 1 - \max_{j\in\{1,\ldots,C\}} \mathrm{sim}(\mathbf{t}_{\mathrm{eval}}, \mathbf{t}_j)\,, \tag{14}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. Based on these measures, we introduce an adaptive coefficient α to dynamically balance general and specialized knowledge. Let $\tilde{\mathbf{t}}_{\mathrm{eval}}$ be the frozen general text embedding, and $\mathrm{TxtAdapt}(\tilde{\mathbf{t}}_{\mathrm{eval}})$ be the adapted embedding with expertise injection. The final text representation is fused as:

$$\mathbf{t}_{\mathrm{eval}} = \alpha \cdot \mathrm{TxtAdapt}(\tilde{\mathbf{t}}_{\mathrm{eval}}) + (1-\alpha)\cdot \tilde{\mathbf{t}}_{\mathrm{eval}}. \tag{15}$$

The adaptive weight α is determined by difficulty:

$$\alpha = (1 - \exp(-\beta \cdot d_{\mathrm{avg}})) \cdot \mathbb{I}(d_{\mathrm{nn}} > \xi)\,, \tag{16}$$

Table 2: Robust accuracy (%) of different models on multiple datasets under different attack methods and perturbation strengths.

Attack	Model	Caltech101		OxfordPets		Flowers102		DTD		EuroSAT	
		ϵ=2/255	ϵ=4/255	ϵ=2/255	ϵ=4/255	ϵ=2/255	ϵ=4/255	ϵ=2/255	ϵ=4/255	ϵ=2/255	ϵ=4/255
PGD	CLIP	2.80	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	C-AVP	46.38	41.10	15.37	10.90	15.77	12.13	7.17	5.75	10.68	8.87
	TeCoA-VPT	45.92	40.32	17.51	11.84	16.24	12.58	7.88	5.34	11.97	9.62
	AdvPT	30.62	29.28	14.84	12.36	18.99	18.20	14.13	12.79	9.07	8.16
	APT	48.98	48.55	22.98	21.67	19.26	17.95	16.36	15.10	11.22	10.09
	AdvMaPLe	52.12	50.17	28.16	26.15	21.07	18.28	9.97	7.87	7.83	6.22
	FAP	50.30	49.07	26.07	24.72	21.19	20.49	18.07	16.36	12.01	11.29
	IADA-RVLM	**52.52**	**51.01**	**28.43**	**26.53**	**24.36**	**22.16**	**20.62**	**19.03**	**13.76**	**12.38**
AutoAttack	CLIP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	C-AVP	42.38	38.13	13.37	8.34	13.19	10.12	5.71	4.29	9.84	7.56
	TeCoA-VPT	40.92	37.30	15.16	11.01	15.04	10.79	6.99	5.00	10.71	8.74
	AdvPT	23.62	20.27	11.14	10.03	15.43	14.62	12.04	10.44	8.18	6.89
	APT	44.98	41.51	21.09	20.28	16.07	13.83	15.02	13.19	9.86	8.89
	AdvMaPLe	47.12	46.17	26.39	**24.96**	19.77	17.28	8.37	5.93	5.47	3.21
	FAP	46.30	44.07	**26.34**	23.72	20.52	19.02	17.85	14.18	10.09	9.35
	IADA-RVLM	**48.88**	**46.91**	26.02	24.03	**23.29**	**21.87**	**19.18**	**18.08**	**11.97**	**10.88**

where β is a scaling factor and ξ is a threshold to avoid unnecessary adaptation when evaluation classes are highly similar to base classes.

This mechanism preserves general semantics for low-difficulty adversarial samples to maintain generalization, while strengthening specialized expertise for high-difficulty samples to improve discriminability. By quantifying class-level semantic difficulty and dynamically adjusting knowledge fusion, IADA-RVLM balances robustness and generalization across different adversarial scenarios.

4 Experimental Results and Analysis

4.1 Datasets and Experimental Setup

We evaluate on datasets with different robust transfer difficulties. Low-difficulty datasets include Caltech101, OxfordPets, and Flowers102, while high-difficulty datasets include EuroSAT and DTD. Following the standard few-shot protocol, each dataset is split into base and novel classes. We adversarially fine-tune on base classes and evaluate on novel classes under a 16-shot setting. Results are averaged over 20 random seeds, and experiments are run on a single NVIDIA A6000 GPU with FP16 mixed precision. We report natural accuracy (Nat) on clean samples and robust accuracy (Rob) under adversarial attacks, and use the Robust–Natural Harmonic Mean (HM) to summarize the robustness–generalization trade-off:

$$\mathrm{HM}(\mathrm{Nat}, \mathrm{Rob}) = \frac{2 \cdot \mathrm{Nat} \cdot \mathrm{Rob}}{\mathrm{Nat} + \mathrm{Rob}}. \tag{17}$$

Our backbone is CLIP ViT-B/16. We use prompt length 2 with prompts inserted into 10 visual layers, and a linear text adapter. The fusion scaling is set to $\beta = 4.0$. We train for 15 epochs with Adadelta and cosine annealing, generate adversarial samples online with 2-step PGD ($\epsilon = 2/255$), and evaluate robustness with 100-step PGD ($\epsilon = 2/255$) and AutoAttack.

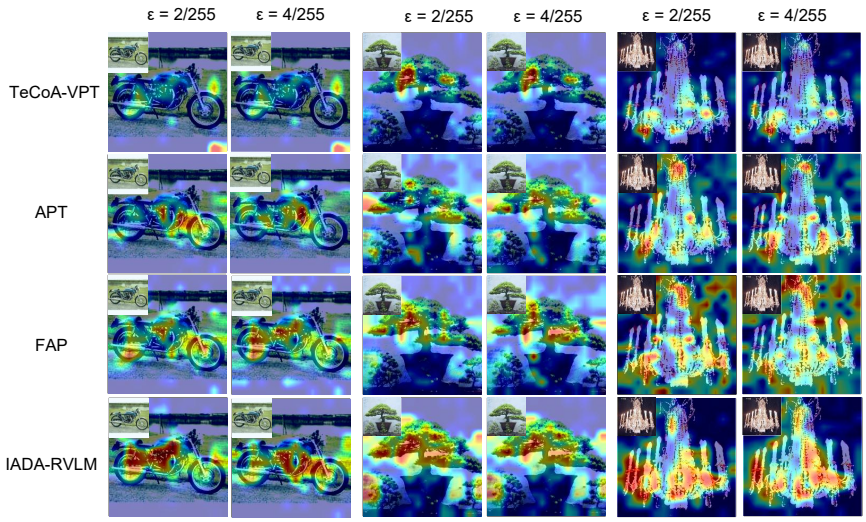


Fig. 5: Grad-CAM visualizations under different attack strengths. Compared with other methods, our method consistently focuses on the target object regions across attack strengths.

4.2 Comparison Experiments

Under the unified evaluation protocol described above, we compare the proposed IADA-RVLM with state-of-the-art robust vision-language models. All methods are evaluated under the same perturbation budgets on novel classes, and we report Nat, Rob, and HM on datasets with different robust transfer difficulty.

As shown in Table 1, IADA-RVLM achieves consistently strong robustness–generalization trade-offs across datasets and obtains the best HM on all benchmarks. It preserves competitive natural accuracy while improving robust accuracy, with more pronounced gains on high-difficulty datasets (e.g., DTD and EuroSAT).

Table 2 reports robust accuracy under different attacks and perturbation strengths. As the perturbation bound increases from $\epsilon = 2/255$ to $\epsilon = 4/255$, all methods degrade, but IADA-RVLM remains the most stable and best-performing across datasets. Under the stronger AutoAttack setting, IADA-RVLM still preserves clear advantages, especially on high-difficulty datasets, indicating good robustness generalization under different attack settings.

To further qualitatively assess robustness under varying perturbation strengths, we visualize model attention regions on adversarial samples using Grad-CAM. We consider two attack strengths, $\epsilon = 2/255$ and $\epsilon = 4/255$, and compare three representative robust fine-tuning strategies, the visual prompt tuning method TeCoA-VPT, the text prompt tuning method APT, and the vision–language joint prompting method FAP against IADA-RVLM. As shown in Figure 5,

Table 3: Ablation study on the expert knowledge adaptation module: comparison of different parameter-efficient fine-tuning strategies in terms of natural accuracy (%) and robust accuracy (%) under 100-step PGD.

Model	Caltech101		EuroSAT	
	Natural	Robust	Natural	Robust
CLIP (fine-tune head only)	72.43	42.78	26.32	9.36
CLIP + TPT	72.51	44.27	26.40	10.28
CLIP + VA	72.35	43.30	26.03	10.00
CLIP + VPT	73.06	45.24	26.85	10.61
CLIP + TA	74.42	48.92	27.41	12.33
CLIP + TPT + VA	72.97	46.29	25.57	10.32
CLIP + VPT + TA	**76.54**	**52.52**	**28.33**	**13.76**

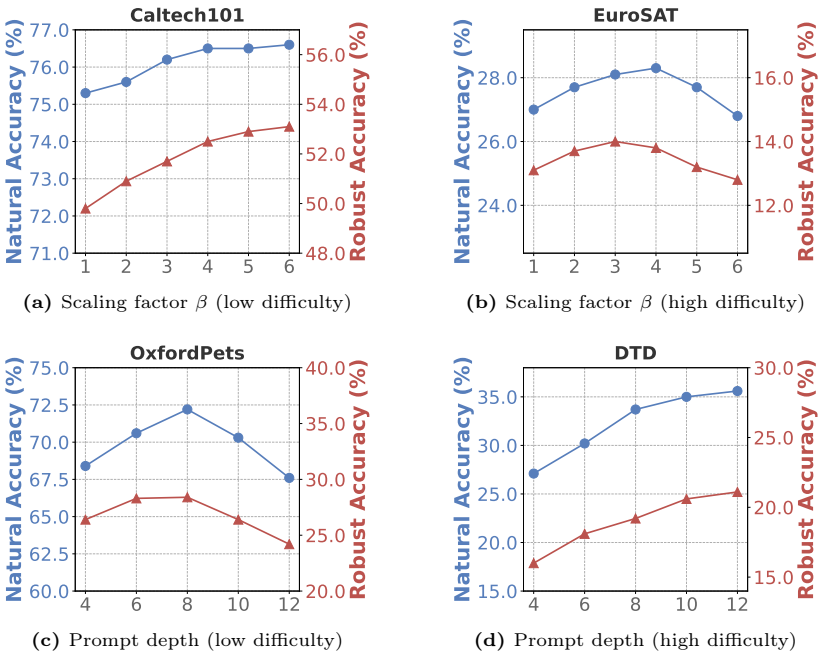
Table 4: Component ablation study on Caltech101 and EuroSAT: natural accuracy (%) and robust accuracy (%) under 100-step PGD.

Inverse-Adv.	Expert Adapt.	Difficulty Adapt.	Caltech101		EuroSAT	
			Natural	Robust	Natural	Robust
			70.20	39.52	22.51	8.25
✓			72.43	42.78	26.32	9.36
	✓		73.04	47.60	25.98	10.96
✓	✓		75.86	50.75	27.25	12.18
	✓	✓	74.75	51.31	27.06	12.82
✓	✓	✓	**76.54**	**52.52**	**28.33**	**13.76**

IADA-RVLM maintains more focused and stable attention on target regions under different perturbation strengths, supporting its improved robustness.

4.3 Ablation Studies

To verify the effectiveness of each component, we conduct ablation studies on Caltech101 and EuroSAT. As shown in Table 3, using only the visual adapter (VA), text prompt tuning (TPT), or their combination brings limited gains. VA tends to erode general knowledge by directly modifying visual features, while TPT requires large updates to text embeddings under few-shot adversarial tuning, increasing overfitting risk. In contrast, visual prompt tuning (VPT) or the text adapter (TA) alone provides moderate robustness improvements, and their combination consistently enhances both natural and robust accuracy. Furthermore, Table 4 shows that integrating inverse-adversarial learning and difficulty-adaptive fusion further strengthens the robustness–generalization trade-off.



(a) Scaling factor β (low difficulty) **(b)** Scaling factor β (high difficulty)

(c) Prompt depth (low difficulty) **(d)** Prompt depth (high difficulty)

Fig. 6: Hyper-parameter analysis under different robust transfer difficulties.

4.4 Hyper-Parameter Analysis

This subsection analyzes the sensitivity of two key hyper-parameters: the fusion scaling factor β and the prompt depth (i.e., the number of prompted visual layers). As shown in Figure 6, larger β tends to help on low-difficulty datasets but can be unstable when overly large on high-difficulty datasets, while deeper prompting is generally more beneficial for high-difficulty datasets. Based on the observed trends, we use $\beta = 4$ and prompt depth 10 in all experiments.

5 Future Work

IADA-RVLM introduces a slight increase in training overhead due to inverse-adversarial sample generation and distribution alignment. Improving computational efficiency and extending the evaluation to more diverse vision-language model backbones remain promising directions for future work.

6 Conclusion

In this paper, we presented IADA-RVLM, an inverse-adversarial and difficulty-adaptive robust vision-language model for robust few-shot classification. To stabilize the optimization trajectory, our method jointly leverages adversarial and

inverse-adversarial samples, where inverse-adversarial samples serve as semantic anchors to guide representations toward high-confidence regions. To improve robustness, we incorporate visual prompt tuning and a text adapter to inject task-specific knowledge into both modalities. To balance robustness and generalization under varying scenarios, we quantify adversarial transfer difficulty via semantic distances and dynamically adjust the fusion of general and specialized knowledge. Extensive experiments across multiple datasets and attack settings demonstrate that IADA-RVLM achieves a more favorable robustness–generalization trade-off than existing robust vision-language tuning approaches.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 12326618 and the Project of Guangdong Provincial Key Laboratory of Information Security Technology under Grant 2023B1212060026.

References

1. Andonian, A., Chen, S., Hamid, R.: Robust cross-modal representation learning with progressive self-distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16430–16441 (2022)
2. Chen, A., Lorenz, P., Yao, Y., Chen, P.Y., Liu, S.: Visual prompting for adversarial robustness. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
3. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014)
4. Dong, J., Moayedi, R.Z., Ong, Y.S., Moosavi-Dezfooli, S.M.: Allies teach better than enemies: Inverse adversaries for robust knowledge distillation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)
5. Dong, J., Moosavi-Dezfooli, S.M., Lai, J., Xie, X.: The enemy of my enemy is my friend: Exploring inverse adversaries for improving adversarial training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24678–24687 (2023)
6. Dong, J., Zhang, C., Qu, X., Ma, Z., Koniusz, P., Ong, Y.S.: Robust superalignment: Weak-to-strong robustness generalization for vision-language models. Advances in Neural Information Processing Systems **38**, 18345–18377 (2025)
7. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop. pp. 178–178. IEEE (2004)
8. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International journal of computer vision **132**(2), 581–595 (2024)
9. Ghiasvand, S., Oskouie, H.E., Alizadeh, M., et al.: Few-shot adversarial low-rank fine-tuning of vision-language models. arXiv preprint arXiv:2505.15130 (2025)

10. Goldblum, M., Fowl, L., Goldstein, T.: Adversarially robust few-shot learning: A meta-learning approach. Advances in Neural Information Processing Systems **33**, 17886–17895 (2020)
11. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
12. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (2019)
13. Huang, H., Erfani, S., Li, Y., Ma, X., Bailey, J.: X-transfer attacks: Towards super transferable adversarial attacks on clip. arXiv preprint arXiv:2505.05528 (2025)
14. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19113–19122 (2023)
15. Li, L., Guan, H., Qiu, J., Spratling, M.: One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24408–24419 (2024)
16. Liu, Y., Ouyang, X., Cui, X.: Gleam: Enhanced transferable adversarial attacks for vision-language pre-training models via global-local transformations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1665–1674 (2025)
17. Luo, B., Guo, J., Yao, Y., Ding, X.: Adversarial style optimization: Enhancing vlm jailbreaks by grpo-based stylistic triggers optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11–19 (2026)
18. Luo, B., Wang, T., Chen, C., Ding, X.: ST-simdiff: Balancing spatiotemporal similarity and difference for efficient video understanding with MLLMs. In: The Fourteenth International Conference on Learning Representations (2026), `https://openreview.net/forum?id=he8kYNcoMA`
19. Luo, B., Wang, T., Chen, H., Ding, X.: Enhancing visual token representations for video large language models via training-free spatial-temporal pooling and gridding. In: The Fourteenth International Conference on Learning Representations (2026), `https://openreview.net/forum?id=MZi9SYPVz5`
20. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
21. Mao, C., Geng, S., Yang, J., Wang, X., Vondrick, C.: Understanding zero-shot adversarial robustness for large-scale models. In: The Eleventh International Conference on Learning Representations (2023)
22. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing. pp. 722–729. IEEE (2008)
23. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)
24. 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: International conference on machine learning. pp. 8748–8763. PmLR (2021)

25. Schlarmann, C., Hein, M.: On the adversarial robustness of multi-modal foundation models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3677–3685 (2023)
26. Schlarmann, C., Singh, N.D., Croce, F., Hein, M.: Robust clip: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In: Proceedings of the 41st International Conference on Machine Learning. pp. 43685–43704 (2024)
27. Xu, Z., Peng, Z., Yang, X., Shen, W.: Fate: Feature-adapted parameter tuning for vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9014–9022 (2025)
28. Yu, C., Han, B., Shen, L., Yu, J., Gong, C., Gong, M., Liu, T.: Understanding robust overfitting of adversarial training and beyond. In: International Conference on Machine Learning. pp. 25595–25610. PMLR (2022)
29. Yu, T., Lu, Z., Jin, X., Chen, Z., Wang, X.: Task residual for tuning vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10899–10909 (2023)
30. Zanella, M., et al.: Low-rank few-shot adaptation of vision-language models. In: CVPR Workshops (2024)
31. Zhang, J., Ma, X., Wang, X., Qiu, L., Wang, J., Jiang, Y.G., Sang, J.: Adversarial prompt tuning for vision-language models. In: European conference on computer vision. pp. 56–72. Springer (2024)
32. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)
33. Zhao, S., Yu, J., Sun, Z., Zhang, B., Wei, X.: Enhanced accuracy and robustness via multi-teacher adversarial distillation. In: European Conference on Computer Vision. pp. 585–602. Springer (2022)
34. 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)
35. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
36. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International journal of computer vision **130**(9), 2337–2348 (2022)
37. Zhou, W., Bai, S., Mandic, D.P., Zhao, Q., Chen, B.: Revisiting the adversarial robustness of vision language models: a multimodal perspective. arXiv preprint arXiv:2404.19287 (2024)
38. Zhou, Y., Xia, X., Lin, Z., Han, B., Liu, T.: Few-shot adversarial prompt learning on vision-language models. Advances in Neural Information Processing Systems **37**, 3122–3156 (2024)
39. Zi, B., Zhao, S., Ma, X., Jiang, Y.G.: Revisiting adversarial robustness distillation: Robust soft labels make student better. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16443–16452 (2021)