Adversarial Prompt Tuning for Vision-Language Models

Jiaming Zhang[1⋆], Xingjun Ma[2], Xin Wang[2], Lingyu Qiu[3], Jiaqi Wang[1], Yu-Gang Jiang[2], and Jitao Sang[1]

[1] Beijing Jiaotong Univisity, Beijing, China
[2] Fudan Univisity, Shanghai, China
[3] Nanjing University of Aeronautics and Astronautics, Nanjing, China

Abstract. With the rapid advancement of multimodal learning, pre-trained Vision-Language Models (VLMs) such as CLIP have demonstrated remarkable capacities in bridging the gap between visual and language modalities. However, these models remain vulnerable to adversarial attacks, particularly in the image modality, presenting considerable security risks. This paper introduces **Adversarial Prompt Tuning (AdvPT)**, a novel technique to enhance the adversarial robustness of image encoders in VLMs. AdvPT innovatively leverages learnable text prompts and aligns them with adversarial image embeddings, to address the vulnerabilities inherent in VLMs without the need for extensive parameter training or modification of the model architecture. We demonstrate that AdvPT improves resistance against white-box and black-box adversarial attacks and exhibits a synergistic effect when combined with existing input denoising defense techniques, further boosting defensive capabilities. Comprehensive experimental analyses provide insights into adversarial prompt tuning, a novel paradigm devoted to improving resistance to adversarial images through textual input modifications, paving the way for future robust multimodal learning research. These findings open up new possibilities for enhancing the security of VLMs. Our code is available at `https://github.com/jiamingzhang94/Adversarial-Prompt-Tuning`.

Keywords: Adversarial defense · Vision-Language models · Prompt tuning

1 Introduction

Large-scale pre-trained Vision-Language Models (VLMs) have demonstrated superb capabilities in generalizing to a wide variety of downstream tasks. These architectures are trained to bridge the gap between visual and language modalities, as demonstrated by the huge amount of web-scale data [26]. With the increasing trend of multimodal learning, there is a growing number of VLMs being released to the public, leading to rapid growth of downstream applications. However,

⋆ This work is done when the author interned at Fudan University.
Corresponding authors: Xingjun Ma and Jitao Sang.

many studies have revealed that VLMs, similar to traditional visual models, are also vulnerable to small adversarial noises, which is a major security threat to deep neural networks (DNNs) [31, 48]. In particular, noise in the image modality is markedly more invisible compared to token replacement in the text modality. Therefore, the imperative task of enhancing the adversarial robustness of the image encoders in VLMs requires an effective solution.

In the image domain, adversarial training (AT) has been proven to be the most effective approach for training robust DNNs against adversarial examples (inputs with adversarial noise) [18]. AT is usually formulated as a min-max optimization problem, which generates adversarial examples at each training iteration to update the image encoder. Therefore, it is computationally expensive, and thus cannot be easily applied to train large VLMs. As such, recent works turn to input pre-processing techniques like diffusion-based purification to improve the adversarial robustness of VLMs [20, 21, 39].

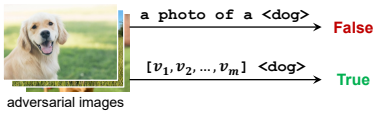


Fig. 1: The defending effect of *AdvPT*: hand-crafted prompts (top) fail to match with adversarial images, whereas prompts constituted by learnable vectors (bottom) enable correct recognition.

Drawing from advances in Natural Language Processing (NLP), we observe a shift from fixed text prompts, such as "`a photo of a <category>`", to learnable prompts in CLIP's text encoder [46, 47]. Such a transition can help enhance image-text matching. Inspired by the idea of prompt tuning, in this work we propose ***Adversarial Prompt Tuning*** (***AdvPT***), a novel approach that improves the adversarial robustness of the image encoders in VLMs using learnable prompts. As depicted in Fig. 1 and Fig. 2, *AdvPT* models the textual prompt with learnable vectors and aligns the clean text embedding with adversarial image embedding to improve adversarial robustness. Specifically, we generate the adversarial images using the image encoder and then compute and save the embeddings of the adversarial images into an adversarial embedding bank. We then discard the image encoder but use the adversarial embedding bank to enhance the adversarial robustness, i.e., we align the clean text embedding with the adversarial image embedding through prompt tuning. This process involves gradient backpropagation through the text encoder to optimize the learnable vectors while preserving the pre-trained parameters. In a nutshell, our *AdvPT* leverages the text encoder's inherent knowledge for rectifying the adversarial embeddings (pre-computed with the image encoder).

We evaluate *AdvPT* against both white-box and black-box adversarial attacks on 8 image datasets, and show that it outperforms the vanilla CLIP (with hand-crafted prompts) by a considerable margin. By focusing on textual input processing and alignment, *AdvPT* opens up a new direction for augmenting adversarial robustness in VLMs. It can also be integrated with image-based defense strategies to further boost the adversarial robustness of the image modality. We

also observe a generalization-robustness trade-off in *AdvPT*, similar to that in traditional AT. We further evaluate the domain transferability of the learnable vectors, testing their performance across various datasets after training on a specific one. Lastly, we conduct an in-depth analysis of the learned vectors and reveal the closest words associated with the vectors, gaining more understanding of the working mechanism of *AdvPT*.

In summary, our main contributions are:

- We propose a novel method *Adversarial Prompt Tuning* (*AdvPT*) to enhance the adversarial robustness of VLMs by aligning the text embeddings with adversarial image embeddings. Specifically, we robustify the image encoder in VLM against adversarial examples using textual prompt modifications.
- We demonstrate the effectiveness of *AdvPT* on various image datasets, showing its superiority over the vanilla CLIP. It can also be combined with input purification methods to further boost the robustness.
- We also provide a set of understandings of the working mechanism of *AdvPT*, the generalization-robustness trade-off, the adaptability of the learned vectors to domain shift, and their linguistic meanings. These understandings can help guide future work to leverage textual input to counter adversarial images.

2 Related Work

2.1 Vision Language Models

VLMs have achieved remarkable success and demonstrated superb capabilities across a wide range of tasks. These models are typically classified into two groups. The first is grounded in large NLP models enhanced with visual modality capabilities, exemplified by GPT-4V [23]. The second group, represented by CLIP [26] and ALIGN [12], treats image and language modalities with equal emphasis. These models acquire joint image-language representations through self-supervised learning from vast data pools. Our study focuses on the latter category of VLMs, specifically on improving the adversarial robustness of their image encoders for image recognition tasks.

2.2 Prompt Learning

The concept of prompt learning originated in the field of NLP. It refers to fine-tuning the prompts instead of model parameters (freezing the model). Research in prompt learning aims to automatically learn more effective prompts instead of using a hand-crafted prompt [15, 17]. This approach has been extended to visual models [13,36] and vision-language models [14,46,47], with the unified objective of enhancing model accuracy through prompt refinement. Our study, while grounded in the CoOp framework [47], diverges in its objective. CoOp represents the initial foray into prompt learning within the visual-language domain, distinguished by its simplicity and rapid processing pipeline. Instead of improving image recognition performance, our focus shifts to leveraging textual input modifications to improve the adversarial robustness of the image encoder.

2.3 Adversarial Defenses

Combating against adversarial images remains an unresolved challenge. Adversarial defenses broadly fall into two camps: model robustification methods and input denoising methods. The former includes methods like AT [18], Fast AT [37], TRADES [42] and MART [35]. This methodology is usually expressed as a min-max optimization problem, with continuous updates to the model parameters across all training iterations. However, this process is computationally demanding, posing difficulties for deployment on VLMs due to the scale of the model and dataset. As a result, the latter approach based on the image process has emerged as a solution suited for VLMs.

The adversarial defense through input image modification is straightforward in its essence. It removes or weakens the impact of adversarial noise through inference-time methods such as input transformations [7,20], smoothing [16,29,44], and rescaling [39]. For example, Xie *et al.* [39] employed random image rescaling to diminish adversarial effects, and Mustafa *et al.* [20] utilized image super-resolution as a defense mechanism. Although somewhat limited in efficacy, these methods are pragmatically valuable for their efficiency. Recently, adversarial purification based on diffusion models has emerged [21,41]. Nie *et al.* [21] introduced the powerful adversarial purification, DiffPure, to address the shortcomings of previous approaches, albeit with increased time complexity. Mao *et al.* [19] identifies that AT of the CLIP on one dataset struggles to impact another dataset, defining this as the zero-shot adversarial robustness problem, and introduced visual prompt tuning [13] to address this.

Our approach deviates from these strategies by not modifying the model nor the input image, presenting a novel defense mechanism against adversarial images. The subsequent sections detail our method and its integration with existing defensive techniques.

3 Revisiting Clip and the Adversarial Robustness of Its Image Encoder

3.1 CLIP

We provide a concise introduction to VLMs, with an emphasis on the CLIP architecture. While our methods are tailored to CLIP, they are potentially extendable to a broader range of VLMs within the contrastive learning framework.

CLIP comprises two distinct encoders: one for images and the other for text. The image encoder aims to distill image embeddings from the input visuals, utilizing either a Convolutional Neural Network (CNN) [8] or a Vision Transformer (ViT) [6] backbone. In contrast, the text encoder relies on a Transformer [32] to generate embeddings from textual data.

During its training phase, CLIP leverages contrastive loss to develop a unified embedding space between visual and language modalities. Upon completion of training, CLIP finds utility in zero-shot image recognition, facilitated through an image-text retrieval mechanism. For example, in the prompt "`a photo of`

`a <class>`", replacing `<class>` with specific categories from a dataset with K classes allows the model to assess the similarity between an image and K textual descriptions.

Denoting input images as x and their corresponding image embeddings from encoder $E(\cdot)$ as e, and considering a set of textual prompts $\{w_i\}_{i=1}^{K}$ as text embeddings produced by text encoder $G(\cdot)$, the prediction probability is mathematically expressed as follows:

$$p(y|e) = \frac{\exp(\text{sim}(e, w_y)/\tau)}{\sum_{i=1}^{K} \exp(\text{sim}(e, w_i)/\tau)}, \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity with a temperature parameter τ.

3.2 Adversarial Robustness of CLIP's Image Encoder

We first introduce our threat model, which describes the assumed knowledge of the adversary, from what inputs they can manipulate to their access to the model architecture and parameters. Our study focuses on the adversarial robustness of image encoders, assuming that the attacker has full knowledge of the model architecture and parameters of image and text encoders, and can perturb the image input. *However, the adversary has no control over the textual input nor knowledge of prompt tuning.* Therefore, text adversarial attacks are also not applicable here [43, 45, 48].

We now introduce the adversarial attacks that target the image encoders. Consider an original input image x, with δ symbolizing adversarial noise. The adversarial example $x' = x + \delta$, once processed by the image encoder E, generates an adversarial embedding e'.

Adversaries can employ two objective functions to impair the accuracy of matching with textual descriptions. The first objective is to make the adversarial embedding e' markedly diverge from the embedding e of the original image, i.e., to maximize the discrepancy between e' and e. The second objective is to ensure the adversarial embedding e' does not align with the corresponding ground-truth textual description embedding w_g, i.e., to maximize the discrepancy between e' and w_g. PGD and AutoAttack are deployed to represent the former and latter objectives, respectively. In this work, we focus on ℓ_∞-norm constrained perturbations, where each δ adheres to $\|\delta\|_\infty \leq \epsilon$, with ϵ denoting the maximum allowable perturbation magnitude.

To defend against adversarial images, existing defense methods generally fall into two categories: model robustification methods and input denoising methods. As mentioned above, model robustification methods like AT struggle to handle VLMs due to efficiency issues. The input denoising operation can be conceptualized as a function h that processes adversarial images, aiming to minimize the disparity between $E(h(x'))$ and $E(x)$.

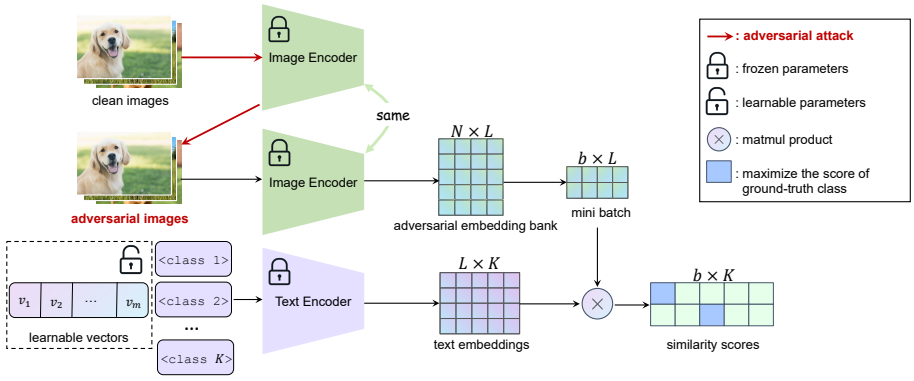


Fig. 2: An overview of the *AdvPT* framework.

4 Adversarial Prompt Tuning

Overview. Our proposed method, *AdvPT*, involves optimizing learnable vectors as text prompts to enhance the robustness against image adversarial attacks. This diverges from previous context optimization approaches [46, 47] aimed at increasing image recognition rates. Fig. 2 provides an framework overview of *AdvPT*. On a K-class dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ of N images and corresponding texts, *AdvPT* begins with feeding the clean images x into the image encoder E to generate its adversarial image x'. The adversarial images are then fed into the image encoder E to obtain the adversarial image embeddings into an adversarial embedding bank $\mathbf{A} \in \mathbb{R}^{N \times L}$, where L is the embedding dimension. The image encoder E is discarded in the subsequent steps. This approach is entirely distinct from traditional defensive methods, which, whether through augmentation (e.g., visual prompt tuning [13,19]) or modification (e.g., AT [18]) of the parameters of CLIP's image encoder branch, rely on **on-the-fly adversarial example generation** during each training epoch. Even with partial parameter tuning (visual prompt), the adversarial example generation necessitates **complete forward and backward propagation** of gradients through the image encoder, resulting in an untenable burden in the context of VLMs.

On the textual side, the prompt for class i is denoted as $[v_1, v_2, \ldots, v_M, c_i]$, with c_i is the embedding representation of the class name. These prompts are then processed by the text encoder G to generate text embeddings $\mathbf{T} \in \mathbb{R}^{L \times K}$. During the fine-tuning process, a mini-batch $\mathbf{B} \in \mathbb{R}^{b \times L}$ with batch size b from $\mathbf{A}$ is used to compute the similarity score $\mathbf{S} = \mathbf{BT} \in \mathbb{R}^{b \times K}$. The objective is to maximize the score of the ground-truth class by optimizing the learnable vectors $V = [v_1, v_2, \ldots, v_M]$, through backpropagation in the text encoder. Overall, the entire process can be roughly divided into two steps: adversarial embedding bank generation and learnable vector optimization. Next, we will introduce the two steps in detail.

4.1 Adversarial Embedding Bank Generation

To improve the image encoder E's adversarial robustness, *AdvPT* first generates adversarial images on encoder E, then re-feeds them into the encoder to obtain and store their adversarial embeddings. Note that *AdvPT* differs greatly from AT, which iteratively generates adversarial examples at each iteration of training and continuously updates the target model on the generated adversarial examples, leading to significant computational costs. Conversely, *AdvPT* fixes the parameters of the image encoder E, channeling focus exclusively on updating the learnable vectors at the input of the text encoder G. This strategy significantly diminishes the number of learnable parameters. With the image encoder E frozen, the generation of the adversarial examples is only a one-pass process. These examples, once processed through E, constitute the adversarial embedding bank $\mathbf{A}$. After this step, the image encoder E is discarded, leaving only the adversarial embedding bank $\mathbf{A}$ for the subsequent prompt tuning.

We employ the PGD attack [18] to generate adversarial images x' on the image encoder $E(\cdot)$ with θ. This process can be formulated as:

$$x' = x'_{(t+1)} = \mathit{\Pi}_{x+\Omega}(x'_{(t)} + \alpha \cdot \text{sign}(\nabla_x J(\theta; x'_{(t)}, x)), \tag{2}$$

where $x'(t)$ represents the adversarial example at iteration t. $\mathit{\Pi}$ is the projection. Ω is the feasible region of δ, which ensures that the perturbed example remains within the allowed limits ϵ. α is the step size for each iteration. $\nabla_x J(\theta; x'_{(t)}, x)$ computes the gradient of the loss function J with respect to the parameters θ of E, wherein J serves as a distance metric quantifying the discrepancy in embeddings between $e' = E(x')$ and $e = E(x)$. In our research, we utilize the Kullback-Leibler Divergence, as in TRADES [42], to serve as our adversarial loss function.

The design of the adversarial embedding bank presents significant advantages. Primarily, it eliminates the need for redundant forward and backward passes through the image encoder, thereby greatly saving computational time. Moreover, the embedding space's lower dimensionality compared to the original image space substantially reduces the required computational memory.

4.2 Learnable Vector Optimization

The next phase in *AdvPT* involves the construction and optimization of the learnable vectors. Specifically, our method seeks to model textual prompts with learnable vectors $V = [v_1, v_2, \ldots, v_M]$, optimized by aligning them with adversarial embeddings, thus rectifying the non-robust features of the images utilized by the model. Initially, the text prompts $[v_1, v_2, \ldots, v_M, c_i]_{i=1}^K$ are fed into the text encoder G, producing text embeddings $\mathbf{T} = [w_1, w_2, \ldots, w_K]^T \in \mathbb{R}^{L \times K}$. In the fine-tuning phase, each iteration retrieves a mini-batch $\mathbf{B} = [e'_1, e'_2, \ldots, e'_b] \in \mathbb{R}^{b \times L}$ from the adversarial embedding bank $\mathbf{A}$. Subsequently, the similarity scores $\mathbf{S} = \mathbf{BT} \in \mathbb{R}^{b \times K}$ can be calculated, with each element representing the predic-

Algorithm 1 Adversarial Prompt Tuning Pipeline

Input: image encoder E, text encoder G, images x and class name c, perturbation restriction ϵ, iteration t
Output: learnable vectors $[v_1, v_2, \ldots, v_M]$
$x' = \text{attack}(x, \epsilon, t; E)$
$\mathbf{A} = E(x')$
Initialize learnable vectors $V = [v_1, v_2, \ldots, v_M]$
for B in iter($\mathbf{A}$) **do**
    Initialize θ_i
    $\mathbf{T} = G([V, c])$
    $\mathbf{S} = \mathbf{BT}$
    Optimize $V \leftarrow$ Maximize $\mathbf{S}$
end for

tion score in the following manner:

$$p(i, j) = p(j|e_i') = \frac{\exp(\text{sim}(e_i', w_j)/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(e_i', w_k)/\tau)}. \tag{3}$$

The learning objective during fine-tuning on the downstream dataset, aimed at maximizing the ground-truth class score, employs the cross-entropy loss function. Notably, at this stage, the image encoder has been discarded, and gradients are backpropagated through the text encoder to update the learnable vectors, while the text encoder is frozen. This procedure is systematically outlined in Algorithm 1.

5 Experiments

In this section, we begin by comparing the adversarial robustness of our proposed approach with hand-crafted prompts under both white-box and black-box adversarial attacks. Second, we compare our method with the state-of-the-art input denoising defensive approaches. Additionally, we investigate the trade-off between generalizability and adversarial robustness in the context of prompt tuning. We also discuss the efficiency between our method and AT. Next, we examine the performance of learnable vectors when trained on a specific dataset but evaluated across various distinct datasets. Finally, we carry out an experimental analysis into interpreting the learnable vectors and perform an exhaustive analysis of hyperparameters.

5.1 Experimental Settings

Datasets. We conduct our study mainly on 8 high-resolution vision datasets: Pets [24], Flowers [22], ImageNet [28], Food101 [1], SUN397 [38], DTD [2], EuroSAT [9], and UCF101 [30]. We adhered to the division of training and testing sets as established in the setup of [47]. For the ImageNet test set, in a

manner consistent with prior studies focusing on adversarial attacks [5,34,40], we use 1,000 images which are randomly sampled (one image per class). Furthermore, to assess the domain generalization capabilities, we employed four variant datasets of ImageNet, namely ImageNetV2 [27], ImageNet-Sketch [33], ImageNet-A [11], and ImageNet-R [10].

Models. Our experiments are centered on the CLIP model. We selected the publicly available version ViT-B/16, and ViT-L/14 [6], which has the largest parameter. Consistent with the vanilla CLIP, we employed hand-crafted prompts as textual input, such as "`a photo of a <class>, a type of pet`" for Pets.

Adversarial Attacks. To evaluate adversarial robustness, we introduced both white-box and black-box adversarial attacks. For white-box adversarial attacks, we employed PGD-40 [18], aimed at maximizing KL Divergence in the embedding space, and AutoAttack [3], aimed at maximizing the contrastive loss between image-text pairs, respectively. Regarding black-box attacks, we implemented black-box attack RAP [25].

Adversarial Defenses. To facilitate comparison with input denoising defenses, we incorporated two distinct categories of defense methods. One is the most effective but relatively time-consuming purification approach based on diffusion model, namely DiffPure [21]. The other is a more immediate but slightly less effective method, including Super resolution [20] and Rescale [39].

Implementation Details. Our methodology builds upon the CoOp framework[4]. Our training process consists of 5 epochs with a batch size of 512 on ImageNet, and 100 epochs with a batch size of 32 on other datasets. The learnable vectors are optimized via SGD, starting with an initial learning rate of 0.002 for ViT-L/14 and 0.005 for ViT-B/16, and adjusted by cosine annealing. The number of learnable vector $M = 32$. To construct the adversarial embedding bank $\mathbf{A}$, we apply the PGD-10 attack with a maximum perturbation of $8/255$ over 10 iterations. For white-box adversarial attack on the test set, we utilize PGD-40 with a maximum perturbation of $16/255$ over 40 iterations. We conduct black-box adversarial attacks on the test set using RAP for 400 iterations. For the selection of the RAP attack surrogate model, we employ ResNet-50 with torchvision weights for ImageNet, and train an additional fully connected layer on downstream datasets. The hyperparameter σ in Super-resolution was set to 0.2. The pre-trained diffusion models in DiffPure is Guided Diffusion [4] and the time step was set to 150.

5.2 Comparison with Vanilla CLIP

We started our evaluation by comparing *AdvPT* with the vanilla CLIP model. Using PGD-40 and RAP, we evaluated adversarial robustness in 8 datasets,

[4] `https://github.com/KaiyangZhou/CoOp`

Table 1: Accuracy (%) under PGD-40 and RAP attacks: The "(↑)" indicates the margin by which *AdvPT* surpasses the vanilla CLIP (hand-crafted prompts).

			Flowers	Pets	Food101	SUN397	DTD	EuroSAT	UCF101	ImageNet
ViT-B/16	vanilla	Clean	71.4	89.1	86.1	62.6	44.4	47.8	66.7	66.1
		PGD	6.4	24.4	14.0	14.7	11.1	22.2	9.1	6.6
		RAP	60.7	79.9	68.3	55.4	33.5	19.2	56.4	28.6
	AdvPT	Clean	87.6	91.3	84.4	70.7	67.9	68.1	77.0	69.1
		PGD	37.4(31.0↑)	41.9(17.5↑)	38.8(24.8↑)	35.7(21.0↑)	39.7(28.6↑)	55.4(33.2↑)	27.2(18.1↑)	19.9(13.3↑)
		RAP	79.0(18.3↑)	81.8(1.9↑)	68.7(0.4↑)	60.0(4.6↑)	50.5(17.0↑)	40.6(21.4↑)	66.0(9.6↑)	30.2(1.6↑)
ViT-L/14	vanilla	Clean	79.3	93.6	91.0	67.6	53.1	58.1	74.2	72.8
		PGD	20.1	50.3	34.3	27.9	20.7	23.3	33.9	28.5
		RAP	70.6	88.2	81.9	62.5	42.5	42.3	67.3	40.2
	AdvPT	Clean	97.6	92.9	90.9	76.4	72.8	79.2	86.5	77.8
		PGD	56.0(35.9↑)	68.7(18.4↑)	54.0(19.7↑)	44.0(16.1↑)	42.0(21.3↑)	62.2(38.9↑)	47.9(14.0↑)	42.9(14.4↑)
		RAP	94.1(23.5↑)	90.4(2.2↑)	82.7(0.8↑)	70.3(7.8↑)	62.4(19.9↑)	50.8(8.5↑)	78.7(11.4↑)	47.6(7.4↑)

Table 2: Accuracy (%) under PGD-40 Attack: The " *+AdvPT* " indicates our method combined with the input denoising defense. The best results are highlighted in **bold**.

		Flowers	Pets	Food101	SUN397	DTD	EuroSAT	UCF101	ImageNet
ViT-B/16	No defense	6.4	24.4	14.0	14.7	11.1	22.2	9.1	6.6
	AdvPT	37.4(31.0↑)	41.9(17.5↑)	38.8(24.8↑)	35.7(21.0↑)	39.7(28.6↑)	55.4(33.2↑)	27.2(18.1↑)	19.9(13.3↑)
	Super	13.8	43.6	58.1	40.5	32.1	43.3	35.4	18.3
	+AdvPT	60.4(46.6↑)	68.3(24.7↑)	69.9(11.8↑)	69.9(29.4↑)	58.2(26.1↑)	**76.7**(33.4↑)	58.4(23.0↑)	34.9(16.6↑)
	DiffPure	59.4	84.1	68.6	55.0	36.9	29.7	60.4	56.6
	+AdvPT	81.9(22.5↑)	86.9(2.8↑)	70.5(1.9↑)	63.9(8.9↑)	60.8(23.9↑)	59.6(29.9↑)	**72.2**(11.8↑)	61.1(4.5↑)
	Rescale	60.1	81.9	79.0	56.9	39.9	40.6	58.6	53.3
	+AdvPT	**87.5**(27.4↑)	**87.4**(5.5↑)	**80.4**(1.4↑)	**67.1**(10.2↑)	**64.4**(24.5↑)	75.4(34.8↑)	72.1(13.5↑)	**61.6**(8.3↑)
ViT-L/14	No defense	20.1	50.3	34.3	27.9	20.7	23.3	33.9	28.5
	AdvPT	56.0(35.9↑)	68.7(18.4↑)	54.0(19.7↑)	44.0(16.1↑)	42.0(21.3↑)	62.2(38.9↑)	47.9(14.0↑)	42.9(14.5↑)
	Super	31.6	67.7	51.0	39.5	34.5	45.3	52.7	40.3
	+AdvPT	74.9(43.3↑)	81.2(13.5↑)	68.2(17.2↑)	55.5(16.0↑)	59.0(24.5↑)	80.1(34.8↑)	70.3(17.6↑)	54.3(14.0↑)
	DiffPure	69.2	90.6	73.8	60.6	46.6	35.6	67.1	64.5
	+AdvPT	92.0(22.8↑)	90.3(0.3↓)	77.2(3.4↑)	70.5(9.9↑)	67.3(20.7↑)	64.5(28.9↑)	79.1(12.0↑)	69.7(5.2↑)
	Rescale	73.2	88.9	83.0	63.0	46.7	46.9	70.2	66.1
	+AdvPT	**94.5**(21.3↑)	**91.0**(2.1↑)	**86.0**(3.0↑)	**73.2**(10.2↑)	**69.8**(23.1↑)	**83.3**(36.4↑)	**82.2**(12.0↑)	**74.8**(8.7↑)

as indicated in Tab. 1. Our findings reveal that: (1) *AdvPT* demonstrates improvements over the vanilla CLIP under both PGD-40 and RAP attacks, with the specific improvement quantified in green. (2) While the primary goal of *AdvPT* is not to enhance generalizability, the empirical finding implies that the enhancement of accuracy emerges as a collateral advantage.

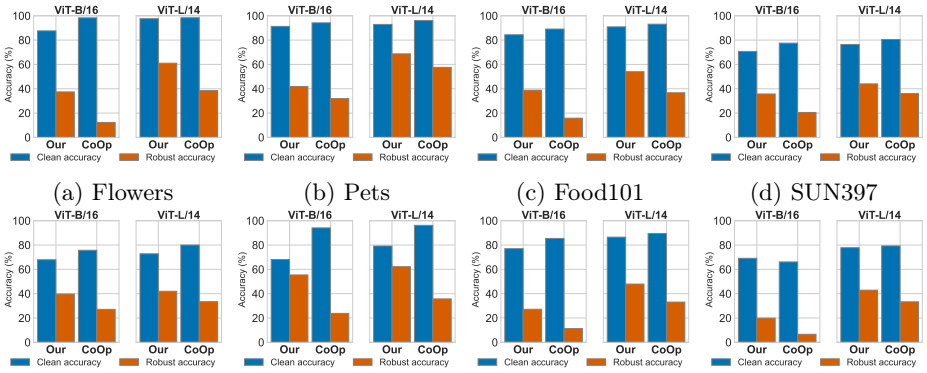
5.3 Comparison with Adversarial Defenses

As described previously, *AdvPT* presents an innovative approach to enhance the robustness of image encoders against adversarial attacks by modifying only the textual input. This method is inherently synergistic with visual-modality input denoising defenses. We evaluated its performance against white-box PGD-40, and observed its compatibility with defenses, as delineated in Tab. 2. Significantly, incorporation of *AdvPT* requires no specialized tuning for the purified images.

Our results show *AdvPT*'s consistent compatibility with benchmark adversarial defenses. Despite a minor 0.3% performance drop in ViT-L14 on Pets, it maintains over 90% accuracy, closely paralleling original example performance,

Table 3: Accuracy (%) under AutoAttack.

		Flowers	Pets	Food101	SUN397	DTD	EuroSAT	UCF101	ImageNet
ViT-B/16	No defense	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	AdvPT	23.3(23.3↑)	7.2(7.2↑)	4.4(4.4↑)	17.6(17.6↑)	28.9(28.9↑)	27.5(27.5↑)	18.5(18.5↑)	11.0(11.0↑)
	DiffPure	54.1	78.9	61.1	51.5	35.1	32.9	56.7	55.5
	+*AdvPT*	**80.3**(26.2↑)	**84.7**(5.8↑)	**65.5**(4.4↑)	**61.4**(9.9↑)	**60.8**(25.7↑)	**59.1**(26.2↑)	**70.8**(14.1↑)	**60.4**(4.9↑)
ViT-L/14	No defense	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	AdvPT	19.2(19.2↑)	3.3(3.3↑)	3.3(3.3↑)	15.5(15.5↑)	25.7(25.7↑)	29.5(29.5↑)	17.0(17.0↑)	9.0(9.0↑)
	DiffPure	63.7	87.5	67.4	51.5	43.4	34.8	64.7	60.7
	+*AdvPT*	**89.3**(25.6↑)	**87.9**(0.4↑)	**72.5**(5.1↑)	**68.5**(17.0↑)	**65.4**(22.0↑)	**63.2**(28.4↑)	**77.8**(13.1↑)	**67.9**(7.2↑)



(a) Flowers (b) Pets (c) Food101 (d) SUN397

(e) DTD (f) EuroSAT (g) UCF101 (h) ImageNet

Fig. 3: *AdvPT* vs. CoOp on generalization and adversarial robustness.

which is acceptable. All improvements are highlighted in green, corroborating the efficacy of the strategy that combines *AdvPT* with input denoising mechanisms.

Remarkably, the synergy of *AdvPT* with baseline defense mechanisms sometimes yielded "$1 + 1 > 2$" contribution. For example, on the ViT-B/16 model applied to the Flowers dataset, *AdvPT* alone increases accuracy by 31.0% (from 6.40% to 37.40%), yet when combined with Super-resolution, it further improves the performance of Super-resolution by 46.60% (from 13.80% to 60.4%). In addition to this, we also introduced AutoAttack, which targets the contrastive loss of image-text pairs. The results in Tab. 3 are consistent with those in Tab. 2, indicating that when combined with the state-of-the-art diffusion model-based defense method, DiffPure, our *AdvPT* achieved enhanced performance. These findings highlight the potential of this innovative synergy strategy to enhance adversarial defense by simultaneously modifying textual and visual inputs, warranting further investigation in future study.

5.4 Generalization-Robustness Trade-off

In this subsection, we discuss the effects of various learning objectives on the learnable vectors within prompt tuning. The primary goal of *AdvPT* is to enhance the adversarial robustness of the image modality in VLMs. In contrast, we explore

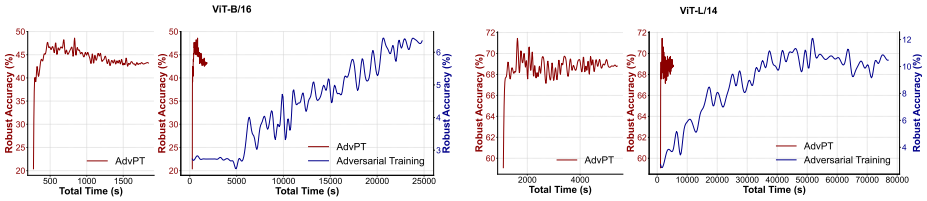


Fig. 4: Efficiency comparison between *AdvPT* and AT on Pets.

whether an objective like CoOp [47], which is fine-tuned on clean images for improved accuracy, affects adversarial robustness differently.

Our comparative analysis of *AdvPT* and CoOp unveils insights into their generalizability and adversarial robustness, as illustrated in Fig. 3. Our findings are twofold: (1) Intriguingly, *AdvPT* significantly outperforms CoOp in adversarial robustness, albeit at a slight cost to generalization. This highlights a potential trade-off between adversarial robustness and generalization in prompt tuning, aligning with conclusions drawn from traditional AT [42]. (2) Although *AdvPT* sacrifices some generalizability, this drawback is mitigated as the model scale increases. Particularly on ViT-L/14, while also enhancing adversarial robustness, the narrowed generalizability gap makes *AdvPT* highly compatible with the ongoing trend towards scaling up models.

5.5 Comparison with Adversarial Training

In this subsection, we compare the efficiency of our method, which focuses on fine-tuning only the prompt, against traditional AT. Specifically, we juxtapose *AdvPT* with PGD-10 AT [18] on the Pets dataset, ensuring that both methods use an equivalent batch size. The comparative results are shown in Fig. 4, including the time taken to compute the adversarial embedding bank $\mathbf{A}$ in the total time reported. We also presented the results of Fast AT [37] in Fig. 5. Although Fast AT is much faster than AT, it still lags significantly behind *AdvPT*.

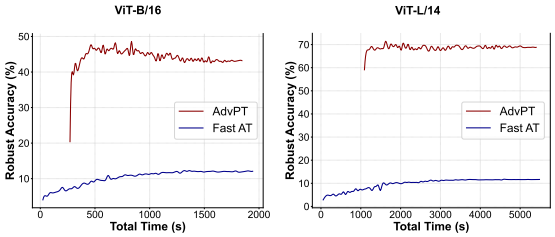


Fig. 5: Efficiency comparison between *AdvPT* and Fast AT on Pets.

Our analysis reveals that *AdvPT* is more time-efficient than AT, requiring at least an order of magnitude less time. Moreover, it demonstrates a superior enhancement in model performance, outperforming AT by at least one order of magnitude in effectiveness. This efficiency advantage, exceeding 100× at least, positions *AdvPT* as a notably superior solution for VLMs.

Table 4: Clean accuracy and robust accuracy (PGD-40) of linear prob CLIP.

		Flowers	Pets	Food101	SUN397	DTD	EuroSAT	UCF101
ViT-B/16	Clean	97.9	91.1	88.4	75.7	77.1	94.3	83.8
	PGD	4.8	10.9	5.2	4.8	13.8	9.2	4.1
ViT-L/14	Clean	99.4	94.2	90.9	79.0	80.1	95.9	88.7
	PGD	6.8	15.7	12.8	7.7	14.5	21.2	8.0

			-V2	-A	-R	-Sketch
ViT-B/16	vanilla CLIP	Clean	60.8	47.7	80.5	46.9
		Robust	6.2	4.7	9.3	5.9
	AdvPT	Clean	62.6	46.3	83.6	45.6
		Robust	16.3	10.1	22.0	9.4
ViT-L/14	vanilla CLIP	Clean	67.9	68.7	91.8	57.2
		Robust	25.6	16.8	34.3	20.7
	AdvPT	Clean	71.1	69.0	92.1	58.5
		Robust	38.5	20.2	43.5	25.8

Fig. 6: *AdvPT* vs. vanilla CLIP on distribution shift.

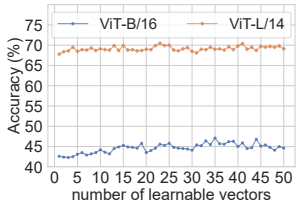


Fig. 7: Effect of number of learnable vector on Pets.

5.6 Comparison with Linear Prob CLIP

In this section, we compare *AdvPT* with linear prob CLIP, which also utilizes additional data, to investigate whether the robustness improvements of *AdvPT* merely result from additional downstream data, as shown in Tab. 4. By comparing with Tab. 1, while it shows an increase in clean accuracy compared to vanilla CLIP, its robustness is reduced, when compared to *AdvPT*. This indicates that merely introducing additional downstream data does not directly contribute to enhanced robustness. Furthermore, it also indicates that the enhancements in robustness are not entirely relevant to improvements in accuracy.

5.7 Evaluation on Domain Shift

A notable advantage of CLIP lies in its adaptability to domain shift. Thus, in this subsection, we evaluate the transferability of *AdvPT* in comparison to the vanilla CLIP in domain shift scenarios. The source dataset utilized is ImageNet, while the target datasets include ImageNetV2, ImageNet-Sketch, ImageNet-A, and ImageNet-R. The results presented in Fig. 6 elucidate that the proposed *AdvPT* outperforms the vanilla CLIP in terms of adversarial robustness, thereby validating its stability across varied domains.

5.8 Further Analysis

Number of Learnable Vector In Sec. 5.4, we observed similarities between adversarial prompt tuning and AT. It is widely acknowledged within the AT framework that a larger count of tunable parameters correlates with enhanced adversarial robustness. To discern whether this correlation persists within *AdvPT*,

Table 5: The nearest words for learnable vectors. N/A means non-Latin characters.

Flowers	Pets	Food101	SUN397	DTD	EuroSAT	UCF101	ImageNet
activated(0.6720)	stores(0.6300)	sii(1.6187)	gaunt(1.4723)	3(0.6263)	ust(0.8010)	laces(1.0643)	N/A(0.6407)
walked(0.7015)	sun(0.6388)	activation(1.6778)	maestro(1.5045)	alization(0.6467)	trip(0.9385)	fa(1.1818)	le(0.6747)
pper(0.7994)	amore(0.6530)	thereal(1.6817)	zoom(1.5162)	cs(0.7361)	vu(1.0143)	deployed(1.2376	telly(0.6995)
bao(0.8742)	favorites(0.6877)	cst(1.6910)	nag(1.5209)	prelude(0.7904)	salam(1.0190)	N/A(1.2625)	hooper(0.7082)
burden(0.8924)	ama(0.6957)	pancreatic(1.8803)	cope(1.5922)	therapists(0.8336)	weymouth(1.1291)	cumbri(1.2966)	naq(0.7121)

we conducted an empirical evaluation of its efficacy under different numbers of learnable vector $M \in [1, 50]$, using the Pets dataset as an example. The empirical results, as illustrated in Fig. 7, suggest that the volume of tunable parameters does not constitute a constraint in *AdvPT*. Instead, unlocking its potential efficacy warrants further investigation.

Interpreting the Learnable Vector In this subsection, we aim to decode what the learnable vectors have captured. However, direct mapping of these learnable vectors to words is infeasible due to the optimization occurring within a continuous space, while word space is discrete. Therefore, we adopt a technique applied in the CoOp experiment, searching vocabulary for the nearest words to the learned vectors by Euclidean distance, as illustrated in Tab. 5. These words are not intuitively understandable, exactly aligning with the non-robust features in adversarial images.

6 Limitations

First, the paper's focus is restricted to image recognition tasks. Exploring the applicability of *AdvPT* to a broader array of tasks, such as Visual Question Answering (VQA) in advanced models like GPT-4V [23], is a worthwhile direction for future research. Second, visual prompts [13, 14] emerge as a promising research avenue, given their extensive trainable parameters, which could enhance adversarial robustness. Yet, it introduces additional branches to the model, thus falling into the model robustification category.

7 Conclusion and Discussion

This study introduces Adversarial Prompt Tuning (*AdvPT*), a novel technique enhancing the adversarial robustness of VLMs such as CLIP. Our approach, focusing on the alignment of learnable text prompts with adversarial image embeddings, represents a significant step forward in securing VLMs against adversarial attacks. Notably, *AdvPT* achieves this heightened security without necessitating extensive model retraining or architectural modifications.

However, we acknowledge that this is an initial foray into a complex domain. Future research should explore the scalability of adversarial prompt tuning across various settings. In conclusion, *AdvPT* presents a promising direction for enhancing VLM's robustness, contributing to the broader endeavor of making AI systems more secure and reliable.

Acknowledgements

This work was supported by National Key R&D Program of China (Grant No. 2022ZD0160103, 2023YFC3310700), National Natural Science Foundation of China (Grant No. 62172094, 62276067) and Science and Technology Commission of Shanghai Municipality (Grant No. 22511106102).

References

1. Bossard, L., Guillaumin, M., Gool, L.V.: Food-101–mining discriminative components with random forests. In: European conference on computer vision. pp. 446–461. Springer (2014)
2. 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)
3. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International conference on machine learning. pp. 2206–2216. PMLR (2020)
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. vol. 34, pp. 8780–8794 (2021)
5. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193 (2018)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Guo, C., Rana, M., Cisse, M., Van Der Maaten, L.: Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117 (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. 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)
10. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8349 (2021)
11. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15262–15271 (2021)
12. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
13. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022)

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, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021)
16. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1778–1787 (2018)
17. Liu, X., Ji, K., Fu, Y., Tam, W.L., Du, Z., Yang, Z., Tang, J.: P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:2110.07602 (2021)
18. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks (2018)
19. Mao, et al.: Understanding zero-shot adversarial robustness for large-scale models. In: ICLR (2023)
20. Mustafa, A., Khan, S.H., Hayat, M., Shen, J., Shao, L.: Image super-resolution as a defense against adversarial attacks. IEEE Transactions on Image Processing **29**, 1711–1724 (2020)
21. Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., Anandkumar, A.: Diffusion models for adversarial purification. In: International Conference on Machine Learning. pp. 16805–16827. PMLR (2022)
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. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
24. 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)
25. Qin, Z., Fan, Y., Liu, Y., Shen, L., Zhang, Y., Wang, J., Wu, B.: Boosting the transferability of adversarial attacks with reverse adversarial perturbation. Advances in Neural Information Processing Systems **35**, 29845–29858 (2022)
26. 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)
27. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400. PMLR (2019)
28. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
29. Salman, H., Sun, M., Yang, G., Kapoor, A., Kolter, J.Z.: Denoised smoothing: A provable defense for pretrained classifiers. Advances in Neural Information Processing Systems **33**, 21945–21957 (2020)
30. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
31. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)

32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
33. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in Neural Information Processing Systems **32** (2019)
34. Wang, X., He, K.: 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)
35. Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., Gu, Q.: Improving adversarial robustness requires revisiting misclassified examples. In: International conference on learning representations (2019)
36. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 139–149 (2022)
37. Wong, E., Rice, L., Kolter, J.Z.: Fast is better than free: Revisiting adversarial training. In: International Conference on Learning Representations (2020)
38. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 3485–3492. IEEE (2010)
39. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects through randomization. In: International Conference on Learning Representations (2018)
40. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2730–2739 (2019)
41. Yoon, J., Hwang, S.J., Lee, J.: Adversarial purification with score-based generative models. In: International Conference on Machine Learning. pp. 12062–12072. PMLR (2021)
42. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International conference on machine learning. pp. 7472–7482. PMLR (2019)
43. Zhang, J., Yi, Q., Sang, J.: Towards adversarial attack on vision-language pre-training models. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5005–5013 (2022)
44. Zhang, Y., Yao, Y., Jia, J., Yi, J., Hong, M., Chang, S., Liu, S.: How to robustify black-box ml models? a zeroth-order optimization perspective. In: International Conference on Learning Representations (2022)
45. Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.M., Lin, M.: On evaluating adversarial robustness of large vision-language models. arXiv preprint arXiv:2305.16934 (2023)
46. 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)
47. 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)
48. Zhou, Z., Hu, S., Li, M., Zhang, H., Zhang, Y., Jin, H.: Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 6311–6320 (2023)