TAPT: Test-Time Adversarial Prompt Tuning for Robust Inference in Vision-Language Models

Xin Wang[1], Kai Chen[1], Jiaming Zhang[2], Jingjing Chen[1], Xingjun Ma[1]*

[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2]Hong Kong University of Science and Technology

Abstract

Large pre-trained Vision-Language Models (VLMs) such as CLIP have demonstrated excellent zero-shot generalizability across various downstream tasks. However, recent studies have shown that the inference performance of CLIP can be greatly degraded by small adversarial perturbations, especially its visual modality, posing significant safety threats. To mitigate this vulnerability, in this paper, we propose a novel defense method called ***Test-Time Adversarial Prompt Tuning (TAPT)*** *to enhance the inference robustness of CLIP against visual adversarial attacks. TAPT is a test-time defense method that learns defensive bimodal (textual and visual) prompts to robustify the inference process of CLIP. Specifically, it is an unsupervised method that optimizes the defensive prompts for each test sample by minimizing a multi-view entropy and aligning adversarial-clean distributions. We evaluate the effectiveness of TAPT on 11 benchmark datasets, including ImageNet and 10 other zero-shot datasets, demonstrating that it enhances the zero-shot adversarial robustness of the original CLIP by at least 48.9% against AutoAttack (AA), while largely maintaining performance on clean examples. Moreover, TAPT outperforms existing adversarial prompt tuning methods across various backbones, achieving an average robustness improvement of at least 36.6%. Code is available at https://github.com/xinwong/TAPT.*

1. Introduction

Vision-Language Models (VLMs) pre-trained on large-scale datasets of image-text pairs have emerged as powerful backbones for numerous applications, including computer vision [17, 34, 57], medical image analysis [15, 51], and robotics [3, 18, 39]. Despite these advancements, studies indicate that even minor adversarial perturbations in input images can significantly degrade the inference performance of VLMs [9, 26, 44, 59, 63], thereby posing critical safety risks across a wide range of downstream applications.

Figure 1. Inference with different prompts. **Top**: inference with hand-crafted prompts fails to recognize the class 'cat'; **Middle**: Inference with fixed prompts tuned by APT methods cannot recognize all adversarial images; **Bottom**: Inference with test-time prompts optimized for each image produces accurate recognitions.

Adversarial training [12, 26, 52, 58] is a general defense strategy that augments training data with adversarial examples crafted to mislead the model. While this method has proven effective [8], it entails costly min-max training, which restricts its practicality—especially for large VLMs, where standard training alone can cost millions of dollars [38]. To address the efficiency limitations of adversarial training, recent research has introduced adversarial prompt tuning (APT) methods [23, 60, 67], which align learnable text prompts with adversarial image embeddings. These differentiable prompts [19, 64, 65] provide a more cost-effective alternative to adversarial training.

Despite its promising results, APT faces three fundamental limitations: (1) The distribution-dependent nature of learnable prompts restricts their generalization to out-of-distribution scenarios and novel tasks. (2) The requirement for task-specific annotated data presents significant challenges for zero-shot applications. (3) While APT methods improve robustness on specific tasks [42], they often compromise the model's overall generalization performance [33, 43]. Given that real-world applications involve an unbounded set of potential tasks, evaluating adversarial robustness solely on predefined downstream tasks proves insufficient. Therefore, achieving zero-shot adversarial robustness—maintaining performance against adversarial attacks on unseen tasks without task-specific training—remains an open challenge.

*Corresponding author: <xingjunma@fudan.edu.cn>

The key to addressing the zero-shot adversarial robustness challenge lies in adaptively identifying a robust prompt that consistently aligns an adversarial image embedding with its correct text embedding for each test sample. To this end, we propose a simple yet effective framework called **Test-Time Adversarial Prompt Tuning (TAPT)**, which dynamically tunes a robust prompt on the fly based on only the provided test sample. Specifically, TAPT learns defensive prompts by minimizing two unsupervised losses: (1) multi-view entropy, which ensures consistent predictions across various augmented views of each test sample, and (2) adversarial-clean embedding alignment, which aligns the means and variances of test sample embeddings with pre-computed adversarial-clean embeddings of a public dataset (e.g., ImageNet) to enhance inference robustness. Notably, TAPT operates during the inference phase without requiring any task-specific training set or annotations. The different inference schemes of training-time defense APT methods and our test-time defense TAPT are illustrated in Figure 1.

We evaluated TAPT against both white-box and black-box adversarial attacks across 11 benchmark datasets. TAPT demonstrated superior performance compared to vanilla CLIP (using hand-crafted prompts) and existing APT methods across three different prompt designs: visual-only prompts, visual-language (V-L) joint prompts, and V-L independent prompts. By dynamically adapting prompts at inference time, TAPT opens a new direction for safeguarding the inference process of pre-trained VLMs. Offering a flexible and efficient solution for zero-shot adversarial robustness while maintaining performance on clean samples, our TAPT method bridges the gap between the need for adversarial robustness and the performance challenges posed by open, real-world environments.

In summary, our main contributions are:

- We propose a novel test-time defense method named **Test-Time Adversarial Prompt Tuning (TAPT)** to robustify the zero-shot inference process of pre-trained VLMs. TAPT adapts the prompt for each test image to achieve robust inference without the need for task-specific tuning for downstream datasets. To the best of our knowledge, TAPT is the first inference-time adversarial defense method for pre-trained VLMs.
- TAPT introduces an adversarial-clean alignment loss that aligns the distribution of a test sample with pre-computed adversarial-clean distributions from a public dataset (ImageNet), thereby improving adversarial robustness while maintaining accuracy on clean samples. It also leverages the advantage of APT by using pre-tuned prompts on ImageNet to further enhance zero-shot adversarial robustness.
- We conduct extensive experiments on 11 datasets, including ImageNet and 10 other zero-shot datasets. Our results demonstrate that TAPT significantly outperforms existing APT baselines against both white-box and black-box attacks. Specifically, TAPT improves zero-shot adversarial robustness against AutoAttack by 36.6% with ViT-B/16, and by 38.0% with ViT-B/32, respectively.

2. Related Work

Here, we briefly review related works on adversarial attacks and defenses for pre-trained VLMs, and test-time adaptation techniques proposed to improve generalization.

Adversarial Attacks on Pre-trained VLMs Adversarial attacks on pre-trained VLMs are broadly categorized as white-box or black-box attacks according to the threat model. In white-box attacks, the attacker has full access to the model parameters and can directly compute adversarial gradients [25, 26, 59, 68]. Black-box attacks, on the other hand, restrict the attacker to querying model outputs [10, 13, 24, 47, 54, 55, 62, 63]. Traditional single-modal attacks designed for vision models, such as PGD [26], DI [54], and AutoAttack [8], can be used directly to attack the image encoders of pre-trained VLMs. Several recent multi-modal attacks have targeted pre-trained VLMs by simultaneously exploiting vulnerabilities in both their image and text encoders. For example, Co-Attack [59] pioneered white-box multi-modal attacks, perturbing both image and text modalities concurrently. SGA [24] extended Co-Attack to the black-box setting, improving the transferability of multi-modal adversarial examples. VLATTACK [55] further generates adversarial examples by fusing perturbations of images and texts from both single-modal and multi-modal.

Adversarial Defenses for Pre-trained VLMs Adversarial training/tuning is a widely used defense strategy for pre-trained VLMs, with existing methods generally classified into adversarial contrastive tuning [28, 36, 49, 50, 52, 66] and adversarial prompt tuning [16, 23, 60, 67]. Adversarial contrastive tuning focuses on enhancing the adversarial robustness of the backbone model. For example, TeCoA [28] examines the impact of fine-tuning and visual prompt tuning on the zero-shot adversarial robustness of VLMs. FARE [36] improves CLIP's image encoder through unsupervised adversarial fine-tuning, enhancing adversarial robustness in models like LLaVA and OpenFlamingo without retraining. PMG-AFT [49] introduces an auxiliary branch to boost zero-shot adversarial robustness, while MMCoA [66] investigates VLM vulnerabilities to multi-modal attacks. In contrast, AdvPT [60] and APT [23] offer an efficient approach to bolster the adversarial robustness of VLMs by tuning only the textual prompts without altering the model parameters. FAP [67] further refines APT by balancing cross-modal consistency between benign and adversarial inputs. While these methods are all training-time defense methods that need to pre-tune the prompt for a specific downstream task, in this work we propose a novel test-time defense method that optimizes the prompt on the

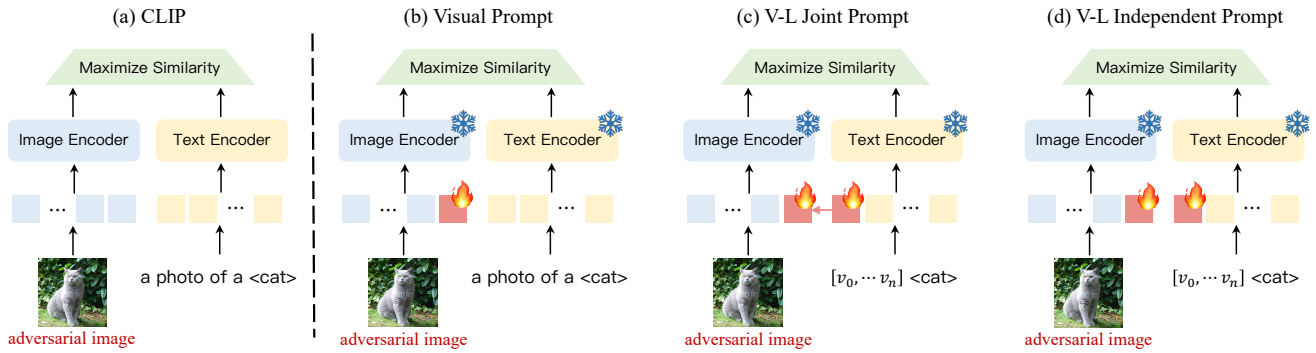

Figure 2. An illustration of CLIP and different adversarial prompt tuning schemes. (a) The original CLIP [34]; (b) - (d) Adversarial prompt tuning with three distinctive prompt designs: Visual Prompt (b), V-L Joint Prompt (c), and V-L Independent Prompt (d).

fly during inference and is task-agnostic.

Test-Time Adaptation Test-time adaptation (TTA) methods enhance the generalization performance of pre-trained models by adapting to individual test samples or batches, addressing distribution shifts between training (source) and testing (target) data. Early TTA methods [29, 37] address domain shift by updating batch normalization statistics based on test batch statistics. Building on this, TENT [46] improves adaptation by updating batch normalization layers to minimize the entropy of predicted probabilities for each test batch, while MEMO [61] extends this by minimizing entropy over multiple augmented input samples. Inspired by TENT, subsequent TTA methods such as CoTTA [48] and EATA [31] enable pre-trained models to adapt to continuously shifting test distributions. More recently, methods like TPT [40], PromptAlign [1], and MTA [56] have focused on tuning prompts exclusively at test time to ensure consistent predictions across various augmented views of a test sample. Kim *et al*. [20] finds that combining TTA with AGL [4] further improves out-of-distribution (OOD) accuracy prediction. In contrast to existing TTA methods which primarily emphasize performance on clean samples, in this work we introduce TAPT, a test-time defense, that specifically enhances zero-shot adversarial robustness against potential attacks.

3. Proposed Method

3.1. Preliminaries

Threat Model We assume a white-box threat model in which the adversary has full knowledge of the target model's architecture and parameters, allowing direct perturbation of test images based on adversarial gradients prior to inference. The defender, or model owner, can deploy any defense strategies to protect against potential adversarial attacks. Specifically, we focus on securing CLIP zero-shot inference, where the defender lacks access to task-specific training data or annotations of the downstream application.

CLIP We denote the CLIP image encoder as $\mathcal{I}$, parameterized by $\theta_{\mathcal{I}}$, and the text encoder as $\mathcal{T}$, parameterized by $\theta_{\mathcal{T}}$. Considering a K-class classification problem, where each image $\mathbf{x}$ is associated with a class label in the format `"a photo of a <class>"`. For a clean sample $\mathbf{x} \in [0,1]^d$ and a target CLIP model comprising $\{\mathcal{I}, \mathcal{T}\}$, a white-box adversarial attack seeks to generate an adversarial example $\mathbf{x}'$ that maximizes the model loss as follows:

$$\mathbf{x}' = \underset{\|\mathbf{x}'-\mathbf{x}\|_\infty \leq \epsilon}{\arg\max}\ \mathcal{L}(\mathcal{I}(\mathbf{x}'), \mathcal{T}(y)), \tag{1}$$

where $\mathcal{L}(\cdot)$ represents the loss function, $\mathbf{x}'$ is the adversarial example, and ϵ denotes the perturbation budget.

Adversarial Prompt Tuning (APT) APT applies adversarial training during the prompt tuning process to enhance adversarial robustness, with recent APT methods primarily developed to defend the visual component of CLIP. Rather than relying on hand-crafted prompts like `"a photo of a <class>"`, APT learns robust prompts from training data to improve adversarial robustness on downstream tasks. As shown in Figure 2, APT can be extended to three distinctive prompt designs: (1) Visual-only (prompts applied solely to the vision branch), (2) Vision-Language (V-L) joint (shared prompts across both branches), and (3) V-L independent (separate prompts for each branch). Specifically, APT optimizes the prompts $\boldsymbol{P} = \{\boldsymbol{P_v}, \boldsymbol{P_t}\} \in \mathbb{R}^{L \times D}$, where $\boldsymbol{P_v}$ and $\boldsymbol{P_t}$ are the visual and textual prompts, respectively, L denotes the number of prompt tokens, and D is the embedding dimension.

Given a downstream training dataset $\mathcal{D}_{\text{train}} = \{(\mathbf{x}, y)\}$, the learnable visual prompt $\boldsymbol{P_v}$ is appended to the visual input tokens, forming the sequence $\{\mathbf{x}, \boldsymbol{P_v}\} = \{\text{CLS}, e_1, e_2, \cdots, e_M, \boldsymbol{P_v}\}$. Similarly, the textual prompt $\boldsymbol{P_t}$ is appended to the text input, forming $\{y, \boldsymbol{P_t}\}$. APT enhances adversarial robustness by aligning the embeddings of clean text with those of adversarial images through a min-max optimization. The corresponding optimization prob-

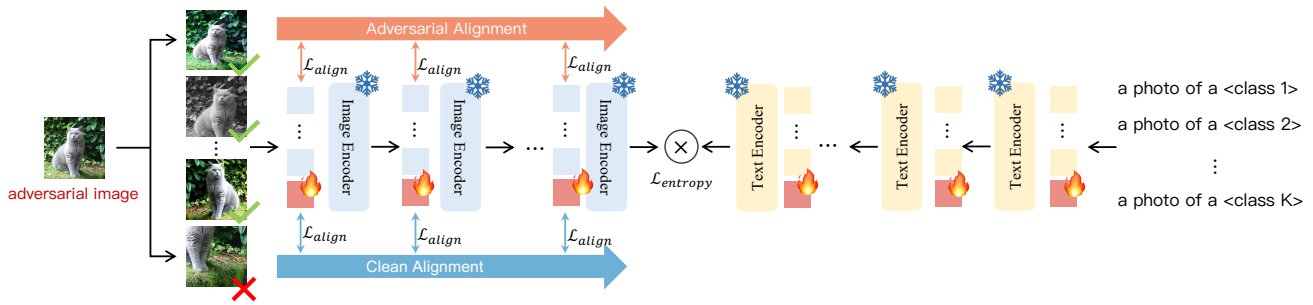

Figure 3. An overview of our proposed TAPT method: Given an adversarial image, TAPT generates multiple augmented views of the image and retains only those views with low entropy in their averaged prediction probabilities. During inference, TAPT then optimizes the prompt by minimizing multi-view entropy across these selected views while aligning their embedding distribution with pre-computed adversarial-clean statistics from a public dataset (ImageNet).

Algorithm 1 Test-Time Adversarial Prompt Tuning

Input: input image $\mathbf{x}$, image encoder $\mathcal{I}$, text encoder $\mathcal{T}$, augmentation function $\mathcal{A}$, entropy threshold τ, pre-computed $\mathcal{D}_{\text{public}}$ statistics $\{\mu_{\text{adv}}, \sigma_{\text{adv}}, \mu_{\text{clean}}, \sigma_{\text{clean}}\}$
Output: learnable prompts $\boldsymbol{P}$
1. Initialize adversarial prompts:
  $\boldsymbol{P} \leftarrow \text{APT}(\mathcal{D}_{\text{public}}, \epsilon; \mathcal{I}, \mathcal{T})$
2. Multi-view entropy-based sample selection:
  Select the top τ entropy from $\mathcal{A}(\mathbf{x})$ to form $\mathcal{H}_\tau(\mathbf{x})$
3. Compute multi-view entropy loss:
  $\mathcal{L}_{\text{entropy}} = -\sum_{i=1}^{K} \tilde{p}(y_i|\mathcal{H}_\tau(\mathbf{x}), \boldsymbol{P}) \log \tilde{p}(y_i|\mathcal{H}_\tau(\mathbf{x}), \boldsymbol{P})$
4. Compute current embedding statistics:
  for each layer l in $\mathcal{I}$ **do**
    $\mu_l = \text{mean}(I_l(\hat{\mathbf{x}}, \boldsymbol{P}))$ for $\hat{\mathbf{x}}$ in $\mathcal{H}_\tau(\mathbf{x})$
    $\sigma_l^2 = \text{std}(\mathcal{I}_l(\hat{\mathbf{x}}, \boldsymbol{P}))$ for $\hat{\mathbf{x}}$ in $\mathcal{H}_\tau(\mathbf{x})$
  end for
5. Adversarial-clean embedding alignment:
  $\mathcal{L}_{\text{adv}} = \frac{1}{L}\sum_{l=1}^{L}\left(\|\mu_l - \mu_{\text{adv},l}\|_1 + \left\|\sigma_l^2 - \sigma_{\text{adv},l}^2\right\|_1\right)$
  $\mathcal{L}_{\text{clean}} = \frac{1}{L}\sum_{l=1}^{L}\left(\|\mu_l - \mu_{\text{clean},l}\|_1 + \left\|\sigma_l^2 - \sigma_{\text{clean},l}^2\right\|_1\right)$
6. Optimize prompts:
  $\mathcal{L}_{\text{TAPT}} = \mathcal{L}_{\text{entropy}} + \alpha\mathcal{L}_{\text{adv}} + (1-\alpha)\mathcal{L}_{\text{clean}}$
  Optimize $\boldsymbol{P} \leftarrow$ Minimize $\mathcal{L}_{\text{TAPT}}$

lem can be formulated as:

$$\arg\min_{\boldsymbol{P}} \mathbb{E}_{\mathcal{D}_{\text{train}}} \max_{\|\mathbf{x}'-\mathbf{x}\|_\infty \leq \epsilon} \mathcal{L}(\mathcal{I}(\mathbf{x}', \boldsymbol{P_v}), \mathcal{T}(y, \boldsymbol{P_t})), \quad (2)$$

where $\mathcal{I}(\mathbf{x}', \boldsymbol{P_v})$ and $\mathcal{T}(y, \boldsymbol{P_t})$ denote the adversarial image embedding and text embedding, respectively. APT methods tune the prompts on the training dataset of each downstream task. At test time, the tuned prompts are fixed to perform inference for different test images.

3.2. Test-Time Adversarial Prompt Tuning (TAPT)

Framework Overview As illustrated in Figure 3, TAPT comprises two main modules: (1) multi-view entropy-based sample selection, and (2) adversarial-clean embedding alignment. The defense procedure of TAPT operates as follows. Given a test image $\mathbf{x} \in \mathcal{D}_{\text{test}}$, TAPT begins by generating M randomly augmented views $\{\mathcal{A}_1(\mathbf{x}), \mathcal{A}_2(\mathbf{x}), \ldots, \mathcal{A}_M(\mathbf{x})\}$ through random augmentations $\mathcal{A}$. The multi-view entropy-based sample selection module then chooses the top-K views with the lowest entropy in their averaged prediction probabilities. Using these selected views, TAPT optimizes the prompt $\boldsymbol{P}$ during inference by minimizing multi-view entropy and enforcing adversarial-clean alignment. The prompt is reset to its initial state before processing each new test sample or batch. The complete procedure of TAPT is outlined in Algorithm 1.

Multi-View Entropy-Based Sample Selection Following prior works [1, 40], we first discard ineffective augmented views $\mathcal{A}(\mathbf{x})$ (e.g., when a random crop removes essential image content) by applying a selection filter with a threshold τ. This filter retains only augmented views with low prediction entropy (high-confidence predictions). Specifically, we define the set of selected augmented views as $\mathcal{H}_\tau(\mathbf{x}) = \{\mathcal{A}_j(\mathbf{x}) | 1 \leq j \leq M, H(\mathcal{A}_j(\mathbf{x})) \leq H_\tau\}$, where H_τ is the entropy threshold corresponding to the top τ lowest entropy values among all M augmentations. TAPT then optimizes prompts by minimizing the multi-view entropy of the averaged prediction probability over these selected views:

$$\mathcal{L}_{\text{entropy}} = -\sum_{i=1}^{K} \tilde{p}(y_i|\mathcal{H}_\tau(\mathbf{x}), \boldsymbol{P}) \log \tilde{p}(y_i|\mathcal{H}_\tau(\mathbf{x}), \boldsymbol{P}), \quad (3)$$

where $\tilde{p}(y_i|\mathcal{H}_\tau(\mathbf{x}), \mathbf{P}) = \frac{1}{|\mathcal{H}_\tau(\mathbf{x})|}\sum_{\hat{\mathbf{x}} \in \mathcal{H}_\tau(\mathbf{x})} p(y_i|\hat{\mathbf{x}}, \mathbf{P})$ denotes the average predicted probability for class y_i over the selected augmented views $\mathcal{H}_\tau(\mathbf{x})$ when using prompt $\boldsymbol{P}$. This optimization encourages the model to achieve consis-

tent predictions by adjusting the prompt specifically for the given instance.

Adversarial-Clean Embedding Alignment An adversarial test image $\mathbf{x}$ can shift the image embeddings produced by the image encoder $\mathcal{I}(\mathbf{x}, \boldsymbol{P})$ compared to those from clean images, potentially misleading the model. To mitigate this vulnerability, we align the mean and variance of the test image embedding with pre-computed statistics from a public dataset $\mathcal{D}_{\text{public}}$. Ideally, this public dataset would originate from the original CLIP pre-training data. However, since this data is unavailable, we use ImageNet as a proxy, given CLIP's strong zero-shot performance on ImageNet with standard prompt tuning [5]. Specifically, we compute the mean and variance of the current embeddings for alignment as follows:

$$\mu_l(\mathcal{H}_\tau(\mathbf{x}); \boldsymbol{P}) = \frac{1}{|\mathcal{H}_\tau(\mathbf{x})|} \sum_{\hat{\mathbf{x}} \in \mathcal{H}_\tau(\mathbf{x})} \mathcal{I}_l(\hat{\mathbf{x}}, \boldsymbol{P}), \tag{4}$$

$$\sigma_l^2(\mathcal{H}_\tau(\mathbf{x}); \boldsymbol{P}) = \frac{\sum_{\hat{\mathbf{x}} \in \mathcal{H}_\tau(\mathbf{x})} (\mathcal{I}_l(\hat{\mathbf{x}}, \boldsymbol{P}) - \mu_l(\mathcal{H}_\tau(\mathbf{x}); \boldsymbol{P}))^2}{|\mathcal{H}_\tau(\mathbf{x})| - 1} \tag{5}$$

where $\mathcal{I}_l(\hat{\mathbf{x}}, \boldsymbol{P})$ represents the embedding vector at layer l for the augmented input $\hat{\mathbf{x}} \in \mathcal{H}_\tau(\mathbf{x})$ given prompt $\boldsymbol{P}$. Here, $\mu_l(\mathcal{H}_\tau(\mathbf{x}); \boldsymbol{P})$ and $\sigma_l^2(\mathcal{H}_\tau(\mathbf{x}); \boldsymbol{P})$ denote the mean and variance of the test sample embeddings at layer l, respectively. We similarly pre-compute the mean and variance of embeddings from the public dataset using the robust prompt $\boldsymbol{P_{adv}}$ (obtained via APT on the public data) and clean prompt $\boldsymbol{P_{clean}}$ (obtained via standard prompt tuning on the public data). These offline statistics are represented by μ_{adv}, σ^2_{adv}, μ_{clean}, and σ^2_{clean}, respectively. We then align the mean and variance of the current embeddings with these pre-computed statistics as follows:

$$\mathcal{L}_{\text{adv}} = \frac{1}{L} \sum_{l=1}^{L} (\|\mu_l - \mu_{adv,l}\|_1 + \|\sigma_l^2 - \sigma_{adv,l}^2\|_1), \tag{6}$$

$$\mathcal{L}_{\text{clean}} = \frac{1}{L} \sum_{l=1}^{L} (\|\mu_l - \mu_{clean,l}\|_1 + \|\sigma_l^2 - \sigma_{clean,l}^2\|_1), \tag{7}$$

$$\mathcal{L}_{\text{TAPT}} = \mathcal{L}_{\text{entropy}} + \alpha \mathcal{L}_{\text{adv}} + (1 - \alpha) \mathcal{L}_{\text{clean}}, \tag{8}$$

where α is a hyperparameter. Setting $\alpha = 0$ aligns the prompt with the clean distribution, which may reduce adversarial robustness, while setting $\alpha = 1$ aligns the prompt with the robust distribution, which may impact performance on clean samples. Our final objective combines the multi-view entropy loss with the adversarial-clean alignment to optimize the prompt during inference for a given test sample. This approach enhances adversarial robustness while preserving accuracy on clean samples. ***Note that, to ensure inference efficiency, TAPT performs only a single step of prompt tuning for each inference.***

4. Experiments

4.1. Experimental Setup

Datasets and Models We experiment on 11 benchmark datasets (ImageNet val set and 10 other zero-shot test datasets): ImageNet [35], Caltech101 [11], DTD [7], EuroSAT [14], Pets [32], Aircraft [27], Food101 [6], Flowers [30], Cars [22], SUN397 [53], and UCF101 [41]. Our experiments focus on the CLIP model, specifically utilizing the ViT-B/16 and ViT-B/32 architectures. Following standard CLIP usage, we used hand-crafted prompts as textual inputs. For example, the prompt "`a photo of a <class>, a type of pet`" was applied for the Pets dataset. A summary of these datasets and their corresponding hand-crafted prompts are provided in the Appendix.

Attack Configuration We evaluate the zero-shot adversarial robustness of CLIP against both white-box and black-box attacks. Specifically, we employ PGD-100 [26] for white-box attacks, DI [54] for black-box attacks, and the more powerful AutoAttack [8]. The hyperparameters for PGD-100 and DI were configured based on the TorchAttacks library [21]. Consistent with [28], we use perturbation budgets of $\epsilon = 1/255$, $2/255$, and $4/255$ for both attacks.

Defense Configuration For existing APT methods, we use the original configurations of APT methods [23, 60] and generate adversarial examples using the PGD-2 attack with step size $\alpha = 1/255$. We then develop more robust versions of APT methods with different prompt designs, including (1) adversarial visual prompt tuning (APT-V), (2) adversarial V-L joint prompt tuning (APT-VLJ), and (3) adversarial V-L independent prompt tuning (APT-VLI).

Implementation Details For our TAPT method, we initialize the defensive prompt using APT on ImageNet, training for 100 epochs with a batch size of 32 and a learning rate of 0.035. To generate augmented views for test-time fine-tuning, we create 63 variations of each test sample using random resized crops and horizontal flips, resulting in 64 images per sample, including the original. From these 64 images, we select the top 10% most confident predictions (with the lowest entropy) and compute the average entropy of their predicted probabilities. For adversarial-clean alignment, we pre-compute embedding statistics from the public dataset (ImageNet) using APT and standard PT, respectively. We then optimize the defensive prompts by minimizing a combined loss of multi-view entropy and adversarial-clean alignment using the AdamW optimizer, with a learning rate of 5×10^{-4} and an adversarial-clean scale factor of $\alpha = 0.5$, on a single NVIDIA A100 GPU.

4.2. Main Results

Zero-Shot Adversarial Robustness We compare our TAPT method with existing APT methods across three

				ImageNet	Caltech101	DTD	EuroSAT	Pets	Aircraft	Food101	Flowers	Cars	SUN397	UCF101	Avg.
CLIP	Vanilla	ViT-B/16	PGD	1.4	21.3	1.4	6.0	5.0	0.0	9.8	1.6	0.7	0.8	1.8	4.5
			DI	6.1	25.8	8.7	0.3	11.1	0.5	9.5	3.0	3.4	4.9	4.7	7.1
			AA	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1
		ViT-B/32	PGD	1.3	22.9	5.0	0.0	2.6	0.0	3.6	1.5	0.2	1.1	1.7	3.6
			DI	6.4	37.2	11.1	0.7	9.9	0.1	7.5	9.3	3.9	8.6	5.8	9.1
			AA	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.1
Visual Only	APT-V	ViT-B/16	PGD	19.4	61.2	18.5	8.0	37.4	3.9	12.1	25.1	9.8	17.2	16.4	20.8
			DI	29.2	67.5	23.5	12.3	47.1	5.0	21.8	30.2	18.1	27.1	22.0	27.6
			AA	14.8	55.9	15.2	2.3	31.3	2.3	8.3	18.0	5.8	12.4	12.7	16.3
		ViT-B/32	PGD	18.9	63.8	20.2	0.6	36.1	2.7	15.0	23.1	9.4	19.1	18.4	20.7
			DI	26.9	68.6	22.0	5.1	46.0	4.8	23.7	26.5	13.9	26.4	24.6	26.2
			AA	8.3	44.9	12.7	0.1	16.4	0.5	5.5	9.0	2.7	7.5	7.8	10.5
	TAPT-V	ViT-B/16	PGD	40.1 (20.7↑)	69.7 (8.5↑)	28.1 (9.6↑)	23.9 (15.9↑)	49.2 (11.8↑)	11.5 (7.6↑)	54.5 (42.4↑)	41.1 (16.0↑)	28.6 (18.8↑)	40.3 (23.1↑)	36.5 (20.1↑)	38.5 (17.7↑)
			DI	46.7 (17.5↑)	75.8 (8.3↑)	33.2 (9.7↑)	33.0 (20.7↑)	56.7 (9.6↑)	12.7 (7.7↑)	61.5 (39.7↑)	46.2 (16.0↑)	35.7 (17.6↑)	45.5 (18.4↑)	40.1 (18.1↑)	44.3 (16.7↑)
			AA	49.2 (34.4↑)	75.7 (19.8↑)	36.6 (21.4↑)	36.9 (34.6↑)	57.1 (25.8↑)	19.4 (17.1↑)	68.8 (60.5↑)	52.9 (34.9↑)	44.0 (38.2↑)	48.9 (36.5↑)	48.1 (35.4↑)	48.9 (32.6↑)
		ViT-B/32	PGD	42.2 (23.3↑)	79.4 (15.6↑)	32.2 (12.0↑)	24.4 (23.8↑)	62.4 (26.3↑)	12.1 (9.4↑)	53.2 (38.2↑)	47.1 (24.0↑)	33.5 (24.1↑)	44.1 (25.0↑)	44.5 (26.1↑)	43.2 (22.5↑)
			DI	46.5 (19.6↑)	81.5 (12.9↑)	33.6 (11.6↑)	24.8 (19.7↑)	66.0 (20.0↑)	13.3 (8.5↑)	57.7 (34.0↑)	48.7 (22.2↑)	37.5 (23.6↑)	47.3 (20.9↑)	47.7 (23.1↑)	45.9 (19.7↑)
			AA	44.1 (35.8↑)	76.3 (31.4↑)	33.5 (20.8↑)	31.4 (31.3↑)	54.9 (38.5↑)	16.4 (15.9↑)	53.6 (48.1↑)	47.3 (38.3↑)	42.4 (39.7↑)	45.6 (38.1↑)	45.0 (37.2↑)	44.6 (34.1↑)
V-L Joint	APT-VLJ	ViT-B/16	PGD	23.9	61.7	18.7	9.7	41.4	3.2	14.1	23.4	12.2	17.7	15.5	22.0
			DI	34.9	72.0	25.0	10.5	52.8	4.2	24.6	32.9	18.9	28.5	24.8	29.9
			AA	16.5	53.2	12.6	5.6	31.3	1.7	8.0	16.0	4.9	11.1	11.1	15.6
		ViT-B/32	PGD	21.4	64.3	14.7	10.4	37.7	2.0	16.5	21.0	8.9	17.5	18.0	21.1
			DI	30.4	69.5	17.0	11.9	47.9	2.9	26.7	26.1	17.3	26.5	26.4	27.5
			AA	10.3	46.2	10.0	3.0	18.5	0.6	6.7	9.5	1.7	7.4	7.2	11.1
	TAPT-VLJ	ViT-B/16	PGD	50.2 (26.3↑)	81.0 (19.3↑)	29.5 (10.8↑)	13.5 (3.8↑)	68.7 (27.3↑)	5.6 (2.4↑)	41.7 (27.6↑)	39.3 (15.9↑)	28.1 (15.9↑)	39.8 (22.1↑)	41.5 (26.0↑)	39.9 (17.9↑)
			DI	51.7 (16.8↑)	82.3 (10.3↑)	30.4 (5.4↑)	14.5 (4.0↑)	69.5 (16.7↑)	5.8 (1.6↑)	43.2 (18.6↑)	40.2 (7.3↑)	30.2 (11.3↑)	41.3 (12.8↑)	42.2 (17.4↑)	41.0 (11.1↑)
			AA	52.4 (35.9↑)	81.5 (28.3↑)	30.2 (17.6↑)	15.4 (9.8↑)	69.1 (37.8↑)	5.8 (4.1↑)	44.8 (36.8↑)	40.0 (24.0↑)	31.0 (26.1↑)	42.1 (31.0↑)	44.3 (33.2↑)	41.5 (25.9↑)
		ViT-B/32	PGD	52.1 (30.7↑)	80.6 (16.3↑)	27.1 (12.4↑)	13.4 (3.0↑)	72.0 (34.3↑)	7.6 (5.6↑)	52.5 (36.0↑)	39.9 (18.0↑)	32.3 (23.4↑)	44.6 (27.1↑)	45.3 (27.1↑)	42.4 (21.3↑)
			DI	52.4 (22.0↑)	80.5 (11.0↑)	27.1 (10.1↑)	13.6 (1.7↑)	71.6 (23.7↑)	7.7 (4.8↑)	52.0 (25.3↑)	40.0 (13.9↑)	33.4 (16.1↑)	44.6 (18.1↑)	45.5 (19.1↑)	42.6 (15.1↑)
			AA	52.4 (42.1↑)	78.6 (32.4↑)	27.4 (17.4↑)	13.1 (10.1↑)	70.0 (51.5↑)	8.6 (8.0↑)	53.0 (46.3↑)	40.5 (31.0↑)	34.7 (33.0↑)	44.9 (37.5↑)	44.8 (37.6↑)	42.5 (31.4↑)
V-L Independent	APT-VLI	ViT-B/16	PGD	24.3	65.3	18.9	10.0	43.6	3.1	14.3	23.6	10.5	18.2	17.4	22.7
			DI	35.1	67.7	19.3	10.6	49.8	4.3	23.6	27.7	16.2	26.4	21.5	27.5
			AA	17.2	57.1	14.4	8.2	35.3	1.5	8.9	16.6	5.1	11.9	12.5	17.2
		ViT-B/32	PGD	21.2	63.2	15.8	10.4	37.6	1.5	16.1	20.2	8.3	16.9	17.6	20.8
			DI	28.9	69.5	20.3	10.9	47.0	2.8	24.8	24.3	14.8	23.8	23.6	26.4
			AA	9.7	45.6	10.6	6.8	17.3	0.4	6.0	8.7	2.0	6.8	6.9	11.0
	TAPT-VLI	ViT-B/16	PGD	50.0 (25.7↑)	79.0 (13.7↑)	32.4 (13.5↑)	36.2 (26.2↑)	67.5 (51.4↑)	13.1 (10.0↑)	65.7 (51.4↑)	49.8 (26.2↑)	39.6 (29.1↑)	48.3 (30.1↑)	47.5 (30.1↑)	48.1 (25.4↑)
			DI	53.8 (18.7↑)	80.5 (12.8↑)	32.3 (13.0↑)	39.6 (29.0↑)	69.4 (19.6↑)	13.3 (9.0↑)	70.6 (47.0↑)	51.1 (23.4↑)	42.5 (26.3↑)	50.3 (23.9↑)	48.2 (26.7↑)	50.1 (22.6↑)
			AA	55.1 (37.9↑)	80.3 (23.2↑)	35.7 (21.3↑)	44.4 (36.2↑)	70.2 (34.9↑)	16.0 (14.5↑)	76.2 (67.3↑)	55.1 (38.5↑)	50.5 (45.4↑)	53.6 (41.7↑)	54.5 (42.0↑)	53.8 (36.6↑)
		ViT-B/32	PGD	48.2 (27.0↑)	82.1 (18.9↑)	30.7 (14.9↑)	31.6 (21.2↑)	68.1 (30.5↑)	4.5 (3.0↑)	64.7 (48.6↑)	44.6 (24.4↑)	41.3 (33.0↑)	47.6 (30.7↑)	49.1 (31.5↑)	46.6 (25.8↑)
			DI	50.0 (21.1↑)	82.7 (13.2↑)	31.5 (11.2↑)	32.7 (21.8↑)	69.4 (22.4↑)	4.4 (1.6↑)	65.6 (40.8↑)	45.1 (20.8↑)	42.8 (28.0↑)	49.1 (25.3↑)	49.9 (26.3↑)	47.6 (21.2↑)
			AA	49.7 (40.0↑)	80.6 (35.0↑)	33.3 (22.7↑)	38.1 (31.3↑)	68.4 (51.1↑)	5.0 (4.6↑)	67.5 (61.5↑)	47.1 (38.4↑)	48.1 (46.1↑)	50.3 (43.5↑)	50.4 (43.5↑)	49.0 (38.0↑)

Table 1. Zero-shot adversarial robustness (%) of different defense methods from ImageNet to downstream datasets, evaluated against PGD [26], DI [54], and AutoAttack (AA) [8] under perturbation budget $\epsilon = 1/255$. The baseline APT methods (APT-V, APT-VLJ, and APT-VLI) were tuned on ImageNet under a 16-shot setting and then assessed on the other 10 datasets. The green upward arrows (↑) highlight the performance improvement of our TAPT over the baselines.

	Source	Target										
	ImageNet	Caltech101	DTD	EuroSAT	Pets	Aircraft	Food101	Flowers	Cars	SUN397	UCF101	**Avg.**
APT-V	60.6	89.7	36.9	26.7	84.8	16.8	63.2	55.7	52.1	58.2	54.4	54.5
TAPT-V	**66.9**	**92.8**	**44.6**	**41.0**	**88.5**	**23.8**	**85.7**	**67.4**	**65.7**	**62.9**	**65.5**	**64.1**
APT-VLJ	64.0	88.5	34.8	16.0	**81.1**	**8.1**	62.6	**53.3**	47.0	53.0	52.8	51.0
TAPT-VLJ	**66.5**	**88.5**	**36.7**	**16.7**	80.6	7.4	**64.6**	50.4	**48.7**	**54.7**	**55.4**	**51.8**
APT-VLI	63.8	88.6	34.1	17.2	80.7	11.6	61.1	51.8	45.9	53.5	52.7	51.0
TAPT-VLI	**65.8**	**90.4**	**39.2**	**48.3**	**82.5**	**16.3**	**85.8**	**60.3**	**62.5**	**60.0**	**64.4**	**61.4**

Table 2. Zero-shot clean accuracy (%) of different defense methods from ImageNet to downstream datasets, including APT baselines (APT-V, APT-VLJ, and APT-VLI) and our TAPT. The backbone is ViT-B/16. The best results are **boldfaced**.

prompt designs: visual-only (V), V-L joint (VLJ), and V-L independent (VLI). Table 1 presents the zero-shot adversarial robustness results under PGD-100, DI, and AutoAttack attacks, with the 'Vanilla' showing results without any defense. It is clear that both white-box and black-box attacks can drastically reduce accuracy (nearly 0%) in the absence of defenses, using only imperceptible noise with $\epsilon = 1/255$. AutoAttack emerges as the most effective attack, achieving a nearly 100% attack success rate on average across all datasets. For defense, our TAPT method consistently achieves the best average performance under AutoAttack across all prompt designs (visual-only, V-L joint, and V-L independent) and ViT architectures (ViT-B/16 and ViT-B/32). With ViT-B/16, TAPT enhances robustness by 32.6% (visual-only), 25.9% (V-L joint), and 36.6% (V-L independent), respectively. Similar improvements are observed with ViT-B/32, where TAPT yields gains of 34.1%, 31.4%, and 38.0% for the respective prompt designs. TAPT also demonstrates superior performance against both PGD-100 and DI attacks.

Furthermore, TAPT with the V-L independent prompt design achieves the highest average zero-shot adversar-

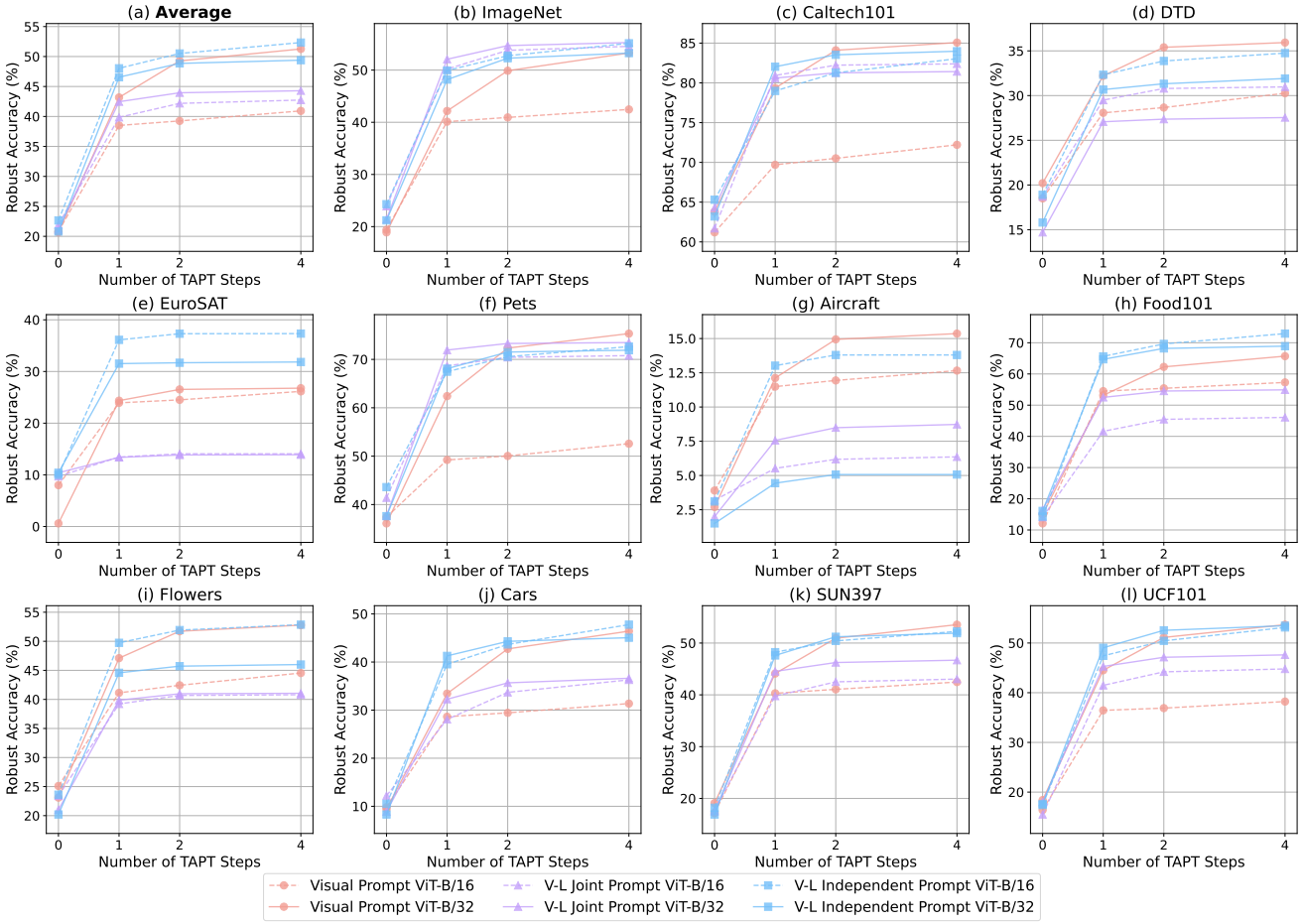

Figure 4. Adversarial robustness (%) of our TAPT method under different test-time adaptation steps (i.e., {0, 1, 2, 4}). The results are reported against the PGD-100 attack on ViT-B/16 and ViT-B/32 architectures.

ial robustness, surpassing both visual-only and V-L joint prompt designs. To summarize, our experiments reveal that: (1) incorporating textual prompts generally improves zero-shot adversarial robustness across various datasets; (2) the V-L joint prompt enhances robustness on the source domain (ImageNet) more effectively than other prompt designs; and (3) the V-L independent prompt design is more effective in enhancing zero-shot adversarial robustness than the V-L joint prompt design, likely due to the challenges in optimizing the interplay between visual and textual prompts.

Zero-Shot Clean Accuracy Table 2 presents the zero-shot clean accuracy of different defense methods from ImageNet to downstream datasets, comparing APT baselines (APT-V, APT-VLJ, and APT-VLI) with our TAPT method on the ViT-B/16 architecture. Note that the APT baselines were trained on ImageNet and subsequently tested on 10 downstream datasets, while for our TAPT, only the robust statistics were computed based on ImageNet. The zero-shot clean accuracy should be compared on all 10 datasets. As shown, TAPT consistently achieves superior clean performance across all 11 datasets, demonstrating significantly stronger generalization capabilities. While a slight performance decrease is observed with TAPT's V-L joint prompt design on Pets, Aircraft, and Flowers, it remains competitive. Overall, TAPT outperforms APT across all three prompt designs (visual-only, V-L joint, and V-L independent), with average improvements of 9.6%, 0.8%, and 10.4%, respectively. These results underscore TAPT's ability to improve adversarial robustness without compromising too much clean accuracy.

4.3. Ablation Studies

Number of TAPT Steps We first examine the effect of the number of test-time adaptation steps on the robust accuracy of TAPT. Figure 4 shows TAPT's robustness performance with varying steps (0, 1, 2, and 4) under the PGD-100 at-

tack. Notably, step $= 0$ represents the baseline performance without TAPT, reflecting the adversarial robustness achieved solely through APT. A clear trend emerges, showing that robust accuracy increases with additional TAPT steps across most datasets and prompt designs. While robustness gains generally stabilize after a few TAPT steps, the improvement from 0 to 1 step is consistently substantial, highlighting the effectiveness of even a single adaptation step. However, datasets vary in sensitivity to the number of TAPT steps; for example, performance stabilizes quickly on datasets like EuroSAT, whereas for datasets such as ImageNet and DTD, further improvements are observed with additional steps. The figure also illustrates that TAPT's benefits are consistent across different prompt designs. We further analyzed the per-sample time overhead of TAPT, finding that the additional inference time per image is 0.095s (visual-only), 0.166s (V-L joint), and 0.165s (V-L independent), respectively. This demonstrates that TAPT not only consistently enhances robust accuracy but also maintains a relatively low time cost.

Different Perturbation Budgets We further assess TAPT's robustness under varying attack strengths, defined by the perturbation budget ϵ. Figure 5 displays zero-shot adversarial robustness across 11 datasets with ϵ values of 1/255, 2/255, and 4/255, and TAPT steps of 1, 2, and 4. As expected, robust accuracy decreases as ϵ increases, indicating stronger attacks. Nonetheless, TAPT consistently enhances robust accuracy across all datasets and ϵ values.

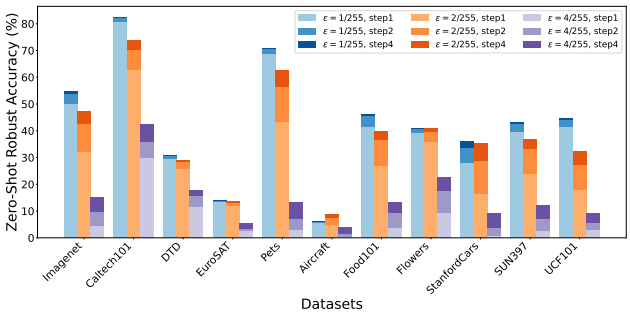

Figure 5. Zero-shot adversarial robustness (y-axis) ofs TAPT under varying perturbation budgets ϵ (1/255, 2/255, and 4/255) and TAPT steps (1, 2, and 4).

TAPT Reset Intervals Our TAPT method resets the prompt to its initial state before processing each test sample, ensuring that each test sample is handled independently. However, we also explored alternative strategies with varied reset intervals. As shown in Table 3, we experimented with reset intervals ranging from 1 to 32, as well as a strategy with no reset during the entire inference process (i.e., "reset=all"). Our findings indicate that continuous prompt adaptation without frequent resets can further improve robust accuracy, suggesting that accumulating information across multiple inputs can be beneficial. However, this continuous strategy introduces a vulnerability to potential poisoning attacks. To mitigate this risk and ensure reliable test-time defense, we recommend the per-sample reset strategy ("reset=1"), which prioritizes robustness against potential attacks over marginal gains from continued TAPT.

Reset Interval	ImageNet	10 Zero-Shot Datasets Avg.
reset=1	49.92	47.85
reset=2	50.20	**47.91**
reset=4	50.69	**47.91**
reset=8	51.14	47.53
reset=16	51.49	46.39
reset=32	**51.62**	44.50
reset=all	0.48	3.79

Table 3. Zero-shot adversarial robustness (%) of TAPT with varying reset intervals on ImageNet and 10 other zero-shot datasets. "reset = N" means the prompt is reset after every N test samples.

5. Limitation

As a test-time defense method, TAPT has certain limitations that warrant further research. Our method primarily addresses attacks in the image modality by aligning adversarial image embeddings with pre-computed public data statistics on a public dataset. Future work could explore additional modality alignment and acceleration techniques to facilitate TAPT's deployment in industrial applications. Moreover, our current focus is limited to image recognition tasks. Extending TAPT to a broader range of tasks, such as visual reasoning and visual question answering in advanced models like GPT-4V [2] and Gemini [45], represents a promising direction for future research.

6. Conclusion

In this paper, we introduced a novel test-time defense method, *Test-Time Adversarial Prompt Tuning (TAPT)*, to enhance the inference robustness of pre-trained VLMs, such as CLIP. TAPT tunes defensive bimodal (textual and visual) prompts for each test sample by leveraging multi-view entropy minimization and adversarial-clean alignment, effectively safeguarding CLIP's zero-shot inference. It also utilizes pre-computed statistics from a public dataset (ImageNet) to defend a wide range of downstream tasks. Comprehensive evaluation across 11 benchmark datasets demonstrates that TAPT effectively enhances zero-shot adversarial robustness against both white-box and black-box attacks, while largely preserving clean accuracy. Compared to training-time adversarial prompt tuning (APT) methods, TAPT offers several advantages: (1) it is unsupervised, (2) enables sample-wise prompt adaptation, (3) delivers superior zero-shot adversarial robustness and clean accuracy, (4) is independent of downstream tasks, and (5) is lightweight. Future research could explore the scalability of test-time defense across different modalities.

Acknowledgements This work is supported by National Key R&D Program of China (Grant No. 2022ZD0160103) and National Natural Science Foundation of China (Grant No. 62276067). The computations in this research were performed using the CFFF platform of Fudan University.

References

[1] Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *NeurIPS*, 2024. 3, 4

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *preprint arXiv:2303.08774*, 2023. 8

[3] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *preprint arXiv:2204.01691*, 2022. 1

[4] Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In *NeurIPS*, 2022. 3

[5] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *preprint arXiv:2203.17274*, 2022. 5

[6] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 5

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 5

[8] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020. 1, 2, 5, 6

[9] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018. 1

[10] Hao Fang, Jiawei Kong, Wenbo Yu, Bin Chen, Jiawei Li, Shutao Xia, and Ke Xu. One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models. *preprint arXiv:2406.05491*, 2024. 2

[11] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshops*, 2004. 5

[12] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. 1

[13] Bangyan He, Xiaojun Jia, Siyuan Liang, Tianrui Lou, Yang Liu, and Xiaochun Cao. Sa-attack: Improving adversarial transferability of vision-language pre-training models via self-augmentation. *preprint arXiv:2312.04913*, 2023. 2

[14] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J-STARS*, 2019. 5

[15] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 2023. 1

[16] Noor Hussein, Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. Promptsmooth: Certifying robustness of medical vision-language models via prompt learning. In *MICCAI*, 2024. 2

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1

[18] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *CVPR*, 2022. 1

[19] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023. 1

[20] Eungyeup Kim, Mingjie Sun, Christina Baek, Aditi Raghunathan, and J Zico Kolter. Test-time adaptation induces stronger accuracy and agreement-on-the-line. In *NeurIPS*, 2024. 3

[21] Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *preprint arXiv:2010.01950*, 2020. 5

[22] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, 2013. 5

[23] Lin Li, Haoyan Guan, Jianing Qiu, and Michael Spratling. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *CVPR*, 2024. 1, 2, 5

[24] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *ICCV*, 2023. 2

[25] Xingjun Ma, Linxi Jiang, Hanxun Huang, Zejia Weng, James Bailey, and Yu-Gang Jiang. Imbalanced gradients: a subtle cause of overestimated adversarial robustness. *Machine Learning*, 2024. 2

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018. 1, 2, 5, 6

[27] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *preprint arXiv:1306.5151*, 2013. 5

[28] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *ICLR*, 2023. 2, 5

[29] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *preprint arXiv:2006.10963*, 2020. 3

[30] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 5

[31] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *ICML*, 2022. 3

[32] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 5

[33] Anibal Pedraza, Oscar Deniz, and Gloria Bueno. On the relationship between generalization and robustness to adversarial examples. *Symmetry*, 2021. 1

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 3

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 5

[36] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *ICML*, 2024. 2

[37] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *NeurIPS*, 2020. 3

[38] Or Sharir, Barak Peleg, and Yoav Shoham. The cost of training nlp models: A concise overview. *preprint arXiv:2004.08900*, 2020. 1

[39] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *CoRL*, 2022. 1

[40] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022. 3, 4

[41] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *preprint arXiv:1212.0402*, 2012. 5

[42] David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *CVPR*, 2019. 1

[43] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *ECCV*, 2018. 1

[44] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2013. 1

[45] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*, 2023. 8

[46] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 3

[47] Haodi Wang, Kai Dong, Zhilei Zhu, Haotong Qin, Aishan Liu, Xiaolin Fang, Jiakai Wang, and Xianglong Liu. Transferable multimodal attack on vision-language pre-training models. In *IEEE S&P*, 2024. 2

[48] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *CVPR*, 2022. 3

[49] Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *CVPR*, 2024. 2

[50] Xin Wang, Kai Chen, Xingjun Ma, Zhineng Chen, Jingjing Chen, and Yu-Gang Jiang. AdvQDet: Detecting query-based adversarial attacks with adversarial contrastive prompt tuning. In *ACM MM*, 2024. 2

[51] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *EMNLP*, 2022. 1

[52] Zeyu Wang, Xianhang Li, Hongru Zhu, and Cihang Xie. Revisiting adversarial training at scale. In *CVPR*, 2024. 1, 2

[53] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 5

[54] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, 2019. 2, 5, 6

[55] Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. In *NeurIPS*, 2023. 2

[56] Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *CVPR*, 2024. 3

[57] Gengyuan Zhang, Jisen Ren, Jindong Gu, and Volker Tresp. Multi-event video-text retrieval. In *ICCV*, 2023. 1

[58] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019. 1

[59] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *ACM MM*, 2022. 1, 2

[60] Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models. In *ECCV*, 2024. 1, 2, 5

[61] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *NeurIPS*, 2022. 3

[62] Peng-Fei Zhang, Zi Huang, and Guangdong Bai. Universal adversarial perturbations for vision-language pre-trained models. In *ACM SIGIR*, pages 862–871, 2024. 2

[63] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *NeurIPS*, 2024. 1, 2

[64] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 1

[65] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 1

[66] Wanqi Zhou, Shuanghao Bai, Qibin Zhao, and Badong Chen. Revisiting the adversarial robustness of vision language models: a multimodal perspective. *preprint arXiv:2404.19287*, 2024. 2

[67] Yiwei Zhou, Xiaobo Xia, Zhiwei Lin, Bo Han, and Tongliang Liu. Few-shot adversarial prompt learning on vision-language models. In *NeurIPS*, 2024. 1, 2

[68] Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *ACM MM*, 2023. 2