ATAC: Augmentation-Based Test-Time Adversarial Correction for CLIP

Linxiang Su András Balogh
University of Szeged
su_linxiang@126.com abalogh@inf.u-szeged.hu

Abstract

Despite its remarkable success in zero-shot image-text matching, CLIP remains highly vulnerable to adversarial perturbations on images. As adversarial fine-tuning is prohibitively costly, recent works explore various test-time defense strategies; however, these approaches still exhibit limited robustness. In this work, we revisit this problem and propose a simple yet effective strategy: Augmentation-based Test-time Adversarial Correction (ATAC). Our method operates directly in the embedding space of CLIP, calculating augmentation-induced drift vectors to infer a semantic recovery direction and correcting the embedding based on the angular consistency of these latent drifts. Across a wide range of benchmarks, ATAC consistently achieves remarkably high robustness, surpassing that of previous state-of-the-art methods by nearly 50% on average, all while requiring minimal computational overhead. Furthermore, ATAC retains state-of-the-art robustness in unconventional and extreme settings and even achieves nontrivial robustness against adaptive attacks. Our results demonstrate that ATAC is an efficient method in a novel paradigm for test-time adversarial defenses in the embedding space of CLIP. Code is available at https://github.com/kylin0421/ATAC.

1. Introduction

Vision–language models (VLMs) trained on web-scale image–text corpora have transformed zero-shot recognition and open-world retrieval, with CLIP emerging as a widely adopted foundation model for image–text alignment [38]. The embedding space of CLIP makes it a particularly useful component in vision-language-action (VLA) systems, where CLIP representations provide a strong perceptual prior that enables agents to generalize to previously unseen environments, tasks, and objects [40, 42, 44]. As such models migrate to safety- and security-critical applications, robustness becomes indispensable: small, human-imperceptible perturbations can reliably induce arbitrary errors in neural networks [2, 5, 16, 20], and CLIP is no exception. Its vulnerability raises concerns about trustworthiness and deployment risks, given its growing influence on machine perception [28, 32, 53].

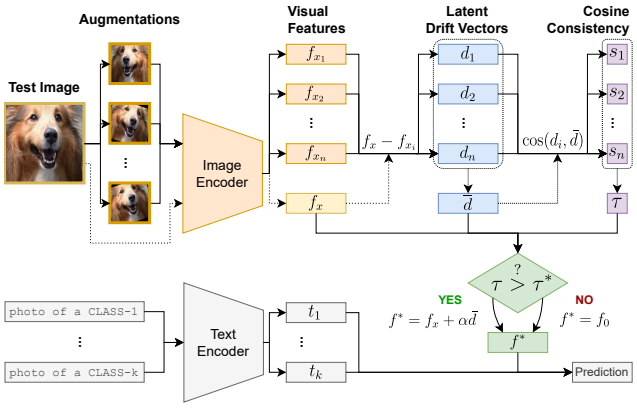

Figure 1. Overview of the ATAC framework. We use the visual features of augmented views to estimate a semantic recovery direction $\bar{d}$, and a cosine consistency gate to control the correction of the visual embedding.

A rich body of research has sought to improve robustness from three complementary directions. Training-time methods (e.g., adversarial training, TRADES) [29, 32, 39] offer stronger worst-case guarantees but are prohibitively costly at the scale of foundation models, and may erode zero-shot generalization [49]. Input-space purification and randomized smoothing [9, 18, 25, 34] reduce attack effectiveness but often trade off clean accuracy and remain susceptible to adaptive attacks. In contrast, *test-time* strategies adapt inference-time behavior without retraining. For CLIP, two main directions have been explored: *prompt-side adaptation* (e.g., test-time prompt tuning) adjusts textual embeddings to counter undesirable and adversarial distribution shifts [41, 43], while *image-side counterattacks* refine inputs at inference to push predictions toward ground-truth classes [49]. Despite rapid progress, both approaches exhibit limitations: gradient-based counterattacks are computationally expensive and rely on sensitive hyperparameters, while prompt-only tuning relies on the unstable assumption that the predictions of augmented views can be aggregated to

counter adversarial attacks.

We take a different route grounded in empirical observations: although adversarial perturbations can flip CLIP's zero-shot prediction, CLIP's *embeddings* remain comparatively stable under standard augmentations (e.g., flip, rotation, color jitter) [11]. We demonstrate that adversarially perturbed samples produce *aligned* augmentation-induced shifts in CLIP's embedding space, while clean samples exhibit *scattered* shifts. Building on this, we propose **A**ugmentation-based **T**est-time **A**dversarial **C**orrection (ATAC), a simple yet powerful test-time defense that operates directly in CLIP's embedding space. ATAC constructs multiple augmented views, computes their embedding *drift vectors* relative to the original, and averages them to estimate a semantic recovery direction. The original embedding is then corrected along this direction by a step α, gated by a cosine-consistency threshold τ that suppresses unnecessary corrections for clean inputs. Conceptually, ATAC bridges the gap between prompt-side tuning and input-space optimization by performing a lightweight, training-free, feature-level correction with negligible overhead.

Across 13 classification benchmarks, ATAC consistently yields remarkable improvements over both test-time defenses and CLIP-specific adversarial fine-tuning methods. Our method improves previous state-of-the-art robustness benchmarks by nearly 50% on average with minimal computational overhead. Further analysis reveals that ATAC is able to effectively exploit a fundamental flaw in untargeted, gradient-based attacks; however, our method also achieves state-of-the-art results when evaluated against attacks that limit or eliminate this flaw. Moreover, ATAC is effective against adaptive attacks from both the perspective of robustness and computational cost. Our results demonstrate the effectiveness of *direct semantic correction in CLIP's embedding space*, a currently underexplored paradigm.

Our contributions are as follows:

- We propose ATAC, a novel test-time adversarial defense method that directly corrects visual embeddings in CLIP's feature space.
- We demonstrate that ATAC achieves state-of-the-art robustness across 13 classification benchmarks, outperforming both adversarial fine-tuning methods and test-time defenses by nearly 50% on average, with minimal computational overhead.
- Further analysis of the effectiveness of ATAC reveals a fundamental flaw of untargeted gradient-based attacks that our method can exploit; however, we show that ATAC still achieves state-of-the-art robustness when this flaw is eliminated.
- Finally, we show that ATAC achieves nontrivial robustness even against adaptive attacks, surpassing other test-time defenses.

2. Related Work

Adversarial robustness in VLMs. While adversarial fragility is well documented in standard deep neural networks [5, 16, 20], recent work has increasingly focused on vision-language models (VLMs) such as CLIP [28, 32, 53]. For VLMs, defenses can be organized into three (often overlapping) families. (i) *Fine-tuning–based robustness* adversarially fine-tunes the image/text encoders or the contrastive objective to enlarge margins [32, 39, 47], but this is costly at the scale of foundation models and can erode zero-shot transfer [49]. (ii) *Prompt-side adaptation* adjusts text and/or visual prompts; this includes *training-time* prompt tuning (e.g., soft prompts) and *test-time* prompt tuning (TPT) that adapts prompts for each input without retraining [24, 45, 48, 52]. In particular, TAPT [48] and R-TPT [41] sit at the intersection of prompt tuning methods and test-time adaptation. (iii) *Test-time defenses beyond prompts* avoid parameter updates by acting on the inference pipeline by applying, for example, stochastic transformations and ensembling to stabilize decisions, or image-side counterattacks that optimize the input at inference to push predictions back to the ground-truth class [9, 37, 49].

Test-time Defenses for VLMs. Vision-language models (VLMs), such as CLIP, inherit strong zero-shot capability and vulnerability to adversarial perturbations in their joint embedding space. Although conventional test-time adaptation (TTA) for VLMs has been extensively studied [1, 14, 43], defending against adversarial attacks at test time remains a relatively new direction. Recent test-time defenses can be broadly categorized into three fronts: *image-side*, *prompt-side*, and *latent-side* approaches.

On the **image side**, TTC [49] formulates a PGD-style counter-attack at inference to escape the "toxic" adversarial basin and recover semantics, achieving solid robustness, albeit with nontrivial computational overhead. Deng *et al.* [12] proposed FPT-Noise, a dynamic scene-aware test-time defense that adaptively injects counterattack noise guided by a feature perception threshold and regulates perturbation strength via scene-aware control, followed by test-time ensembling to suppress residual noise. Liu *et al.* [27] further introduced Self-Calibrated Consistency (SCC), enforcing semantic and spatial consistency across augmented views to correct adversarially perturbed embeddings, substantially improving CLIP's zero-shot robustness without retraining.

On the **prompt side**, APT [24] introduces prompt tuning into adversarial defense, but requires optimization through backpropagation. TAPT [48] learns bimodal (visual and textual) defensive prompts for each test sample through multi-view entropy minimization and distribution alignment, while R-TPT [41] reformulates the classical objective to further enhance robustness. Beyond adversarial robustness, more extensive test-time adaptations or parameter-

efficient fine-tuning for CLIP (e.g., TPT and prompt learning variants [1, 14, 43], CLIP-Adapter [15]) demonstrate that lightweight test-time interventions—either on features or prompts-can substantially alter VLM behavior without retraining.

On the **latent side**, CLIPure [51] follows a purification-based approach, denoising CLIP embeddings directly in the latent space. Building upon a stochastic differential equation (SDE) framework that bridges attack and purification, it models latent likelihoods via diffusion priors (CLIPure-Diff) or cosine similarity (CLIPure-Cos).

In this work, we introduce a novel latent-side test-time defense that, instead of purification, optimizing pixels, or tuning prompts, performs semantic correction by leveraging augmentation-induced drift vectors to realign adversarial embeddings toward their true semantics.

3. Preliminaries and Notation

Zero-shot classification with CLIP. CLIP is a vision-language model comprising two modules: an image encoder E_I and a text encoder E_T. These modules have been trained on 400M image-text pairs to align images to their corresponding texts via cosine similarity.

Given a k-class image classification problem with labels $y_1, \ldots, y_k$, text prompts $\{TP(y_i)\}_{i=1}^k$ are constructed that represent each class, e.g., "`A photo of a {label}.`" Let $t_i = E_T(TP(y_i))$, denote the encoded representation of the text prompts, and $f_x = E_I(x)$ denote the encoded representation of an image x. Then, CLIP predicts the label that maximizes the cosine similarity between its embedding and the embedding of the image:

$$\arg\max_i \frac{\exp(\cos(f_x, t_i)/p_{temp})}{\sum_{j=1}^k \exp(\cos(f_x, t_j)/p_{temp})}, \tag{1}$$

where $\cos(\cdot)$ represents the cosine similarity operation and p_{temp} is a temperature parameter, typically set to 0.01.

Adversarial attacks on CLIP. CLIP is highly vulnerable to adversarial attacks [32]. The goal of the attacker is to find a small perturbation δ, such that δ is ϵ-bounded in the L_p-ball, i.e., $\|\delta\|_p \leq \epsilon$, and the image $x + \delta$ is misclassified by CLIP. For the sake of simplicity, we omit the projection of $x + \delta$ to the pixel space in all notations.

In a white-box setting, the attacker has access to the model, its gradients, and the ground truth label y_c of the image x. To find an adversarial perturbation, the attacker solves the optimization problem

$$\delta^* = \arg\max_{\delta \in \mathcal{S}} \mathcal{L}(x + \delta, y_c), \tag{2}$$

where $S = \{\delta : \|\delta\|_p \leq \epsilon\}$, $\mathcal{L}(\cdot)$ measures the classification loss of the model, and the adversarial input is $x_a = x + \delta^*$.

Eq. (2) can be indirectly approximated by the Projected Gradient Descent (PGD) [30] algorithm:

$$x^t = \prod_{x+S} (x^{t-1} + \gamma \text{sgn}(\nabla_x \mathcal{L}(x, y_c))) \quad (t = 1 \ldots T), \tag{3}$$

where T is the number of attack steps, γ is the step size, x^T is the adversarial example, and $x^0 = x$, or, for PGD with random start, $x^0 = x + \delta_0$, where $\delta_0 \sim \mathcal{U}(-\epsilon, \epsilon)$.

4. The ATAC Framework

In this section, we introduce **A**ugmentation-based **T**est-time **A**dversarial **C**orrection (ATAC), a novel test-time adversarial defense method for CLIP.

4.1. Motivation

Let us begin by highlighting three key results from related work along with their shortcomings that motivate our method.

Test-time counterattacks. Although adversarial finetuning has been shown to substantially increase the robustness of CLIP, it requires costly training procedures. To mitigate this, Xing *et al.* [49] explore test-time counterattacks that enable CLIP to defend itself without additional training. Their counterattack, also implemented by an adversarial attack, aims to maximize the L_2 distance between the original and the counterattacked image in the embedding space, thereby allowing the input to escape the adversarial "toxic region" in the embedding space, where it was moved by the original (malicious) attack.

However, this approach has two key limitations. Firstly, during inference, it relies on computationally costly counterattacks with sensitive hyperparameters. Secondly, the objective is only aimed at maximizing embedding shift, without any emphasis on restoring the original semantics of the input. Therefore, even if the counterattack succeeds in moving the embedding away from the adversarial region, this movement may not be towards the semantically correct direction.

Augmentations against adversarial attacks. Although recent work shows that the embedding space of CLIP remains relatively stable towards common image transformations (e.g., flipping, rotations, and color jittering) [11], extensive studies show that adversarial attacks are highly sensitive to such augmentations [20, 26, 33, 50]. In other words, *augmentations can mitigate the effect of adversarial attacks*.

This observation motivates test-time transformation ensembling (TTE) [37], where predictions over augmented views are aggregated to form the final prediction. However, as shown in Tab. 2, TTE yields limited robustness gains, leading us to hypothesize that augmentations alone are not enough to mitigate the effects of adversarial attacks.

Robust test-time prompt tuning. R-TPT [41] uses numerous augmentations to find views of the input image with low-

entropy predictions. In order to ignore adversarial or outlier views, they ensemble these predictions using a weighting scheme based on feature-level nearest neighbors to obtain the corrected prediction.

However, only aggregating predictions may allow low-entropy incorrect predictions of adversarial views to still mislead the method. Moreover, their weighting scheme is not an explicit way of recovering the semantics of the original input, as feature representations of incorrectly classified views can form tight clusters.

4.2. Our Method

To address the limitations of previous work, we propose **A**ugmentation-based **T**est-time **A**dversarial **C**orrection (ATAC). Our method uses the latent representations of augmented views to explicitly estimate a semantic recovery direction for adversarial inputs, and a cosine-consistency gate to control the correction process and avoid the over-correction of clean samples. Fig. 1 shows an overview of our method.

Formally, given an input image x, we first apply n different transformations that yield the augmented views $x_1, \ldots, x_n$, and, using the image encoder of CLIP, get their encoded representations $f_{x_1}, \ldots, f_{x_n}$ along with the encoded representation of the original input f_x. We then compute the latent drifts $d_1, \ldots, d_n$ and the mean drift $\bar{d}$ as

$$\{d_i = f_x - f_{x_i}\}_{i=1}^{n}, \quad \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i. \tag{4}$$

We assess the *directional consistency* of the latent drift vectors with the mean drift as

$$\tau = \frac{1}{n}\sum_{i=1}^{n} \cos(d_i, \bar{d}), \tag{5}$$

where $\cos$ is the cosine similarity operation. Lower τ values indicate scattered drift vectors, while higher values indicate that drift vectors point towards the same direction in the latent space. Since augmentations can mitigate the effect of adversarial attacks (as discussed in Sec. 4.1), we expect adversarial inputs to yield directionally consistent drift vectors. In this case, $\bar{d}$ represents a *semantic recovery direction* pointing towards the correct decision region from f_x, which we use to recover the original semantics of the attacked image.

On the other hand, we expect clean inputs to yield low τ values, as their representations, along with the representations of their augmented views, are scattered in the correct decision region. To avoid modifying the latent representations of clean images, we introduce a gating threshold τ^*, and perform the test-time correction

$$f^* = f_x + \alpha\bar{d} \quad \text{if} \quad \tau > \tau^*, \tag{6}$$

Method	Inference time (s)
CLIP (no defense) [38]	3.63±1.67
TTC [49]	20.67±1.96
R-TPT [41]	916.33±111.06
ATAC (ours)	18.41±0.40

Table 1. Inference times of the original CLIP and test-time defense methods on 1000 images, using a single RTX A6000 GPU. Results are averaged over 6 datasets, with standard deviations indicated.

otherwise we keep the original image embedding, i.e., $f^* = f_x$. We then normalize[1] f^* and treat it as the visual embedding for downstream tasks, such as classification.

Our method builds upon the robust embedding space of CLIP and the ability of augmentations to mitigate adversarial attacks. We address the limitations of related works by explicitly estimating a semantic recovery direction in CLIP's embedding space, using the representations of augmented views. ATAC is training-free and only requires n forward passes at inference, making it highly efficient compared to other test-time defenses, as shown in Tab. 1.

5. Results

5.1. Experimental Setup

Datasets. We closely follow the experimental setup of [49] and conduct our experiments on 13 datasets, which include general object recognition datasets CIFAR10 [22], CIFAR100 [22], STL10 [8], Caltech101 [13] and Caltech256 [17], fine-grained recognition datasets OxfordPets [36], Flowers102 [35], Food101 [4], StanfordCars [21], the scene recognition dataset Country211 [38], and domain-specific datasets FGVCAircraft [31], EuroSAT [19], DTD [7]. Similar to [49], we used the preprocessing pipeline of CLIP [38] for all datasets.

Implementation Details. Following [49], we used the official pre-trained CLIP ViT-B/32 [38] in our implementation. In all experiments, we used the cosine-consistency threshold $\tau^* = 0.85$ and the correction step size $\alpha = 7$. We used $n = 5$ augmentations in our framework, namely horizontal flip and rotations with degrees $\pm 15°$ and $\pm 30°$. We present ablations over these parameters in Sec. 5.3 and the Appendix.

Adversaries. We present our results against the PGD adversary (Eq. (3)) under the L_∞ norm. Unless otherwise specified, we used $\epsilon = 4/255$, $T = 10$, $\gamma = 1/255$ and random start in all our attacks. We present results against further adversaries, such as the Carlini-Wagner attack [5] and AutoAttack [10], in the Appendix, with similarly excellent results.

[1] Note that, when $f^* = f_x$ (i.e., f_x is not corrected), this normalization has no effect, since f_x is already normalized.

(%)		CLIP	Adversarial Finetuning				Test-time Defense				Δ
			CLIP-FT	TeCoA	PMG-AFT	FARE	TTE	TTC	R-TPT	**ATAC (ours)**	
CIFAR10	Rob.	0.43	2.75	11.7	15.59	5.42	3.47 ± 2.77	28.51 ± 0.36	33.84 ± 1.55	$\mathbf{91.80 \pm 0.09}$	+90.37
	Acc.	85.12	**84.90**	65.15	71.45	78.46	84.74 ± 0.40	81.18 ± 0.07	82.19 ± 1.03	79.39	−5.63
CIFAR100	Rob.	0.05	0.67	9.25	10.80	4.54	1.37 ± 0.96	9.06 ± 0.11	18.52 ± 1.08	$\mathbf{86.27 \pm 0.42}$	+86.22
	Acc.	57.14	**59.51**	36.30	41.51	47.38	58.61 ± 0.25	56.34 ± 0.20	52.69 ± 0.94	52.65	−4.49
STL10	Rob.	0.16	3.75	31.83	35.40	17.59	32.56 ± 11.76	52.40 ± 0.34	76.33 ± 2.46	$\mathbf{98.30 \pm 0.16}$	+98.14
	Acc.	96.40	94.49	81.69	84.35	89.11	$\mathbf{96.26 \pm 0.04}$	95.83 ± 0.03	96.09 ± 0.24	93.94	−2.46
Caltech101	Rob.	0.59	4.81	21.00	25.03	10.13	30.19 ± 7.92	36.66 ± 0.25	68.11 ± 0.24	$\mathbf{86.76 \pm 0.11}$	+86.17
	Acc.	85.66	83.63	64.41	69.06	76.58	85.84 ± 0.09	86.15 ± 0.08	$\mathbf{86.62 \pm 0.62}$	82.00	−3.66
Caltech256	Rob.	0.12	1.41	11.76	13.68	5.09	23.23 ± 7.77	27.25 ± 0.08	54.45 ± 0.71	$\mathbf{90.86 \pm 0.11}$	+90.74
	Acc.	81.72	78.53	52.05	53.32	67.22	$\mathbf{82.48 \pm 0.08}$	76.59 ± 0.12	77.67 ± 0.47	79.10	−2.62
OxfordPets	Rob.	0.00	1.66	3.71	5.10	0.30	3.18 ± 2.94	24.64 ± 0.53	44.15 ± 1.08	$\mathbf{87.46 \pm 0.19}$	+87.46
	Acc.	87.44	84.14	53.94	56.66	70.10	$\mathbf{88.13 \pm 0.13}$	64.70 ± 0.33	84.46 ± 0.62	85.15	−2.29
Flowers102	Rob.	0.00	0.13	3.81	4.26	0.62	3.52 ± 2.51	13.60 ± 0.33	32.46 ± 0.47	$\mathbf{85.69 \pm 0.16}$	+85.69
	Acc.	65.46	53.37	27.78	28.88	41.01	65.20 ± 0.23	63.24 ± 0.21	62.92 ± 0.85	**64.29**	−1.17
FGVCAircraft	Rob.	0.00	0.00	0.12	0.06	0.03	0.43 ± 0.43	6.40 ± 0.38	7.20 ± 0.62	$\mathbf{50.02 \pm 0.31}$	+50.02
	Acc.	20.10	14.04	3.51	3.24	7.77	$\mathbf{20.18 \pm 0.35}$	15.99 ± 0.04	19.14 ± 0.62	19.65	−0.45
StanfordCars	Rob.	0.00	0.00	0.41	0.40	0.04	1.46 ± 1.21	12.84 ± 0.20	20.76 ± 1.78	$\mathbf{70.80 \pm 0.08}$	+70.80
	Acc.	52.02	42.11	15.18	16.79	32.09	$\mathbf{52.73 \pm 0.31}$	41.52 ± 0.15	61.75 ± 0.24	51.41	−0.61
Country211	Rob.	0.00	0.00	0.19	0.24	0.02	0.24 ± 0.15	2.44 ± 0.15	0.42 ± 0.24	$\mathbf{65.55 \pm 0.25}$	+65.55
	Acc.	15.25	12.07	3.66	3.34	6.58	14.66 ± 0.14	11.99 ± 0.01	13.40 ± 0.62	**16.45**	+1.20
Food101	Rob.	0.00	0.04	1.35	2.12	0.24	5.31 ± 4.09	17.89 ± 0.13	39.97 ± 1.25	$\mathbf{96.11 \pm 0.07}$	+96.11
	Acc.	83.88	64.86	21.90	27.97	41.98	$\mathbf{83.96 \pm 0.01}$	80.00 ± 0.07	83.41 ± 0.47	82.87	−1.01
EuroSAT	Rob.	0.00	0.00	10.71	10.36	7.34	0.11 ± 0.09	13.57 ± 0.12	6.46 ± 1.03	$\mathbf{66.57 \pm 0.06}$	+66.57
	Acc.	42.59	27.64	17.53	19.19	18.22	44.38 ± 1.62	$\mathbf{53.24 \pm 0.09}$	21.83 ± 1.22	37.28	−5.31
DTD	Rob.	0.11	0.00	5.16	5.21	2.50	7.16 ± 2.32	11.40 ± 0.28	26.97 ± 1.03	$\mathbf{76.06 \pm 0.58}$	+75.95
	Acc.	40.64	36.49	20.11	17.29	28.03	$\mathbf{41.35 \pm 0.29}$	35.69 ± 0.08	42.66 ± 0.41	38.46	−2.18
Avg.	Rob.	0.11	1.17	8.54	9.87	4.14	8.63 ± 3.23	19.74 ± 0.05	33.05 ± 0.47	$\mathbf{80.94 \pm 0.02}$	+80.83
	Acc.	62.57	56.60	35.63	37.93	46.50	$\mathbf{62.96 \pm 0.13}$	58.65 ± 0.06	60.37 ± 0.36	60.20	−2.37

Table 2. Classification accuracy (%) on both adversarial images (Rob.) under a 10-step PGD attack with $\epsilon = 4/255$ and clean images (Acc.) across 13 datasets. We report the mean and standard deviation for test-time methods over 3 runs. The best robust and clean accuracies among adversarial defenses are indicated in bold, with the second best robust accuracies underlined for ease of comparison. The last column reports the gains of our method w.r.t. original CLIP without any finetuning or test-time operations.

Baselines. We compare ATAC with several test-time defenses for CLIP, as well as adversarial fine-tuning methods. Among test-time defenses, we use Test-time Transformation Ensembling (TTE) [37] with 9 augmentations (horizontal flip, 4 crops, and horizontal flip with 4 crops), Test-time Counterattacks (TTC) [49] with a PGD-style counterattack of 5 steps, $\epsilon_{ttc} = 4/255$, $\tau^*_{ttc} = 0.2$, and $\beta = 2$, as well as Robust Test-time Prompt Tuning (R-TPT) [41] with 64 augmentations, and a 1-step Adam optimizer with a learning rate of 0.005.

To compare ATAC with adversarial fine-tuning methods, we use the baselines of [49]: TeCoA [32], PMG-AFT [47], and FARE [39] that were adversarially fine-tuned using the Tiny ImageNet dataset [23] with an attack budget of $\epsilon = 4/255$, as well as the regular CLIP image encoder (CLIP-FT) fine-tuned on Tiny ImageNet without adversarial training.

5.2. ATAC for Adversarial Robustness

We evaluate all defense methods across 13 datasets against the PGD attack and present the results in Tab. 2. For most datasets, Tiny ImageNet-based adversarial fine-tuning yields minimal robustness gains, showing a significant limitation of this approach and the superiority of test-time defenses.

Among test-time defenses, *ATAC clearly stands out* in

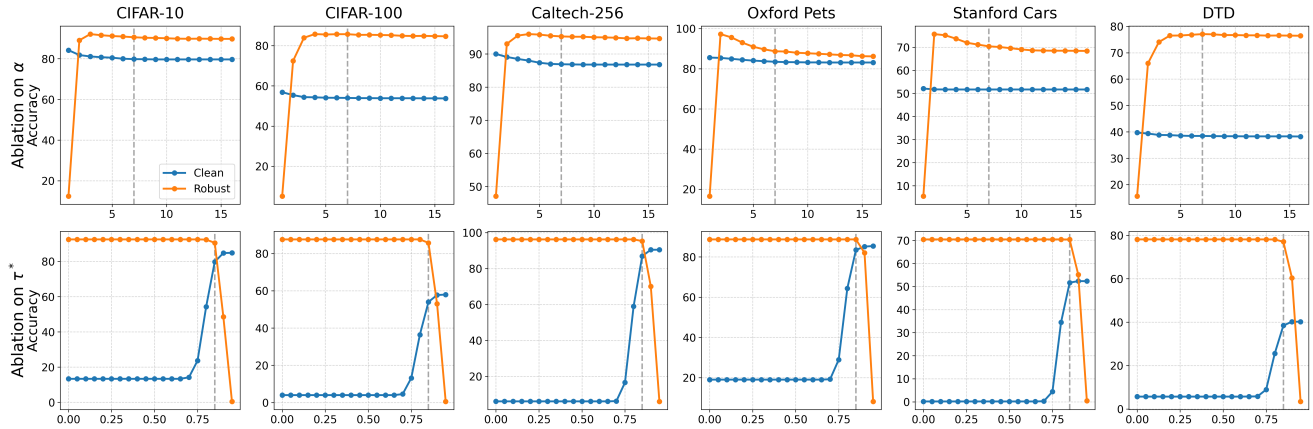

Figure 2. Ablations on the correction step size α (top row) and the cosine-consistency threshold τ^* (bottom row) on 6 datasets. Blue lines show accuracy on clean images and orange lines represent accuracy on adversarial images in each parameter setting. Vertical gray lines represent the parameter settings adopted in our experiments.

terms of robustness by consistently achieving the highest robust accuracy across all methods. ATAC significantly improves the robustness of undefended CLIP up to 90% in some cases, without any fine-tuning or costly test-time optimization. Furthermore, in all cases, ATAC substantially outperforms the robustness of previous state-of-the-art test-time defenses such as R-TPT and TTC by as much as 70%, and nearly 50% on average.

As expected, adversarial defense methods usually incur a penalty when it comes to their performance on clean samples. This phenomenon, known as the robustness-accuracy trade-off, is also present with ATAC: in order to gain robustness, our method suffers a minor loss in accuracy as a result of correcting some clean examples. For ATAC, the decrease in clean accuracy is in line with previous methods. However, the increase in robustness is outstanding, and the resulting robust accuracy sometimes eclipses clean performance. We further investigate this phenomenon in Sec. 6.

5.3. Ablation Study

In this section, we investigate the effect of the cosine-consistency threshold τ^* and the correction step size α separately, while keeping the other parameters of ATAC unchanged. We conduct ablations on six datasets and follow the experimental setup described in Sec. 5.1. Each parameter setting is evaluated on 2000 samples drawn from each dataset. We present our results in Fig. 2, along with further ablations on the augmentations used in ATAC in the Appendix.

The effect of α is minimal across all datasets, particularly when it is sufficiently large (≥ 3). This is likely a result of the normalization of the corrected visual embedding. This result further demonstrates that the effectiveness of ATAC is based primarily on the correct estimation of the semantic recovery direction and is insensitive to the value of α.

However, the value of τ^* is crucial. When τ^* is low (i.e., ≤ 0.7), clean samples are not protected from the correction mechanism, leading to extremely low clean accuracies. As τ^* increases, fewer clean samples are corrected unnecessarily, but the ratio of corrected adversarial samples also decreases, leading to an increase in clean accuracy and a decrease in robust accuracy. We found $\tau^* = 0.85$ to yield the best balance between robustness and accuracy across all datasets, which is further supported by our analysis of augmentations and τ distributions in the Appendix.

6. On the Robustness-Accuracy Trade-off

Our main results in Sec. 5 show an unusual phenomenon: the robust accuracy of ATAC often exceeds not only its clean accuracy, but also that of CLIP. This is generally considered impossible in the adversarial robustness literature. In this section, we explore a possible source of this phenomenon and how it relates to the effectiveness of our method.

We hypothesize that this phenomenon stems from a weakness of the attack objective, namely that the attack relies too much on the ground truth label. As gradient-based, untargeted attacks maximize the loss of the true label, they move the embedding of the input away from the correct decision region along the "path of least resistance". However, the embedding shift introduced by the attack carries hidden directional information pointing away from the true class.

Our method is able to exploit this hidden label-dependent directional information by estimating its reverse using augmentations. Continuing our hypothesis, if one were to reduce the amount of label-dependent information the attacker could use, estimating a semantic recovery direction would be more difficult, leading to a degradation in the performance of ATAC.

Dataset	PGD $\epsilon = 4/255$		PGD $\epsilon = 127.5/255$		Early-stopped PGD		Unsupervised PGD		Targeted PGD	
	CLIP	ATAC	CLIP	ATAC	CLIP	ATAC	CLIP	ATAC	CLIP	ATAC
CIFAR-10	0.00	91.80	0.00 (+0.00)	98.25 (+6.45)	0.00 (+0.00)	59.70 (-32.10)	17.48 (+17.48)	54.95 (-36.85)	0.05 (+0.05)	53.50 (-38.30)
CIFAR-100	0.00	86.27	0.00 (+0.00)	95.05 (+8.78)	0.00 (+0.00)	44.52 (-41.75)	5.55 (+5.55)	21.52 (-64.75)	0.07 (+0.07)	18.02 (-68.25)
STL-10	0.05	98.30	0.00 (-0.05)	98.72 (+0.42)	0.05 (+0.00)	92.25 (-6.05)	25.10 (+25.05)	84.88 (-13.42)	0.47 (+0.42)	88.20 (-10.10)
Flowers102	0.00	85.69	0.00 (+0.00)	92.15 (+6.46)	0.00 (+0.00)	59.62 (-26.07)	3.45 (+3.45)	30.07 (-55.62)	0.25 (+0.25)	24.70 (-60.99)
FGVAircraft	0.00	50.02	0.00 (+0.00)	68.08 (+18.73)	0.00 (+0.00)	18.96 (-30.39)	0.63 (+0.63)	10.56 (-38.79)	0.00 (+0.00)	8.28 (-41.74)
DTD	0.11	76.06	0.00 (-0.11)	80.16 (+3.09)	0.11 (+0.00)	49.73 (-27.34)	8.03 (+7.92)	23.09 (-53.98)	1.06 (+0.95)	19.95 (-56.11)
Avg.	0.03	81.36	0.00 (-0.03)	88.74 (+8.45)	0.03 (+0.00)	54.13 (-26.16)	10.04 (+10.01)	37.51 (-42.78)	0.21 (+0.18)	35.44 (-45.92)

Table 3. Robust accuracies of undefended CLIP and ATAC under specially designed attack scenarios. Values in parentheses indicate the change compared to the robust accuracies of the original methods against the untargeted $\epsilon = 4/255$ PGD attack, with values in red and blue representing increased and decreased robustness values, respectively.

6.1. Experimental Setup

In order to test whether ATAC *can* and *does* exploit the label-dependent directional information introduced by the attacks, we design four extreme settings.

Increasing Label-Dependent Information. We use PGD with $\epsilon = 127.5/255$ to maximize the amount of label-dependent information introduced by the attack. We design this setting to test the reverse of our hypothesis, i.e., introducing more label-dependent information leads to a more consistent estimation of the semantic recovery direction.

Early Stopping. We use PGD with early stopping, i.e., we stop the optimization of Eq. (3) at the earliest t where x^t is misclassified. This reduces the aforementioned shift in the embedding space, making the drift vectors more scattered, leading to inconsistent estimates of the recovery direction.

Unsupervised Attack. We use a PGD attack that aims to maximize the L_2-distance between the visual features of the original and the attacked images. Although this attack introduces large shifts and thereby more consistent drift vectors, the estimated recovery direction is not guaranteed to recover the original semantics due to the attack objective completely omitting label supervision.

Targeted Attack. We use a targeted PGD attack that aims to create a perturbation δ that minimizes $\mathcal{L}(x + \delta, y_t)$ for a target label $y_t \neq y_c$. Due to not having access to the true label, this attack cannot exploit the "path of least resistance" in the embedding space, creating smaller embedding shifts similar to the early-stopped attack.

6.2. Results

Tab. 3 shows the robustness of CLIP and ATAC in the four extreme settings across six datasets. We used 4000 samples from the test set of each dataset for all evaluations. The results clearly show that, as the true label-based supervision is limited, the attacks become harder to correct for our method, resulting in lower robust accuracies. In fact, the phenomenon that robust accuracy eclipses clean performance completely disappears. In contrast, when the attack budget ϵ is large, and therefore the influence of the true label is increased, the semantic recovery direction becomes easier to estimate, and ATAC achieves even higher robust accuracies. This confirms our hypothesis that unsupervised, gradient-based attacks introduce easy-to-estimate shifts in the embedding space by relying too much on the ground truth label. These results also show that ATAC can and indeed does exploit this hidden directional information.

On the other hand, even in scenarios where the attack has limited or no access to the true labels, ATAC still achieves robust accuracies that are comparable and in most cases superior to those of all competitive baselines shown in Tab. 2. This result shows that ATAC does not exclusively rely on the directional information injected by the attacks, further demonstrating the effectiveness of our method even in extreme scenarios.

7. Robustness Against Adaptive Attacks

A proper evaluation of adaptive or test-time adversarial defense methods, especially ones that include non-differentiable components, must take *adaptive attacks* into account [6, 46]. To this end, we design two adaptive attacks specifically tailored against ATAC. These attacks have full access to all components of our method, including the augmentations used, the gating threshold τ^*, and the correction step size α. Following the guiding principle of [46], our attacks adapt to all non-differentiable aspects of the defense.

Due to space limitations, we only provide a high-level intuition behind our attacks, and give a detailed overview along with their pseudocodes in the Appendix (**??**).

Lure Adaptive Attack. This attack jointly optimizes the adversarial perturbation so that (i) the latent drift vectors are aligned, thereby increasing τ and activating the correction

Dataset	Robust accuracy (%)						Running time (s)			
	PGD $\epsilon = 4/255$		Lure $\epsilon = 4/255$		Avoid $\epsilon = 4/255$		Lure $\epsilon = 4/255$		Avoid $\epsilon = 4/255$	
	TTC	ATAC	TTC	ATAC	TTC	ATAC	TTC	ATAC	TTC	ATAC
CIFAR-100	9.06	86.27	0.83 (-8.23)	5.83 (-80.44)	1.67 (-7.39)	11.25 (-75.02)	0.150	0.650	0.114	0.652
Caltech256	27.25	90.86	4.58 (-22.67)	15.83 (-75.03)	8.75 (-18.50)	47.50 (-43.36)	0.149	0.656	0.111	0.658
OxfordPets	24.64	87.46	0.00 (-24.64)	10.83 (-76.63)	1.67 (-22.97)	45.83 (-41.63)	0.149	0.651	0.114	0.655
StanfordCars	12.84	70.80	0.00 (-12.84)	4.17 (-66.63)	0.00 (-12.84)	18.33 (-52.47)	0.147	0.652	0.122	0.652
EuroSAT	13.57	66.57	0.83 (-12.74)	6.25 (-60.32)	1.67 (-11.90)	9.17 (-57.40)	0.138	0.649	0.114	0.650
DTD	11.40	76.06	0.42 (-10.98)	3.33 (-72.73)	0.83 (-10.57)	15.42 (-60.64)	0.136	0.647	0.114	0.650
Avg.	16.46	76.67	1.28 (-15.18)	7.21 (-69.46)	2.78 (-13.68)	24.25 (-52.42)	0.145	0.651	0.115	0.653

Table 4. Comparison of TTC and ATAC against their respective adaptive attacks on 6 datasets. We report robust accuracies of both defenses in %, as well as the running times of each attack in seconds per sample. Values in blue indicate the decrease in robust accuracy compared to the non-adaptive PGD baseline.

mechanism, and (ii) the adversarial loss is maximized over the whole pipeline. To achieve the latter, we use Expectation over Transformation (EOT) [3].

Avoid Adaptive Attack. Contrary to the Lure attack, this attack aims to avoid activating the correction mechanism by reducing τ. In addition, it also uses EOT to jointly maximize the loss over the original CLIP model.

Comparison to TTC. In order to compare our method with related baselines, we conduct experiments with adaptive attacks against TTC [49]. We evaluate the adaptive attack proposed in their paper that aims to reduce the L_2 distance between the visual embeddings of the image and its counterattacked variant. We also implement another adaptive attack for TTC that aims to jointly maximize the adversarial loss and the L_2 distance between the visual embeddings of the attacked and counterattacked images. For ease of comparison, we dub the former the Lure strategy and the latter the Avoid strategy against TTC.

7.1. Results

We evaluate both TTC and ATAC against their corresponding adaptive attacks, with a budget of $\epsilon = 4/255$ and present the results in Tab. 4. Although both defenses lose most of their robustness against these attacks, as expected when attacks are tailored against test-time defenses [6], the robust accuracy retained by ATAC is significantly higher. This shows that even in worst-case conditions, *ATAC still achieves nontrivial robustness*, further underscoring the effectiveness of our method. Furthermore, adaptive attacks require considerably more compute against ATAC, which is an additional benefit of our method in real-life worst-case scenarios.

Interestingly, the Avoid strategies perform worse against both defenses. This is only surprising for ATAC, where avoiding the correction mechanism would benefit a well-crafted adversarial attack. However, as demonstrated in Sec. 6, the objective of increasing the loss in an untargeted manner is at odds with creating less consistent drifts that would prevent correction, which explains why this attack is more difficult. On the other hand, we hypothesize that the Lure strategy succeeds because its adversarial perturbations are less mitigated by augmentations.

8. Conclusions and Future Work

In this paper, we introduce ATAC, a novel test-time adversarial defense method that efficiently estimates a semantic recovery direction in CLIP's feature space, using the visual features of augmented views.

Through a rigorous experimental analysis, we show that our method achieves state-of-the-art robustness on 13 classification benchmarks, beating the previous best methods by an average of nearly 50% in robust accuracy. We further demonstrate that unsupervised, gradient-based attacks overly rely on true-label supervision, inducing consistent shifts in CLIP's feature space, allowing ATAC to estimate a consistent recovery direction. However, even when this flaw is eliminated, ATAC still achieves state-of-the-art robustness. Furthermore, our method achieves nontrivial robustness against adaptive attacks at a comparatively large computational cost for the attacker, further underscoring the usability of ATAC in worst-case and real-life scenarios.

As the combination of adversarial defenses for CLIP is gaining traction, we hope future works can explore the combination of ATAC with other test-time defenses and adversarial fine-tuning methods. Moreover, future work could extend the novel paradigm pioneered by ATAC: correcting adversarial samples in feature space rather than image space.

9. Acknowledgements

The research was supported by the AI Competence Centre of the Cluster of Science and Mathematics of the Centre of Excellence for Interdisciplinary Research, Development and Innovation of the University of Szeged, and by project 2024-1.2.3-HU-RIZONT-2024-00017, implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the 2024-1.2.3-HU-RIZONT funding scheme.

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. *Advances in Neural Information Processing Systems*, 36:80396–80413, 2023. 2, 3

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pages 484–501. Springer, 2020. 1

[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018. 8

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 4

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 1, 2, 4

[6] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness, 2019. 7, 8

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2013. 4

[8] Adam Coates, A. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, 2011. 4

[9] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019. 1, 2

[10] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 4

[11] Ashim Dahal, Saydul Akbar Murad, and Nick Rahimi. Embedding shift dissection on clip: Effects of augmentations on vlm's representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4814–4818, 2025. 2, 3

[12] Jia Deng, Jin Li, Zhenhua Zhao, and Shaowei Wang. Fpt-noise: Dynamic scene-aware counterattack for test-time adversarial defense in vision-language models, 2025. 2

[13] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:594–611, 2006. 4

[14] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning, 2023. 2, 3

[15] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. 3

[16] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. 1, 2

[17] Gregory Griffin, Alex Holub, Pietro Perona, et al. Caltech-256 object category dataset. Technical report, Technical Report 7694, California Institute of Technology Pasadena, 2007. 4

[18] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations, 2018. 1

[19] Patrick Helber, Benjamin Bischke, Andreas R. Dengel, and Damian Borth. 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*, 12:2217–2226, 2017. 4

[20] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019. 1, 2, 3

[21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 4

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[23] Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24408–24419, 2024. 2

[25] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2018. 1

[26] Blerta Lindqvist. Delving into the pixels of adversarial samples, 2021. 3

[27] Jiaxiang Liu, Jiawei Du, Xiao Liu, Prayag Tiwari, and Mingkun Xu. Self-calibrated consistency can fight back for adversarial robustness in vision-language models. *arXiv preprint arXiv:2510.22785*, 2025. 2

[28] Xingjun Ma, Yifeng Gao, Yixu Wang, Ruofan Wang, Xin Wang, Ye Sun, Yifan Ding, Hengyuan Xu, Yunhao Chen, Yunhao Zhao, et al. Safety at scale: A comprehensive survey of large model and agent safety. *Foundations and Trends® in Privacy and Security*, 8(3-4):254–469, 2025. 1, 2

[29] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1

[30] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. 3

[31] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *ArXiv*, abs/1306.5151, 2013. 4

[32] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *The Eleventh International Conference on Learning Representations*. 1, 2, 3, 5

[33] Chengzhi Mao, Mia Chiquier, Hao Wang, Junfeng Yang, and Carl Vondrick. Adversarial attacks are reversible with natural supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 661–671, 2021. 3

[34] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, pages 16805–16827. PMLR, 2022. 1

[35] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. 4

[36] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 4

[37] Juan C Pérez, Motasem Alfarra, Guillaume Jeanneret, Laura Rueda, Ali Thabet, Bernard Ghanem, and Pablo Arbeláez. Enhancing adversarial robustness via test-time transformation ensembling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 81–91, 2021. 2, 3, 5

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 1, 4

[39] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *ICML*, 2024. 1, 2, 5

[40] Nur Muhammad (Mahi) Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. 1

[41] Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29958–29967, 2025. 1, 2, 3, 4, 5

[42] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Proceedings of the 5th Conference on Robot Learning*, pages 894–906. PMLR, 2022. 1

[43] 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. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 1, 2, 3

[44] Sumedh Anand Sontakke, Jesse Zhang, Séb Arnold, Karl Pertsch, Erdem Biyik, Dorsa Sadigh, Chelsea Finn, and Laurent Itti. RoboCLIP: One demonstration is enough to learn robot policies. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[45] Baoshun Tong, Kaiyu Song, and Hanjiang Lai. Test-time alignment-enhanced adapter for vision-language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2

[46] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33:1633–1645, 2020. 7

[47] Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24502–24511, 2024. 2, 5

[48] Xin Wang, Kai Chen, Jiaming Zhang, Jingjing Chen, and Xingjun Ma. Tapt: Test-time adversarial prompt tuning for robust inference in vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19910–19920, 2025. 2

[49] Songlong Xing, Zhengyu Zhao, and Nicu Sebe. Clip is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of clip. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15172–15182, 2025. 1, 2, 3, 4, 5, 8

[50] Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L Yuille. Adversarial attacks beyond the image space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4302–4311, 2019. 3

[51] Mingkun Zhang, Keping Bi, Wei Chen, Jiafeng Guo, and Xueqi Cheng. Clipure: Purification in latent space via clip for adversarially robust zero-shot classification. In *The Thirteenth International Conference on Learning Representations*. 3

[52] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337 – 2348, 2021. 2

[53] Wanqi Zhou, Shuanghao Bai, Qibin Zhao, and Badong Chen.

Revisiting the adversarial robustness of vision language models: a multimodal perspective. *CoRR*, 2024. 1, 2