

Diversifying Counterattacks: Orthogonal Exploration for Robust CLIP Inference

Chengze Jiang¹, Minjing Dong², Xinli Shi¹, Jie Gui^{*1,3,4}

¹ Southeast University, Nanjing, China

² Department of Computer Science, City University of Hong Kong, Hong Kong

³ Engineering Research Center of Blockchain Application, Supervision And Management (Southeast University), Ministry of Education, China

⁴ Purple Mountain Laboratories, China

czjiang@seu.edu.cn, minjdong@cityu.edu.hk, xinli_shi@seu.edu.cn, guijie@seu.edu.cn

Abstract

Vision-language pre-training models (VLPs) demonstrate strong multimodal understanding and zero-shot generalization, yet remain vulnerable to adversarial examples, raising concerns about their reliability. Recent work, Test-Time Counterattack (TTC), improves robustness by generating perturbations that maximize the embedding deviation of adversarial inputs using PGD, pushing them away from their adversarial representations. However, due to the fundamental difference in optimization objectives between adversarial attacks and counterattacks, generating counterattacks solely based on gradients with respect to the adversarial input confines the search to a narrow space. As a result, the counterattacks could overfit limited adversarial patterns and lack the diversity to fully neutralize a broad range of perturbations. In this work, we argue that enhancing the diversity and coverage of counterattacks is crucial to improving adversarial robustness in test-time defense. Accordingly, we propose Directional Orthogonal Counterattack (DOC), which augments counterattack optimization by incorporating orthogonal gradient directions and momentum-based updates. This design expands the exploration of the counterattack space and increases the diversity of perturbations, which facilitates the discovery of more generalizable counterattacks and ultimately improves the ability to neutralize adversarial perturbations. Meanwhile, we present a directional sensitivity score based on averaged cosine similarity to boost DOC by improving example discrimination and adaptively modulating the counterattack strength. Extensive experiments on 16 datasets demonstrate that DOC improves adversarial robustness under various attacks while maintaining competitive clean accuracy.

Code — <https://github.com/bookman233/DOC>

Arxiv version — <https://arxiv.org/abs/2511.09064>

Introduction

Vision-language pre-training models (VLPs) have emerged as powerful multimodal systems, demonstrating strong zero-shot generalization (Zhang et al. 2024b; Yang et al. 2025; Laurençon et al. 2024). Among them, CLIP is a representative VLP that aligns visual and textual representations

through contrastive learning and achieves impressive performance in vision tasks (Radford et al. 2021; Jiao et al. 2023). While recent research primarily focuses on improving the performance of CLIP models (Zhou et al. 2023), their adversarial robustness receives comparatively less attention (Dong et al. 2023). Recent studies reveal that CLIP is vulnerable to adversarial examples, *i.e.*, human-imperceptible perturbations that can mislead predictions of the model (Yu, Zhang, and Xu 2024; Zhang, Zhou, and Li 2024; Yang, Jeong, and Yoon 2024). This vulnerability raises concerns about the reliability of CLIP (Li et al. 2024b; Zhang et al. 2025; Ge et al. 2023). Since an increasing number of CLIP models are deployed in security-related downstream tasks, enhancing their adversarial robustness has become an urgent research priority (Wortsman et al. 2022).

One representative solution is adversarial fine-tuning, which improves adversarial robustness by fine-tuning the pretrained CLIP model using adversarial examples (Mao et al. 2022; Schlarmann et al. 2024). Another approach is adversarial prompt tuning, which introduces learnable text tokens into the embedding space and uses a small validation set to better align prompt embeddings with those of adversarial images (Li et al. 2024a; Sheng et al. 2025). Although these methods improve the adversarial robustness of CLIP, they still present notable limitations. First, adversarial fine-tuning introduces significant computational overhead, which grows with the size of the dataset (Alfarra et al. 2022; Zhang et al. 2024d). In contrast, prompt tuning requires only a few labeled examples to adjust the prompt, thereby reducing the computational cost (Wang et al. 2025). However, it operates in the learned embedding space rather than the human-interpretable textual domain, causing the learned prompts to lose semantic interpretability (Raman et al. 2023). Most importantly, although CLIP benefits from large-scale pretraining that gives it impressive generalization ability (Radford et al. 2021; Hu et al. 2022), fine-tuning its model weights can diminish this generalization (Wang et al. 2024b). Recently, Test-Time Counterattack (TTC) is presented as a parameter-free and data-agnostic defense that leverages the expressive power of CLIP to improve adversarial robustness (Xing, Zhao, and Sebe 2025). TTC fixes the adversarial input as an anchor and optimizes a counterattack using PGD (Madry et al. 2018) to maximize the l_2 distance between the adversarial input and its counterattacked variants, thereby pushing

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

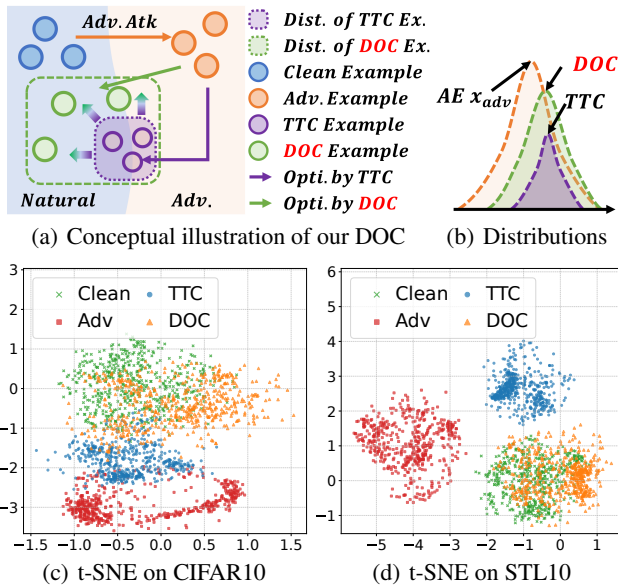


Figure 1: (a)-(b) Conceptual illustration of our methodology. We propose to generate more diverse counterattacks to neutralize adversarial perturbations. (c)-(d) t-SNE of example embeddings obtained by TTC and our DOC.

adversarial input away from the adversarial neighborhood.

While TTC presents promising progress, there exists a fundamental mismatch between the optimization objectives of adversarial attack and counterattack. Specifically, adversarial attacks aim at maximizing the loss (defined in equation (2)), while counterattacks aim at maximizing the distance between adversarial and counterattack examples (defined in equation (3)). This mismatch could even be further amplified regarding the optimization strategy in TTC since it uses PGD to generate counterattacks and could overfit to the surrogate objective easily, which can hardly approximate the accurate adversarial perturbation distribution. Ultimately, this mismatch hinders the counterattack from effectively neutralizing the underlying adversarial perturbations. Thus, in the absence of label supervision at test time, refining the optimization strategy of counterattacks becomes crucial to alleviate overfitting induced by the mismatch of inherent optimization objectives. A natural and direct approach is to augment the optimization process to increase counterattack diversity, enabling broader exploration of the adversarial perturbation space and enhancing the ability to neutralize a wide range of potential threats (as shown in Fig. 1(a) and (b)). Therefore, improving counterattack diversity to more effectively defend against adversarial threats of CLIP remains an open and valuable research challenge.

Consequently, we introduce Directional Orthogonal Counterattack (DOC), which augments each optimization step of counterattack with a randomized component orthogonal to the primary gradient direction and incorporates a momentum-based update. This design expands the counterattack search space to increase distribution diversity, allowing the counterattack to escape narrow local optima and

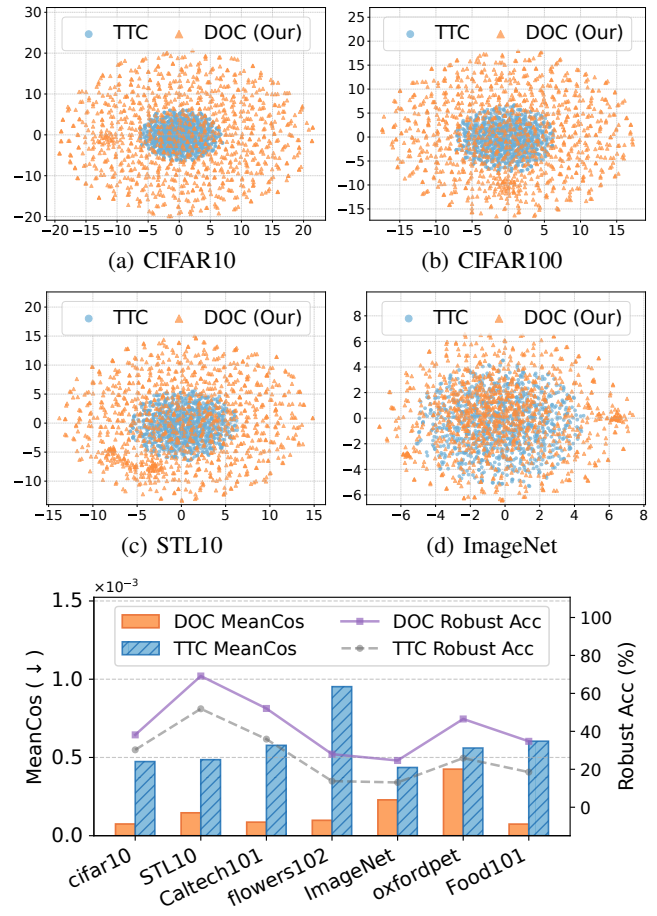


Figure 2: (a)-(c) t-SNE visualizations of counterattacks generated by TTC and our DOC. (Bottom) Comparison of mean cosine similarity of counterattack and robust accuracy under PGD-10 with $\epsilon_{\text{atk}} = 4/255$. More details on 15 datasets are presented in **Supplementary Materials**.

more effectively neutralize adversarial effects in an unsupervised setting (as shown in Fig. 1). As further illustrated in Fig. 2, t-SNE visualizations and mean cosine similarity (MeanCos, where lower values indicate higher diversity (Schwinn et al. 2022; Zhu et al. 2023)) show that DOC generates more diverse counterattacks compared to TTC, resulting in improved adversarial robustness of CLIP. Furthermore, DOC introduces a directional sensitivity score, defined as the cosine similarity between the original image embedding and its randomly perturbed versions, which guides the adaptive modulation of counterattack strength. Comprehensive evaluations on 16 datasets confirm that the components of DOC jointly improve the test-time robustness of CLIP models while preserving competitive clean accuracy. The main contributions are summarized as follows:

- We propose DOC to more effectively neutralize adversarial perturbations by expanding the counterattack search space and increasing diversity through the incorporation of orthogonal components and momentum.
- We introduce the directional sensitivity score via cosine

similarity, which determines the necessity of a counterattack and enables fine-grained control over its strength.

- Experiments on 16 datasets show that DOC outperforms state-of-the-art test-time defenses in adversarial robustness while maintaining competitive clean accuracy.

Related Works

Adversarial Robustness

Deep neural networks are vulnerable to adversarial attacks (Cui et al. 2024; Jiang et al. 2025; Xia et al. 2024). To mitigate this vulnerability, adversarial training is recognized as one of the most effective defenses (Tong et al. 2024; Xhonneux et al. 2024; Kuang et al. 2024). However, it imposes significant computational costs and often struggles with overfitting (Wang et al. 2024c; Jia et al. 2024). In parallel, test-time defenses have attracted increasing attention because they do not require modifying model parameters (Croce et al. 2022), including adversarial purification (Nie et al. 2022) and loss-based adjustment (Wu et al. 2021; Alfarra et al. 2022). Despite their progress, existing test-time defenses remain susceptible to attacks designed to circumvent their mechanisms. For example, Hedge Defense (HD) optimizes test-time perturbations by maximizing the loss across all classes (Wu et al. 2021). While promising, HD relies on classification-oriented objectives and assumes access to supervised information or adversarially trained backbones. Although adversarial defense methods have made progress, most existing approaches focus on unimodal supervised settings and face challenges when generalizing to modern vision-language models, which rely on multimodal embedding architectures and do not depend on supervised information for inference.

Adversarial Robustness of VLPs

VLPs demonstrate strong zero-shot generalization capabilities (Zhang et al. 2024a; Yang et al. 2024) but remain vulnerable to adversarial attacks (Tu, Deng, and Gedeon 2023; Zhang et al. 2025). Therefore, various defense strategies are presented to improve the robustness of VLPs. Among them, adversarial fine-tuning trains the model with adversarial examples to strengthen robustness (Mao et al. 2022; Gong et al. 2025). TeCoA demonstrates transferability across tasks (Mao et al. 2022), and PMG-AFT adds CLIP-guided regularization to relieve overfitting (Wang et al. 2024c). Another approach is adversarial prompt tuning (Zhang et al. 2024c), which adjusts input prompts and learns optimized prompt tokens to better align text and image features under adversarial conditions (Wang et al. 2025; Sheng et al. 2025). Despite these advances, existing methods require supervised training, access to downstream tasks, or rely on prompt engineering, which risks undermining the generalization of models or introducing additional training processes (Mou, Zhang, and Ye 2024). To address this limitation, recent work by Liu *et al.* introduces TTC, which neutralizes adversarial perturbations by counterattack, achieving defense without changing model parameters or using prompt engineering (Xing, Zhao, and Sebe 2025). However, a challenge is that the distributional shift between adversarial and clean examples

makes using the adversarial embedding as an anchor risk overfitting to the local adversarial structure. Motivated by this, we aim to enhance counterattack diversity to broaden the search space and improve the neutralization of adversarial noise, thereby boosting CLIP’s adversarial robustness.

Methodology

Background and Preliminaries

Background CLIP is a representative foundation VLP that achieves impressive zero-shot performance through large-scale pretraining on paired image-text data (Cao et al. 2024), which comprises an image encoder $I_\theta : \mathcal{X} \rightarrow \mathbb{R}^d$ and a text encoder $T_\phi : \mathcal{T} \rightarrow \mathbb{R}^d$, parameterized by θ and ϕ , respectively (Gao et al. 2024a). For inference, given an input image $\mathbf{x} \in \mathcal{X}$ and a textual prompt $\mathbf{t}_i \in \mathcal{T}$ representing the i -th class, CLIP computes their cosine similarity as follows:

$$s(\mathbf{x}, \mathbf{t}_i) = \frac{\langle I_\theta(\mathbf{x}), T_\phi(\mathbf{t}_i) \rangle}{\|I_\theta(\mathbf{x})\| \cdot \|T_\phi(\mathbf{t}_i)\|}, \quad (1)$$

where \mathbf{t}_i denotes the textual prompt for the i -th class (Radford et al. 2021). The similarity across all candidate classes is normalized to yield the predicted class distribution as $P(y = i | \mathbf{x}) = \exp(s) / \sum_j \exp(s)$. The predicted label is determined as the class with the highest probability.

Adversarial Vulnerability of VLPs To evaluate the adversarial robustness of VLPs, an adversary obtains adversarial perturbation δ_{adv} , bounded by an ℓ_p norm, such that the adversarial example $\mathbf{x}_{\text{adv}} = \mathbf{x} + \delta_{\text{adv}}$ leads to incorrect predictions (Gao et al. 2024b; Guo et al. 2024). The objective of an adversarial attack is typically formulated as the following constrained maximization problem (Zhao et al. 2023):

$$\delta_{\text{atk}} = \arg \max_{\delta} \mathcal{L}(\mathbf{x} + \delta, y), \quad \text{s.t.} \quad \|\delta_{\text{atk}}\|_p \leq \epsilon_{\text{atk}}, \quad (2)$$

where y denotes the label, \mathcal{L} is the loss function, and ϵ_{atk} is the adversarial perturbation budget (Wang et al. 2024a). By optimizing the objective (2), various adversarial attacks can generate perturbations δ_{adv} that are injected into the original input to create adversarial examples that mislead the VLPs.

Test-Time Counterattacks for CLIP Recently, TTC is presented as a learning-free defense that operates during inference, which generates a counterattack perturbation δ_{ca} that neutralizes potential adversarial perturbations in the input (Xing, Zhao, and Sebe 2025). Formally, TTC maximizes the embedding distance between the adversarial example \mathbf{x}_{adv} and the counterattack example $\mathbf{x}_{\text{ca}} = \mathbf{x}_{\text{adv}} + \delta_{\text{ca}}$ as

$$\delta_{\text{ca}} = \arg \max_{\|\delta_{\text{ca}}\|_p \leq \epsilon_{\text{ca}}} \|I_\theta(\mathbf{x}_{\text{adv}} + \delta_{\text{ca}}) - I_\theta(\mathbf{x}_{\text{adv}})\|, \quad (3)$$

where ϵ_{ca} denotes the budget of counterattack perturbation. To approximate the maximization problem (3), TTC adopts PGD to update counterattack perturbation δ_{ca} as follows:

$$\delta_{\text{ca}}^{t+1} = \Pi \left[\delta_{\text{ca}}^t + \alpha \cdot \text{sign} \left(\nabla_{\mathbf{x}_{\text{adv}}} \mathcal{L}(\mathbf{x}_{\text{adv}}, \delta_{\text{ca}}^t) \right) \right], \quad (4)$$

where $\mathcal{L} = \|I_\theta(\mathbf{x}_{\text{adv}} + \delta_{\text{ca}}) - I_\theta(\mathbf{x}_{\text{adv}})\|$, $\Pi(\cdot)$ denotes the projection operation, and α signifies the step size.

Directional Orthogonal Counterattack

Orthogonal Gradient Augmentation Crafting counterattacks using PGD presents a fundamental challenge due to the intrinsic differences between adversarial attacks and counterattacks. Specifically, while adversarial attacks maximize loss with respect to class labels as in equation (2), counterattacks operate without label supervision and aim to push the adversarial input away from its corrupted embedding as in equation (3). On this basis, using PGD (4), which relies on gradients with respect to the adversarial input to generate counterattacks, restricts the optimization to a narrow region as defined in equation (3), and fails to explore the adversarial space that truly requires neutralization, as described in equation (2). Furthermore, since ground-truth labels are unavailable at test time, addressing the mismatch in optimization objectives hinges critically on improving the counterattack strategy. Consequently, we propose enhancing the diversity of counterattacks to discover more generalizable solutions by exploring a broader region of adversarial space, which mitigates overfitting and better counteracts the underlying adversarial perturbation distribution.

Therefore, we introduce randomized exploration along directions orthogonal to the primary gradient, coupled with the momentum-based update strategy. This design expands the counterattack search space, enabling it to escape narrow local optima and explore regions beyond the reach of standard PGD, thereby more effectively approximating and neutralizing a broader range of adversarial perturbations. As shown in Fig. 1 (c)-(d), DOC generates more dispersed and generalized counterattacks, guiding adversarial examples closer to the distribution of clean examples and enhancing robustness. Specifically, we first compute the normalized gradient:

$$\mathbf{g} = \frac{\nabla_{\mathbf{x}_{\text{adv}}} \mathcal{L}(I_{\theta}(\mathbf{x}_{\text{adv}} + \delta_{\text{ca}}^t), I_{\theta}(\mathbf{x}_{\text{adv}}))}{\|\nabla_{\mathbf{x}_{\text{adv}}} \mathcal{L}(I_{\theta}(\mathbf{x}_{\text{adv}} + \delta_{\text{ca}}^t), I_{\theta}(\mathbf{x}_{\text{adv}}))\|}. \quad (5)$$

Rather than updating solely along the gradient direction, we introduce an orthogonal component to expand the search region for counterattacks. Given the gradient (5) and a vector $\mathbf{r} \sim \mathcal{N}(0, 1)$, we compute the orthogonal component as

$$\mathbf{r}^{\perp} = \frac{\mathbf{r} - \langle \mathbf{r}, \mathbf{g} \rangle \mathbf{g}}{\|\mathbf{r} - \langle \mathbf{r}, \mathbf{g} \rangle \mathbf{g}\|}, \quad (6)$$

where orthogonal projection ensures $\langle \mathbf{r}^{\perp}, \mathbf{g} \rangle = 0$. We then form the composite update direction \mathbf{d} by combining the gradient direction and the orthogonal component as

$$\mathbf{d} = \mathbf{g} + \lambda \cdot \mathbf{r}^{\perp}, \quad (7)$$

where λ controls the strength of the orthogonal injection. To further alleviate the overfitting of counterattack perturbations and enhance their generalization, we adopt a momentum-based update scheme as follows:

$$\mathbf{m}_t = \mu \cdot \mathbf{m}_{t-1} + (1 - \mu) \cdot \mathbf{d}, \quad (8)$$

where $\mu \in [0, 1)$ is the momentum factor. Finally, the iterative role of our counterattack perturbation is presented as

$$\delta_{\text{ca}}^{t+1} = \Pi(\delta_{\text{ca}}^t + \alpha \cdot \text{sign}(\mathbf{m}_t)). \quad (9)$$

Compared to standard PGD, our method expands the counterattack search space and enhances perturbation diversity, enabling better generalization to a wider range of potential adversarial perturbations and thereby improving robustness.

Algorithm 1: Implementation of DOC

Input: CLIP model I_{θ} ; Input example \mathbf{x} ; Counterattack perturbation budget ϵ_{ca} ; Sample time M ; Step size α ; Counterattack steps T ; Hyperparameters λ , τ , and γ .
Output: Counterattack perturbation δ_{ca} .
 /* Directional Sensitivity Score */
 1: **for** $m = 1$ to M **do**
 2: $\boldsymbol{\eta}^m \leftarrow \mathcal{U}(-\epsilon_{\text{ca}}, \epsilon_{\text{ca}})$.
 3: $\mathbf{x}_{\text{input}}^m = \mathbf{x}_{\text{input}} + \boldsymbol{\eta}^m$.
 4: $\tau_{\text{cos}} \leftarrow \tau_{\text{cos}} + \cos(I_{\theta}(\mathbf{x}_{\text{input}}^m), I_{\theta}(\mathbf{x}_{\text{input}}))$.
 5: **end for**
 6: $\hat{\tau}(\mathbf{x}_{\text{input}}) \leftarrow 1 - \tau_{\text{cos}}/M$ as Eq. (10).
 7: $w \leftarrow \text{Eq. (11)}$.
 /* Orthogonal Gradient Aug */
 8: Initialize $\mathbf{m}_0 \leftarrow 0$, $\delta_{\text{ca}}^0 \sim \mathcal{U}(-\epsilon_{\text{ca}}, \epsilon_{\text{ca}})$.
 9: **for** $t = 1$ to T **do**
 10: Normalized gradient $\mathbf{g} \leftarrow \text{Eq. (5)}$.
 11: $\mathbf{r} \sim \mathcal{N}(0, 1)$.
 12: $\mathbf{r}^{\perp} \leftarrow \text{Eq. (6)}$.
 13: $\mathbf{d} \leftarrow \mathbf{g} + \lambda \cdot \mathbf{r}^{\perp}$ as Eq. (7).
 14: $\mathbf{m}_t \leftarrow \mu \cdot \mathbf{m}_{t-1} + (1 - \mu) \cdot \mathbf{d}$ as Eq. (8).
 15: $\delta_{\text{ca}}^t \leftarrow \Pi(\delta_{\text{ca}}^{t-1} + \alpha \cdot \text{sign}(\mathbf{m}_t))$ as Eq. (9).
 16: **end for**
 17: $\delta_{\text{ca}} \leftarrow w \cdot \delta_{\text{ca}} + (1 - w) \cdot \delta_{\text{ca}}^0$.

Counterattack with Directional Sensitivity Score Counterattacks require identifying whether an input is a clean or an adversarial example to determine the need for countermeasures. Prior work addresses this by leveraging pseudo-stability, based on the observation that adversarial examples tend to exhibit larger embedding shifts under random perturbations (Wu et al. 2021; Xing, Zhao, and Sebe 2025). This is measured by the ℓ_2 distance between the input example and its noisy counterpart, but it raises two concerns. First, two embeddings may have similar directions but differ in scale, which can inflate the ℓ_2 distance despite semantic similarity. Second, relying on a single noisy sample introduces randomness, making the decision process unstable.

Correspondingly, we adopt cosine similarity to measure pseudo-stability, focusing on directional alignment and being invariant to scaling. Furthermore, we average the similarity over multiple random perturbations to mitigate stochastic effects and improve decision robustness. Specifically, for the input example $\mathbf{x}_{\text{input}}$ with unknown status as clean or adversarial, we generate M noisy versions $\mathbf{x}_{\text{input}}^m = \mathbf{x}_{\text{input}} + \boldsymbol{\eta}^m$, where $\boldsymbol{\eta}^m \sim [\epsilon_{\text{ca}} \cdot \text{sign}(\mathcal{N}(0, 1))]$ as follows:

$$\hat{\tau}(\mathbf{x}_{\text{input}}) = 1 - \frac{1}{M} \sum_{m=1}^M \cos(I_{\theta}(\mathbf{x}_{\text{input}}^m), I_{\theta}(\mathbf{x}_{\text{input}})), \quad (10)$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity. A lower $\hat{\tau}(\mathbf{x})$ indicates that perturbed embeddings remain directionally aligned, suggesting the input is clean. Conversely, a higher score reflects directional inconsistency, indicating a potential adversarial example. To improve sample discriminability, we apply a soft gating function instead of a hard threshold, which avoids abrupt binary decisions and mitigates sen-

Dataset	Acc	CLIP	Adversarial Fine-Tuning						Test-Time Defence				Δ_o	Δ_{\uparrow}
			TeCoA ¹	TeCoA ⁴	PMG ¹	PMG ⁴	FARE ¹	FARE ⁴	Anti	HD	TTC	DOC		
CIFAR10	Robust	0.00	7.72	11.83	10.16	15.79	2.02	5.47	0.32	1.82	30.25	38.14	38.14	7.89
	Clean	85.08	64.64	65.15	70.68	71.45	74.46	78.46	83.44	78.23	81.32	81.25	-3.83	-2.19
CIFAR100	Robust	0.00	6.39	9.39	7.71	11.12	2.87	4.59	0.22	0.96	9.46	15.46	15.46	6.00
	Clean	57.16	35.94	36.30	40.32	41.51	46.67	47.38	53.96	52.86	56.11	55.96	-1.20	2.00
STL10	Robust	0.04	24.10	31.91	28.49	35.77	10.05	17.72	2.25	3.80	51.89	69.16	69.12	17.27
	Clean	96.41	87.40	81.69	88.56	84.35	91.76	89.11	95.47	89.50	96.03	95.83	-0.58	0.36
ImageNet	Robust	0.00	1.65	3.07	2.07	3.71	0.16	0.83	0.15	0.04	13.07	24.64	24.64	11.57
	Clean	59.72	34.89	27.76	36.12	28.51	48.79	40.48	54.29	54.54	32.36	41.91	-17.81	-12.63
Caltech101	Robust	0.60	15.70	21.41	19.50	26.01	5.14	10.29	3.14	1.62	35.90	52.05	51.45	16.15
	Clean	85.69	71.64	64.41	75.43	69.06	80.95	76.58	83.99	82.33	85.99	86.54	0.85	0.55
Caltech256	Robust	0.13	8.26	12.14	10.57	13.88	2.17	5.39	1.44	0.55	26.38	43.08	42.95	16.70
	Clean	81.72	61.11	52.05	62.20	53.32	73.28	67.22	79.40	79.12	75.96	79.24	-2.48	-0.16
OxfordPets	Robust	0.00	0.95	3.96	1.77	5.19	0.22	0.32	0.10	0.00	25.89	46.52	46.52	20.63
	Clean	87.35	62.06	53.94	65.85	56.66	79.37	70.10	80.53	80.91	60.70	74.05	-13.30	-6.86
Flowers102	Robust	0.00	1.84	3.88	2.55	4.95	0.03	0.62	0.05	0.00	13.77	27.99	27.99	14.22
	Clean	65.43	36.71	27.78	36.97	28.88	48.04	41.01	62.80	58.22	63.23	64.48	-0.95	1.25
FGVCAircraft	Robust	0.00	0.03	0.15	0.03	0.09	0.00	0.04	0.00	0.00	7.77	11.19	11.19	3.42
	Clean	20.07	5.43	3.51	5.43	3.24	10.80	7.77	15.64	16.36	15.96	18.15	-1.92	1.79
StanfordCars	Robust	0.00	0.15	0.47	0.15	0.61	0.01	0.04	0.00	0.00	12.66	24.57	24.57	11.91
	Clean	52.07	20.91	15.18	25.36	16.79	38.68	32.09	36.14	44.28	41.54	48.51	-3.56	4.23
SUN397	Robust	0.00	1.30	2.31	1.90	3.37	0.13	0.65	0.11	0.00	13.43	16.71	16.71	3.28
	Clean	58.50	36.69	28.16	37.98	29.93	52.42	43.57	55.99	53.17	46.68	47.15	-11.35	-8.84
Country211	Robust	0.00	0.05	0.22	0.12	0.34	0.00	0.03	0.00	0.00	2.72	4.98	4.98	2.26
	Clean	15.22	4.75	3.66	4.64	3.34	9.25	6.58	11.60	11.72	12.07	13.46	-1.76	1.39
Food101	Robust	0.00	0.56	1.43	1.03	2.19	0.06	0.34	0.07	0.64	18.52	34.74	34.74	16.22
	Clean	83.86	30.00	21.90	36.62	27.97	55.24	41.98	75.95	80.30	79.86	82.46	-1.40	2.16
EuroSAT	Robust	0.00	9.81	10.82	9.62	10.52	0.00	7.58	0.03	0.49	14.24	14.49	14.49	0.25
	Clean	42.57	16.36	17.53	18.14	19.19	21.10	18.22	36.81	39.08	53.09	52.92	10.35	-0.17
DTD	Robust	0.11	4.20	5.19	4.31	5.30	0.90	2.89	0.37	0.16	11.91	19.68	19.57	7.77
	Clean	40.43	25.16	20.11	21.76	17.29	31.97	28.03	38.55	34.89	36.12	36.44	-3.99	-2.11
PCAM	Robust	0.00	20.95	44.13	12.87	36.38	0.64	3.74	0.25	12.04	51.61	52.95	52.95	1.34
	Clean	52.95	49.96	49.98	12.87	49.80	52.53	50.17	52.61	50.38	53.11	53.84	-0.89	0.73
Average	Robust	0.06	6.48	10.15	7.05	10.95	1.53	3.78	0.53	1.38	21.22	31.02	30.96	9.80
	Clean	61.51	40.23	35.57	39.93	37.58	50.96	46.23	57.32	56.62	55.63	58.26	-3.25	0.94

Table 1: Clean and robust accuracy under PGD-10 with $\epsilon_{\text{atk}} = 4/255$ on 16 datasets. Adversarial fine-tuning methods are trained on Tiny ImageNet, with superscripts indicating the attack budget used during fine-tuning. Δ_o indicates the improvement over the original CLIP, and Δ_{\uparrow} denotes the gain over the previous best. Bold indicates the best performance.

sitivity to threshold hyperparameters as follows:

$$w = \sigma(\gamma \cdot (\tau - \hat{\tau}(x))) \in (0, 1), \quad (11)$$

where τ denotes the predefined threshold, γ controls the sharpness, and $\sigma(\cdot)$ is the sigmoid function. Therefore, the final counterattack perturbation δ_{ca} is generated as $\delta_{\text{ca}} = w \cdot \delta_{\text{ca}} + (1 - w) \cdot \delta_{\text{ca}}^0$ with noise $\delta_{\text{ca}}^0 \sim \mathcal{U}(-\epsilon_{\text{ca}}, \epsilon_{\text{ca}})$.

Compared to the ℓ_2 norm, our directional sensitivity score based on cosine similarity provides more reliable indicators of adversarial perturbations, as it is less affected by irrelevant scaling in high-dimensional feature spaces. Meanwhile, rather than applying hard binarization, we employ an adaptive mechanism to modulate counterattack strength, enabling finer discrimination between inputs and more flexible responses. Additionally, averaging over multiple random perturbations mitigates the instability of single-sample estimates and improves the stability of counterattack decisions.

Experiments and Analysis

Experiment Settings

Datasets for Evaluation We conduct systematic experiments and analyses across 16 datasets. For general object classification, we include CIFAR-10/100 (Krizhevsky, Hinton et al. 2009), STL-10 (Coates, Ng, and Lee 2011), ImageNet (Deng et al. 2009), Caltech-101 (Fei-Fei, Fergus, and Perona 2006), and Caltech-256 (Griffin et al. 2007). For fine-grained classification, we consider Oxford Pets (Parkhi et al. 2012), Flowers-102 (Nilsback and Zisserman 2008), Food-101 (Bossard, Guillaumin, and Van Gool 2014), and Stanford Cars (Krause et al. 2013). For scene recognition, we use SUN397 (Xiao et al. 2010) and Country211 (Radford et al. 2021). In addition, we incorporate domain-specific datasets, including FGVC Aircraft (Maji et al. 2013), EuroSAT (Helber et al. 2019), DTD (Cimpoi et al. 2014), and PatchCame-

Method	CIFAR10	CIFAR100	STL10	ImageNet	Caltech101	Caltech256	OxfordPets	Flower102	FGVCAircraft	StanfordCars	SUN397	Country211	Food101	EuroSAT	DTD	PCAM	Avg. Rob.	Avg. Acc.	
CLIP	0.00	0.00	0.03	0.00	0.07	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.11	0.94	0.07	61.51
HD	1.68	0.00	1.71	0.01	0.23	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.11	0.07	5.04	0.56	54.85	
TTC	30.15	8.64	53.08	11.98	34.83	25.15	24.45	12.85	6.66	11.38	12.74	2.21	16.46	14.66	12.39	52.07	20.61	55.63	
DOC	35.68	12.10	66.42	20.91	48.07	39.16	41.89	25.11	10.11	20.20	14.08	3.66	29.26	14.41	17.13	52.73	28.18	58.34	
Δ_{CLIP}	35.68	12.10	66.39	20.91	48.00	39.08	41.89	25.11	10.11	17.84	13.87	3.66	29.26	14.41	17.02	51.79	27.69	-3.17	
Δ_{\uparrow}	5.53	3.46	13.34	8.93	13.24	14.01	17.44	12.26	3.45	8.82	1.34	1.45	12.80	-0.25	4.74	0.66	7.58	2.71	

Table 2: Performance of DOC under CW attack with a perturbation budget of $\epsilon_{\text{atk}} = 4/255$. Δ_{CLIP} denotes the improvement over the original CLIP, and Δ_{\uparrow} indicates the gain over the previous best performance. The best performance is shown in bold.

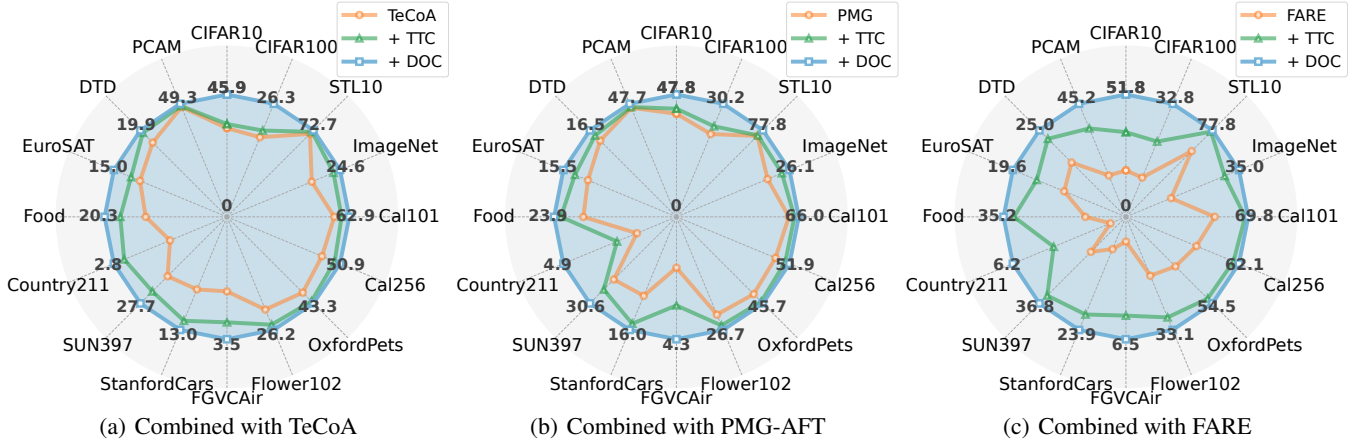


Figure 3: Performance of DOC combined with adversarial fine-tuning, including TeCoA (Mao et al. 2022), PMG-AFT (Wang et al. 2024c), and FARE (Schlarman et al. 2024). Robust accuracy is evaluated on 16 datasets using PGD-10 with $\epsilon_{\text{atk}} = 1/255$.

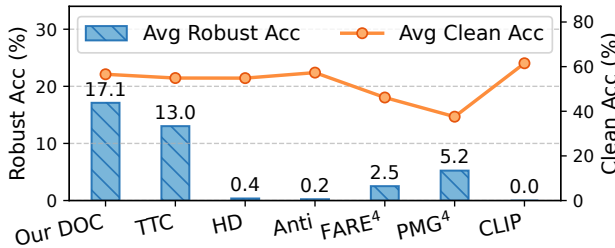


Figure 4: Performance of DOC and other baselines under AutoAttack with a perturbation budget of $\epsilon_{\text{atk}} = 4/255$. Clean and robust accuracy is averaged across 16 datasets.

lyon (PCAM) (Bejnordi et al. 2017).

Baselines for Comparison As research on improving the zero-shot adversarial robustness of VLPs via test-time defense is still in its early stages and available methods are limited, we primarily compare our DOC with the state-of-the-art approach, TTC (Xing, Zhao, and Sebe 2025). We further include representative test-time defenses, covering Anti-Adversary (Anti) (Alfarra et al. 2022) and Hedge Defense (HD) (Wu et al. 2021). Although our method targets test-time defense, we also compare it with three adversarial fine-tuning approaches, including TeCoA (Mao et al. 2022),

PMG-AFT (PMG) (Wang et al. 2024c), and FARE (Schlarman et al. 2024), which fine-tune CLIP on Tiny ImageNet. **Implementation Details** The counterattack budget is set to $\epsilon_{\text{ca}} = 4/255$, following prior work (Xing, Zhao, and Sebe 2025). We evaluate adversarial robustness under PGD (Madry et al. 2018), CW (Carlini and Wagner 2017), and AutoAttack (AA) (Croce and Hein 2020) with $\epsilon_{\text{atk}} = 4/255$ bounded by ℓ_{∞} norm. The counterattack is performed with a batch size of 256 and 4 steps using a default step size of $\alpha_{\text{ttc}} = 3/255$. All experiments are conducted on a single NVIDIA 4090 GPU, using PyTorch 3.9.13 and CUDA 12.0. We adopt the publicly available CLIP model ViT-B/32 provided in Hugging Face as the backbone, and freeze model parameters throughout the evaluation to ensure a consistent inference-only setting (Radford et al. 2021). Additional results under alternative settings are provided in the **Supplementary Materials**.

Main Results

Adversarial Robustness under PGD We evaluate our method and baselines under PGD-10 across 16 datasets, and the results are shown in Table 1. While adversarial fine-tuning methods improve robustness, they significantly degrade clean accuracy, and this degradation becomes more severe as the fine-tuning perturbation budget increases. More-

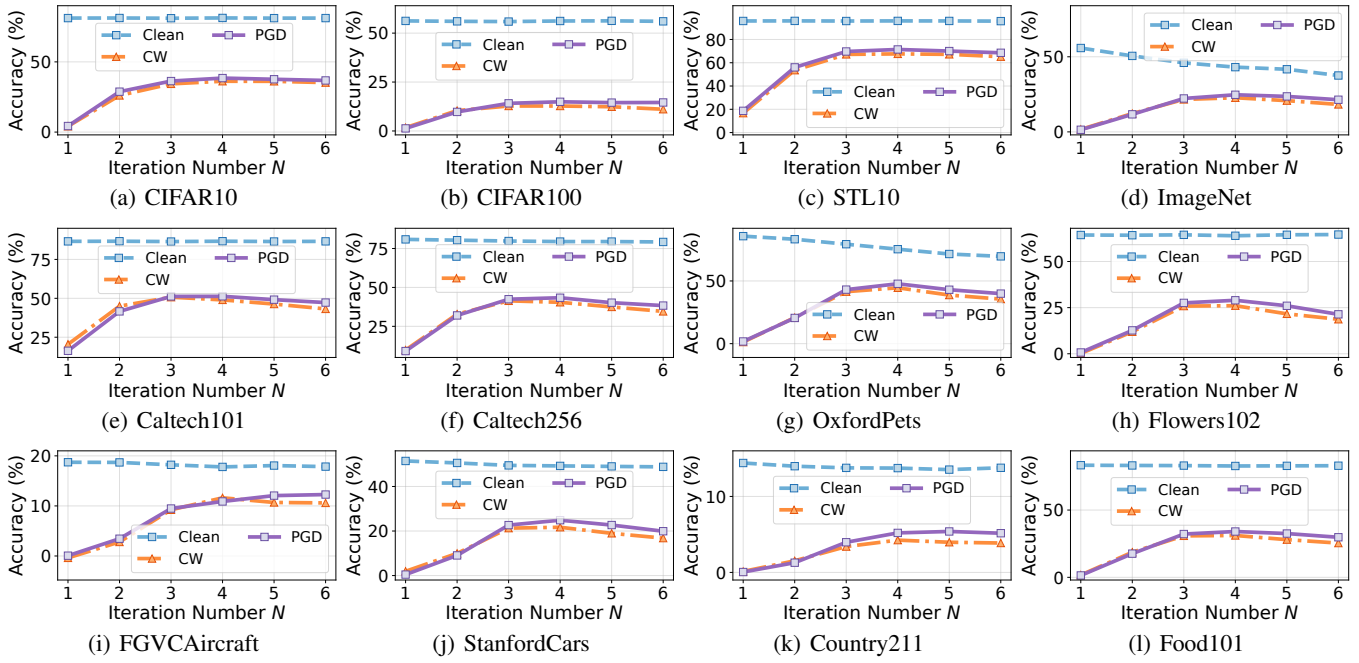


Figure 5: Performance with numbers of counterattack steps N . Robust accuracy is evaluated by PGD-10 and CW with the perturbation budget $\epsilon_{\text{atk}} = 4/255$. Results on remaining datasets are presented in **Supplementary Materials**.

over, adversarial fine-tuning requires access to source data and incurs additional computational overhead. In contrast, our DOC achieves significant improvements in adversarial robustness while maintaining competitive clean accuracy. Specifically, DOC outperforms the state-of-the-art TTC, improving the average robust accuracy by 9.80%, and retains a higher clean accuracy. Furthermore, compared to the original CLIP model, DOC improves robust accuracy by over 30% with minimal impact on clean performance, demonstrating its competitiveness as a test-time defense.

Adversarial Robustness under CW and AutoAttack

We further evaluate the robustness of our DOC against stronger attacks, including CW and AutoAttack. The corresponding results are reported in Table 2 (CW) and Fig. 4 (AutoAttack). Specifically, our method consistently outperforms prior approaches, achieving average improvements of over 7.58% under CW and 4.1% under AutoAttack across 16 datasets. Compared to TTC, which also leverages CLIP’s pretrained features for counterattack generation, DOC introduces enhancements such as directional sensitivity discrimination and orthogonal-guided optimization, leading to consistent and better defense performance, with gains observed on nearly all datasets. Importantly, these improvements are achieved without additional training costs, making DOC practical for real-world deployment.

Combining DOC with Adversarial Fine-Tuning

Although our DOC is designed as a test-time defense, it can be integrated as a plug-in module to further enhance adversarially fine-tuned models. We follow the settings in (Xing, Zhao, and Sebe 2025) and report the results in Fig.

3. When applied to adversarially finetuned models, covering TeCoA, PMG-AFT, and FARE, DOC consistently improves adversarial robustness, which brings an improvement of 4% – 5% over the baselines. Notably, when combined with FARE, DOC achieves an average robust accuracy increase of over 18% compared to the original CLIP. Interestingly, we observe that the magnitude of robustness gains varies across fine-tuning methods. This likely stems from the fact that adversarial fine-tuning can reduce the model’s embedding space sensitivity to input perturbations, which, while improving robustness, may also compromise the representational adopted for effective counterattack generation. Despite this, DOC remains effective in most cases, underscoring its adaptability and ability to leverage both pre-trained and fine-tuned encoder representations. Overall, DOC can serve as a lightweight enhancement to adversarial fine-tuning, without introducing additional training costs.

Ablation Study

We conduct ablation experiments to evaluate the contribution of each component in DOC. Table 3 reports the average clean and robust accuracy across 16 datasets under $\epsilon_{\text{atk}} = 4/255$ with five random seeds (1-5). We adopt TTC as the baseline. Enabling DSS alone improves clean accuracy over the baseline by better distinguishing between clean and adversarial examples, which suppresses unnecessary perturbations on clean inputs, reducing the risk of amplifying benign variations into adversarial directions. Using OGA alone yields larger gains in robust accuracy, supporting our design motivation that diversity counterattack directions help neutralize adversarial perturbations more effectively without su-

DSS	OGA	Clean	PGD	CW	AA
✗	✗	55.66 \pm 0.08	21.43 \pm 0.07	20.70 \pm 0.11	21.97 \pm 0.16
✓	✗	58.23 \pm 0.05	23.37 \pm 0.06	22.27 \pm 0.07	22.66 \pm 0.11
✗	✓	55.38 \pm 0.12	31.83 \pm 0.10	29.02 \pm 0.12	26.07 \pm 0.19
✓	✓	58.27 \pm 0.09	31.04 \pm 0.08	28.15 \pm 0.13	25.89 \pm 0.18

Table 3: Ablation study results of our DOC. Clean and robust accuracy is reported as the average across 16 datasets. DSS and OGA denote the directional sensitivity score and the orthogonal gradient augmentation, respectively.

pervised information. Combining DSS and OGA achieves the best balance by improving both robustness and clean accuracy, which confirms DOC provides a reliable discrimination mechanism to prevent over-correction on clean examples and better neutralize adversarial perturbations.

Hyperparameter Selection and Discussion

Due to page limitations, we analyze the key hyperparameter, counterattack steps N , while results for other parameters are included in the **Supplementary Materials**. Specifically, we use the default settings and an adversarial perturbation budget of $\epsilon_{\text{atk}} = 4/255$. As shown in Fig. 5, increasing N consistently improves robustness up to $N = 3$ and saturates around $N = 3$ or $N = 4$. This trend suggests that even a small number of counterattack steps can yield substantial adversarial robustness gains, and that selecting an appropriate N enables sufficient exploration of the adversarial perturbation space to effectively suppress adversarial effects. Importantly, clean accuracy remains stable, confirming that our DOC improves robustness not by sacrificing clean performance. The consistent robustness gains across both low-resolution and fine-grained datasets certify the competitiveness of our DOC in improving adversarial robustness.

Conclusion and Future

This work revisits the optimization strategy for counterattacks in test-time defense and identifies that vanilla PGD-based updates lack perturbation diversity, limiting their effect in neutralizing diverse adversarial patterns. Accordingly, we present Directional Orthogonal Counterattack (DOC), which enhances diversity by expanding the perturbation space through orthogonal exploration and momentum-based optimization, thereby better counteracting potential adversarial perturbation. In addition, DOC incorporates a directional sensitivity score computed via averaged cosine similarity, offering a stable and more discriminative criterion to adaptively guide counterattack strength.

Although developed on CLIP, our method does not rely on specific network architectures, label supervision, or training data. Instead, our DOC exploits the model’s intrinsic representational capacity, enabling straightforward transfer to other multimodal systems, including large-scale vision-language models. More importantly, we show that enhancing counterattack diversity substantially improves adversarial robustness, offering a promising direction for lightweight and scalable multimodal defenses.

Acknowledgments

This work was supported in part by the grant of the National Natural Science Foundation of China under Grant 62172090, in part by the Start-up Research Fund of Southeast University under Grant RF1028623097, in part by the Start-up Grant (No. 9610680) of the City University of Hong Kong, and in part by the Young Scientist Fund (No. 62406265) of NSFC. We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations.

References

- Alfarra, M.; Pérez, J. C.; Thabet, A.; Bibi, A.; Torr, P. H.; and Ghanem, B. 2022. Combating adversaries with anti-adversaries. In *AAAI*, 5992–6000.
- Bejnordi, B. E.; Veta, M.; Van Diest, P. J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J. A.; Hermsen, M.; Manson, Q. F.; Balkenhol, M.; et al. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22): 2199–2210.
- Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101—mining discriminative components with random forests. In *ECCV*, 446–461. Springer.
- Cao, M.; Bai, Y.; Zeng, Z.; Ye, M.; and Zhang, M. 2024. An empirical study of clip for text-based person search. In *AAAI*, volume 38, 465–473.
- Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *SP*, 39–57. IEEE.
- Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *CVPR*, 3606–3613.
- Coates, A.; Ng, A.; and Lee, H. 2011. An analysis of single-layer networks in unsupervised feature learning. In *AIS-TATS*, 215–223. JMLR.
- Croce, F.; Goyal, S.; Brunner, T.; Shelhamer, E.; Hein, M.; and Cemgil, T. 2022. Evaluating the adversarial robustness of adaptive test-time defenses. In *ICML*, 4421–4435. PMLR.
- Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2206–2216. PMLR.
- Cui, X.; Aparcedo, A.; Jang, Y. K.; and Lim, S.-N. 2024. On the robustness of large multimodal models against image adversarial attacks. In *CVPR*, 24625–24634.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. IEEE.
- Dong, J.; Koniusz, P.; Zhang, Y.; Zhu, H.; Liu, W.; Qu, X.; and Ong, Y.-S. 2023. Improving Zero-Shot Adversarial Robustness in Vision-Language Models by Closed-form Alignment of Adversarial Path Simplices. In *ICML*.
- Fei-Fei, L.; Fergus, R.; and Perona, P. 2006. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4): 594–611.

- Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024a. Clip-adapter: Better vision-language models with feature adapters. *Int. J. Comput. Vis.*, 132(2): 581–595.
- Gao, S.; Jia, X.; Ren, X.; Tsang, I.; and Guo, Q. 2024b. Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory. In *ECCV*, 442–460. Springer.
- Ge, Y.; Ren, J.; Gallagher, A.; Wang, Y.; Yang, M.-H.; Adam, H.; Itti, L.; Lakshminarayanan, B.; and Zhao, J. 2023. Improving zero-shot generalization and robustness of multimodal models. In *CVPR*, 11093–11101.
- Gong, S.; Haoyu, L.; Dou, Q.; and Farnia, F. 2025. Boosting the visual interpretability of clip via adversarial fine-tuning. In *ICLR*.
- Griffin, G.; Holub, A.; Perona, P.; et al. 2007. Caltech-256 object category dataset. Technical report, Technical Report 7694, California Institute of Technology Pasadena.
- Guo, Q.; Pang, S.; Jia, X.; Liu, Y.; and Guo, Q. 2024. Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. *IEEE Trans. Inf. Forensics Security*.
- Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 12(7): 2217–2226.
- Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2022. Scaling up vision-language pre-training for image captioning. In *CVPR*, 17980–17989.
- Jia, X.; Zhang, Y.; Wei, X.; Wu, B.; Ma, K.; Wang, J.; and Cao, X. 2024. Improving fast adversarial training with prior-guided knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(9): 6367–6383.
- Jiang, C.; Wang, J.; Dong, M.; Gui, J.; Shi, X.; Cao, Y.; Tang, Y. Y.; and Kwok, J. T.-Y. 2025. Improving Fast Adversarial Training via Self-Knowledge Guidance. *IEEE Trans. Inf. Forensics Security*.
- Jiao, S.; Wei, Y.; Wang, Y.; Zhao, Y.; and Shi, H. 2023. Learning mask-aware clip representations for zero-shot segmentation. *NeurIPS*, 36: 35631–35653.
- Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *ICCV*, 554–561.
- Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.
- Kuang, H.; Liu, H.; Lin, X.; and Ji, R. 2024. Defense against adversarial attacks using topology aligning adversarial training. *IEEE Trans. Inf. Forensics Security*, 19: 3659–3673.
- Laurençon, H.; Tronchon, L.; Cord, M.; and Sanh, V. 2024. What matters when building vision-language models? *NeurIPS*, 37: 87874–87907.
- Li, L.; Guan, H.; Qiu, J.; and Spratling, M. 2024a. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *CVPR*, 24408–24419.
- Li, X.; Zhang, W.; Liu, Y.; Hu, Z.; Zhang, B.; and Hu, X. 2024b. Language-driven anchors for zero-shot adversarial robustness. In *CVPR*, 24686–24695.
- Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*.
- Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.
- Mao, C.; Geng, S.; Yang, J.; Wang, X.; and Vondrick, C. 2022. Understanding Zero-shot Adversarial Robustness for Large-Scale Models. In *ICLR*.
- Mou, Y.; Zhang, S.; and Ye, W. 2024. Sg-bench: Evaluating llm safety generalization across diverse tasks and prompt types. *NeurIPS*, 37: 123032–123054.
- Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion Models for Adversarial Purification. In *ICML*, 16805–16827. PMLR.
- Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *ICVGIP*, 722–729. IEEE.
- Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *CVPR*, 3498–3505. IEEE.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.
- Raman, M.; Maini, P.; Kolter, Z.; Lipton, Z. C.; and Pruthi, D. 2023. Model-tuning Via Prompts Makes NLP Models Adversarially Robust. In *ACL*.
- Schlarmann, C.; Singh, N. D.; Croce, F.; and Hein, M. 2024. Robust CLIP: Unsupervised Adversarial Fine-Tuning of Vision Embeddings for Robust Large Vision-Language Models. In *ICML*, 43685–43704. PMLR.
- Schwinn, L.; Bungert, L.; Nguyen, A.; Raab, R.; Pulsmeier, F.; Precup, D.; Eskofier, B.; and Zanca, D. 2022. Improving robustness against real-world and worst-case distribution shifts through decision region quantification. In *ICML*, 19434–19449. PMLR.
- Sheng, L.; Liang, J.; Wang, Z.; and He, R. 2025. R-TPT: Improving Adversarial Robustness of Vision-Language Models through Test-Time Prompt Tuning. In *CVPR*, 29958–29967.
- Tong, K.; Jiang, C.; Gui, J.; and Cao, Y. 2024. Taxonomy driven fast adversarial training. In *AAAI*, volume 38, 5233–5242.
- Tu, W.; Deng, W.; and Gedeon, T. 2023. A closer look at the robustness of contrastive language-image pre-training (clip). *NeurIPS*, 36: 13678–13691.
- Wang, H.; Dong, K.; Zhu, Z.; Qin, H.; Liu, A.; Fang, X.; Wang, J.; and Liu, X. 2024a. Transferable multimodal attack on vision-language pre-training models. In *SP*, 1722–1740. IEEE.
- Wang, H.; Liu, F.; Jiao, L.; Wang, J.; Hao, Z.; Li, S.; Li, L.; Chen, P.; and Liu, X. 2024b. Vilt-clip: Video and language

- tuning clip with multimodal prompt learning and scenario-guided optimization. In *AAAI*, volume 38, 5390–5400.
- Wang, S.; Zhang, J.; Yuan, Z.; and Shan, S. 2024c. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *CVPR*, 24502–24511.
- Wang, X.; Chen, K.; Zhang, J.; Chen, J.; and Ma, X. 2025. Tapt: Test-time adversarial prompt tuning for robust inference in vision-language models. In *CVPR*, 19910–19920.
- Wortsman, M.; Ilharco, G.; Kim, J. W.; Li, M.; Kornblith, S.; Roelofs, R.; Lopes, R. G.; Hajishirzi, H.; Farhadi, A.; Namkoong, H.; et al. 2022. Robust fine-tuning of zero-shot models. In *CVPR*, 7959–7971.
- Wu, B.; Pan, H.; Shen, L.; Gu, J.; Zhao, S.; Li, Z.; Cai, D.; He, X.; and Liu, W. 2021. Attacking adversarial attacks as a defense. *arXiv preprint arXiv:2106.04938*.
- Xhonneux, S.; Sordoni, A.; Günnemann, S.; Gidel, G.; and Schwinn, L. 2024. Efficient adversarial training in llms with continuous attacks. *NeurIPS*, 37: 1502–1530.
- Xia, S.; Yang, W.; Yu, Y.; Lin, X.; Ding, H.; Duan, L.; and Jiang, X. 2024. Transferable adversarial attacks on sam and its downstream models. *NeurIPS*, 37: 87545–87568.
- Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 3485–3492. IEEE.
- Xing, S.; Zhao, Z.; and Sebe, N. 2025. Clip is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of clip. In *CVPR*, 15172–15182.
- Yang, C.; An, Z.; Huang, L.; Bi, J.; Yu, X.; Yang, H.; Diao, B.; and Xu, Y. 2024. Clip-kd: An empirical study of clip model distillation. In *CVPR*, 15952–15962.
- Yang, H.; Jeong, J.; and Yoon, K.-J. 2024. Prompt-driven contrastive learning for transferable adversarial attacks. In *ECCV*, 36–53. Springer.
- Yang, S.; Chen, Y.; Tian, Z.; Wang, C.; Li, J.; Yu, B.; and Jia, J. 2025. Visionzip: Longer is better but not necessary in vision language models. In *CVPR*, 19792–19802.
- Yu, L.; Zhang, H.; and Xu, C. 2024. Text-guided attention is all you need for zero-shot robustness in vision-language models. *NeurIPS*, 37: 96424–96448.
- Zhang, B.; Zhang, P.; Dong, X.; Zang, Y.; and Wang, J. 2024a. Long-clip: Unlocking the long-text capability of clip. In *ECCV*, 310–325. Springer.
- Zhang, C.; Wang, S.; Li, X.; Zhu, Y.; Qi, H.; and Huang, Q. 2025. Enhancing the Robustness of Vision-Language Foundation Models by Alignment Perturbation. *IEEE Trans. Inf. Forensics Security*.
- Zhang, D.-C.; Zhou, Z.; and Li, Y.-F. 2024. Robust test-time adaptation for zero-shot prompt tuning. In *AAAI*, volume 38, 16714–16722.
- Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024b. Vision-language models for vision tasks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(8): 5625–5644.
- Zhang, J.; Ma, X.; Wang, X.; Qiu, L.; Wang, J.; Jiang, Y.-G.; and Sang, J. 2024c. Adversarial prompt tuning for vision-language models. In *ECCV*, 56–72. Springer.
- Zhang, M.; Bi, K.; Chen, W.; Guo, J.; and Cheng, X. 2024d. CLIPure: Purification in Latent Space via CLIP for Adversarially Robust Zero-Shot Classification. In *ICLR*.
- Zhao, Y.; Pang, T.; Du, C.; Yang, X.; Li, C.; Cheung, N.-M. M.; and Lin, M. 2023. On evaluating adversarial robustness of large vision-language models. *NeurIPS*, 36: 54111–54138.
- Zhou, Z.; Lei, Y.; Zhang, B.; Liu, L.; and Liu, Y. 2023. Zeg-clip: Towards adapting clip for zero-shot semantic segmentation. In *CVPR*, 11175–11185.
- Zhu, H.; Ren, Y.; Sui, X.; Yang, L.; and Jiang, W. 2023. Boosting adversarial transferability via gradient relevance attack. In *ICCV*, 4741–4750.