TIMA: Text-Image Mutual Awareness for Balancing Zero-Shot Adversarial Robustness and Generalization Ability

Fengji Ma[1], **Hei Victor Cheng**[2*], **Chenxing Li**[3], **Li Liu**[1 *]
[1]Hong Kong University of Science and Technology (Guangzhou)
[2]Aarhus University
[3]Tencent AI Lab
avrillliu@hkust-gz.edu.cn, hvc@ece.au.dk

Abstract

Achieving zero-shot adversarial robustness without sacrificing generalization remains challenging for foundation models such as CLIP, especially under large adversarial perturbations. Through empirical analyses, we identify three critical yet overlooked issues: (1) Logit margins exhibit a stable offset between small and large adversarial perturbations, suggesting that explicitly adjusting margins could improve robustness against unseen large perturbations. (2) A significant negative correlation exists between logit margin and inter-class semantic similarity, indicating that semantic structures are insufficiently leveraged by existing methods. (3) Existing methods for adjusting text embeddings disrupt the intrinsic semantic consistency established by pre-trained models, undermining generalization capability. Motivated by these findings, we propose a novel Text-Image Mutual Awareness (TIMA) framework, including a Text-Aware Image (TAI) tuning module with an Adaptive Semantic-Aware Margin (ASAM) to explicitly calibrate logit margins, and an Image-Aware Text (IAT) tuning module with Semantic Consistent Minimum Hyperspherical Energy (SCMHE) to preserve semantic consistency. Comprehensive experiments validate that TIMA significantly outperforms existing approaches by effectively addressing the identified limitations.

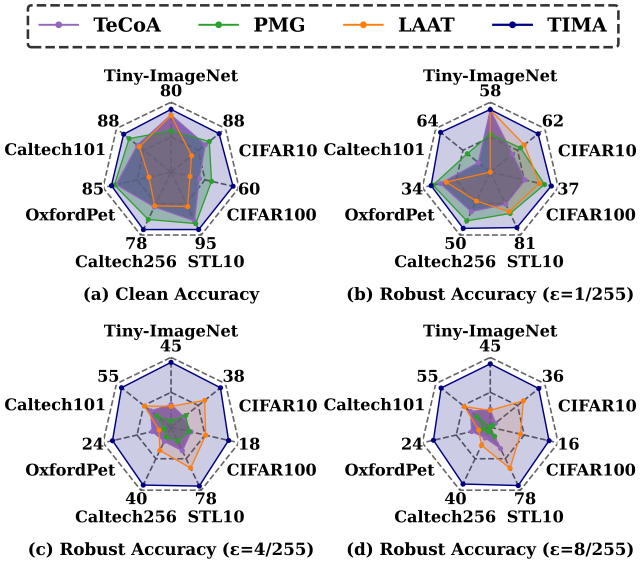


Figure 1: Zero-shot accuracies of compared methods and our proposed method. We show zero-shot robust accuracy of different methods under different perturbation radii. Better view by zooming in.

1 Introduction

Large-scale foundation models (Radford et al. 2021; Jia et al. 2021; Ahn et al. 2022; Ramesh et al. 2022) have recently garnered significant attention due to their excellent zero-shot generalization ability. Contrastive Language-Image Pre-training (CLIP) (Radford et al. 2021), the most widely used vision-language model, demonstrates the ability to accurately classify new classes with only simple text captions, even those that have not been encountered before. However, recent studies (Mao et al. 2022; Zhao et al. 2023; Yang and Mirzasoleiman 2023; Michels et al. 2023; Noever and Noever 2021; Wang et al. 2023; Inkawhich, McDonald, and Luley 2023; Li et al. 2025a) revealed that while these foundation models exhibit strong generalization performance, they are extremely vulnerable to adversarial perturbation. Notably, existing adversarial methods (Madry et al. 2017; Carlini and Wagner 2017; Croce and Hein 2020) have effectively attacked these foundation models, resulting in performance degradation of up to 90% on different datasets, even with minor attacks. Meanwhile, recent works (Li et al. 2025b; Li, Chen, and Hu 2025; Li et al. 2024a) on robust recognition and detection further emphasize the practical importance of robustness in real-world systems: PartImageNet++ (Li et al. 2024a) scales up part-based models for robust recognition on ImageNet (Deng et al. 2009), and adversarially robust object detectors are advanced by improving backbones (Li, Chen, and Hu 2025) and employing patch-based composite adversarial training against physically realizable attacks (Li et al. 2025b).

Traditionally, time-intensive and resource-heavy adversarial training is needed to achieve adversarial robustness. To bypass retraining across new tasks or datasets and to harvest the full potential of foundation models, the concept of **zero-shot adversarial robustness** in foundation models (Mao et al. 2022; Li et al. 2024b; Wang et al. 2024) has emerged as a new

*Corresponding authors

and urgent research topic. The objective of zero-shot adversarial robustness is twofold: to transfer adversarial robustness in a zero-shot learning manner, and to preserve zero-shot generalization capabilities of foundational models. Here, adversarial robustness refers to a model's ability to withstand adversarial attacks with various perturbations. Zero-shot generalization, on the other hand, refers to the model's capacity to achieve high accuracy on previously unseen classes.

In pursuit of a zero-shot adversarial robust CLIP model, the pioneering work TeCoA (Mao et al. 2022) proposed aligning adversarial image embeddings with their corresponding text embeddings. This eliminates the need for retraining in subsequent visual tasks. Building on this idea, PMG (Wang et al. 2024) introduced additional constraints from pre-trained models and clean samples into the objective function, encouraging the preservation of generalized pre-trained information. Unlike the above two methods, which use only an adversarial fine-tuned image encoder, LAAT (Li et al. 2024b) proposed an algorithm to expand the distance between fixed textual embeddings in hyperspherical space, showing promise against larger perturbation radii. However, existing methods fail to achieve zero-shot adversarial robustness while facing large perturbations (as shown in Fig. 1) that are unseen during adversarial training.

To understand these limitations, our comprehensive empirical analyses (in Sec. 3) have identified three important insights: **1)** We observe a positive stable offset in logit margins when transitioning from small adversarial perturbations (e.g., $\varepsilon = 1/255$) to a range of larger, unseen perturbations (such as $\varepsilon = 2/255, 4/255$, and $8/255$). Specifically, we find that the differences in logit margins under varying adversarial attack perturbation bounds can be approximated by a stable constant across different training epochs. This phenomenon indicates that explicitly calibrating logit margins during training could substantially enhance robustness against larger, unseen adversarial perturbations. **2)** We discover a significant negative correlation between the logit margin under adversarial perturbations and the semantic similarity between classes. Our analysis shows that categories with higher semantic similarity suffer from larger logit margin reductions, suggesting that existing methods fail to effectively leverage inherent semantic relationships to improve robustness. **3)** We identify that current methods aimed at increasing inter-class distances among text embeddings (e.g., LAAT (Li et al. 2024b)) inadvertently disrupt the intrinsic semantic consistency originally learned by pre-trained CLIP models. Specifically, expansions in text embedding distances are found to significantly diminish semantic coherence, thereby undermining zero-shot generalization performance.

Motivated by these insights, we propose a novel framework named *Text-Image Mutual Awareness (TIMA)*, as shown in Fig. 2, designed to simultaneously address these challenges. TIMA comprises two complementary modules: a) Text-Aware Image (TAI) tuning module: Motivated by the above insight 1) and 2), we introduce the *Adaptive Semantic-Aware Margin (ASAM)*, explicitly calibrating decision boundaries according to class-level semantic similarity. By adaptively modulating logit margins during adversarial fine-tuning, TAI significantly improves robustness against large adversarial perturbations. b) Image-Aware Text (IAT) tuning module: Motivated by the above insight 3), we propose the *Semantic Consistent Minimum Hyperspherical Energy (SC-MHE)* method. It uniformly enlarges inter-class distances among text embeddings on a hypersphere while employing semantic consistency regularization to preserve intrinsic semantic consistency from pre-trained embeddings, effectively maintaining zero-shot generalization capabilities.

The main contributions can be summarized as:

- This work proposes a novel TIMA framework, containing TAI and IAT. TIMA maximizes the contrastive and interactive information between text and image modalities, achieving a balance between zero-shot robustness and generalization, leading to effective multimodal training.
- Motivated by insights from empirical analyses, by introducing an adaptive margin, we propose a new ASAM that provably increases inter-class distance and enhances zero-shot adversarial robustness. Notably, ASAM within the TAI tuning module is designed as a plug-and-play component, enabling seamless integration into any existing adversarial fine-tuning framework.
- The proposed SC-MHE in the IAT tuning module achieves a trade-off between increasing text embedding inter-class distances and maintaining semantic information by semantic consistency regularization to prevent degradation in generalization performance.
- Extensive experiments show TIMA's superiority in zero-shot robust accuracy and clean accuracy in multiple datasets and under different perturbations, establishing its efficacy compared with state-of-the-art (SOTA) methods. Notably, TIMA demonstrates adversarial robustness to minor perturbation ($\varepsilon = 1/255$) despite clean-data-only training.

2 Related Works

Adversarial Robustness and Margin. Prior works have sought to improve robustness by maximizing the margin from the decision boundary, using techniques like uniform loss objectives (Ding et al. 2018) or adaptive perturbation bounds per sample (Fazlyab et al. 2024; Xu et al. 2023). In contrast, our approach operates with a fixed training perturbation and instead improves robustness by identifying and enforcing a sample-level adaptive margin.

Zero-Shot Adversarial Robustness. Existing methods for zero-shot adversarial robustness fall into three categories. The first, including TeCoA (Mao et al. 2022) and PMG (Wang et al. 2024), fine-tunes the image encoder but struggles against large, unseen perturbations. The second, such as LAAT (Li et al. 2024b), adjusts text embeddings but can disrupt semantic consistency. A third category of training-free methods has recently emerged, which perform test-time defenses in either the feature (Tong et al. 2025) or pixel space (Xing, Zhao, and Sebe 2025). In contrast, our TIMA framework introduces a Text-Image Mutual Awareness approach, co-optimizing image and text encoders via adaptive semantic-aware margins and semantic-consistent energy minimization to balance robustness and generalization.

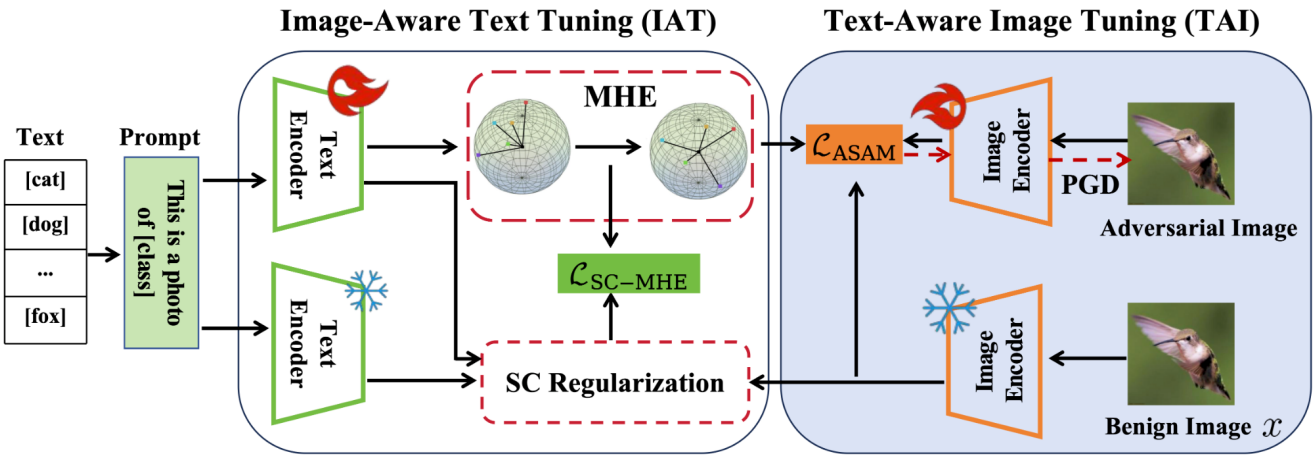


Figure 2: Framework of our TIMA, including IAT and TAI tuning modules. SC-Regularization means semantic consistency regularization (Eq.(7)). $\mathcal{L}_{\text{ASAM}}$ and $\mathcal{L}_{\text{SC-MHE}}$ represent Adaptive Semantic-Aware Margin (ASAM) and Semantic Consistent Minimum Hyperspherical Energy (SC-MHE). Adversarial images are generated by Projected Gradient Descent (PGD).

3 Methodology

We begin by presenting the background and preliminaries on contrastive adversarial training and zero-shot adversarial robustness. Next, we describe the experimental observations that motivate our approach. Finally, we introduce the TIMA framework in detail, including its key components, as illustrated in Fig. 2.

Preliminaries

We adopt the CLIP model as the foundation for our study, as it is a large-scale pre-trained vision-language model. The proposed method aims to train a foundational model that balances zero-shot robustness and generalization. Given an image-text pair $(x, \mathcal{T})$, where x represents an input image and $\mathcal{T}$ represents a text prompt with a fixed template "This is a photo of {}", CLIP uses the image encoder ϕ and the text encoder ψ encode both the image and the text in fixed-dimensional embeddings $z = \phi(x)$ and $t = \psi(\mathcal{T})$.

Text-Guided Contrastive Adversarial Training. To generate the adversarial image x^ε from a clean image x, TeCoA (Mao et al. 2022) employs the Projected Gradient Descent (PGD) method (Madry et al. 2017), aiming to find an x^ε such that $||x^\varepsilon - x||_\infty \leq \varepsilon$ which maximizes a chosen loss function (typically the cross-entropy loss w.r.t. the true label y_c). This is achieved through an iterative process involving gradient ascent and projection steps. The resulting adversarial image x^ε is then used to compute the loss $\mathcal{L}$:

$$\mathcal{L}(x^\varepsilon, \mathcal{T}, y) = -\sum_{i,j}\left[y_{ij}\log\frac{\exp(\mathrm{s}(z_i^\varepsilon, t_j)/\tau)}{\sum_k \exp(\mathrm{s}(z_i^\varepsilon, t_k)/\tau)}\right], \quad (1)$$

where $z_i^\varepsilon = \phi(x_i^\varepsilon)$ represents the adversarial image embedding of the i-th adversarial image example x_i^ε encoded by the image encoder ϕ. And $y_{ij} = 1$ if the image-text pair $(x_i, \mathcal{T}_j)$ is positive, otherwise, $y_{ij} = 0$. τ is the temperature parameter, and s indicates calculating the cosine similarity of the two embeddings.

Logit Margin. We define the logit margin, a core concept for our analysis. For an adversarial image embedding z_i^ε of an image x_i with ground-truth class c_i, its logit margin γ_i^ε is the difference between the cosine similarity with the most confusing class text embedding t_k and the correct class embedding t_{c_i}:

$$\gamma_i^\varepsilon = \max_k \ \mathrm{s}(z_i^\varepsilon, t_k) - \mathrm{s}(z_i^\varepsilon, t_{c_i}). \quad (2)$$

The logit margin captures how adversarial perturbations shift image-text alignment away from the true class. To quantify the margin's shift under stronger attacks, we further define the **perturbation-induced logit margin shift**. Specifically, we focus on the shift from a small training perturbation ($\varepsilon_{\text{small}} = 1/255$) to a larger test-time perturbation ε:

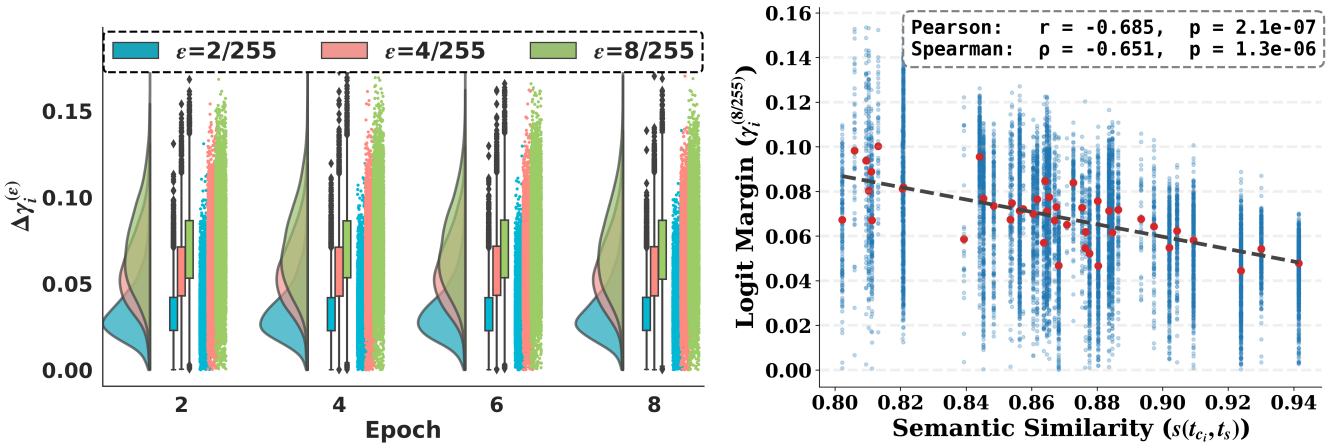
$$\Delta\gamma_i^{(\varepsilon)} := \gamma_i^{(\varepsilon)} - \gamma_i^{(1/255)}, \quad (3)$$

which will be referred to throughout the rest of the paper as the perturbation-induced logit margin shift.

Motivation

To clarify the link between logit margin (Eq.(2) and Eq.(3)) and robustness under large unseen perturbations, we analyze the behavior of the SOTA method TeCoA, focusing on logit margin shift, semantic preservation, and robustness. We also examine why LAAT, despite its robustness, shows poor generalization. These insights motivate our proposed approach.

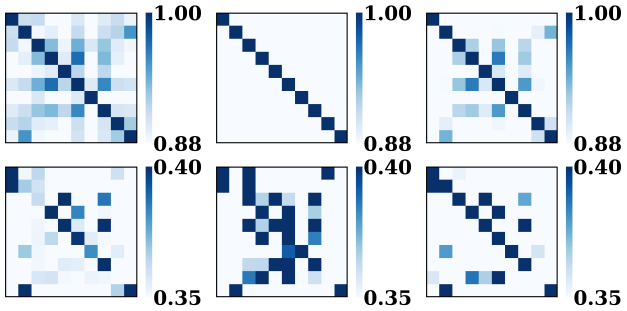
Perturbation-Induced Logit Margin Shift Analysis. We analyze the logit margin shift $\Delta\gamma_i$ caused by perturbations, which measures how the model's class separation changes from small training-time perturbations ($\varepsilon = 1/255$) to larger ones at test time ($\varepsilon = 2, 4, 8/255$). The results show that $\Delta\gamma_i$ remains stable across training epochs, suggesting a consistent positive gap between margins under large and small perturbations. This indicates that reducing margins under small perturbations can help improve robustness to larger ones. In



(a) Perturbation- Induced Logit Margin Shift ($\Delta\gamma_i$) *vs.* Epoch

(b) Logit Margin *vs.* Semantic Similarity

Figure 3: Motivation for ASAM. (a) demonstrates the distribution of perturbation-induced logit margin shift $\Delta\gamma_i$ between large unseen ($\varepsilon = 2, 4, 8/255$) and small seen ($\varepsilon = 1/255$) perturbations, revealing a relatively stable offset during training. This stable margin offset indicates potential for improving robustness against unseen larger perturbations by adjusting decision boundaries during training. (b) illustrates the relationship between semantic similarity and logit margin under strong perturbation (*e.g.* $\varepsilon = 8/255$). Each blue dot represents a sample x_i, where the x-axis shows the semantic similarity between the ground-truth class c_i and the most confusing class $s = \arg\max_k [s(z_i^\varepsilon, t_k) - s(z_i^\varepsilon, t_{c_i})]$, and the y-axis shows the corresponding logit margin $\gamma_i^\varepsilon = s(z_i^\varepsilon, t_s) - s(z_i^\varepsilon, t_{c_i})$. Red points denote median logit margins aggregated across samples with similar semantic similarity.



(a) CLIP (b) LAAT (c) Ours

Figure 4: Cosine similarity matrices on CIFAR-10. **Top row:** text-text similarity. **Bottom row:** clean image-text similarity. Our SC-MHE (c) better preserves the semantic structure of the original CLIP model (a) compared to LAAT (b).

addition, some samples show large $\Delta\gamma_i$ values throughout training, meaning they are more sensitive to perturbation strength. These may be unclear or rare cases, highlighting the need for a sample-wise adaptive margin.

Assumption 1. The image encoder ϕ is L-Lipschitz, implying its Jacobian's spectral norm is bounded: $\sup_{x' \in \mathcal{N}} \|J_\phi(x')\|_2 \leq L$.

Assumption 2. The Jacobian of the encoder J_ϕ is M-Lipschitz continuous: $\forall x_1, x_2 \in \mathcal{N}$, $\|J_\phi(x_1) - J_\phi(x_2)\|_2 \leq M\|x_1 - x_2\|_2$.

Proposition 1. *Based on Assumption 1 and 2, there exists* $K \leq 2M$ *such that* $\|\nabla\gamma(x)\| \leq 2L$ *and, for any* $0 < \varepsilon_1 < \varepsilon_2$,

$$\left|\Delta\gamma - (\varepsilon_2 - \varepsilon_1)\,\|\nabla\gamma(x)\|\right| \leq \frac{K}{2}(\varepsilon_2^2 - \varepsilon_1^2).$$

So the logit margin shift $\Delta\gamma$ is radius-ordered and deviation-bounded. As the bound grows, $\Delta\gamma$ increases in a stable, bounded-variation manner—exactly the trend in Fig.3(a). Full proof is in the *Appendix*.

Correlation Between Logit Margin and Semantic Similarity. In adversarially trained models (e.g., TeCoA), we further study the relationship between class-wise text embedding similarity and logit margins under strong perturbations. Fig. 3(b) plots the cosine similarity between the ground-truth class embedding t_{c_i} and its most similar class t_s (x-axis), and the corresponding logit margin $\gamma^{(8/255)}$ under $\varepsilon = 8/255$ (y-axis). Each blue point is a sample, and red points are class medians. We observe a strong negative correlation (Pearson $r = -0.685$, Spearman $\rho = -0.651$, both $p \ll 0.001$), with the medians showing an approximately linear trend. This motivates our margin in Eq.(4).

Based on observations from Fig. 3 (a) and (b), our **ASAM** design follows **three principles**: (1) adjusting margins under small perturbations improves robustness to larger ones; (2) the margin should reflect class-level semantic similarity; (3) it should adapt to individual samples.

Semantic Inconsistency in Text Embedding Adjustment. While prior methods that adjust text embeddings, such as LAAT (Li et al. 2024b), successfully increase inter-class distances, they disrupt the intrinsic semantic consistency of the pre-trained CLIP model. As illustrated in Fig. 4(b), this damages the inherent semantic relationships between similar

classes (e.g., cats and dogs, cars and trucks), thereby degrading zero-shot generalization performance on clean data and motivating our proposed semantic-aware regularization.

Text-Image Mutual Awareness

Adaptive Semantic-Aware Margin. Based on the above analysis, we introduce ASAM to explicitly adjust logit margins. The decision boundary in TeCoA's objective (Eq.(1)) is $s(z_i^\varepsilon, t_k) = s(z_i^\varepsilon, t_{c_i})$. To improve robustness, we enforce a semantic-aware negative margin:

$$s(z_i^\varepsilon, t_k) - s(z_i^\varepsilon, t_{c_i}) = -m \cdot (1 - s(t_k, t_{c_i})), \quad (4)$$

where $m > 0$ is a scaling factor. This constraint imposes a larger separation for more semantically similar classes, directly addressing the vulnerability observed in Fig. 3(b).

Furthermore, to achieve sample-wise adaptivity, we make the margin conditional on the sample's inherent ambiguity. Using a frozen image encoder ϕ_{frozen} to get the clean embedding $\hat{z}_i = \phi_{\text{frozen}}(x_i)$, we define the adaptive margin $\mathcal{M}_{ik}$ as:

$$\mathcal{M}_{ik} = \begin{cases} m \cdot \mathrm{s}(t_{c_i}, t_k), & \text{if } \mathrm{s}(\hat{z}_i, t_k) \geq \eta \cdot \mathrm{s}(\hat{z}_i, t_{c_i}) \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where η is a threshold. This mechanism selectively penalizes only the samples that are inherently easy to confuse.

Finally, we integrate this adaptive margin into the contrastive loss function to obtain the final ASAM loss:

$$\mathcal{L}_{\text{ASAM}} = -\sum_{i,j} y_{ij} \log \frac{\exp([\mathrm{s}(z_i^\varepsilon, t_j) - \mathcal{M}_{ij}]/\tau)}{\sum_k \exp([\mathrm{s}(z_i^\varepsilon, t_k) - \mathcal{M}_{ik}]/\tau)}, \quad (6)$$

where y_{ij} is the one-hot label.

Semantic Consistent Minimum Hyperspherical Energy. To address the semantic degradation caused by prior text embedding adjustments (*e.g.,* LAAT), we propose the Semantic Consistent Minimum Hyperspherical Energy (SC-MHE) method as the core of our IAT tuning module.

SC-MHE aims to uniformly distribute text embeddings on a hypersphere to maximize inter-class separation, while simultaneously preserving the semantic structure learned by the pre-trained model. This is achieved by a loss function with two components: 1) MHE, which encourages uniform spacing by maximizing the Euclidean distance $d(\cdot, \cdot)$ between text embeddings t. 2) Semantic Consistency Regularization, which prevents semantic drift during optimization. We introduce Semantic Consistency Regularization $\mathcal{R}$ to align the predictive distribution of the tuned text embeddings (t) with that of the frozen, original embeddings ($\hat{t}$) on clean images as follows:

$$\mathcal{R}_j = \sum_i p_{ij}^{tr} \log\left(\frac{p_{ij}^{tr}}{p_{ij}^{st}}\right), \quad (7)$$

where $p_{ij}^{st} = \frac{\exp(\mathrm{s}(\hat{z}_i, t_j)/\tau)}{\sum_k \exp(\mathrm{s}(\hat{z}_i, t_k)/\tau)}$ and $p_{ij}^{tr} = \frac{\exp(\mathrm{s}(\hat{z}_i, \hat{t}_j)/\tau)}{\sum_k \exp(\mathrm{s}(\hat{z}_i, \hat{t}_k)/\tau)}$.

The combined SC-MHE loss function is:

$$\mathcal{L}_{\text{SC-MHE}} = \sum_j \left[\sum_{k \neq j} \frac{1}{1 + d^\alpha(t_j, t_k)} + \lambda_{\mathrm{T}} \sum_i p_{ij}^{tr} \log \frac{p_{ij}^{tr}}{p_{ij}^{st}} \right] \quad (8)$$

where $\alpha = 2$ and λ_{T}=1. Due to space limits, we justify why MHE improves zero-shot robustness in *Appendix*.

4 Experiments

Experimental Setup

Dataset and Baseline. To evaluate zero-shot adversarial robustness and generalization, we consider 14 datasets spanning multiple recognition tasks, including CIFAR10, CIFAR100 (Krizhevsky, Hinton et al. 2009), STL10 (Coates, Ng, and Lee 2011), Caltech101 (Fei-Fei, Fergus, and Perona 2006), Caltech256 (Griffin et al. 2007) for generic classification; OxfordPets (Parkhi et al. 2012), Food101 (Bossard, Guillaumin, and Van Gool 2014), Flowers (Nilsback and Zisserman 2008), StandfordCars (Krause et al. 2013), FGVC (Maji et al. 2013) for fine-grained classification; SUN397 (Xiao et al. 2010) for scene recognition; and DTD (Cimpoi et al. 2014) for texture recognition. We compare with TeCoA (Mao et al. 2022), PMG (Wang et al. 2024), and LAAT (Li et al. 2024b) under the zero-shot setting. All methods are fine-tuned on Tiny-ImageNet (Deng et al. 2009) with matched training settings for fair comparison. We evaluate zero-shot robustness under PGD (Madry et al. 2017), CW (Carlini and Wagner 2017), AutoAttack (Croce and Hein 2020), and BIM (Kurakin, Goodfellow, and Bengio 2018).

Implementation Details. We use the CLIP-B/32 architecture with the prompt "This is a photo of a {}". The model is fine-tuned on Tiny-ImageNet for 10 epochs (batch size 512, learning rate 10^{-4}) using an SGD optimizer (momentum 0.9). Adversarial training employs a 2-step PGD attack ($l_\infty, \varepsilon = 1/255$, step size 1/255). The source code is included in the *supplementary materials*.

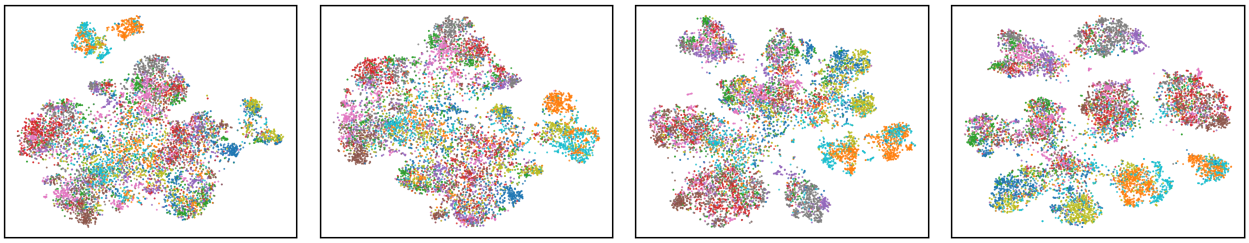
Result and Analysis

Compared with SOTA Methods. Tab. 1 shows the zero-shot robust accuracy under PGD attack and zero-shot clean accuracy of SOTA methods and our proposed approach TIMA under the ℓ_∞ setting, with the maximum adversarial perturbation set to $\varepsilon = 1/255, 8/255$. Regarding the zero-shot robust accuracy, under small bound attacks ($\varepsilon = 1/255$), zero-shot robust accuracy of TIMA exceeds TeCoA by 4.69%, PMG by 2.80%, and LAAT by 6.64%. When facing a large attack ($\varepsilon = 8/255$), zero-shot robust accuracy of our proposed TIMA exceeds TeCoA by 13.30%, PMG by 16.22%, and LAAT by 9.95%. Our proposed method shows a greater performance improvement over other methods under an 8/255 perturbation bound attack compared to the performance improvement observed under a 1/255 perturbation bound attack. Besides zero-shot adversarial robustness, on zero-shot generalization (clean accuracy), TIMA improves by 5.50% compared with TeCoA, 3.08% compared with PMG, and 11.53% compared with LAAT. From Tab. 1, we conclude that our method also has a favorable perfect performance both in zero-shot adversarial robustness and zero-shot generalization. For results on adversarial fine-tuning on ImageNet, please refer to *Appendix*. In addition to clean and robust accuracy, we report training cost and inference efficiency to assess overall computation in the *Appendix*.

Adversarial Evaluation Across Multiple Attacks. As illustrated in Fig. 6, we conducted a comprehensive robustness evaluation across a continuous spectrum of perturbation magnitudes (from $\varepsilon = 1/255$ to $8/255$) under BIM attack. The

Eval.	Compared Methods	Zero-shot datasets														Average (%)
		TinyIN	CIFAR10	CIFAR100	STL-10	ImageNet	OxfordPets	Food	SUN	DTD	CalTech101	CalTech256	StandfordCars	FGVC	Flowers	
clean	CLIP	60.74	89.06	62.32	97.16	59.12	85.82	83.15	57.67	40.52	85.47	82.02	52.17	8.40	65.04	67.76
	TeCoA	76.00	76.10	46.00	91.39	43.87	75.33	43.11	46.15	29.73	78.03	69.08	30.71	7.03	40.82	53.81
	PMG	70.98	78.47	50.31	92.10	46.08	76.62	53.45	45.85	31.12	81.98	73.07	35.79	5.08	46.29	56.23
	LAAT	75.84	68.08	42.72	87.75	42.72	51.21	34.81	43.92	18.14	77.12	68.73	21.06	7.20	29.68	47.78
	TIMA (Ours)	**77.75**	**84.23**	**57.90**	**93.61**	**46.55**	**78.80**	**55.82**	**53.37**	**32.55**	**84.59**	**76.27**	**37.37**	**9.18**	**42.38**	**59.31**
$\varepsilon = 1/255$	CLIP	2.72	9.57	4.55	35.40	3.50	2.72	3.95	1.02	2.50	28.99	20.27	0.66	0.00	1.37	8.37
	TeCoA	56.63	54.91	30.15	75.14	12.98	29.76	9.59	13.86	10.77	45.26	35.13	3.95	0.00	13.82	28.00
	PMG	52.52	56.67	34.19	76.89	**15.17**	31.45	**13.10**	**16.30**	10.82	50.11	41.03	5.03	0.00	15.24	29.89
	LAAT	56.72	57.50	33.15	76.54	13.30	28.35	12.83	13.53	8.51	32.90	28.42	0.71	**0.66**	1.56	26.05
	TIMA (Ours)	**56.75**	**60.53**	**35.56**	**79.58**	15.16	**31.92**	13.06	15.66	**14.36**	**61.88**	**45.77**	**6.75**	0.59	**20.12**	**32.69**
$\varepsilon = 8/255$	CLIP	1.99	5.03	0.42	31.43	1.11	4.91	10.10	0.36	0.05	22.23	13.62	0.36	0.00	0.36	6.57
	TeCoA	15.29	7.06	3.95	40.93	1.85	6.54	1.45	1.10	0.53	23.11	15.03	0.12	0.00	0.00	8.35
	PMG	4.13	5.51	2.07	30.06	0.54	2.34	1.13	0.42	0.37	17.26	12.13	0.03	0.00	0.00	5.43
	LAAT	14.67	23.44	9.34	58.59	1.04	4.11	2.67	1.62	0.32	28.73	18.61	0.22	**0.12**	0.36	11.70
	TIMA (Ours)	**41.31**	**32.45**	**14.38**	**73.84**	**10.21**	**20.05**	**9.55**	**7.48**	**2.29**	**50.04**	**37.06**	**1.18**	0.00	**3.32**	**21.65**

Table 1: Clean and adversarial evaluation of CLIP model under PGD attack. The best accuracies are bold.



(a) TeCoA (b) PMG (c) LAAT (d) TIMA (Ours)

Figure 5: T-SNE visualizations of image embeddings on the CIFAR-10 dataset under large perturbation.

ε	Compared Methods	Zero-shot datasets														Average (%)
		TinyIN	CIFAR10	CIFAR100	STL-10	ImageNet	OxfordPets	Food	SUN	DTD	CalTech101	CalTech256	StandfordCars	FGVC	Flowers	
0	TeCoA	76.61	88.77	65.02	96.31	51.06	82.36	71.63	57.83	36.01	84.77	79.98	45.16	10.05	46.34	63.71
	TIMA (Ours)	78.65	90.41	67.40	96.08	54.23	84.74	75.61	61.90	37.66	85.80	81.90	47.02	13.12	48.71	65.95
$1/255$	TeCoA	0.98	3.71	1.76	5.08	0.04	0.00	0.00	0.09	0.00	0.00	0.33	0.00	0.00	0.13	0.87
	TIMA (Ours)	16.43	26.30	12.66	37.70	4.74	10.33	10.18	3.09	2.49	19.57	18.53	1.16	0.13	0.78	11.72

Table 2: Clean and adversarial evaluation under PGD attack on various datasets *without adversarial training*.

results consistently demonstrate that our proposed TIMA method (green triangles) maintains a significant and stable performance advantage over baseline methods like TeCoA and PMG across all tested datasets. Notably, this performance gap widens dramatically as the perturbation bound ε increases. While the robust accuracy of baseline methods drops sharply and tends to collapse at larger perturbations (especially $\varepsilon \geq 4/255$), TIMA exhibits a much more gradual performance drop, showcasing its superior robustness against stronger attacks. This superior performance directly validates the core motivation of our work: by identifying and addressing the stable logit margin offset between small and

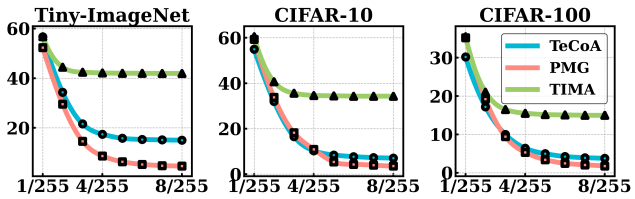


Figure 6: Adversarial evaluation under different perturbation bounds across different datasets at inference time.

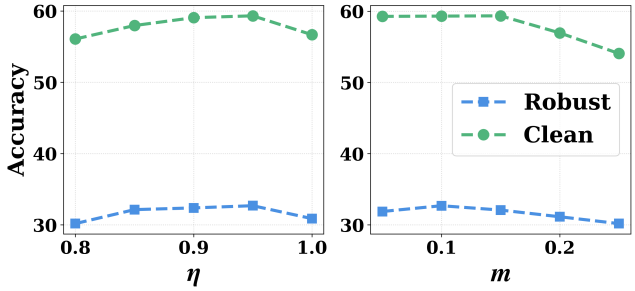


Figure 7: ASAM Sensitivity to Hyperparameters m and η.

large perturbations, our novel ASAM and SC-MHE module explicitly calibrates decision boundaries during training. This enables the model to effectively generalize the robustness learned from small perturbations to unseen, larger perturbations at inference time, thus achieving a more comprehensive and reliable defense across a wide range of attack perturbation magnitudes. For results on more datasets under CW, BIM, and AutoAttack, please refer to *Appendix*.

Visualization Analysis. The t-SNE visualizations in Fig. 5 demonstrate the superiority of our method under strong adversarial perturbations (CIFAR-10, $\varepsilon = 8/255$). Baseline methods, particularly TeCoA, suffer from severe feature collapse, with embeddings from nearly all classes tangled together. While PMG and LAAT form some discernible clusters, they still exhibit significant inter-class overlap and intra-class dispersion. In contrast, TIMA learns a much cleaner feature space, maintaining both tight intra-class clusters and clear inter-class boundaries. This provides strong evidence of TIMA's superior ability to preserve discriminative features and semantic structure, thereby enhancing robustness.

Adversarial Robustness without Adversarial Training. A key finding, detailed in Tab. 2, is the emergent adversarial robustness achieved by TIMA even when trained exclusively on clean data. Without any exposure to adversarial examples during fine-tuning, TIMA achieves an average robust accuracy of 11.72% against ℓ_∞ PGD attacks ($\varepsilon = 1/255$), a substantial +10.85% improvement over the TeCoA baseline. This phenomenon is not accidental but an emergent property of our framework's design. The ASAM module, by enforcing semantic-aware margins even on clean samples, acts as an implicit robustness regularizer, creating more resilient decision boundaries. Concurrently, the SC-MHE module optimizes the geometry of the class prototype space by uniformly separating text embeddings while preserving semantic consistency. The synergy between these two components fundamentally enhances the model's intrinsic robustness, demonstrating that TIMA improves the model's core structure rather than merely learning to defend specific attack patterns.

Methods	ε	CIFAR10	CIFAR100	STL10	ImageNet
TeCoA	0	76.10	46.00	91.39	43.87
w/ ASAM	0	80.41	51.87	92.40	42.75
w/ SC-MHE	0	76.74	45.92	91.94	44.50
TIMA (Ours)	0	**84.23**	**57.90**	**93.61**	**46.55**
TeCoA	$8/255$	7.06	3.95	40.93	1.85
w/ ASAM	$8/255$	31.56	**17.44**	67.11	**12.77**
w/ SC-MHE	$8/255$	23.62	14.08	48.97	5.64
TIMA (Ours)	$8/255$	**32.45**	14.38	**73.84**	10.21

Table 3: Ablation Study of the proposed ASAM and SC-MHE module. "w/" donates TeCoA with different modules.

Guidelines and Sensitivity Analysis of η and m in ASAM. We conduct a sensitivity analysis of the ASAM hyperparameters η and m (see Fig. 7). As shown in the figure, both clean and robust accuracies remain stable when $m \in [0.05, 0.15]$ and $\eta \in [0.85, 0.95]$, and our default setting ($m = 0.10$, $\eta = 0.95$) lies in this stable region. Detailed analysis and selection strategies are provided in *Appendix*.

Ablation Study

Tab. 3 presents the results of an ablation study on two core parts, ASAM and SC-MHE, showing the zero-shot clean accuracy and zero-shot robust accuracy across multiple test datasets under a large perturbation bound ($\varepsilon = 8/255$). Besides these results, Fig. 4 (c) shows the relationship between clean image embeddings and tuned text embeddings, which presents less semantic inconsistency than those of LAAT.

5 Conclusion

In this work, we propose the Text-Image Mutual Awareness (TIMA) framework to balance zero-shot robustness and generalization. Its TAI module uses an adaptive margin (ASAM) to enhance robustness, while its IAT module employs semantic-consistent energy minimization (SC-MHE) to preserve generalization. We show that co-optimizing logit margins and semantic alignments is crucial for building robust, generalizable models.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62471420), Guang-Dong Basic and Applied Basic Research Foundation (2025A1515012296), and 2025 Tencent AI Lab Rhino-Bird Program. Hei Victor Cheng is supported in part by the Aarhus Universitets Forskningsfond project number AUFF 39001 and the NordForsk Nordic University Cooperation on Edge Intelligence (Grant No. 168043).

References

Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Fu, C.; Gopalakrishnan, K.; Hausman, K.; et al. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. *arXiv preprint arXiv:2204.01691*.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–Mining Discriminative Components with Random Forests. In *European Conference on Computer Vision (ECCV)*, 446–461. Springer.

Carlini, N.; and Wagner, D. 2017. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57. IEEE.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing Textures In The Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3606–3613.

Coates, A.; Ng, A.; and Lee, H. 2011. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 215–223. JMLR Workshop and Conference Proceedings.

Croce, F.; and Hein, M. 2020. Reliable Evaluation of Adversarial Robustness with An Ensemble of Diverse Parameter-Free Attacks. In *International Conference on Machine Learning (ICML)*, 2206–2216. PMLR.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255. IEEE.

Ding, G. W.; Sharma, Y.; Lui, K. Y. C.; and Huang, R. 2018. MMA Training: Direct Input Space Margin Maximization through Adversarial Training. In *International Conference on Learning Representations (ICLR)*.

Fazlyab, M.; Entesari, T.; Roy, A.; and Chellappa, R. 2024. Certified Robustness via Dynamic Margin Maximization and Improved Lipschitz Regularization. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2006. One-Shot Learning of Object Categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4): 594–611.

Griffin, G.; Holub, A.; Perona, P.; et al. 2007. Caltech-256 Object Category Dataset. Technical report, Technical Report 7694, California Institute of Technology Pasadena.

Inkawhich, N.; McDonald, G.; and Luley, R. 2023. Adversarial Attacks on Foundational Vision Models. *arXiv preprint arXiv:2308.14597*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision. In *International Conference on Machine Learning (ICML)*, 4904–4916. PMLR.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d Object Representations For Fine-Grained Categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 554–561.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning Multiple Layers of Features From Tiny Images.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial Examples in The Physical World. In *Artificial Intelligence Safety and Security*, 99–112. Chapman and Hall/CRC.

Li, H.; Wan, H.; Zhang, L.; Jiu, M.; Li, S.; Xu, M.; and Khan, M. H. 2025a. Towards Robust Multimodal Domain Generalization via Modality-Domain Joint Adversarial Training. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACMMM)*, 180–188.

Li, X.; Chen, H.; and Hu, X. 2025. On the Importance of Backbone to the Adversarial Robustness of Object Detectors. *IEEE Transactions on Information Forensics and Security (TIFS)*.

Li, X.; Liu, Y.; Dong, N.; Qin, S.; and Hu, X. 2024a. Partimagenet++ Dataset: Scaling Up Part-Based Models For Robust Recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 396–414. Springer.

Li, X.; Zhang, W.; Liu, Y.; Hu, Z.; Zhang, B.; and Hu, X. 2024b. Language-Driven Anchors For Zero-Shot Adversarial Robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24686–24695.

Li, X.; Zhu, Y.; Huang, Y.; Zhang, W.; He, Y.; Shi, J.; and Hu, X. 2025b. PBCAT: Patch-Based Composite Adversarial Training Against Physically Realizable Attacks on Object Detection. *arXiv preprint arXiv:2506.23581*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv preprint arXiv:1706.06083*.

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 *The Eleventh International Conference on Learning Representations (ICLR)*.

Michels, F.; Adaloglou, N.; Kaiser, T.; and Kollmann, M. 2023. Contrastive Language-Image Pretrained (CLIP) Models are Powerful Out-of-Distribution Detectors. *arXiv preprint arXiv:2303.05828*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated Flower Classification Over a Large Number of Classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 722–729. IEEE.

Noever, D. A.; and Noever, S. E. M. 2021. Reading Isn't Believing: Adversarial Attacks On Multi-Modal Neurons. *arXiv preprint arXiv:2103.10480*.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and Dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (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 *International Conference on Machine Learning (ICML)*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Tong, B.; Lai, H.; Pan, Y.; and Yin, J. 2025. On the Zero-shot Adversarial Robustness of Vision-Language Models: A Truly Zero-shot and Training-free Approach. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, 19921–19930.

Wang, J.; Hu, X.; Hou, W.; Chen, H.; Zheng, R.; Wang, Y.; Yang, L.; Huang, H.; Ye, W.; Geng, X.; et al. 2023. On the Robustness of Chatgpt: An Adversarial and Out-Of-Distribution Perspective. *arXiv preprint arXiv:2302.12095*.

Wang, S.; Zhang, J.; Yuan, Z.; and Shan, S. 2024. Pre-trained Model Guided Fine-Tuning for Zero-Shot Adversarial Robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24502–24511. IEEE.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun Database: Large-Scale Scene Recognition From Abbey to Zoo. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 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 *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, 15172–15182.

Xu, Y.; Sun, Y.; Goldblum, M.; Goldstein, T.; and Huang, F. 2023. Exploring and Exploiting Decision Boundary Dynamics for Adversarial Robustness. *arXiv preprint arXiv:2302.03015*.

Yang, W.; and Mirzasoleiman, B. 2023. Robust Contrastive Language-Image Pretraining against Adversarial Attacks. *arXiv preprint arXiv:2303.06854*.

Zhao, Y.; Pang, T.; Du, C.; Yang, X.; Li, C.; Cheung, N.-M.; and Lin, M. 2023. On Evaluating Adversarial Robustness of Large Vision-Language Models. *arXiv preprint arXiv:2305.16934*.