When CLIP Sees More, It Fights Back Harder: Multi-View Guided Adaptive Counterattacks for Test-Time Adversarial Robustness

Supplementary Material

6. Limitations

Although MAC demonstrates consistent robustness improvements across 20 diverse datasets without any retraining, it is inherently dependent on the choice of the augmentation distribution and the multi-view configuration. In our experiments, we adopt a set of low-level image transformations and a small number of views, which may not be optimal for all visual domains or attack types. Consequently, the quality and diversity of the generated views can bound the achievable robustness, particularly in specialized modalities (e.g., medical or remote-sensing imagery) where appropriate semantics-preserving transformations differ from those of natural images. Exploring richer, domain-aware, or even learnable augmentation strategies that can automatically adapt the multi-view generation process to dataset- or task-specific characteristics is an important direction for future work.

7. Datasets

As our method operates in a test-time setting, all experiments are performed strictly on the test splits without using training data. We evaluate our method on **20 datasets**:

- **Caltech101** [15]: Object category dataset with 101 classes and 2,465 test images.
- **DTD** [10]: Texture classification dataset with 47 classes and 1,692 test images.
- **Flower102** [44]: Flower species dataset with 102 classes and 2,463 test images.
- **Pets** [46]: Pet image dataset with 37 classes and 3,669 test images.
- **UCF101** [64]: Human action frame-based dataset with 101 classes and 3,783 test images.
- **Aircraft** [39]: Fine-grained aircraft recognition dataset with 100 classes and 3,333 test images.
- **EuroSAT** [18]: Satellite imagery dataset with 10 classes and 8,100 test images.
- **Cars** [31]: Fine-grained car model classification dataset with 196 classes and 8,041 test images.
- **SUN397** [78]: Scene recognition dataset with 397 classes and 19,850 test images.
- **Food101** [7]: Food image dataset containing 101 classes and 30,300 test images.
- **ImageNet** [14]: Large-scale object classification dataset with 1,000 classes and 50,000 test images.
- **ImageNet-A** [21]: Natural adversarial images with 200 classes and 7,500 test images.
- **ImageNet-V2** [50]: Re-collected ImageNet samples with 1,000 classes and 10,000 test images.
- **ImageNet-R** [20]: Artistic renditions of ImageNet classes with 200 classes and 30,000 test samples.
- **ImageNet-S** [71]: Sketch images with 1,000 classes and 50,889 test images.
- **CIFAR10** [32]: 10-class natural image dataset with 10,000 test images.
- **CIFAR100** [32]: 100-class natural image dataset with 10,000 test images.
- **STL10** [11]: Image classification dataset with 10 classes and 8,000 test images.
- **Caltech256** [17]: Object recognition dataset with 256 classes and 30,607 test images.
- **Country211** [48]: Geographic location prediction dataset with 211 classes and 21,100 test images.

8. Implementation Details

Backbones and Baselines. We use pretrained CLIP ResNet50, CLIP ViT-B/32, CLIP ViT-B/16, and CLIP ViT-L/14 as the vision-language backbones, and all CLIP parameters are frozen without any tuning process. The CLIP text prompt is formulated as "a photo of [CLASS]". We use CLIP ViT-B/32 as the default setting for all experiments, unless mentioned otherwise. For fair comparisons, all compared methods are evaluated under the same CLIP model and strong attack scenario, using their officially released implementations and default hyperparameters.[1] Among the compared methods, we conduct TAPT [74] experiments with CLIP ViT-B/16, as the method provides the necessary precomputed values exclusively for this backbone.

Augmentation and counterattack. Inspired by AugMix [19], the augmentation distribution $\mathcal{T}$ applies a mild random affine transformation (rotation $\leq 5°$, translation $\leq 4\%$, scale jitter $\pm 8\%$) to every image, followed by additional augmentations, including Gaussian blur with standard deviation 0.5-1.2, additive Gaussian noise with standard deviation ≤ 0.02, and color jitter (brightness, contrast, and saturation ± 0.08; hue ± 0.02), each applied independently with a probability of 0.5. For counterattack, we set a reasonable ℓ_∞ budget of $\epsilon^{(\mathrm{ca})} = 8$ with short iterations (K=4) and step size of 2. Using this setting, we empirically confirm that the counterattack perturbations remain

[1] MTA: https://github.com/MaxZanella/MTA; TTC: https://github.com/Sxing2/CLIP-Test-time-Counterattacks; TAPT: https://github.com/xinwong/TAPT; R-TPT: https://github.com/TomSheng21/R-TPT

Table 5. Evaluation of various adversarial finetuning and test-time adaptation methods on five datasets using CLIP-ViT-B/32, reporting both clean accuracy (Acc.) and adversarial accuracy under PGD-10 attack with ϵ = 4.0 (Rob.). The highest score is highlighted in bold. In the adversarial fine-tuning methods, the superscripts represent the attack budgets used to generate adversarial images during adversarial fine-tuning.

Dataset	Metric	CLIP	Adversarial Fine-tuning							Test-time Defense					
			CLIP-FT	TeCoA^1	TeCoA^4	PMG-AFT^1	PMG-AFT^4	FARE^1	FARE^4	RN	TTE	Anti-adv	HD	TTC	Ours
CIFAR10	Acc.	**85.12**	84.90	64.61	65.15	70.69	71.45	74.44	78.46	81.18	84.74	83.44	78.23	81.18	83.84
	Rob.	0.43	2.75	7.69	11.7	10.20	15.59	1.94	5.42	0.00	3.47	0.32	1.67	28.51	**35.96**
CIFAR100	Acc.	57.14	**59.51**	35.96	36.30	40.32	41.51	46.67	47.38	56.34	58.61	53.96	52.86	56.34	56.70
	Rob.	0.05	0.67	6.54	9.25	7.60	10.80	2.64	4.54	0.00	1.37	0.22	0.00	9.06	**18.36**
STL10	Acc.	**96.40**	94.49	87.40	81.69	88.56	84.35	91.72	89.11	95.85	96.26	95.47	89.50	95.83	96.17
	Rob.	0.16	3.75	24.80	31.83	28.49	35.40	9.99	17.59	0.06	32.56	2.25	3.39	52.40	**73.49**
Caltech256	Acc.	81.72	78.53	61.14	52.05	62.24	53.32	73.32	67.22	81.25	**82.48**	79.40	79.12	76.59	81.57
	Rob.	0.12	1.41	8.29	11.76	10.65	13.68	2.18	5.09	0.16	23.23	1.44	0.34	27.25	**52.56**
Country211	Acc.	**15.25**	12.07	4.75	3.66	4.64	3.34	9.26	6.58	14.80	14.66	11.60	11.72	11.99	14.45
	Rob.	0.00	0.00	0.05	0.19	0.12	0.24	0.00	0.02	0.00	0.24	0.00	0.00	2.44	**3.78**
Average	Acc.	67.13	65.90	50.77	47.77	53.29	50.79	59.08	57.75	65.88	**67.35**	64.77	62.29	64.39	66.55
	Rob.	0.15	1.72	9.47	12.95	11.41	15.14	3.35	6.53	0.04	12.17	0.85	1.08	23.93	**36.83**

Algorithm 1 MAC: Multi-View Guided Adaptive Counterattack

Require: CLIP image encoder f, text encoder g, class prompts $\{\phi(c_j)\}$, test image x

Ensure: Robust prediction $\hat{y}$

1: Construct a multi-view $v = [v_0, \ldots, v_{N-1}]$ by applying $N-1$ stochastic augmentations to x.
2: Obtain view embeddings $z_i = f(v_i)$ for all i.
3: Initialize perturbations $\delta_i \sim \mathcal{U}[-\epsilon^{(\mathrm{ca})}, \epsilon^{(\mathrm{ca})}]$ and update δ_i through K steps of projected gradient ascent on $\|f(v_i+\delta_i) - z_i\|_2^2$ with step size α and ℓ_∞-projection $\|\delta_i\|_\infty \le \epsilon^{(\mathrm{ca})}$, yielding δ_i for each view.
4: Sample an augmentation T, compute normalized embeddings $\bar{z}_i = z_i/\|z_i\|_2$ and $\bar{z}'_i = f(T(v_i))/\|z_i\|_2$, set corruption degree $d_i = \|\bar{z}'_i - \bar{z}_i\|_2$, and obtain a soft weight $w_i = \sigma\big((d_i - \tau_{\mathrm{thres}})/\tau_{\mathrm{temp}}\big) \in [0,1]$ for all i.
5: Form adaptively counterattacked views $\tilde{v}_i = v_i + w_i\delta_i$ for all i.
6: Compute CLIP similarities between $f(\tilde{v}_i)$ and $\{g(\phi(c_j))\}$, average scores across all views, and set $\hat{y} = \arg\max_j \bar{s}_j$ where $\bar{s}_j$ is the aggregated similarity for class j.

visually imperceptible, as illustrated in Fig. 11.

Algorithm Algorithm 1 provides an overview of our proposed MAC method, detailing how multi-view guided counterattacks and corruption-aware soft weighting are incorporated during test-time inference.

Attack settings. For all attack settings, perturbations are applied to image tensors normalized to the $[0,1]$ range, where the perturbation budget is expressed in the $[0,1]$ domain as $\epsilon^{(\mathrm{atk})}/255$. We enable a random initialization and use one restart for the attack to ensure strong adversarial pressure. The detailed configurations are as follows:

- **PGD attack.** We use an L_∞-bounded PGD attack with a perturbation budget of $\epsilon^{(\mathrm{atk})} = 4$. The step size is set to $\epsilon^{(\mathrm{atk})}/4$, following common practice. We run 100 iterations for CLIP ViT models and 10 iterations for CLIP ResNet models.
- **DI^2-FGSM attack.** We use an L_∞-bounded DI^2-FGSM attack with the same $\epsilon^{(\mathrm{atk})}$, step size, and number of iterations as PGD. All diversity-transformation parameters follow standard settings, with a resize rate of 0.9, diversity probability of 0.5, and decay set to 0.0.
- **CW attack.** We adopt the L_2-bounded Carlini-Wagner attack with c=3.0, κ=0, 500 optimization steps, and a learning rate of 0.01.
- **AutoAttack.** We use the L_∞-bounded standard AutoAttack configuration with the same $\epsilon^{(\mathrm{atk})}$ as PGD, keeping all internal thresholds and settings at default values of their library.
- **MAC-adaptive attack.** To evaluate robustness under fully adaptive white-box threats, we design a pipeline-aware PGD attack that explicitly incorporates the MAC defense into the attack objective. At each iteration, the attacker forms a perturbed image, generates N views using the augmentation distribution $\mathcal{T}$, and then applies the multi-view guided counterattack with corruption-aware soft weighting to these views. The attack maximizes the cross-entropy loss between the aggregated logits over the counterattacked views and the ground-truth label, thereby increasing the model's prediction error and effectively implementing an expectation-over-transformations (EOT) objective [4]. To backpropagate through MAC, including its non-differentiable operations, we employ a BPDA-style straight-through estimator [3, 6]: the forward pass runs the full MAC pipeline, while the backward pass

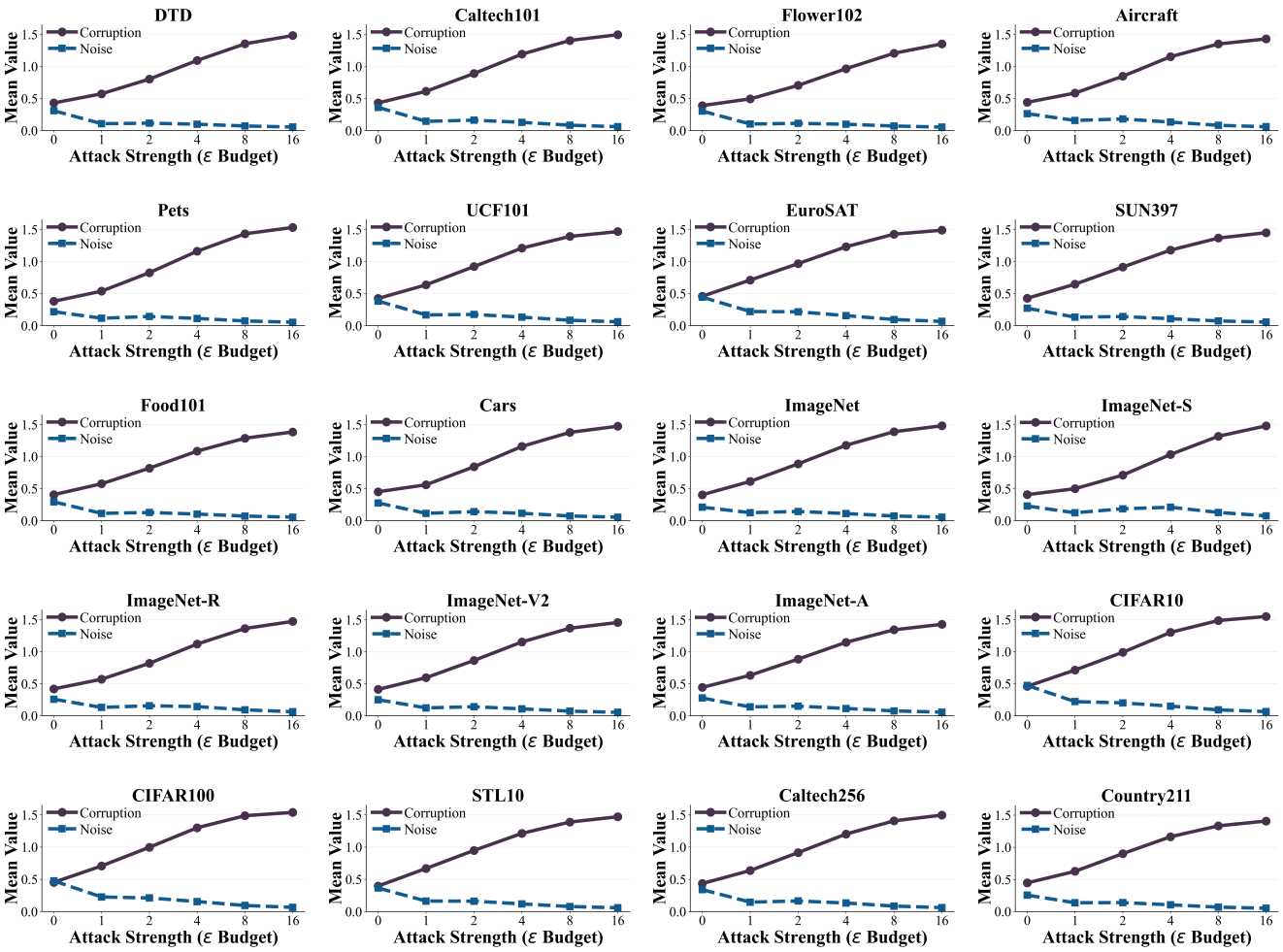

Figure 9. Comparison of the mean corruption degree of our MAC and the mean noise-driven deviation of TTC [81] under varying attack strengths $\epsilon^{(\mathrm{atk})}$ across 20 datasets. The corruption degree increases monotonically with stronger attacks and exhibits consistent trends across all datasets, whereas the noise-driven deviation shows non-monotonic behaviors. This demonstrates that our corruption degree provides a more reliable and stable indicator of corruption severity than the noise-driven deviation.

treats it as the identity transformation. The perturbation δ is updated using an L_∞-bounded PGD procedure with the same step size, budget, and number of iterations as in the standard PGD setting, resulting in a strictly defense-aware MAC-adaptive attack.

9. Experimental Results

Comparisons with adversarial fine-tuning methods. We compare our method with adversarial fine-tuning approaches and test-time defenses under strong PGD attack settings reported in [81]. All performance values for the compared methods are taken from the supplementary material of [81], where further implementation details can also be found. As shown in Tab. 5, adversarial fine-tuning approaches such as CLIP-FT [81], TeCoA [40], PMG-AFT [73], and FARE [56] offer moderate robustness gains but incur substantial degradation in clean accuracy and still yield limited adversarial performance. Test-time defense methods, including RN [81], TTE [47], Anti-adv [1], and HD [76], preserve high clean accuracy but provide only modest robustness improvements.

In contrast, MAC achieves the highest robustness across all datasets in the benchmark, reaching **36.83%** average adversarial accuracy while maintaining competitive clean accuracy (**66.55%**). These results demonstrate that MAC achieves a substantially better clean-robustness trade-off than both adversarial fine-tuning and prior test-time defense methods, without requiring any model fine-tuning.

Comparisons with TTC across additional datasets. As shown in Tab. 5, MAC also surpasses TTC on the five datasets, improving the average adversarial accuracy from **23.93%** to **36.83%**. When these results are considered together with the broader evaluations in Tab. 1 and Tab. 2,

Table 6. Comparison of clean accuracy (Acc.) and adversarial robustness (Rob.) of TTC and our MAC across different architectures, such as CLIP RN50, ViT-B/32, ViT-B/16, and ViT-L/14, and counterattack perturbation budgets $\epsilon^{(\text{ca})} \in \{4, 8, 16\}$. Results are averaged over ten fine-grained recognition datasets.

CLIP Architecture	CLIP [48] Acc.	CLIP [48] Rob.	$\epsilon^{(\text{ca})}$	TTC [81] Acc.	TTC [81] Rob.	MAC (Ours) Acc.	MAC (Ours) Rob.
RN50	54.5	0.1	4	52.0	0.8	**53.6**	**22.0**
			8	50.1	2.9	**53.4**	**34.8**
			16	47.6	23.9	**53.3**	**38.7**
ViT-B/32	58.9	0.0	4	56.6	6.8	**58.7**	**33.7**
			8	56.2	3.9	**58.7**	**45.2**
			16	55.8	1.3	**58.6**	**46.2**
ViT-B/16	62.7	0.0	4	60.5	4.8	**62.3**	**48.7**
			8	59.8	0.2	**62.3**	**58.6**
			16	58.7	11.1	**62.3**	**58.8**
ViT-L/14	69.9	0.0	4	68.6	4.6	**69.5**	**57.3**
			8	68.5	0.8	**69.4**	**64.3**
			16	67.2	25.8	**69.3**	**64.3**

MAC is consistently superior to TTC across all *20 datasets* evaluated. Taken collectively, these findings establish MAC as a strictly stronger counterattack mechanism than TTC across fine-grained, large-scale, and OOD benchmarks.

Stability across architectures and perturbation budgets. We analyze the behavior of TTC and MAC across different architectures and counterattack perturbation budgets $\epsilon^{(\text{ca})}$, as shown in Tab. 6. While TTC exhibits inconsistent behavior across perturbation levels, where its clean accuracy and adversarial robustness fluctuate significantly within the same architecture as $\epsilon^{(\text{ca})}$ varies, MAC demonstrates both stable clean performance and steadily increasing robustness as the $\epsilon^{(\text{ca})}$ increases. This suggests that TTC's noise-driven hard gating mechanism struggles to reliably distinguish between clean and adversarial inputs, leading to unstable accuracy and robustness. In contrast, MAC leverages the estimated corruption degree as a reliable indicator of corruption severity and incorporates it into a corruption-aware soft weighting scheme, which adaptively handles various input conditions, including clean, weakly perturbed, and strongly perturbed images. As a result, MAC achieves consistently higher and more stable performance than TTC across various architectures and counterattack strengths. We further investigate the underlying cause of this stability gap in the following paragraph by directly comparing MAC's corruption degree with TTC's noise-driven deviation.

Comparing the corruption degree of MAC and the noise-driven deviation of TTC. For each image, we use the *corruption degree* as an indicator of corruption severity to adaptively scale the counterattack intensity. Images with a degree above the threshold receive proportionally stronger counterattacks, while those below the threshold are only lightly updated. In contrast, TTC [81] employs a *noise-driven deviation* as a binary trigger: only images with deviation below a fixed threshold are counterattacked.

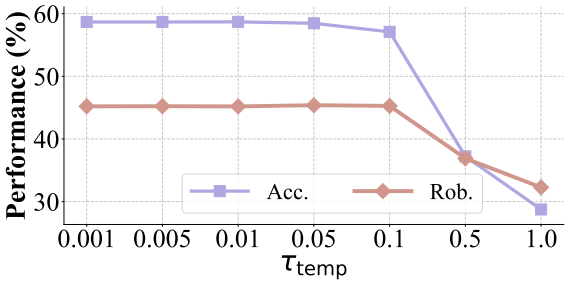

Figure 10. Analysis of the temperature parameter τ_{temp}, averaged over ten fine-grained recognition datasets. Moderate temperature values yield stable and near-optimal performance, while large temperatures reduce adaptivity and degrade overall performance.

To examine how faithfully each metric reflects actual corruption, we compare the behaviors of corruption degree and noise-driven deviation across 20 datasets under varying attack strengths, as shown in Fig. 9. The corruption degree of MAC increases *monotonically* with attack strength and exhibits consistently similar curve shapes across all datasets, demonstrating that it serves as a stable and reliable corruption-severity indicator. In contrast, the noise-driven deviation of TTC shows strongly *non-monotonic* trends. As the attack strength increases, the deviation does not consistently decrease; instead, it may rise, fall, or plateau depending on the dataset. This instability is especially observed in distribution-shifted benchmarks such as ImageNet-S, where the deviation increases from $\epsilon^{(\text{atk})}$=1 to $\epsilon^{(\text{atk})}$=4, but then decreases again at $\epsilon^{(\text{atk})}$=8. Such irregular patterns indicate that noise-driven deviation fails to reliably measure corruption severity.

This discrepancy arises because noise-driven deviation of TTC is insensitive to structured, geometric, and semantic variations introduced by adversarial perturbations, whereas our augmentation-based corruption degree captures these structural changes by leveraging affine transformations, blur, and color jitter. This fundamental difference directly explains the stability gap observed in Tab. 6: MAC's corruption degree provides a reliable signal that supports stable behavior across architectures and counterattack perturbation budgets, whereas TTC's deviation leads to unstable clean accuracy and robustness. Consequently, incorporating corruption degree into MAC yields substantial robustness improvements (**+37.7%** on average) compared to incorporating noise-driven deviation into MAC, as validated by the ablation results in Fig. 8.

Impact of temperature parameter τ_{temp}. We further analyze the effect of the temperature parameter τ_{temp}, which controls the softness of the corruption-aware soft weighting. As shown in Fig. 10, large temperatures ($\tau_{\text{temp}} > 0.1$)

Table 7. Evaluation of various test-time adaptation methods on eight fine-grained recognition datasets using CLIP-ViT-L/14, reporting both clean accuracy (Acc.) and adversarial accuracy under PGD-100 attack with $\epsilon = 4.0$ (Rob.). The highest score is highlighted in bold. Δ Rob. represents the robust accuracy gain of our method over the best existing tuning-free method.

Category	Method	Metric	Caltech101	DTD	Flower102	Pets	UCF101	Aircraft	EuroSAT	Cars	Average
Baseline	CLIP [48]	Acc.	95.2	52.4	**76.2**	93.1	73.7	30.0	**55.1**	76.8	**69.1**
		Rob.	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Tuning-based	R-TPT [61]	Acc.	95.7	**54.0**	**76.2**	**93.7**	74.3	31.7	44.3	77.2	68.4
		Rob.	88.2	38.0	55.6	72.9	55.6	17.2	20.4	49.1	49.6
Tuning-free	MTA [85]	Acc.	**95.8**	53.4	76.1	**93.7**	**74.7**	**32.7**	47.8	**78.4**	**69.1**
		Rob.	83.1	27.2	44.2	64.9	47.5	8.0	7.5	36.6	39.9
	TTC [81]	Acc.	93.5	51.0	74.5	92.2	72.7	27.9	45.9	74.6	66.5
		Rob.	9.9	5.3	6.9	6.4	2.5	0.4	0.1	3.2	4.3
	MAC (Ours)	Acc.	94.1	52.1	75.0	93.6	73.1	28.0	47.5	74.6	67.3
		Rob.	**92.7**	**45.3**	**70.4**	**87.0**	**68.7**	**31.3**	**25.6**	**69.8**	**61.4**
	Δ Rob.		**+9.6**	**+18.1**	**+26.2**	**+22.1**	**+21.2**	**+23.3**	**+18.1**	**+33.2**	**+21.5**

Table 8. Effect of removing each augmentation for multi-view generation. Results are averaged over ten fine-grained recognition datasets.

Augmentation Configuration	Acc.	Rob.
MAC w/o Affine Transform	58.6	6.0
MAC w/o Color Jitter	58.6	44.9
MAC w/o Gaussian Noise	58.4	44.7
MAC w/o Gaussian Blur	**58.8**	42.3
MAC	58.7	**45.2**

overly smooth the transition between clean and corrupted regimes, reducing the adaptivity of the counterattack and leading to noticeable drops in both clean and adversarial accuracy. In contrast, temperatures below 0.1 produce consistently strong performance, where the soft weighting provides a good balance between sensitivity and smoothness. Overall, MAC is not highly sensitive to the exact choice of τ_{temp} as long as it is not set high, and a moderate value yields stable and near-optimal results across datasets.

Comparisons with state-of-the-art on CLIP ViT-L/14. We further evaluate MAC and the compared methods on the strong backbone, CLIP ViT-L/14, across eight fine-grained recognition datasets. As shown in Tab. 7, MAC achieves the highest adversarial robustness on every dataset, reaching an average robust accuracy of **61.4%**. Interestingly, on the Aircraft dataset, MAC even achieves higher adversarial accuracy (31.3%) than clean accuracy (28.0%), suggesting that the counterattack can substantially restore corrupted representations and yield exceptional robustness in this domain. Across all datasets, MAC consistently improves robustness, with gains of up to **+33.2** points over the best existing tuning-free method. Furthermore, despite being a *tuning-free* method, MAC surpasses the strongest tuning-based baseline, R-TPT, by a considerable margin (**+11.8** points on average) while maintaining competitive clean accuracy. These results demonstrate that MAC delivers state-of-the-art robustness on CLIP ViT-L/14 without requiring any model tuning.

Ablation study of augmentation components. To understand the contribution of each augmentation in MAC, we remove individual components and evaluate the performance, as summarized in Tab. 8. Removing the affine transform causes a severe robustness collapse from 45.2% to 6.0%, showing that geometric variation is essential for generating multi-view inputs and estimating corruption degree. Other augmentations, such as color jitter, Gaussian noise, and Gaussian blur, provide moderate but meaningful gains, reducing robustness by 0.3 to 2.9 points when removed. Overall, the full set of augmentations yields the best performance, indicating that diverse geometric and photometric variations jointly strengthen MAC's multi-view guided counterattack and corruption degree estimation.

Visual appearance preservation under MAC To qualitatively assess how MAC affects the visual appearance of input images, we examine how the multi-view adversarial inputs are transformed after applying MAC. As illustrated in Fig. 11, we compare four visual states of each sample: the clean image, the adversarial input, the multi-view adversarial inputs generated by our augmentation distribution, and the multi-view adversarial inputs after MAC is applied. Because PGD perturbations are constrained to be imperceptible, the adversarial input shows no perceptible differences from the clean inputs, even though its embedding is significantly distorted. The multi-view adversarial inputs further include additional augmented variants (*e.g.*, affine transforms, blur, and color jitter) that reflect the diverse views MAC operates on. After applying our multi-view counterattack with corruption-aware soft weighting, the resulting images remain visually indistinguishable from both the clean and adversarial inputs; no noticeable color inconsistencies or structural distortions are introduced. Despite this per-

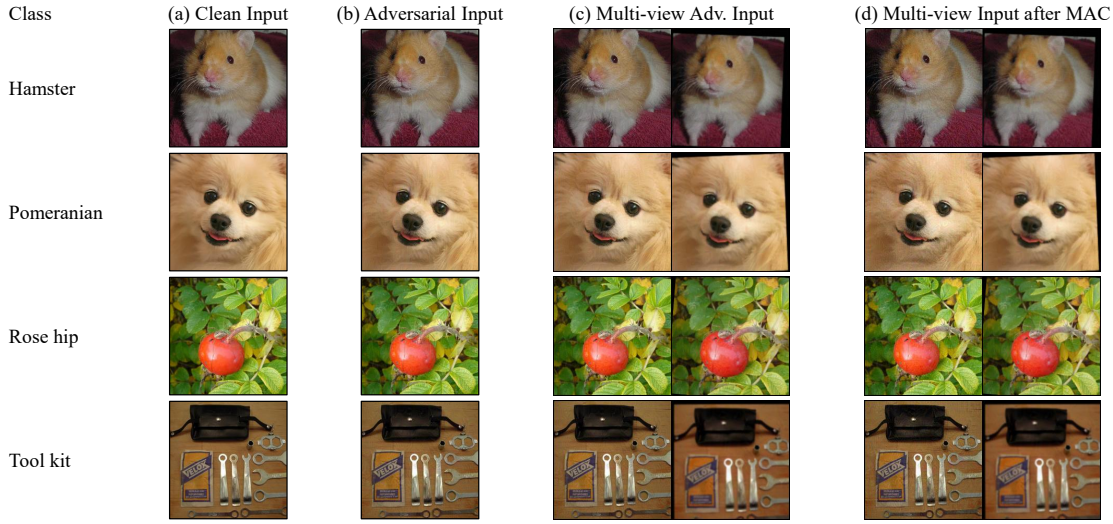

Figure 11. Qualitative visualization of MAC's visual preservation. For each example, we show four visual states: the clean image, the adversarial input, the multi-view adversarial inputs produced by our augmentation distribution, and the multi-view adversarial inputs after applying MAC. MAC not only preserves the visual quality of the input images but also effectively restores their embeddings toward the clean feature space, as shown in Fig. 5. This demonstrates that MAC preserves perceptual appearance while mitigating adversarial corruption.

ceptual invariance, the underlying representations are substantially restored toward the clean feature space, as shown in Fig. 5. This demonstrates that MAC effectively mitigates adversarial corruption in the feature space while fully preserving the perceptual quality of the input images.

Since the attacker's configuration is unknown to the defense model, MAC adopts counterattack settings that are independent of the attacker. Although a large $\epsilon^{(\mathrm{ca})}$ does not affect external inputs as MAC's perturbations operate internally within the defense pipeline, we simply choose a conservative ℓ_∞ budget of $\epsilon^{(\mathrm{ca})} = 8$ with a small number of iterations (K=4). This configuration reliably preserves visual quality while still providing strong robustness.

References

[1] Motasem Alfarra, Juan C Pérez, Ali Thabet, Adel Bibi, Philip HS Torr, and Bernard Ghanem. Combating adversaries with anti-adversaries. In *AAAI*, 2022. 3

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *ECCV*, 2020. 2

[3] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018. 7, 2

[4] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *ICML*, 2018. 2

[5] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In *IJCAI*, 2021. 2

[6] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 7, 2

[7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 5, 1

[8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 2017. 2, 7

[9] Xinyu Chen, Haotian Zhai, Can Zhang, Xiupeng Shi, and Ruirui Li. Multi-cache enhanced prototype learning for test-time generalization of vision-language models. In *ICCV*, 2025. 3

[10] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 5, 1

[11] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011. 5, 1

[12] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, 2019. 2

[13] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020. 2, 7

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5, 1

[15] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR workshop*, 2004. 5, 1

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 2, 3

[17] Gregory Griffin, Alex Holub, Pietro Perona, et al. Caltech-256 object category dataset. Technical report, Technical Report 7694, California Institute of Technology Pasadena, 2007. 5, 1

[18] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 5, 1

[19] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *ICLR*, 2020. 6, 1

[20] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 5, 1

[21] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 5, 1

[22] Fanding Huang, Jingyan Jiang, Qinting Jiang, Hebei Li, Faisal Nadeem Khan, and Zhi Wang. Cosmic: Clique-oriented semantic multi-space integration for robust clip test-time adaptation. In *CVPR*, 2025. 3

[23] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, 2018. 2

[24] Chanhwi Jeong, Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Test-time prompt tuning for zero-shot depth completion. In *ICCV*, 2025. 3

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

[26] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *CVPR*, 2024. 3

[27] Sunoh Kim, Taegil Ha, Kimin Yun, and Jin Young Choi. Swag-net: Semantic word-aware graph network for temporal video grounding. In *ACM CIKM*, 2022. 2

[28] Sunoh Kim, Jungchan Cho, Joonsang Yu, YoungJoon Yoo, and Jin Young Choi. Gaussian mixture proposals with pull-push learning scheme to capture diverse events for weakly supervised temporal video grounding. In *AAAI*, 2024. 1

[29] Sunoh Kim, Daeho Um, HyunJun Choi, and Jin Young Choi. Learnable negative proposals using dual-signed cross-entropy loss for weakly supervised video moment localization. In *MM*, 2024. 2

[30] Sunoh Kim, Kimin Yun, and Daeho Um. Finding optimal video moment without training: Gaussian boundary optimization for weakly supervised video grounding. *IEEE Transactions on Multimedia*, 2026. 2

[31] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshops*, 2013. 5, 1

[32] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5, 1

[33] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018. 2

[34] Lin Li, Haoyan Guan, Jianing Qiu, and Michael Spratling. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *CVPR*, 2024. 1, 3

[35] Xiao Li, Wei Zhang, Yining Liu, Zhanhao Hu, Bo Zhang, and Xiaolin Hu. Language-driven anchors for zero-shot adversarial robustness. In *CVPR*, 2024. 1

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[37] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008. 7

[38] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 2, 3, 4, 6

[39] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5, 1

[40] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *ICLR*, 2023. 1, 3

[41] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016. 2

[42] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, 2017. 2

[43] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *ICML*, 2022. 2

[44] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian conference on computer vision, graphics & image processing*, 2008. 5, 1

[45] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, 2016. 2

[46] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 5, 1

[47] Juan C Pérez, Motasem Alfarra, Guillaume Jeanneret, Laura Rueda, Ali Thabet, Bernard Ghanem, and Pablo Arbeláez. Enhancing adversarial robustness via test-time transformation ensembling. In *ICCV*, 2021. 3

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 3, 5, 6, 7, 4

[49] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 1

[50] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 5, 1

[51] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, 2020. 2

[52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1

[53] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *NeurIPS*, 2019. 2

[54] Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. In *NeurIPS*, 2020.

[55] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *ICLR*, 2018. 2

[56] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *ICML*, 2024. 1, 3

[57] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *NeurIPS*, 2020. 2

[58] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *NeurIPS*, 2019. 2

[59] Ashshak Sharifdeen, Muhammad Akhtar Munir, Sanoojan Baliah, Salman Khan, and Muhammad Haris Khan. O-tpt: Orthogonality constraints for calibrating test-time prompt tuning in vision-language models. In *CVPR*, 2025. 3

[60] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multimodal language models. In *ICLR*, 2023. 3

[61] Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *CVPR*, 2025. 2, 3, 5, 6, 7

[62] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022. 2, 3

[63] Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. In *ACL*, 2022. 1, 2

[64] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2012. 5, 1

[65] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 2019. 2

[66] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020. 3

[67] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 2, 3

[68] Devavrat Tomar, Guillaume Vray, Behzad Bozorgtabar, and Jean-Philippe Thiran. Tesla: Test-time self-learning with automatic adversarial augmentation. In *CVPR*, 2023. 3

[69] Baoshun Tong, Hanjiang Lai, Yan Pan, and Jian Yin. On the zero-shot adversarial robustness of vision-language models: A truly zero-shot and training-free approach. In *CVPR*, 2025. 3

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

[71] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 5, 1

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

[73] Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *CVPR*, 2024. 1, 3

[74] Xin Wang, Kai Chen, Jiaming Zhang, Jingjing Chen, and Xingjun Ma. Tapt: Test-time adversarial prompt tuning for robust inference in vision-language models. In *CVPR*, 2025. 2, 3, 6, 7, 1

[75] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *ICLR*, 2020. 2

[76] Boxi Wu, Heng Pan, Li Shen, Jindong Gu, Shuai Zhao, Zhifeng Li, Deng Cai, Xiaofei He, and Wei Liu. Attacking adversarial attacks as a defense. *arXiv preprint arXiv:2106.04938*, 2021. 2, 3

[77] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *NeurIPS*, 2020. 2

[78] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 5, 1

[79] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferabil-

ity of adversarial examples with input diversity. In *CVPR*, 2019. 2, 7

[80] Peng Xie, Yequan Bie, Jianda Mao, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. Chain of attack: On the robustness of vision-language models against transfer-based adversarial attacks. In *CVPR*, 2025. 2

[81] Songlong Xing, Zhengyu Zhao, and Nicu Sebe. Clip is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of clip. In *CVPR*, 2025. 2, 3, 4, 5, 6, 7, 8

[82] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *CVPR*, 2023. 1

[83] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *ICLR*, 2021. 2

[84] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 1

[85] Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *CVPR*, 2024. 3, 6, 7, 5

[86] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019. 2

[87] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2024. 1

[88] Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models. In *ECCV*, 2024. 1, 3

[89] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. In *ECCV*, 2022. 3

[90] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *NeurIPS*, 2023. 3

[91] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 1

[92] Lihua Zhou, Mao Ye, Shuaifeng Li, Nianxin Li, Xiatian Zhu, Lei Deng, Hongbin Liu, and Zhen Lei. Bayesian test-time adaptation for vision-language models. In *CVPR*, 2025. 3

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

[94] Xingyu Zhu, Beier Zhu, Shuo Wang, Kesen Zhao, and Hanwang Zhang. Enhancing clip robustness via cross-modality alignment. In *NeurIPS*, 2025. 3