BACKDOOR AS PROBE: TEST-TIME ADVERSARIAL DEFENSE FOR CLIP

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adversarial defense improves the robustness of vision-language foundation models such as CLIP without retraining. However, adversarial activation shifts are typically treated as distortions to suppress, rather than signals to exploit. We turn these shifts into defense signals by repurposing the trigger-to-target mechanism of backdoors. The key is to implant a defender-controlled backdoor as a probe that is weakly activated by clean inputs but strongly activated by adversarial shifts. Based on this insight, we propose *Backdoor as Probe* (BaP), a test-time adversarial defense for CLIP. BaP constructs the probe through a closed-form model edit to a selected MLP layer. It projects the average adversarial activation shift and a defender-specified semantic direction onto the layer's low-energy input and output activation subspaces to obtain the trigger and target directions, respectively. At inference time, adversarial inputs produce measurable responses along the target direction for detection. BaP then selectively rectifies detected inputs by optimizing a small perturbation that steers their representations away from adversarial shifts and toward the clean subspace. Experiments across 16 benchmarks show that BaP improves average robust accuracy from 1.0% to 52.3% while retaining clean accuracy, achieving performance comparable to state-of-the-art methods with up to a $5.7\times$ inference speedup. BaP further shows the generalization to adversarial attacks on large vision-language models.

1 INTRODUCTION

CLIP has become a widely used foundation model for vision-language models (VLMs) (Radford et al., 2021; Li et al., 2023; Bai et al., 2023). However, it remains vulnerable to adversarial perturbations (Cui et al., 2024; Nie et al., 2026b). To improve its robustness, adversarial fine-tuning (AFT) (Mao et al., 2023; Wang et al., 2024b) offers a straightforward approach by training the model on adversarial samples, but it requires substantial training data and often compromises clean accuracy. Recent test-time defenses (Xing et al., 2025; Liu et al., 2026) provide a more efficient solution by rectifying adversarial samples during inference. Nevertheless, existing test-time defenses primarily treat adversarially induced activation shifts as distortions to suppress (Pérez et al., 2021). Yet these shifts encode how adversarial representations depart from clean ones and can therefore serve as signals. This motivates a different question: can we leverage the shifts as defense signals?

To answer this question, we revisit backdoors from a defender's perspective. Backdoors are typically viewed as threats (Gu et al., 2017). An attacker implants a trigger-to-target mapping that causes the model to produce an attacker-specified output while retaining normal behavior on clean inputs. Although designed for malicious purposes, this selective mapping can instead couple an attack-sensitive activation with a defender-controlled response, thereby serving as a diagnostic probe. Prior studies have explored related ideas for adversarial defense in discriminative classifiers (He et al., 2016). Trapdoor implants class-specific honeypots that attract adversarial optimization toward recognizable activation signatures (Shan et al., 2020), while AI-Shielder combines a controlled backdoor with label-space mapping to recover predictions from adversarial inputs (Zhu et al., 2026). However, both methods are class-specific, making them difficult to extend directly to open-vocabulary CLIP models.

Repurposing backdoors for CLIP is not well studied due to three challenges. First, the implanted probe must be class-agnostic so that it remains effective in the open-vocabulary scenario. Second, it must remain inactive on clean inputs while producing a strong response to adversarial inputs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

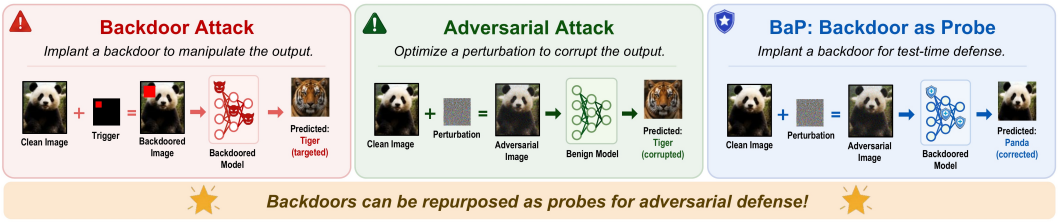


Figure 1: Comparison of backdoor attacks, adversarial attacks, and BaP. Backdoor attacks implant triggers to induce attacker-specified behavior, whereas adversarial attacks perturb inputs to corrupt model predictions. BaP instead implants a defender-controlled probe for test-time defense.

Third, the probe must be implanted efficiently, as the backdoor methods for CLIP commonly rely on large-scale fine-tuning (Liu & Zhang, 2025) which incurs high computational cost.

To address these challenges, we propose ***B**ackdoor **a**s **P**robe* (**BaP**), a test-time adversarial defense for CLIP. Inspired by backdoor-editing methods (Li et al., 2024b; Wang et al., 2024a), BaP constructs a defender-controlled probe within a selected MLP layer. Using clean calibration data, BaP identifies low-energy subspaces of the layer's input and output activations. It first projects the average activation shift between paired adversarial and clean examples onto the input-side low-energy subspace, yielding the trigger direction. It then projects a defender-specified semantic direction onto the output-side low-energy subspace, yielding the target direction. BaP applies a closed-form model edit that maps the trigger direction to the target direction without training. Because the trigger is derived from activation shifts rather than class labels, the resulting probe is class-agnostic. At inference time, BaP measures each input's response along the implanted target direction and uses its magnitude as a detection metric. For inputs identified as suspicious, BaP optimizes a small perturbation that first moves its representation away from the current input feature and then pulls it toward the clean subspace. Fig. 1 contrasts conventional backdoor attacks and adversarial attacks with our method.

Experiments across 16 zero-shot image-classification datasets show that BaP improves average robust accuracy from 1.0% to 52.3% while retaining clean accuracy. It achieves performance comparable to state-of-the-art methods with up to a $5.7\times$ inference speedup. BaP also generalizes across different CLIP backbones and attack algorithms, and further shows the generalization against adversarial attacks on large vision-language models.

Our main contributions are summarized as follows.

- We introduce BaP, a test-time defense that repurposes a defender-controlled backdoor as an internal diagnostic probe. To our knowledge, BaP is the first backdoor-based adversarial defense for open-vocabulary models.
- We show that constraining the trigger and target directions to low-energy activation subspaces yields an effective probe. BaP implants the probe through a closed-form edit and uses the target-response magnitude for selective test-time detection and rectification.
- We conduct experiments across 16 zero-shot classification datasets. BaP substantially improves adversarial robustness while preserving clean accuracy, and generalizes across CLIP backbones and attack algorithms.

2 Preliminaries and Related Work

Zero-Shot Classification with CLIP. By leveraging large-scale image–text pretraining, CLIP exhibits strong open-vocabulary zero-shot classification capabilities (Radford et al., 2021). Formally, let f_θ and g_ϕ denote the image and text encoders of CLIP, respectively. Given an image x and a set of class-specific text prompts $\{t_c\}_{c=1}^{C}$, CLIP predicts the class $\hat{c}$ whose text representation has the highest cosine similarity with the image representation:

$$\hat{c} = \arg\max_c \left\langle \frac{f_\theta(x)}{|f_\theta(x)|_2}, \frac{g_\phi(t_c)}{|g_\phi(t_c)|_2} \right\rangle. \tag{1}$$

Adversarial Attacks on CLIP. Despite its strong zero-shot generalization, CLIP remains vulnerable to adversarial perturbations (Madry et al., 2018). Given a clean image x with label y, an attacker

constructs an adversarial sample $x_{\text{adv}} = x + \delta$ by introducing an imperceptible perturbation δ constrained by $||\delta||_\infty \leq \epsilon_{\text{adv}}$. For zero-shot classification, the attack can be formulated as

$$\max_{||\delta||_\infty \leq \epsilon_{\text{adv}}} \mathcal{L}_{\text{CE}} \left(\{ \left\langle \frac{f_\theta(x+\delta)}{|f_\theta(x+\delta)|_2}, \frac{g_\phi(t_c)}{|g_\phi(t_c)|_2} \right\rangle \}_{c=1}^{C}, y \right), \tag{2}$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss over the zero-shot classification logits. Standard attacks such as projected gradient descent (PGD) (Madry et al., 2018) and AutoAttack (Croce & Hein, 2020) iteratively optimize the δ using input gradients while projecting it back onto a ℓ_∞-norm ball.

Adversarial Defenses for CLIP. Adversarial fine-tuning (AFT) improves CLIP robustness by updating model parameters on adversarial samples. TeCoA (Mao et al., 2023) is the first work to improve the zero-shot adversarial robustness of CLIP through text-guided contrastive adversarial training. Subsequent methods, including PMG-AFT (Wang et al., 2024b), FARE (Schlarmann et al., 2024), and Sim-CLIP+ (Hossain & Imteaj, 2024), extend AFT to broader settings and mitigate catastrophic forgetting. Another line of work focuses on adversarial prompt tuning. These methods learn robust visual or textual prompts while keeping the pretrained backbone frozen. Representative methods include APT (Li et al., 2024a), AdvPT (Zhang et al., 2024), FAP (Zhou et al., 2024), and COAPT (Wang et al., 2025b). Inspired by test-time adaptation (Shu et al., 2022; Abdul Samadh et al., 2023), recent methods improve robustness during inference without retraining the entire model. TAPT (Wang et al., 2025a) dynamically optimizes defensive visual and textual prompts for each test input, whereas R-TPT (Sheng et al., 2025) combines point-wise entropy minimization with reliability-weighted multi-view aggregation. C-TPT (Yoon et al., 2024), D-TPT (Han & Hwang, 2025), and COLA (Zhu et al., 2025) further develop this test-time prompt-tuning paradigm. Diffusion-based purification provides another solution but incurs substantial computational overhead (Zhang et al., 2025). Most relevant to our work are transformation-based test-time defenses. TTC (Xing et al., 2025) generates a reverse perturbation that moves the input away from the adversarial feature induced by an attack. CSR (Nie et al., 2026a) performs spectral contrastive rectification using low-frequency features as positive anchors and the unrectified input feature as a negative anchor. ET3 (Mirza et al., 2026) provides a lightweight transformation by directly minimizing CLIP's input energy. Unlike these approaches that regard the adversarial activation shifts as distortions to suppress, BaP leverages these shifts as trigger signals and associate them with an observable target.

Backdoors for Adversarial Defense. Previous studies have explored whether deliberately implanted backdoors can be repurposed for adversarial defense. Trapdoor embeds class-specific trapdoors that attract optimization-based attacks toward predefined activation signatures, enabling adversarial inputs to be detected (Shan et al., 2020). AI-Shielder implants label-dependent defensive backdoors and exploits a secret source-to-target class mapping to recover predictions from adversarial inputs (Zhu et al., 2026). Table 1 summarizes their capabilities. Despite these advances, existing methods only defend specific class sets. This dependence limits their applicability to open-vocabulary CLIP and motivates us to conduct class-agnostic detection and rectification.

Table 1: Comparison of backdoor-based adversarial defenses.

Method	Adversarial detection	Adversarial rectification	Open vocabulary
Trapdoor	✓	✗	✗
AI-Shielder	✗	✓	✗
BaP	✓	✓	✓

3 METHOD

3.1 THREAT MODEL AND METHOD OVERVIEW

Given a pretrained CLIP model consisting of an image encoder f_θ and a text encoder g_ϕ, we consider an attacker who perturbs a clean image x into $x^{\text{adv}} = x + \delta$, where $\|\delta\|_\infty \leq \epsilon$. The defender has white-box access to the model and is allowed to edit the model. At inference time, the defender aims to detect and rectify adversarial samples while preserving the model's performance on clean samples.

As shown in Fig. 2, BaP introduces a defender-controlled diagnostic probe inspired by backdoor mechanisms. The probe is carefully designed to produce a large response to adversarial samples but a small response to clean samples. The probe is then implanted into an MLP layer of the model through an edit. At inference time, the defender distinguishes adversarial samples by computing the

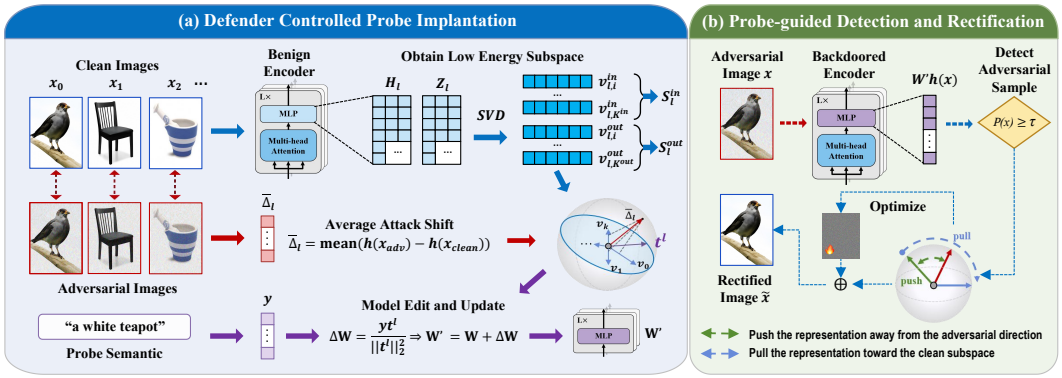


Figure 2: Overview of BaP. ***(a) Defender-controlled probe implantation.*** BaP constructs an attack-sensitive direction in a low-energy clean subspace and implants it through an edit. ***(b) Probe-guided detection and rectification.*** The implanted probe detects suspicious inputs and selectively guides their representations toward the clean subspace.

activation response of the probe. A perturbation is then optimized to rectify the adversarial sample by moving it away from the current adversarial feature and pulling it toward the clean subspace.

3.2 Defender-Controlled Probe Implantation

To achieve efficient and accurate probe implantation, inspired by backdoor-editing methods (Li et al., 2024b; Wang et al., 2024a), we implant the probe by applying an edit to the `fc2` weight W_l of the MLP at layer l of the encoder. Previous studies have shown that MLPs in Transformer architectures encode concept-specific representations that can be manipulated to control responses (Meng et al., 2022; Tamayo et al., 2024). Formally, let $h_l(x) \in \mathbb{R}^d$ denote the activation of the `CLS` token of image x at the input to this `fc2`. Given a clean calibration set $\mathcal{X} = \{x_i\}_{i=1}^N$, we collect the input activation matrix and the output activation matrix:

$$\begin{aligned} H_l &= [h_l(x_1), \ldots, h_l(x_N)]^\top \in \mathbb{R}^{N \times d_{in}}, \\ Z_l &= [W_l h_l(x_1), \ldots, W_l h_l(x_N)]^\top \in \mathbb{R}^{N \times d_{out}}, \end{aligned} \tag{3}$$

and perform singular value decomposition (SVD):

$$H_l = U_l \Sigma_l V_l^\top, \quad Z_l = U_l^{\text{out}} \Sigma_l^{\text{out}} (V_l^{\text{out}})^\top. \tag{4}$$

Let $\mathcal{I}_l^{\text{in}}$ denote the index set of the right singular vectors corresponding to the smallest K^{in} singular values in H_l, and let $\mathcal{I}_l^{\text{out}}$ denote the index set of the right singular vectors corresponding to the smallest K^{out} singular values. We define the input-side low-energy subspace and the output-side low-energy subspace as

$$\mathcal{S}_l^{\text{in}} = \text{span}\{v_{l,j}^{in} : j \in \mathcal{I}_l^{\text{in}}\}, \quad \mathcal{S}_l^{\text{out}} = \text{span}\left\{v_{l,j}^{\text{out}} : j \in \mathcal{I}_l^{\text{out}}\right\}. \tag{5}$$

These subspaces characterize the directions with the lowest activation energy of clean samples at the input and output of `fc2`. Constructing the probe within these subspaces therefore helps reduce interference with clean representations.

For the clean samples and their corresponding adversarial pairs $\{(x_i, x_i^{\text{adv}})\}_{i=1}^N$, we compute the average attack activation shift:

$$\bar{\Delta}_l = \frac{1}{N} \sum_{i=1}^{N} \left[h_l(x_i^{\text{adv}}) - h_l(x_i)\right]. \tag{6}$$

Let $\sigma_{l,j}^{in}$ denote the singular value of H_l associated with $v_{l,j}^{in}$, such that ${\sigma_{l,j}^{in}}^2$ measures the clean activation energy along $v_{l,j}^{in}$. For each $j \in \mathcal{I}_l^{\text{in}}$, we define:

$$a_j = \frac{{v_{l,j}^{in}}^\top \bar{\Delta}_l}{{\sigma_{l,j}^{in}}^2}, \qquad t_l = \nu \frac{\sum_{j \in \mathcal{I}_l^{\text{in}}} a_j v_{l,j}^{in}}{\left\| \sum_{j \in \mathcal{I}_l^{\text{in}}} a_j v_{l,j}^{in} \right\|_2}. \tag{7}$$

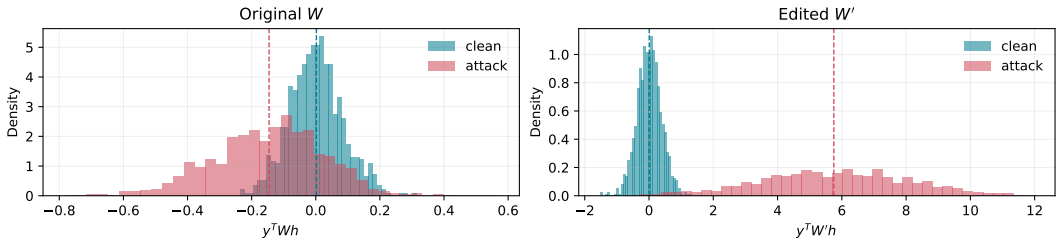


Figure 3: Visualization of the probe response $p(x)$ before and after the edit. It contains 1,000 clean and 1,000 adversarial images from STL-10 using CLIP ViT-B/16. The edited probe amplifies adversarial activation shifts into a separable response space.

This weighting emphasizes directions with large attack-induced shifts and low clean activation energy, where $\nu = 5$ is the predefined norm of the probe direction. In addition to the input-side probe direction, we further construct an output-side target direction for carrying and reading the probe response. Let p be a target semantic prompt specified by the defender, such as "a white teapot," and let its normalized text feature be $q = \frac{g_\phi(p)}{\|g_\phi(p)\|_2}$. Let A be the visual projection matrix of the CLIP vision encoder, and define the semantics-induced hidden-space direction as $y_0 = A^\dagger q$, where $A^\dagger$ denotes the Moore–Penrose pseudoinverse of A (Penrose, 1955).

We project the semantic direction y_0 onto this subspace as $\bar{y}_l = P_{\mathcal{S}_l^{\text{out}}}(y_0)$, where $P_{\mathcal{S}_l^{\text{out}}}(\cdot)$ is the orthogonal projection function, and normalize it to obtain the final target shift vector $y_l = \gamma \frac{\bar{y}_l}{\|\bar{y}_l\|_2}$, where γ is the predefined norm of the target shift. Finally, BaP binds the input-side trigger direction t_l to the output-side semantic target direction y_l through the following closed-form edit:

$$W_l' = W_l + \Delta W_l, \qquad \Delta W_l = \frac{y_l\, t_l^\top}{\|t_l\|_2^2}. \tag{8}$$

This edit satisfies $\Delta W_l t_l = y_l$. Therefore, when the input representation produces a strong response along the attack-sensitive direction t_l, the edited `fc2` produces a corresponding additional response along the output direction y_l specified by the defender. We denote the encoder after implantation by $f_{\theta'}$ and provide a more detailed explanation of the edit in Appendix B.

3.3 Probe-Based Adversarial Sample Detection

Let $\hat{y}_l = \frac{y_l}{\|y_l\|_2}$ be the unit output-side direction. For an input x, BaP obtains the input representation $h_l(x)$ of `fc2` at layer l and directly reads its response along $\hat{y}_l$ from the edited `fc2` output:

$$p(x) = \hat{y}_l^\top W_l' h_l(x). \tag{9}$$

Fig. 3 shows the distribution of the probe response $p(x)$ for 1,000 clean and 1,000 adversarial images from STL-10 using CLIP ViT-B/16. Before probe implantation, the two distributions overlap substantially. After implantation, clean responses remain concentrated near zero, whereas adversarial responses shift toward much larger values. This separation demonstrates that the edit produces a clean-inactive yet attack-sensitive probe.

In the end, the detection method is:

$$G(x) = \mathbb{I}[p(x) \geq \tau]. \tag{10}$$

Here, τ denotes the detection threshold. If and only if $G(x) = 1$, BaP identifies the input as a suspicious sample and performs the subsequent rectification process.

3.4 Adversarial Sample Rectification

BaP rectifies each flagged input in two stages, escape and repair. Let ξ denote the rectification perturbation, where $\|\xi\|_\infty \leq \epsilon_c$. We initialize ξ_0 and define the escape loss:

$$\mathcal{L}_{\text{esc}}(\xi) = \|f_{\theta'}(x+\xi) - f_{\theta'}(x)\|_2^2. \tag{11}$$

BaP performs two steps of projected gradient ascent:

$$\xi_{r+1} = \xi_r + \alpha \operatorname{sign}\left(\nabla_{\xi_r} \mathcal{L}_{\text{esc}}(\xi_r)\right), \qquad r = 0, 1, \tag{12}$$

where $\alpha = 2/255$. This stage moves the representation away from its current adversarial state.

BaP then guides the representation toward the clean activation distribution. Let $\mathcal{M}_l^{\text{clean}}$ denote the principal subspace of the clean activation matrix H_l, and let $P_{\mathcal{M}_l^{\text{clean}}}(\cdot)$ denote the orthogonal projection onto this subspace. We define

$$e_l(\cdot) = h_l'(\cdot) - P_{\mathcal{M}_l^{\text{clean}}}\left(h_l'(\cdot)\right) \tag{13}$$

as the residual component outside this subspace. We precompute a global clean direction v_l^{clean} as the normalized mean residual direction over the calibration set $\mathcal{X}$. The repair loss is:

$$\mathcal{L}_{\text{rep}}(\xi) = \underbrace{\|e_l(x+\xi)\|_2^2}_{\mathcal{L}_e} - \lambda \underbrace{\cos(e_l(x+\xi), v_l^{\text{clean}})}_{\mathcal{L}_u}, \tag{14}$$

where $\cos(\cdot,\cdot)$ is the cosine similarity. The first term reduces the distance from the clean subspace, while the second aligns the residual with the clean direction. Starting from ξ_2, BaP performs one projected gradient descent step:

$$\xi_3 = \xi_2 - \alpha \operatorname{sign}\left(\nabla_{\xi_2} \mathcal{L}_{\text{rep}}(\xi_2)\right), \qquad \tilde{x} = x + \xi_3. \tag{15}$$

The edited model then classifies $\tilde{x}$. Appendix A provides the complete algorithm.

4 Experiments

4.1 Experimental Setup

Datasets and Models. Following previous work, we evaluate BaP on a comprehensive benchmark consisting of 16 datasets. These datasets cover general object recognition, including ImageNet (Deng et al., 2009), CIFAR-10/100 (Krizhevsky, 2009), STL10 (Coates et al., 2011), Caltech-101/256 (Fei-Fei et al., 2004; Griffin et al., 2007); fine-grained classification, including OxfordPets (Parkhi et al., 2012), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), StanfordCars (Krause et al., 2013); scene recognition, including SUN397 (Xiao et al., 2010), Country211 (Radford et al., 2021); and domain-specific applications, including FGVCAircraft (Maji et al., 2013), EuroSAT (Helber et al., 2019), DTD (Cimpoi et al., 2014), PCAM (Veeling et al., 2018). For zero-shot image classification, we use the prompt "a photo of {}". We adopt CLIP ViT-B/16 as the default backbone and further evaluate BaP on CLIP ViT-B/32 and CLIP ViT-L/14. We also test the performance on CLIP ViT-L/14@336, which is the vision encoder of LLaVA (Liu et al., 2023).

Baselines. We compare BaP with state-of-the-art test-time defenses, including TTE (Pérez et al., 2021), HD (Wu et al., 2021), Anti-Adv (Alfarra et al., 2022), LPF (Ziyadinov & Tereshonok, 2023), TTC (Xing et al., 2025), R-TPT (Sheng et al., 2025) and ET3 (Mirza et al., 2026). All baselines are implemented with their original hyperparameter settings to ensure a fair comparison.

Implementation Details. Unless otherwise specified, we set the adversarial perturbation budget to $1/255$, with 10 attack steps for PGD and 50 attack steps for AutoAttack. Specifically, we use the targeted APGD variant of the AutoAttack. All experiments are conducted on 8 NVIDIA RTX 4090 GPUs. For BaP, we estimate the adversarial shift using $N = 1000$ PGD adversarial samples generated on ImageNet with a perturbation budget of $1/255$ and 10 attack steps. The rectification noise budget is set to $\epsilon_c = 4/255$, with a step size of $\alpha = 2/255$. We set the edited layer to $l = 6$, with $\nu = 5$, $\gamma = 40$, $K^{in} = 256$, $K^{out} = 32$, and $\lambda = 0.05$. For all datasets, we use a fixed detection threshold of $\tau = 1.93$.

4.2 Main Results

Results on 16 Datasets. Table 2 reports zero-shot adversarial robustness across 16 datasets. Compared with the original CLIP model, BaP improves the average robust accuracy from 1.0% to 52.3%. Among test-time defenses, R-TPT preserves the highest clean accuracy of 63.3%, but incurs high per-sample inference latency, as shown in Table 3. In contrast, BaP achieves the highest average

Table 2: Top-1 zero-shot accuracy (%) under 10-step PGD with $\ell_\infty = 1/255$. "Clean" and "Rob." denote accuracies on clean and adversarial samples, respectively. The final two columns report the performance of our BaP compared to the original CLIP.

	Dataset	Original		Test-Time Defense																Δ	
		CLIP		R-TPT		LPF		HD		Anti-Adv		TTE		TTC		ET3		BaP (Ours)			
Type	Name	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
General	ImageNet	63.9	0.0	66.7	51.0	58.1	30.5	59.7	4.1	61.5	23.9	66.2	23.2	40.9	27.8	58.6	10.2	60.3	39.6	-3.6	+39.6
	CIFAR10	88.1	0.5	81.6	69.2	89.0	40.4	84.1	11.8	82.8	63.5	85.5	29.8	90.0	28.2	77.7	29.2	86.8	59.1	-1.3	+58.6
	CIFAR100	59.6	0.1	51.8	36.4	63.4	19.6	57.6	7.9	51.5	34.9	60.4	14.2	63.1	11.1	50.0	13.9	57.5	34.1	-2.1	+34.0
	STL10	97.5	4.8	96.8	92.7	96.9	77.6	96.8	34.0	97.2	83.1	97.6	75.6	96.4	51.1	92.8	52.1	96.7	93.4	-0.8	+88.6
	Caltech101	83.5	1.3	86.1	80.9	81.3	65.6	82.6	28.4	82.3	58.2	87.2	61.2	75.8	31.3	80.3	41.1	80.6	76.2	-2.9	+74.9
	Caltech256	83.5	1.5	88.0	80.5	82.8	66.5	80.4	20.9	81.6	55.8	87.3	59.4	73.4	41.8	78.2	34.4	79.7	74.8	-3.8	+73.3
Fine-G	OxfordPets	88.9	0.0	85.8	68.3	79.4	41.7	84.2	4.0	86.5	36.7	84.8	12.5	78.8	27.0	79.5	16.0	83.4	65.7	-5.5	+65.7
	Flowers102	66.0	0.0	65.8	48.8	59.8	33.1	63.3	3.5	63.5	25.3	65.3	5.5	55.8	23.3	59.8	12.6	62.8	55.8	-3.2	+55.8
	Food101	84.8	0.0	86.6	68.9	79.6	38.6	86.0	1.1	83.9	29.3	85.3	24.8	57.5	33.2	79.8	10.4	81.2	66.1	-3.6	+66.1
	StanfordCars	65.2	0.0	68.6	45.4	54.0	16.9	57.8	1.4	62.5	13.7	59.0	14.4	46.6	20.2	56.9	5.7	58.8	56.4	-6.4	+56.4
Scene	SUN397	63.6	0.2	64.1	53.1	59.2	31.3	59.7	4.0	62.5	22.7	65.4	20.2	47.5	26.2	57.0	9.2	60.4	50.0	-3.2	+49.8
	Country211	17.0	0.0	19.2	8.9	14.7	2.5	15.1	0.0	15.2	1.9	15.8	0.3	10.2	4.4	13.3	1.1	16.7	14.0	-0.3	+14.0
Domain	FGVCAircraft	23.1	0.0	23.8	16.7	17.8	7.2	18.8	1.3	20.4	6.0	23.3	4.9	13.8	10.6	18.3	1.5	20.7	22.4	-2.4	+22.4
	EuroSAT	42.9	0.0	29.6	22.3	41.6	4.9	41.7	8.5	38.7	25.5	42.3	8.3	45.6	10.2	40.0	12.8	42.2	42.0	-0.7	+42.0
	DTD	42.3	0.1	44.2	36.2	40.5	26.7	40.4	8.5	40.6	22.1	41.9	20.6	35.7	22.1	37.4	14.9	41.2	34.7	-1.1	+34.6
	PCAM	48.4	7.4	54.6	39.2	48.6	48.4	48.4	36.1	48.5	48.2	44.3	4.3	48.2	23.3	48.9	48.0	48.6	52.8	+0.2	+45.4
All	Avg.	63.6	1.0	**63.3**	51.2	56.7	34.5	61.0	11.0	61.2	34.4	63.2	23.7	55.0	24.5	58.0	19.6	61.1	**52.3**	-2.5	+51.3

Table 3: Efficiency analysis on an NVIDIA GeForce RTX 4090 GPU.

(a) End-to-end inference latency (ms/image).

Time	CLIP	R-TPT	HD	Anti-Adv	TTE	TTC	ET3	BaP
T_{Clean}	3.37	176.45	184.43	32.39	5.43	33.24	23.17	**3.75**
$T_{Rob.}$	3.37	176.20	189.34	32.44	**5.41**	35.13	23.58	57.37
$T_{Avg.}$	3.37	176.33	186.89	32.42	**5.42**	34.19	23.38	30.56

(b) Runtime of BaP stages (ms).

Stage	**Time**
Rank-one Implantation	2.70
Sample Detection	3.56
Sample Rectification	53.94

robust accuracy of 52.3%, outperforming R-TPT while retaining 61.1% clean accuracy. These results validate the effectiveness of repurposing backdoors as internal probes for test-time adversarial defense. The detection ROC curves for all datasets are provided in Appendix G.

Efficiency Analysis. Table 3a compares the inference latency of test-time defenses. Benefiting from selective rectification, BaP processes clean inputs in only 3.75 ms, while adversarial inputs require 57.37 ms. Its average latency is 30.56 ms, making it 5.7× faster than R-TPT and 6× faster than HD. Although TTE and ET3 are faster, their robust accuracies are only 23.7% and 19.6%, respectively. Table 3b further reports the runtime of BaP stages. Notably, the closed-form edit implants the probe in only 2.70 ms. Sample detection requires only 3.56 ms per image, while the 53.94 ms rectification is invoked only for suspicious inputs. Overall, BaP balances adversarial robustness with inference efficiency, supporting practical deployment in real-world scenarios.

Results under Different Attack Objectives. To evaluate the performance of BaP under different attack objectives, we test it against cross-modal, targeted, and label-free attacks with a perturbation budget of $\ell_\infty = 4/255$. As shown in Table 4, BaP achieves the highest robust accuracy across all seven attack configurations. Although the implanted probe is only trained on ImageNet with PGD attack of $\ell_\infty = 1/255$, the consistent improvements across objectives indicate that the probe captures transferable adversarial behavior rather than being tied to a specific attack. Appendix C shows more details.

Table 4: ImageNet top-1 accuracy (%) across attack objectives at $\ell_\infty = 4/255$ with 50 steps. 'AA' denotes 'AutoAttack'.

Method	**Clean**	**Cross-modal**		**Targeted**			**Label-free**	
		PGD	AA	PGD	DLR	AA	PGD	AA
CLIP	63.9	0.0	0.0	0.0	0.0	0.0	0.1	0.0
TTC	40.9	2.7	0.3	25.4	6.8	6.2	10.2	1.6
ET3	58.6	3.3	0.9	24.5	16.5	2.3	15.9	8.5
BaP	**60.3**$^{\downarrow 3.6}$	**34.6**$^{\uparrow 34.6}$	**35.1**$^{\uparrow 35.1}$	**36.1**$^{\uparrow 36.1}$	**39.3**$^{\uparrow 39.3}$	**36.9**$^{\uparrow 36.9}$	**42.4**$^{\uparrow 42.3}$	**40.0**$^{\uparrow 40.0}$

Results under Stronger Attacks. Table 5 evaluates the defenses under a larger perturbation budget of $\ell_\infty = 4/255$ using PGD and AutoAttack. Under PGD, BaP achieves the highest accuracy on the General and Domain benchmarks, reaching 57.2% and 29.1%, respectively, while remaining competitive on Fine-Grained and Scene recognition. Under AutoAttack, BaP performs best on the General benchmarks, whereas R-TPT remains stronger on Fine-Grained, Scene, and Domain benchmarks. These results indicate that BaP remains effective under stronger attacks.

Table 5: Comparison of zero-shot classification accuracy under stronger attacks at $\ell_\infty = 4/255$.

Method	General			Fine-Grained			Scene			Domain		
	Clean	PGD	AutoAttack	Clean	PGD	AutoAttack	Clean	PGD	AutoAttack	Clean	PGD	AutoAttack
CLIP	79.4	0.0	0.0	76.2	0.0	0.0	40.3	0.0	0.0	39.2	0.0	0.0
R-TPT	78.5	36.2	28.4	**76.7**	**46.4**	**42.8**	**41.7**	**26.7**	**26.4**	38.1	27.5	**23.2**
TTE	**80.7**	13.4	11.5	73.6	1.7	0.4	40.6	1.2	1.4	38.0	13.4	11.5
LPF	78.6	37.4	30.8	68.2	11.3	8.2	37.0	8.0	6.9	37.1	18.6	16.7
HD	76.9	0.5	0.1	72.8	0.0	0.0	37.4	0.0	0.0	37.3	0.1	0.0
Anti-Adv	76.2	22.6	2.7	74.1	1.7	0.0	38.9	1.1	0.2	37.1	9.8	1.7
TTC	73.3	14.7	0.6	59.7	2.0	0.0	28.9	1.2	0.0	35.8	9.3	0.3
ET3	72.9	7.7	0.3	69.0	0.9	0.0	35.1	0.4	0.0	36.1	8.7	0.1
BaP	76.9$^{\downarrow 2.5}$	**57.2**$^{\uparrow 57.2}$	**35.3**$^{\uparrow 35.3}$	71.6$^{\downarrow 4.6}$	43.1$^{\uparrow 43.1}$	17.4$^{\uparrow 17.4}$	38.6$^{\downarrow 1.7}$	22.4$^{\uparrow 22.4}$	8.2$^{\uparrow 8.2}$	**38.2**$^{\downarrow 1.0}$	**29.1**$^{\uparrow 29.1}$	13.8$^{\uparrow 13.8}$

Table 6: Comparison of zero-shot classification accuracy on CLIP-B/32 and CLIP-L/14 under 10-step PGD at $\ell_\infty = 1/255$. More detailed results are provided in Appendix F.

Method	CLIP-B/32								CLIP-L/14							
	General		FG		Scene		Domain		General		FG		Scene		Domain	
	clean	rob	clean	rob	clean	rob	clean	rob	clean	rob	clean	rob	clean	rob	clean	rob
CLIP	76.7	4.0	71.8	0.3	38.8	0.4	35.9	6.6	83.7	4.0	84.2	0.3	45.3	0.2	46.8	0.3
R-TPT	72.9	41.9	**71.4**	45.4	38.7	29.6	35.7	27.2	84.2	**76.8**	**83.5**	**70.3**	47.1	37.5	42.4	**37.3**
LPF	74.8	38.0	62.0	18.9	36.2	11.3	33.7	17.4	84.0	66.9	78.8	51.7	44.2	26.2	44.5	28.5
HD	76.0	15.9	68.6	4.9	35.1	3.1	35.1	13.6	82.7	38.3	79.7	9.3	43.0	6.2	44.2	15.9
Anti-Adv	75.3	39.1	70.3	12.6	37.3	7.0	33.2	19.9	82.5	67.9	82.4	48.0	45.2	22.0	43.5	30.9
TTE	**77.5**	47.3	68.6	27.9	**39.3**	9.6	**36.8**	19.3	**86.1**	58.5	81.6	32.3	**47.8**	12.2	43.7	23.6
TTC	74.5	43.5	66.5	27.5	32.9	17.5	34.2	22.4	80.5	33.3	70.3	39.7	35.1	19.7	42.0	15.5
ET3	69.4	34.4	62.0	15.6	35.0	8.8	31.3	20.5	81.1	43.3	77.9	16.4	43.4	8.8	43.2	22.0
BaP	74.8$^{\downarrow 1.9}$	**62.4**$^{\uparrow 58.4}$	66.0$^{\downarrow 5.8}$	**54.9**$^{\uparrow 54.6}$	37.5$^{\downarrow 1.3}$	**31.0**$^{\uparrow 30.6}$	34.8$^{\downarrow 1.1}$	**32.7**$^{\uparrow 26.1}$	82.1$^{\downarrow 1.6}$	62.4$^{\uparrow 58.4}$	79.8$^{\downarrow 4.4}$	70.1$^{\uparrow 69.8}$	44.4$^{\downarrow 0.9}$	**40.7**$^{\uparrow 40.5}$	**44.9**$^{\downarrow 1.9}$	35.5$^{\uparrow 35.2}$

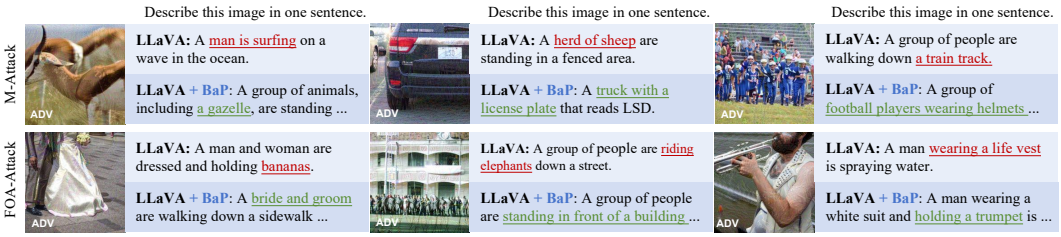


Figure 4: Qualitative results of BaP against M-Attack and FOA-Attack on image captioning.

Results on Different Backbones. Table 6 evaluates the transferability of BaP to CLIP ViT-B/32 and ViT-L/14. On ViT-B/32, BaP achieves the highest robust accuracy across all four dataset categories. When applied to ViT-L/14, BaP obtains the best robustness on Scene recognition at 40.7% and achieves performance competitive with R-TPT. These results show that BaP remains effective across different CLIP architectures.

Results on Attacks against LVLMs. Recent attacks on large vision-language models (LVLMs) have demonstrated their effectiveness against open-ended generation tasks. To evaluate BaP beyond classification, we assess the robustness of LLaVA on image captioning under M-Attack (Li et al., 2025) and FOA-Attack (Jia et al., 2025). Following their original attack settings, we randomly sample 100 images from COCO (Lin et al., 2014) and generate adversarial examples with an ℓ_∞ budget of $16/255$. Captioning performance is evaluated using GPTScore (Li et al., 2025), where the detailed prompt is provided in Appendix E. Table 9 compares BaP with TTC and ET3. Benefiting from its selective rectification mechanism, BaP preserves a clean score of 99.9, only 0.1 below the undefended model. Under M-Attack and FOA-Attack, BaP improves the robust score from 21.1 to 24.6 and from 19.1 to 23.6, respectively, outperforming both competing defenses. Fig. 4 further provides qualitative comparisons. Under both attacks, the undefended LLaVA produces captions containing attack-induced concepts, whereas BaP recovers key semantics consistent with the visual content. The ROC curves for the detection are provided in Appendix G.

Table 9: Defense performance against M-Attack and FOA-Attack.

Method	M-Attack		FOA-Attack	
	Clean	Rob.	Clean	Rob.
Origin Model	100.0	21.1	100.0	19.1
+TTC defense	84.4	23.8	84.4	20.0
+ET3 defense	87.2	19.4	87.2	16.7
+BaP defense	**99.9**$^{\downarrow 0.1}$	**24.6**$^{\uparrow 3.5}$	**99.9**$^{\downarrow 0.1}$	**23.6**$^{\uparrow 4.5}$

Table 7: Rectification-objective ablation across the 16 datasets.

$\mathcal{L}_e$	$\mathcal{L}_u$	$\mathcal{L}_{\text{esc}}$	**Clean**	**Rob.**
✓			**62.3**	21.3
✓	✓		**62.3**	22.3
✓	✓	✓	61.1	**52.3**

Table 8: The adaptive PGD at $\ell_\infty = 1/255$ with 10 steps.

β		**0**	**0.01**	**0.1**	**1**
w/o BaP	Clean	63.9	63.9	63.9	63.9
	Rob.	0.0	0.1	0.3	2.6
w/ BaP	Clean	60.3	60.3	60.3	60.3
	Rob.	39.6	34.0	**27.2**	34.7

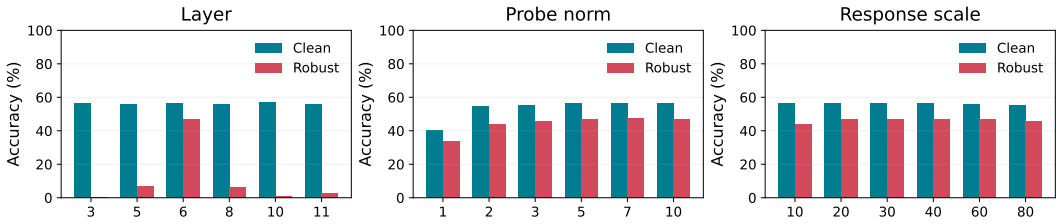


Figure 5: The sensitivity to the edited layer l, probe norm ν, and response scale γ. More ablation results are in Appendix D.

4.3 Ablation Study

Ablation of the Rectification Objective. Table 7 ablates the three objectives used for adversarial rectification. Using only the clean-subspace residual loss $\mathcal{L}_e$ yields robust accuracy of 21.3%. Adding the clean-direction alignment loss $\mathcal{L}_u$ provides an improvement to 22.3%. Incorporating the escape loss $\mathcal{L}_{\text{esc}}$ increases robust accuracy to 52.3%, with a limited clean accuracy reduction from 62.3% to 61.1%. Combining all three objectives achieves the best overall performance.

Hyperparameter Sensitivity. Fig. 5 studies the effects of the edited layer l, probe norm ν, and response scale γ. The edited layer has the largest influence on robustness. Placing the probe at Layer 6 yields the highest robust accuracy, whereas earlier or later layers perform considerably worse despite similar clean accuracy. Increasing the probe norm from 1 to moderate values improves both clean and robust accuracy, after which the performance gradually saturates. In contrast, BaP remains stable across response scales ranging from 10 to 80, with the best overall performance obtained around 40. Based on these results, we use Layer 6, a probe norm of 5, and a response scale of 40 as the default configuration.

4.4 Robustness to an Adaptive Attack

We evaluate BaP against an adaptive attack that jointly maximizes the cross-entropy loss of the edited model and suppresses the probe response. Given an adversarial input x_{adv}, the attack maximizes

$$\mathcal{L}_{\text{adapt}}(x_{\text{adv}}, y) = \text{CE}(f_{\theta'}(x_{\text{adv}}), y) - \beta\,\text{ReLU}(p(x_{\text{adv}}) - \tau)\,, \tag{16}$$

where $p(\cdot)$ denotes the probe response of Eq. 9, τ is the detection threshold, and β controls the strength of regularization term. Table 8 reports the results on ImageNet dataset. Increasing β from 0 to 0.1 reduces the robust accuracy of BaP from 39.6% to 27.2%, showing that explicitly targeting the detector weakens the defense. Nevertheless, at $\beta = 0.1$, BaP retains robust accuracy of 27.2%, compared with 0.3% without BaP. Further increasing β to 1 raises the robust accuracy of BaP to 34.7%. It is noted that the robust accuracy without BaP increases from 0.0% to 2.6%. This suggests that a larger β weakens the attack, allowing BaP to achieve higher robust accuracy.

5 Conclusion

We propose BaP, a test-time defense that repurposes a controlled backdoor as an internal probe for CLIP. Through a closed-form edit, BaP treats adversarial activation shifts as triggers and maps the induced responses into a defender-specified target space. The resulting probe detects and rectifies adversarial inputs while largely preserving clean performance. Experiments across 16 benchmarks demonstrate its robustness, efficiency, and transferability across backbones and attacks. We hope BaP motivates further exploration of backdoor-inspired probes for improving adversarial robustness.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation supporting qualitative and thematic data analysis. We did not use generative AI to generate research hypotheses, design the proposed method, derive scientific conclusions, or produce experimental results. In addition, we used generative AI tools for recommended-disclosure tasks, including language polishing of author-written text and assistance with routine coding. All AI-assisted text was reviewed for consistency with the authors' intended meaning, and AI-assisted code was inspected and tested by the authors. We take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Although BaP is inspired by backdoor mechanisms, it is intended for adversarial defense. Its defender-controlled probe is used to detect and rectify adversarial inputs. It should be applied only to models for which the defender has authorization. We believe this work advances the defensive use of backdoor-inspired mechanisms and contributes to more robust vision-language models.

REPRODUCIBILITY STATEMENT

BaP is executed on Ubuntu 20.04.3 LTS with an Intel(R) Xeon(R) Platinum 8358P CPU @ 2.60GHz. The machine is equipped with 1.0 TB of RAM and 8 NVIDIA RTX 4090 GPUs. Our experiments are conducted using CUDA 12.2, Python 3.10.0, and PyTorch 2.2.0. We provide all source code to facilitate reproduction. The code is available at anonymous Github on `https://anonymous.4open.science/r/BaP`. All configuration files for BaP are included in the repository.

REFERENCES

Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H. Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:80396–80413, 2023.

Motasem Alfarra, Juan C Pérez, Ali Thabet, Adel Bibi, Philip HS Torr, and Bernard Ghanem. Combating adversaries with anti-adversaries. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, pp. 5992–6000, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, pp. 446–461. Springer, 2014.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

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

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24625–24634. IEEE, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, 2009.

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 *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pp. 178–178. IEEE, 2004.

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

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Identifying vulnerabilities in the machine learning model supply chain. In *Proceedings of the Neural Information Processing Symposium Workshop Mach. Learning Security (MLSec)*, pp. 1–5, 2017.

Jisu Han and Wonjun Hwang. D-tpt: Dimensional entropy maximization for calibrating test-time prompt tuning in vision-language models. *arXiv preprint arXiv:2510.09473*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

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*, 12(7):2217–2226, 2019.

Md Zarif Hossain and Ahmed Imteaj. Securing vision-language models with a robust encoder against jailbreak and adversarial attacks. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 6250–6259. IEEE, 2024.

Xiaojun Jia, Sensen Gao, Simeng Qin, Tianyu Pang, Chao Du, Yihao Huang, Xinfeng Li, Yiming Li, Bo Li, and Yang Liu. Adversarial attacks against closed-source MLLMs via feature optimal alignment. In *The Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 554–561, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Lin Li, Haoyan Guan, Jianing Qiu, and Michael Spratling. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24408–24419, 2024a.

Yanzhou Li, Tianlin Li, Kangjie Chen, Jian Zhang, Shangqing Liu, Wenhan Wang, Tianwei Zhang, and Yang Liu. Badedit: Backdooring large language models by model editing. In *International Conference on Learning Representations (ICLR)*, 2024b.

Zhaoyi Li, Xiaohan Zhao, Dong-Dong Wu, Jiacheng Cui, and Zhiqiang Shen. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of GPT-4.5/4o/o1. In *The Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Liangsheng Liu, Si Chen, Jiamin Wu, Weiwei Feng, Zhixin Cheng, Xiaotian Yin, Wenfei Yang, and Tianzhu Zhang. Adversarial attacks already tell the answer: Directional bias-guided test-time defense for vision-language models. In *The International Conference on Learning Representations (ICLR)*, 2026.

Zhaoyi Liu and Huan Zhang. Stealthy backdoor attack in self-supervised learning vision encoders for large vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. doi: 10.1109/CVPR52734.2025.02333.

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

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

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *The International Conference on Learning Representations (ICLR)*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems (NeurIPS)*, 35:17359–17372, 2022.

Mujtaba Hussain Mirza, Antonio D'Orazio, Odelia Melamed, and Iacopo Masi. A provable energy-guided test-time defense boosting adversarial robustness of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8598–8609, June 2026.

Sen Nie, Jie Zhang, Zhuo Wang, Shiguang Shan, and Xilin Chen. Contrastive spectral rectification: Test-time defense towards zero-shot adversarial robustness of CLIP. In *International Conference on Machine Learning (ICML)*, 2026a.

Sen Nie, Jie Zhang, Jianxin Yan, Shiguang Shan, and Xilin Chen. V-attack: Targeting disentangled value features for controllable adversarial attacks on lvlms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42257–42267, June 2026b.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3498–3505. IEEE, 2012.

R. Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.

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 *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 81–91. IEEE, 2021.

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

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

Shawn Shan, Emily Wenger, Bolun Wang, Bo Li, Haitao Zheng, and Ben Y Zhao. Gotta catch'em all: Using honeypots to catch adversarial attacks on neural networks. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pp. 67–83, 2020.

Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29958–29967, 2025.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:14274–14289, 2022.

Daniel Tamayo, Aitor Gonzalez-Agirre, Javier Hernando, and Marta Villegas. Mass-editing memory with attention in transformers: A cross-lingual exploration of knowledge. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics (ACL)*, pp. 5831–5847, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pp. 210–218. Springer, 2018.

Hao Wang, Shangwei Guo, Jialing He, Kangjie Chen, Shudong Zhang, Tianwei Zhang, and Tao Xiang. Eviledit: Backdooring text-to-image diffusion models in one second. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3657–3665, 2024a.

Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24502–24511. IEEE, 2024b.

Xin Wang, Kai Chen, Jiaming Zhang, Jingjing Chen, and Xingjun Ma. Tapt: Test-time adversarial prompt tuning for robust inference in vision-language models. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19910–19920, 2025a.

Zhangyun Wang, Ni Ding, and Aniket Mahanti. Learning robust vision-language models from natural latent spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

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.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492. IEEE, 2010.

Songlong Xing, Zhengyu Zhao, and Nicu Sebe. Clip is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of clip. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15172–15182. IEEE, 2025.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *International Conference on Learning Representations (ICLR)*, 2024.

Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models. In *European conference on computer vision (ECCV)*, pp. 56–72. Springer, 2024.

Mingkun Zhang, Keping Bi, Wei Chen, Jiafeng Guo, and Xueqi Cheng. Clipure: Purification in latent space via CLIP for adversarially robust zero-shot classification. In *The International Conference on Learning Representations (ICLR)*, 2025.

Yiwei Zhou, Xiaobo Xia, Zhiwei Lin, Bo Han, and Tongliang Liu. Few-shot adversarial prompt learning on vision-language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:3122–3156, 2024.

Hong Zhu, Shengzhi Zhang, and Kai Chen. Ai-shielder: Exploiting backdoors to defend against adversarial attacks. *IEEE Transactions on Dependable and Secure Computing (TDSC)*, 23(1): 1244–1259, 2026.

Xingyu Zhu, Beier Zhu, Shuo Wang, Kesen Zhao, and Hanwang Zhang. Enhancing clip robustness via cross-modality alignment. *The Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Vadim Ziyadinov and Maxim Tereshonok. Low-pass image filtering to achieve adversarial robustness. *Sensors*, 23(22):9032, 2023.