# Prompt Ensemble Image Purification for Test-time Adversarial Robustness of CLIP

**Yuxuan Zhang**[1], **Shuchang Wang**[2], **Zhenbo Shi**[2], **Xiaoman Liu**[2], **Jiajun Hu**[2], **Zhidong Yu**[2], **Wei Song**[1*], **Wei Yang**[2]

[1]School of Artificial Intelligence and Computer Science, Jiangnan University
[2]School of Computer Science and Technology, University of Science and Technology of China

## Abstract

Vision-Language Models like CLIP exhibit remarkable zero-shot capabilities but remain vulnerable to adversarial attacks. Existing defenses, such as adversarial fine-tuning or test-time defense, either incur high computational costs that lead to cause catastrophic forgetting, or struggle against strong adversarial attacks. In this work, we reveal that adversarial perturbations are highly overfitted to the decision boundary of a single canonical text prompt. By introducing several semantic prompt variants, we identify a *Prompt Sensitivity Gap*: adversarial examples exhibit significantly higher prediction variance across prompts compared to benign images. Motivated by this insight, we propose Prompt Ensemble Image Purification (PEIP), an efficient test-time defense framework. PEIP features a dual-objective purification loop that jointly suppresses prediction variance to dismantle adversarial alignment and reinforces the most responsive class across prompt variants to facilitate semantic recovery. Furthermore, to accelerate inference, we introduce a statistically calibrated Early Exit mechanism that bypasses benign images based on their initial prompt sensitivity. Extensive experiments across 16 classification benchmarks, multiple CLIP architectures, and CLIP-based zero-shot semantic segmentation tasks demonstrate that PEIP achieves state-of-the-art adversarial robustness while preserving exact zero-shot accuracy and high efficiency.

## 1 Introduction

Vision-Language Models (VLMs), especially CLIP [36], have exhibited remarkable zero-shot capabilities in image classification. Despite this success, CLIP remains critically vulnerable to adversarial examples [14, 53, 48], where imperceptible perturbations can induce misclassifications, thereby hindering its deployment in safety-critical scenarios.

To counter such threats, adversarial fine-tuning has been proposed as an intuitive defense strategy [39, 37, 28], which leverages adversarial samples to adapt the CLIP model. However, applying this approach to large-scale VLMs is computationally expensive [16] and may significantly degrade performance on clean inputs due to catastrophic forgetting [30]. Consequently, test-time defenses [43, 2, 41, 35] have emerged as a promising training-free alternative that purifies input images during inference. Nevertheless, existing image purification methods still struggle to defend against strong attacks such as PGD [26] and AutoAttack [8], incur high optimization costs, and fail to exploit the inherent property of the text modality to disrupt adversarial alignments.

In this work, we explore how text prompt variety impacts adversarial robustness in CLIP. Current attacks typically optimize image perturbations against a single canonical text prompt (e.g., "`A photo of a {class}`") [43, 38]. However, our analysis reveals that this causes adversarial examples to

*Wei Song is the corresponding author. E-mail: songwei@jiangnan.edu.cn

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

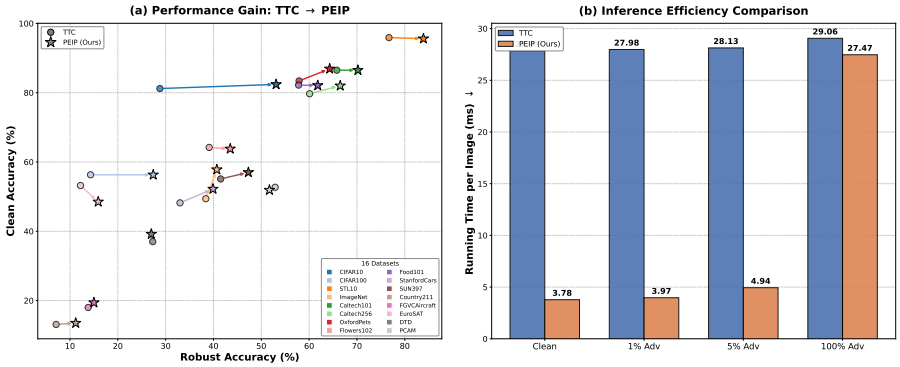



Figure 1: PEIP vs. TTC (existing top-performing image purification method). Higher robust accuracy with preserved clean accuracy in most cases, and faster inference when adversarial samples are rare.

highly overfit to the specific decision boundary of that prompt. We demonstrate that simply using different text prompts effectively rotates the decision boundary in the embedding space. Because adversarial perturbations are finely tuned to the original boundary, this rotation disrupts their alignment to some extent, leading to fluctuations in their predictions. We formalize this phenomenon as the *Prompt Sensitivity Gap*: while clean images naturally remain stable predictions across different prompts, adversarial examples exhibit relatively higher variance.

Motivated by this insight, we propose Prompt Ensemble Image Purification (PEIP), an inference-efficient defense that exploits the Prompt Sensitivity Gap through a two-phase pipeline. (i) **Dual-Objective Purification**: PEIP performs an iterative image purification procedure that jointly optimizes a Variance Suppression Objective and a Semantic Restoring Objective. The former breaks adversarial alignment by pulling the compromised prediction toward an ensemble consensus, while the latter boosts the most responsive class across prompt variants to recover semantic meaning. (ii) **Early Exit Mechanism**: Recognizing that most real-world inputs are benign, we introduce a detection mechanism that compares the initial prompt sensitivity with a statistically calibrated threshold. Rather than being tuned via grid search, this threshold is derived from the $\chi^2$ distribution of the sample variance, enabling principled control of the false positive rate. As a result, most of the clean samples can bypass the purification process, preserving exact zero-shot accuracy while achieving up to a $7\times$ speedup in mixed-data scenarios (e.g., $1\%$ or $5\%$ adversarial samples).

Extensive experiments on 16 datasets and multiple CLIP architectures show that PEIP achieves state-of-the-art adversarial robustness across different attacks with varying budgets. As shown in Figure 1, PEIP also enables faster inference in practical scenarios where adversarial samples are rare. Furthermore, PEIP generalizes to CLIP-based zero-shot semantic segmentation, significantly improving the robustness of two representative models on three benchmark datasets.

Our main contributions are summarized as follows:

- We explore the use of the text modality to guide image purification in CLIP and theoretically establish the Prompt Sensitivity Gap as a principled measure of adversarial vulnerability.
- We propose PEIP, a test-time defense framework for CLIP that dismantles adversarial noises via a dual-objective image purification loop. It further incorporates an Early Exit mechanism to preserve zero-shot accuracy for benign inputs while yielding high efficiency.
- Extensive experiments across 16 benchmarks and multiple CLIP architectures demonstrate that PEIP achieves top-performing adversarial robustness. Moreover, PEIP generalizes well to dense prediction tasks, such as CLIP-based zero-shot semantic segmentation.

## 2 Related Work

### 2.1 Adversarial Attack on CLIP

Despite the strong zero-shot classification capability of CLIP, it remains vulnerable to imperceptible adversarial attacks. Classical attack methods, such as PGD [26], CW [5], and AutoAttack [8],

introduce small perturbations to the image modality, typically optimizing them against a single canonical prompt via gradient-based procedures to mislead predictions [43]. Another line of work aims to disrupt the cross-modal alignment between images and texts. Co-Attack [48] pioneered the joint perturbation of both visual and textual inputs, while AdvCLIP [53] explores universal perturbations capable of misleading a wide range of downstream tasks. Several studies also focus on improving attack transferability in black-box settings [14, 25, 44]. In this paper, since CLIP is fully open-sourced, we focus on designing defense strategies against general image-modality attacks in the white-box setting, following the most closely related work TTC [43].

### 2.2 Adversarial Defense on CLIP

To defend against attacks on CLIP, two main strategies have been proposed. Adversarial fine-tuning leverages adversarial samples to improve the robustness of the vanilla CLIP model. Specifically, TeCoA [28] and PMG-AFT [39] introduce novel objectives to boost adversarial robustness and preserve clean performance. FARE [37] adopts unsupervised learning to enhance robustness transferability. In addition, several approaches fine-tune prompts to improve adversarial robustness [47, 24, 46]. Despite these advances, such approaches require costly fine-tuning and may decrease the clean accuracy due to catastrophic forgetting [10, 11].

Motivated by test-time adaptation [45, 20, 33], test-time defenses that do not require retraining CLIP have attracted increasing attention. One line of work optimizes text prompts to align adversarial visual features, such as R-TPT [38] and TAPT [40]. Another line focuses on image purification [41, 35, 30, 2], which aims to purify images before prediction. The most closely related work, TTC [43], also follows the image purification paradigm, aiming to guide features out of adversarial regions. However, these approaches still exhibit limited robustness against general attacks, and the purification process is often costly, leading to high latency in practical scenarios. In this paper, we enhance image purification with prompt ensembles and introduce an Early Exit strategy to accelerate the process.

## 3 Motivation

In this section, we analyze the geometric interplay between prompt variations and adversarial examples in CLIP. We demonstrate that minor semantic shifts in prompts rotate the decision boundary within the joint embedding space. Because adversarial perturbations overfit to the original boundary normal, such rotations amplify their margin fluctuations, a phenomenon we formalize as the Prompt Sensitivity Gap.

### 3.1 CLIP Classification Formulation

We consider CLIP comprising an image encoder $f : \mathbb{R}^n \to \mathbb{S}^d$ and a text encoder $g : \mathcal{P} \times \mathcal{C} \to \mathbb{S}^d$, where $\mathcal{P}$ is the space of textual prompt templates and $\mathcal{C}$ is the set of class labels. Both encoders map inputs to the $d$-dimensional unit hypersphere.

Let $t_p(c) = g(p, c)$ denote the text embedding for class $c \in \mathcal{C}$ under a specific prompt template $p \in \mathcal{P}$. The classification logit for class $c$ given an image $x$ is defined as the inner product:

$$z_c(x, p) = \langle f(x), t_p(c) \rangle \tag{1}$$

The model outputs the prediction $\hat{y} = \arg\max_{c \in \mathcal{C}} z_c(x, p)$.

### 3.2 Prompt Variant and Boundary Rotation

For a given input image $x$ with ground-truth class $y$, let $k \in \mathcal{C} \setminus \{y\}$ denote a competitive rival class. The classification margin under prompt $p$ is defined as the logit difference:

$$m_{y,k}(x, p) = z_y(x, p) - z_k(x, p) = \langle f(x), t_p(y) - t_p(k) \rangle \tag{2}$$

Let $w_{y,k}(p) = t_p(y) - t_p(k)$ represent the classification normal vector separating classes $y$ and $k$. The margin can be compactly written as $m_{y,k}(x, p) = \langle f(x), w_{y,k}(p) \rangle$. The decision boundary between $y$ and $k$ is therefore characterized by the hyperplane $\mathcal{H}_{y,k}(p) = \{v \in \mathbb{R}^d \mid \langle v, w_{y,k}(p) \rangle = 0\}$.

Let $p_0$ denote the canonical prompt (e.g., "`A photo of a {class}.`"). We consider a semantic prompt variant $p_i \in \mathcal{P}_{aug}$ (e.g., "`A painting of a {class}.`"), which induces a corresponding shift in the text embeddings, yielding a perturbed classification vector:

$$w_{y,k}(p_i) = w_{y,k}(p_0) + \Delta w_{y,k}^{(i)} \tag{3}$$

**Proposition 1** (Margin Perturbation)**.** *For any image $x$, the variation in the classification margin induced by a prompt perturbation $p_i$ is strictly a function of the image embedding's projection onto the boundary shift $\Delta w_{y,k}^{(i)}$:*

$$m_{y,k}(x, p_i) - m_{y,k}(x, p_0) = \langle f(x), \Delta w_{y,k}^{(i)} \rangle \tag{4}$$

**Proposition 2** (Prompt-Induced Decision Boundary Rotation)**.** *Given a canonical prompt $p_0$ and a semantic subsitution $p_i$, the decision boundary hyperplanes $\mathcal{H}_{y,k}(p_0)$ and $\mathcal{H}_{y,k}(p_i)$ intersect at a dihedral angle $\theta$ given by:*

$$\theta = \arccos\left(\frac{\langle w_{y,k}(p_0), w_{y,k}(p_i)\rangle}{\|w_{y,k}(p_0)\|\|w_{y,k}(p_i)\|}\right) \tag{5}$$

The detailed proofs of Propositions 1 and 2 are provided in Appendix A.

### 3.3 Adversarial Margin Amplification

Consider an adversarial image $x_{adv} = x + \delta$ crafted to cross the decision boundary $\mathcal{H}_{y,k}(p_0)$, minimizing the margin $m_{y,k}(x_{adv}, p_0) < 0$ subject to $\|\delta\|_p \leq \epsilon$. To deceive the classifier, the adversarial feature perturbation $\Delta f = f(x_{adv}) - f(x)$ is optimized to be highly anti-aligned with the classification normal vector $w_{y,k}(p_0)$. We formalize this via an orthogonal decomposition:

$$\Delta f = -\alpha w_{y,k}(p_0) + \zeta \tag{6}$$

where $\alpha > 0$ represents the magnitude of the adversarial shift, and $\zeta$ is a residual vector orthogonal to $w_{y,k}(p_0)$, and $\langle \zeta, w_{y,k}(p_0)\rangle = 0$.

**Theorem 1** (Adversarial Margin Amplification under Boundary Rotation)**.** *Let $x_{adv}$ be an adversarial example optimized under prompt $p_0$ with attack strength $\alpha > 0$. Under a prompt perturbation $p_i$, the margin shift decomposes as:*

$$m_{y,k}(x_{adv}, p_i) - m_{y,k}(x_{adv}, p_0) = \langle f(x), \Delta w_{y,k}^{(i)} \rangle - \alpha \langle w_{y,k}(p_0), \Delta w_{y,k}^{(i)} \rangle + \langle \zeta, \Delta w_{y,k}^{(i)} \rangle \tag{7}$$

**Corollary 1** (Prompt Sensitivity Gap)**.** *Let $\mathcal{P}_{test} = \{p_0\} \cup \mathcal{P}_{aug}$ be a distribution of semantic prompts, including the canonical prompt and several variants. Define the prompt sensitivity $\mathcal{S}(x)$ as the variance of the margin across $\mathcal{P}_{test}$:*

$$\mathcal{S}(x) = \mathrm{Var}_{p_i \sim \mathcal{P}_{test}}\left(m_{y,k}(x, p_i)\right) \tag{8}$$

*Assuming the prompt-induced boundary shift $\Delta w_{y,k}^{(i)}$ has isotropic variance properties, there exists a mathematical gap between the expected sensitivities of adversarial and clean examples. Because the attack strength $\alpha$ is substantial, this gap is strictly lower-bounded by the adversarial penalty:*

$$\mathbb{E}_x[\mathcal{S}(x_{adv})] - \mathbb{E}_x[\mathcal{S}(x)] \geq \alpha^2 \mathrm{Var}_{p_i}\left(\langle w_{y,k}(p_0), \Delta w_{y,k}^{(i)} \rangle\right) \tag{9}$$

As to the proofs of Theorem 1 and Corollary 1, please refer to Appendix A for more details.

*Remark.* The sensitivity gap does not imply that prompt variants are completely immune to the attack, as their textual embedding similarity still makes them partially vulnerable. Rather, it simply reflects that the canonical prompt degrades significantly more on the adversarial images.

## 4 Prompt Ensemble Image Purification for CLIP

### 4.1 Variance Suppression Objective

The theoretical insight indicates that predictions across prompt variants remain relatively stable compared to the canonical prompt. This observation motivates us to disrupt the adversarial alignment

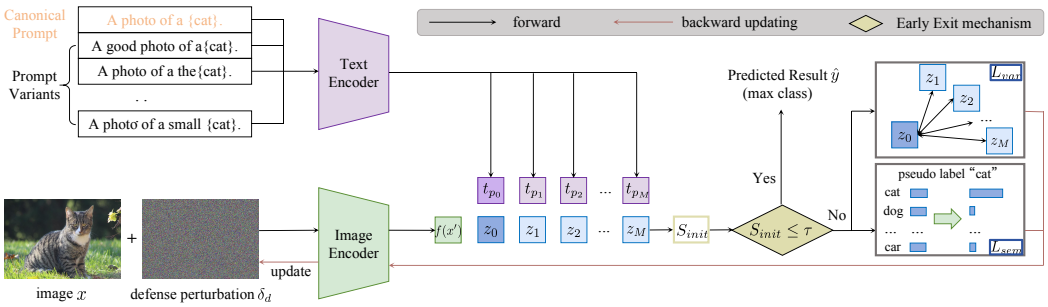



Figure 2: Overview of PEIP with Early Exit Strategy.

by penalizing the prediction discrepancy between the compromised canonical prompt and its variants. Let $\mathcal{P}_{aug} = \{p_1, \ldots, p_M\}$ denote a diverse ensemble of semantic prompt variants, excluding the canonical prompt $p_0$. For a test image $x$, let $P_i(x) = \text{softmax}(\langle f(x), t_{p_i} \rangle)$ be the predictive probability over $C$ classes under prompt $p_i$. To harness the inherent robustness of the variants, we aggregate their predictions to construct a stable semantic anchor, defined as the ensemble consensus:

$$\bar{P}_{aug}(x) = \frac{1}{M} \sum_{i=1}^{M} P_i(x) \tag{10}$$

where $\bar{P}_{aug}(x)$ serves as a fixed anchor and remains constant during the purification loop of $x$. To break the adversarial alignment, we design a Variance Suppression Objective $\mathcal{L}_{var}$ as a distribution matching penalty. We force the compromised canonical prediction $P_0(x)$ to align with the relatively robust initial ensemble consensus $\bar{P}_{aug}(x)$ by minimizing their KL divergence:

$$\mathcal{L}_{var}(x) = \text{KL}\left(\bar{P}_{aug}(x) \| P_0(x)\right) = \sum_{c=1}^{C} \bar{P}_{aug}(c|x) \log\left(\frac{\bar{P}_{aug}(c|x)}{P_0(c|x)}\right) \tag{11}$$

Minimizing this targeted divergence generates a gradient that explicitly neutralizes the adversarial shift $\Delta f$ optimized for $p_0$. Acting as a regularized expulsive force, this formulation smoothly rotates the image embedding away from the spurious adversarial vector, forcing $P_0(x)$ to shed its pathological overfitting and conform to the prompt-invariant predictions.

### 4.2 Semantic Restoration Objective

Although minimizing $\mathcal{L}_{var}$ mitigates adversarial alignment, the shared textual similarity among prompt variants still leaves them partially vulnerable to the attack. This adversarial impact often renders the consensus distribution highly uncertain or even misleading. To restore natural semantic information, we propose a Semantic Restoration Objective $\mathcal{L}_{sem}$. Motivated by the Prompt Sensitivity Gap that prompt ensembling inherently suffers less degradation than the canonical prompt, we select the pseudo-label $\tilde{y}$ by identifying the category with the maximum probability rebound from the canonical prediction $P_0(x)$ to the ensemble consensus $\bar{P}_{aug}(x)$:

$$\tilde{y} = \arg\max_{c} \left(\bar{P}_{aug}(c|x) - P_0(c|x)\right) \tag{12}$$

We then force $P_0(x)$ to align with this anchor via a standard cross-entropy objective:

$$\mathcal{L}_{sem}(x) = -\log P_0(\tilde{y}|x) \tag{13}$$

### 4.3 Dual Purification Objective

Since the pseudo-label $\tilde{y}$ is not always correct (empirically $\sim$93% accurate for PGD-based adversarial images but only $\sim$57% for benign images, refer to Appendix E), directly optimizing $\mathcal{L}_{sem}$ alone leads to degraded clean accuracy due to erroneous pseudo-labels. To address this issue, $\mathcal{L}_{var}$ acts as a regularizer against confirmation bias. When $\mathcal{L}_{sem}$ amplifies an incorrect prediction, it induces a large $\mathcal{L}_{var}$ penalty due to inconsistency with the cross-prompt consensus, thereby suppressing such

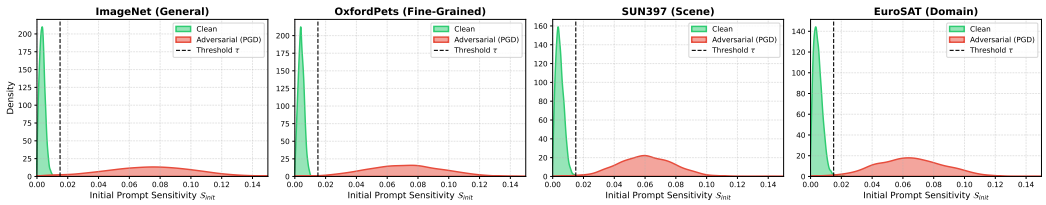



Figure 3: Distributions of $S_{init}$. A pronounced gap exists between benign and adversarial samples across diverse prompt ensembles. Our formulation of $\tau$ precisely thresholds these populations.

updates. As a result, $\mathcal{L}_{sem}$ encourages confident predictions to recover semantic meaning, while $\mathcal{L}_{var}$ constrains the optimization to remain consistent with the ensemble consensus.

Combining these two objectives, we propose Prompt Ensemble Image Purification (PEIP), a test-time defense framework to resist adversarial samples on CLIP, as illustrated in Figure 2. Accordingly, the optimal defensive perturbation $\delta_d^*$ is obtained via projected gradient descent:

$$\delta_d^* = \arg \min_{\|\delta_d\|_p \leq \epsilon_d} [\mathcal{L}_{var}(x + \delta_d) + \lambda \mathcal{L}_{sem}(x + \delta_d)] \tag{14}$$

where $\lambda > 0$ controls the trade-off between prompt variance suppression and semantic restoration.

### 4.4 Test-Time Defense with Early Exit Strategy

Prior to the purification loop, we perform a forward pass across the prompt ensemble to compute the initial sensitivity. During zero-shot inference, both the ground-truth $y$ and the rival class $k$ are unknown, making exact margin variance computation infeasible. To capture the worst-case probability fluctuation without relying on unbounded logit values, we leverage the monotonic mapping of the softmax function to define the initial sensitivity as the maximum variance across all classes:

$$\mathcal{S}_{init} = \max_{c \in \mathcal{C}} \mathrm{Var}_{p_i \in \mathcal{P}_{test}}(\mathrm{softmax}(z_c(x, p_i))) \tag{15}$$

As theoretically predicted by Corollary 1 and empirically validated in Figure 3, a pronounced distribution gap exists: clean images consistently exhibit minimal variance, whereas adversarial examples trigger significantly higher $\mathcal{S}_{init}$. This distinct separation enables an Early Exit mechanism to bypass optimization for benign inputs. Rather than tuning the threshold $\tau$ via ad-hoc grid search, we derive it precisely from statistical principles.

For benign images, the predictive probability exhibits relatively minimal functional variation across prompts. We can empirically model the fluctuation of the top-class probability under prompt perturbations as a zero-mean Gaussian process, i.e., $\Delta P_c \sim \mathcal{N}(0, \sigma_c^2)$, where $\sigma_c^2$ is the intrinsic population variance of the benign predictive probabilities. Let $M_{\text{total}} = M + 1$ denote the total number of evaluated prompts. Statistically, the scaled unbiased sample variance $\mathcal{S}_{init}$ follows a Chi-Square ($\chi^2$) distribution with $M_{\text{total}} - 1$ degrees of freedom:

$$\frac{(M_{\text{total}} - 1)\mathcal{S}_{init}}{\sigma_c^2} \sim \chi^2_{M_{\text{total}}-1} \tag{16}$$

To statistically guarantee that the Early Exit mechanism reliably identifies benign images without excessive false alarms, we control the theoretical false positive rate at a predefined significance level $\alpha$ (e.g., $\alpha = 0.01$). Consequently, the threshold must be exactly calibrated to the $(1 - \alpha)$-quantile of the $\chi^2$ distribution. Thus, the exact formulation for $\tau$ is:

$$\tau(M_{\text{total}}) = \sigma_c^2 \cdot \frac{\chi^2_{1-\alpha, M_{\text{total}}-1}}{M_{\text{total}} - 1} \tag{17}$$

The resulting inference flow is highly efficient: if $\mathcal{S}_{init} \leq \tau$, the input is deemed benign, and the ensemble prediction is yielded instantly. Only suspicious inputs ($\mathcal{S}_{init} > \tau$) trigger the subsequent iterative purification loop. The complete procedure of PEIP is detailed in Appendix B.

Table 1: Comparison of top-1 zero-shot classification accuracy (%) evaluated against 10-step PGD attacks ($\ell_\infty = 1/255$). "Clean" and "Rob." denote the accuracy on benign and adversarial inputs, respectively. The rightmost columns highlight the improvements of our PEIP over vanilla CLIP.

| | Dataset | (%) | CLIP | Adversarial Fine-tuning | | | | Test-time Defence | | | | | | | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | CLIP-FT | TeCoA | PMG-AFT | FARE | R-TPT | TAPT | TTE | Anti-adv | HD | TTC | PEIP (ours) | |
| General | CIFAR10 | Rob. | 0.7 | 3.3 | 33.6 | 40.7 | 19.7 | 27.9±0.1 | 25.8±0.1 | 41.4±6.1 | 12.4±0.1 | 17.2±0.5 | 28.8±0.2 | **53.1**±0.7 | +52.4 |
| | | Clean | 85.1 | **84.9** | 64.6 | 70.7 | 74.4 | 81.2±0.1 | 81.7±0.2 | 84.7±0.4 | 83.5±0.1 | 78.2±0.2 | 81.2±0.1 | 82.4±0.3 | -2.7 |
| | CIFAR100 | Rob. | 0.3 | 0.9 | 19.0 | 22.5 | 11.4 | 15.7±0.7 | 20.2±0.5 | 20.1±4.0 | 5.7±0.0 | 3.9±0.1 | 14.3±0.3 | **27.4**±0.3 | +27.1 |
| | | Clean | 57.1 | **59.5** | 36.0 | 40.3 | 46.7 | 56.8±0.2 | 57.1±0.1 | 58.6±0.3 | 54.0±0.2 | 52.9±0.2 | 56.3±0.2 | 56.3±0.1 | -0.8 |
| | STL10 | Rob. | 11.0 | 12.7 | 70.1 | 73.1 | 59.1 | 65.2±0.4 | 73.3±0.3 | 78.5±3.8 | 37.4±0.4 | 39.0±0.3 | 76.7±0.2 | **83.9**±0.4 | +72.9 |
| | | Clean | 96.4 | 94.5 | 87.4 | 88.6 | 91.7 | 96.2±0.1 | 94.7±0.4 | **96.3**±0.0 | 95.5±0.1 | 89.5±0.1 | 95.9±0.0 | 95.6±0.1 | -0.8 |
| | ImageNet | Rob. | 1.2 | 0.9 | 18.9 | 21.4 | 14.0 | 37.8±0.3 | 34.3±0.2 | 31.0±4.4 | 8.7±0.1 | 6.6±0.1 | 38.4±0.1 | **40.7**±0.2 | +39.5 |
| | | Clean | 59.7 | 54.2 | 34.9 | 36.1 | 48.8 | 59.7±0.1 | 58.5±0.2 | **60.0**±0.1 | 54.3±0.1 | 54.5±0.1 | 49.4±0.0 | 57.8±0.0 | -1.9 |
| | Caltech101 | Rob. | 14.7 | 14.2 | 55.5 | 61.1 | 50.7 | **79.8**±0.1 | 68.4±0.2 | 67.6±3.9 | 34.8±0.2 | 31.5±0.2 | 65.8±0.1 | 70.2±0.3 | +55.5 |
| | | Clean | 85.7 | 83.6 | 71.7 | 75.5 | 81.0 | **86.7**±0.1 | 85.1±0.3 | 85.8±0.1 | 84.0±0.1 | 82.3±0.0 | 86.5±0.1 | 86.5±0.0 | +0.8 |
| | Caltech256 | Rob. | 8.5 | 6.8 | 43.2 | 45.9 | 38.8 | 62.3±0.0 | 55.8±0.1 | 60.1±4.0 | 25.4±0.2 | 23.5±0.1 | 60.1±0.0 | **66.5**±0.2 | +58.0 |
| | | Clean | 81.7 | 78.5 | 61.1 | 62.2 | 73.3 | 81.3±0.0 | 80.7±0.1 | **82.5**±0.1 | 79.4±0.1 | 79.1±0.0 | 79.7±0.0 | 82.0±0.3 | -0.3 |
| Fine-Grained | OxfordPets | Rob. | 1.0 | 2.1 | 38.4 | 41.2 | 31.1 | **74.2**±0.0 | 59.4±0.1 | 50.3±7.3 | 20.4±0.2 | 12.0±0.2 | 57.9±0.2 | 64.3±0.1 | +63.3 |
| | | Clean | 87.4 | 84.1 | 62.1 | 65.9 | 79.4 | 84.6±0.1 | 85.1±0.3 | **88.1**±0.1 | 80.6±0.4 | 80.9±0.1 | 83.4±0.2 | 86.9±0.1 | -0.5 |
| | Flowers102 | Rob. | 1.1 | 0.5 | 21.9 | 23.4 | 17.1 | **47.5**±0.0 | 40.4±0.1 | 35.9±4.7 | 7.2±0.4 | 7.3±0.1 | 39.1±0.3 | 43.5±0.5 | +42.4 |
| | | Clean | 65.5 | 53.4 | 36.8 | 37.0 | 48.0 | 60.6±0.2 | 61.0±0.1 | **65.2**±0.2 | 62.7±0.1 | 58.2±0.1 | 64.2±0.2 | 63.8±0.0 | -1.7 |
| | Food101 | Rob. | 0.7 | 0.4 | 13.9 | 18.6 | 11.7 | 48.9±0.3 | 52.5±0.0 | 43.9±7.0 | 13.1±0.2 | 8.0±0.1 | 57.8±0.2 | **61.8**±0.3 | +61.1 |
| | | Clean | 83.9 | 64.9 | 30.0 | 36.6 | 55.3 | 83.6±0.0 | 82.7±0.1 | **84.0**±0.0 | 75.8±0.2 | 80.3±0.1 | 82.2±0.0 | 82.1±0.1 | -1.8 |
| | StanfordCars | Rob. | 0.0 | 0.1 | 8.8 | 11.7 | 6.8 | 37.6±0.3 | 34.1±0.1 | 22.4±4.2 | 4.4±0.3 | 2.7±0.1 | 33.0±0.1 | **39.9**±0.4 | +39.4 |
| | | Clean | 52.0 | 42.1 | 20.9 | 25.4 | 38.7 | 52.4±0.1 | 51.5±0.2 | **52.7**±0.3 | 36.2±0.3 | 44.3±0.0 | 48.2±0.2 | 52.2±0.6 | +0.2 |
| Scene | SUN397 | Rob. | 1.1 | 0.9 | 19.4 | 22.6 | 14.9 | 41.7±0.0 | 44.2±0.1 | 30.8±4.4 | 8.1±0.0 | 6.4±0.1 | 41.5±0.0 | **47.3**±0.1 | +46.2 |
| | | Clean | 58.5 | 55.7 | 36.7 | 38.0 | 52.4 | 57.6±0.1 | **59.2**±0.1 | 59.1±0.1 | 56.0±0.0 | 53.2±0.0 | 55.1±0.1 | 57.0±0.2 | -1.5 |
| | Country211 | Rob. | 0.0 | 0.0 | 1.8 | 2.1 | 0.9 | 9.3±0.0 | 6.5±0.1 | 3.1±0.9 | 0.7±0.1 | 0.5±0.0 | 7.1±0.0 | **11.2**±0.5 | +11.2 |
| | | Clean | 15.3 | 12.1 | 4.8 | 4.6 | 9.3 | 13.1±0.0 | **14.9**±0.2 | 14.7±0.2 | 11.6±0.1 | 11.7±0.1 | 13.1±0.1 | 13.5±0.4 | -1.8 |
| Domain | FGVCAircraft | Rob. | 0.0 | 0.0 | 2.5 | 2.2 | 1.4 | 12.6±0.0 | 13.0±0.0 | 6.2±1.4 | 1.3±0.1 | 1.3±0.1 | 13.8±0.4 | **15.0**±0.2 | +15.0 |
| | | Clean | 20.1 | 14.0 | 5.3 | 5.6 | 10.9 | 17.5±0.2 | 18.4±0.3 | **20.2**±0.4 | 15.9±0.2 | 16.4±0.0 | 18.0±0.2 | 19.4±0.3 | -0.7 |
| | EuroSAT | Rob. | 0.0 | 0.0 | 12.0 | 12.6 | 10.7 | 14.5±0.3 | 14.3±0.6 | 6.9±2.1 | 2.2±0.0 | 4.6±0.1 | 12.2±0.2 | **15.9**±0.1 | +15.9 |
| | | Clean | 42.6 | 27.6 | 16.6 | 18.5 | 21.9 | 37.0±0.1 | 40.5±0.5 | 44.4±1.6 | 36.8±0.2 | 39.1±0.1 | **53.2**±0.1 | 42.5±0.0 | -0.1 |
| | DTD | Rob. | 3.0 | 2.4 | 17.6 | 15.0 | 15.6 | 19.7±0.6 | 12.6±0.2 | 23.9±2.3 | 5.6±0.1 | 11.6±0.2 | **27.3**±0.3 | 27.0±0.1 | +24.0 |
| | | Clean | 40.6 | 36.5 | 25.2 | 21.8 | 32.1 | 35.8±0.1 | 40.1±0.2 | **41.3**±0.3 | 38.9±0.2 | 34.9±0.4 | 37.0±0.2 | 39.2±0.3 | -1.4 |
| | PCAM | Rob. | 0.1 | 1.1 | 48.2 | 46.2 | 16.2 | 41.3±0.7 | 44.6±0.2 | 10.6±3.2 | 5.0±0.1 | 44.7±0.2 | **52.9**±0.2 | 51.7±0.4 | +51.6 |
| | | Clean | 52.0 | 47.2 | 50.0 | 50.0 | 52.5 | **53.1**±0.2 | 51.6±0.1 | 51.0±0.1 | 52.5±0.0 | 50.4±0.0 | **52.7**±0.1 | 51.9±0.0 | -0.1 |
| All | **Avg.** | Rob. | 2.7 | 2.9 | 26.5 | 28.8 | 20.0 | 39.8±0.2 | 37.5±0.2 | 33.3±4.0 | 12.0±0.0 | 13.8±0.1 | 39.2±0.0 | **45.0**±0.2 | +42.3 |
| | | Clean | 61.5 | 55.8 | 40.3 | 42.3 | 51.0 | 59.8±0.1 | 60.2±0.2 | **61.8**±0.1 | 57.4±0.1 | 56.6±0.0 | 59.8±0.1 | 60.5±0.1 | -0.6 |

Table 2: Comparison of top-1 zero-shot classification accuracy (%) evaluated against stronger adversarial attacks, specifically 50-step PGD, CW, and AutoAttack ($\ell_\infty = 4/255$).

| Method | General | | | | Fine-Grained | | | | Scene | | | | Domain | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Clean | PGD | CW | AA | Clean | PGD | CW | AA | Clean | PGD | CW | AA | Clean | PGD | CW | AA |
| CLIP | 77.6 | 0.3 | 0.5 | 0.0 | 72.2 | 0.0 | 0.2 | 0.0 | 36.9 | 0.2 | 0.1 | 0.1 | 38.8 | 0.0 | 0.0 | 0.0 |
| CLIP-FT | 75.9 | 2.1 | 2.5 | 0.2 | 61.1 | 0.3 | 0.4 | 0.0 | 33.9 | 0.2 | 0.2 | 0.0 | 31.3 | 0.5 | 0.8 | 0.0 |
| TeCoA | 59.3 | 15.4 | 17.1 | 8.4 | 37.5 | 7.2 | 8.5 | 4.6 | 20.8 | 3.1 | 3.6 | 0.7 | 24.3 | 6.9 | 6.8 | 3.6 |
| PMG-AFT | 62.2 | 23.5 | 21.3 | 9.2 | 41.2 | 20.8 | 22.1 | 8.9 | 21.3 | 9.2 | 8.4 | 2.7 | 24.0 | 13.2 | 13.1 | 6.7 |
| FARE | 69.3 | 15.7 | 18.1 | 6.5 | 55.4 | 10.8 | 12.0 | 4.8 | 30.9 | 3.6 | 3.6 | 0.9 | 29.4 | 5.2 | 4.5 | 1.8 |
| R-TPT | 77.0 | 23.7 | 26.4 | 14.6 | 70.3 | 26.7 | 25.4 | 11.0 | 35.4 | 14.4 | 15.0 | 5.8 | 35.9 | 13.7 | 17.2 | 6.0 |
| TAPT | 76.3 | 24.0 | 25.1 | 13.9 | 70.1 | 23.0 | 20.7 | 7.4 | **37.1** | 15.0 | 11.9 | 6.6 | 37.6 | 11.7 | 15.0 | 8.6 |
| TTE | **78.0** | 26.2 | 26.8 | 12.7 | **72.5** | 16.5 | 19.1 | 8.3 | 36.9 | 10.3 | 11.4 | 4.7 | 39.2 | 4.6 | 5.8 | 1.3 |
| Anti-Adv | 75.1 | 12.5 | 13.0 | 3.7 | 63.8 | 4.9 | 7.0 | 0.8 | 33.8 | 2.2 | 2.3 | 0.8 | 36.0 | 1.0 | 1.1 | 0.2 |
| HD | 72.8 | 14.2 | 13.3 | 4.6 | 65.9 | 2.9 | 4.4 | 0.6 | 32.5 | 2.7 | 3.0 | 0.4 | 35.2 | 8.3 | 10.1 | 3.5 |
| TTC | 74.8 | 25.9 | 27.3 | 15.1 | 69.5 | 24.8 | 24.9 | 11.9 | 34.1 | 15.1 | 12.8 | 8.3 | **40.2** | 16.9 | 15.7 | 8.5 |
| PEIP (ours) | 76.8$^{\downarrow 0.8}$ | **35.1**$^{\uparrow 34.8}$ | **35.4**$^{\uparrow 34.9}$ | **20.8**$^{\uparrow 20.8}$ | 71.3$^{\downarrow 0.9}$ | **28.6**$^{\uparrow 28.6}$ | **29.3**$^{\uparrow 29.1}$ | **18.4**$^{\uparrow 18.4}$ | 35.3$^{\downarrow 1.6}$ | **19.6**$^{\uparrow 19.4}$ | **19.1**$^{\uparrow 19.0}$ | **11.4**$^{\uparrow 11.3}$ | 38.3$^{\downarrow 0.5}$ | **24.7**$^{\uparrow 24.7}$ | **24.4**$^{\uparrow 24.4}$ | **13.5**$^{\uparrow 13.5}$ |

# 5 Experiments

## 5.1 Experiment Setup

**Datasets and Models**. In alignment with the experimental protocols established in [39, 43], we evaluate our approach across a diverse suite of 16 benchmarks. These datasets span several visual scenarios [30]: General object recognition (CIFAR10/100 [22], STL10 [7], ImageNet [9], and Caltech101/256 [15, 17]), Fine-Grained identification (OxfordPets [34], Flowers102 [32], Food101 [3], and StanfordCars [21]), Scene understanding (SUN397 [42], Country211 [36]), and Domain-specific datasets (FGVCAircraft [27], EuroSAT [19], DTD [6], and PCAM [12]). As to the models, we not only adopt classic CLIP ViT-B/16 as default, but also employ ViT-B/32 and ViT-L/14 for more comprehensive evaluations.

**Implementation Details**. Unless explicitly stated, we set the perturbation budget to $1/255$ under $\ell_\infty$ constraint, and the number of attack iterations 10 and 50 for PGD [26], as well as 50 for CW [5], and AutoAttack (AA) [8]. For our PEIP, we set the defense perturbation $\epsilon_d = 4/255$, aligning with previous test-time approaches [43, 30], we set iteration $K = 3$, balanced factor $\lambda = 1$, and prompt number $M = 4$. Following prior works [43, 28, 31], we conduct our adversarial attacks on the canonical prompt $p_0$. The detailed procedure for obtaining prompt variants is provided in Appendix C. To calibrate the Early Exit threshold $\tau$, we estimate the natural variance $\sigma_c^2$ a priori using 2,000 randomly sampled benign images drawn from four datasets, each randomly selected from a distinct data type. Averaged over three independent trials, this offline estimation yields a deterministic threshold of $\tau = 0.016$. We randomly select 2,000 samples for each dataset for comprehensive

Table 3: Comparison of running time per image (ms) (↓) with varying proportions of adversarial samples on a single RTX 5090.

| Adv. % | CLIP | R-TPT | TAPT | TTE | Anti-Adv | HD | TTC | PEIP |
|---|---|---|---|---|---|---|---|---|
| Clean | 3.09 | 149.48 | 181.73 | 4.92 | 26.51 | 110.20 | 27.83 | **3.78** |
| 1% Adv. | 3.09 | 149.33 | 181.28 | 4.99 | 26.47 | 110.16 | 27.98 | **3.97** |
| 5% Adv. | 3.09 | 149.04 | 182.40 | 5.02 | 26.55 | 110.33 | 28.13 | **4.94** |
| 50% Adv. | 3.09 | 149.83 | 181.50 | **5.18** | 26.61 | 110.37 | 28.55 | 14.21 |
| 100% Adv. | 3.09 | 149.21 | 181.31 | **4.95** | 26.54 | 110.20 | 29.06 | 27.47 |

Table 4: Ablation of each component on clean and adversarial performance. "E.E" denotes Early Exit.

| $\mathcal{L}_{var}$ | $\mathcal{L}_{sem}$ | E.E. | ImageNet | | OxfordPets | | SUN397 | | EuroSAT | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Clean | Rob. | Clean | Rob. | Clean | Rob. | Clean | Rob. |
| | | | **59.7** | 1.2 | **87.4** | 1.0 | **58.5** | 1.1 | **42.6** | 0.0 |
| ✓ | | | 55.2 | 27.4 | 83.6 | 40.5 | 56.4 | 35.3 | 42.0 | 12.7 |
| | ✓ | | 28.5 | 38.8 | 47.1 | 59.9 | 31.5 | 47.1 | 22.4 | 15.2 |
| ✓ | ✓ | | 51.3 | **41.2** | 80.2 | **64.8** | 52.8 | **47.9** | 37.3 | **16.3** |
| ✓ | ✓ | ✓ | 57.8 | 40.7 | 86.9 | 64.3 | 57.0 | 47.3 | 42.5 | 15.9 |

evaluation, and each experiment is conducted by 3 times. Moreover, following prior works [28, 43], we adopt the single-prompt evaluation protocol using "A photo of a {class}".

**Comparison Approaches.** To benchmark the efficacy of our proposed method, we conduct a comparative analysis against a representative set of state-of-the-art defenses. Following the taxonomy in robust vision, these baselines are categorized into: (i) Adversarial Fine-tuning methods, including CLIP-FT [43], TeCoA [28], PMG-AFT [39], and FARE [37]; as well as (ii) Test-Time Defenses, including R-TPT [38], TAPT [40], TTE [35], Anti-Adv [2], HD [41], and TTC [43]. For a fair comparison, all baselines are reproduced using their official configurations.

### 5.2 Main Results

**General Performance on 16 datasets.** We evaluate the performance of CLIP ViT-B/16 under various defense strategies using clean inputs and the commonly used 10-step PGD attack with perturbation $1/255$. Table 1 reports the results across 16 datasets. As shown, adversarial fine-tuning methods suffer from significantly reduced clean accuracy due to catastrophic forgetting, with average drops of 21.2% for TeCoA and 19.2% for PMG-AFT. In contrast, most test-time defense approaches fail to achieve strong robustness against adversarial attacks. For example, although the best image purification approach TTC achieves 39.2% accuracy on adversarial samples, it is still far from that of clean samples on vanilla CLIP. By comparison, our PEIP achieves 60.9% clean accuracy and 45.0% adversarial robustness. It reduces the accuracy of the vanilla CLIP model by only 1.0%, while surpassing the previous best defense method by 5.8% average points in adversarial robustness.

**Performance under stronger attacks.** To further evaluate the defense capability of PEIP, we adopt stronger adversarial attacks, including 50-step PGD, CW, and AutoAttack, with a larger perturbation budget of $\ell_\infty = 4/255$. As shown in Table 2, PEIP consistently achieves the best performance against these stronger attacks in most cases. In particular, under the General scene, PEIP improves robustness by +9.2% and +8.1% under PGD and CW attacks, respectively, compared with the previous top-performing purification method TTC.

**Performance on more backbones of CLIP.** To evaluate the defense generality of PEIP across different CLIP architectures, we further conduct experiments on CLIP ViT-B/32 and ViT-L/14. The results are exhibited in Appendix D, PEIP also achieves stronger robustness.

### 5.3 Ablation Study

To evaluate the efficacy of each component, we conduct ablation studies across multiple datasets, i.e., ImageNet, OxfordPets, SUN397, and EuroSAT, under a 10-step PGD attack ($\ell_\infty = 1/255$).

**Objective Components.** We conduct ablation studies (Table 4) to analyze each component of PEIP. Using only $\mathcal{L}_{var}$ disrupts adversarial alignment on the canonical prompt, but the ensemble consensus may still retain adversarial bias due to the textual embedding similarity among prompt variants, thus limiting the robustness. In contrast, using only $\mathcal{L}_{sem}$ provides a strong signal for semantic recovery but significantly degrades clean accuracy. This is because the pseudo-label $\tilde{y}$, derived from the probability rebound, is reliable for adversarial samples (∼92%), but less reliable for clean samples (∼57%, refer to Appendix F), where small fluctuations can lead to incorrect labels. Combining the two objectives balances this trade-off. When $\mathcal{L}_{sem}$ amplifies an incorrect $\tilde{y}$ on clean images, it incurs a large KL divergence with the consensus, suppressing such updates. As a result, $\mathcal{L}_{sem}$ increases confidence on the most responsive category across prompts (typically the ground-truth class for

Table 5: Extension to zero-shot semantic segmentation. We report clean accuracy (mIoU%) and adversarial robustness (mIoU%) under 10-step PGD ($\ell_\infty = 1/255$) for MaskCLIP and ClearCLIP across three benchmarks.

| Method | MaskCLIP | | | | | | ClearCLIP | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pascal VOC | | Pascal Context | | COCO Stuff | | Pascal VOC | | Pascal Context | | COCO Stuff | |
| | Clean | Rob. | Clean | Rob. | Clean | Rob. | Clean | Rob. | Clean | Rob. | Clean | Rob. |
| Vanilla | 70.6 | 37.3 | 26.7 | 10.8 | 17.3 | 7.4 | 78.1 | 39.6 | 36.2 | 17.9 | 23.8 | 8.7 |
| +TTC defense | 67.4 | 41.1 | 24.8 | 13.1 | **16.7** | 9.0 | 75.5 | 45.0 | 35.4 | 21.6 | 23.0 | 11.2 |
| +PEIP defense | **69.1**$^{\downarrow 1.5}$ | **48.5**$^{\uparrow 11.2}$ | **25.0**$^{\downarrow 1.7}$ | **15.7**$^{\uparrow 4.9}$ | **16.7**$^{\downarrow 0.6}$ | **12.1**$^{\uparrow 4.7}$ | **76.0**$^{\downarrow 2.1}$ | **54.4**$^{\uparrow 14.8}$ | **35.7**$^{\downarrow 0.5}$ | **24.9**$^{\uparrow 7.0}$ | **23.2**$^{\downarrow 0.6}$ | **15.3**$^{\uparrow 6.6}$ |

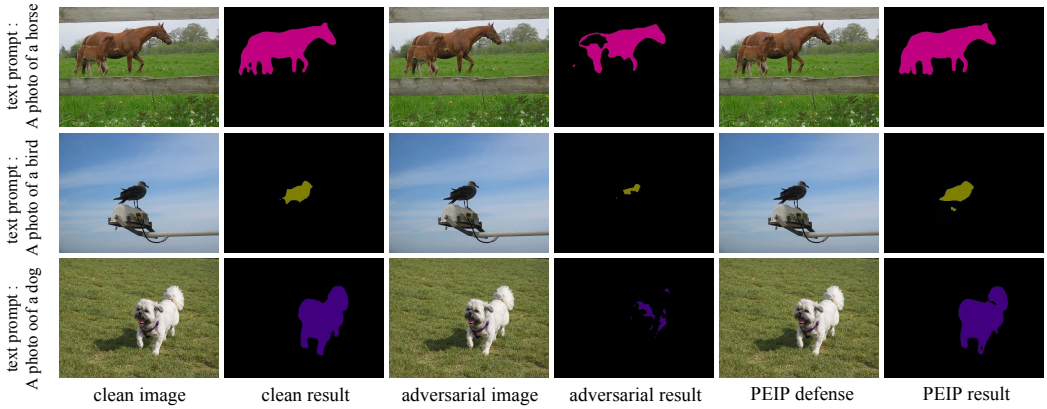



Figure 4: Qualitative results of zero-shot semantic segmentation on MaskCLIP. Adversarial perturbations degrade predicted masks, while PEIP defends against the attack and restores accurate segmentation.

adversarial samples), while $\mathcal{L}_{var}$ enforces consistency with prompt-invariant semantics. This leads to improved adversarial robustness (e.g., $41.2\%$ on ImageNet) while maintaining strong clean accuracy.

Test-time optimization on benign inputs may decrease clean performance (e.g., -8.4% on ImageNet). The Early Exit mechanism alleviates this by bypassing stable benign samples, preserving near-baseline clean accuracy with minimal robustness loss. Moreover, since real-world data is dominated by benign samples, selectively applying purification only to suspected adversarial inputs significantly reduces computation. Under simulated settings (1% and 5% adversarial samples with PGD), PEIP achieves up to $7\times$ speedup over TTC (refer to Table 3), demonstrating its practical efficiency.

In addition, we also explore the iteration number $K$, the influence of the balance factor $\lambda$, and the prompt ensemble size $M$. Please see Appendix F for more details.

### 5.4 Against Adaptive Attack

To evaluate PEIP under adaptive attacks, we consider a fully informed white-box adversary with complete knowledge of the defense, including its prompt ensemble, purification objectives, random initialization, and inference procedure. Rather than attacking vanilla CLIP, the adversary optimizes perturbations against the final defended prediction by backpropagating through the complete PEIP purification pipeline.

We instantiate this threat model using cross-entropy-based Adaptive PGD with 10 steps ($\ell_\infty = 1/255$) and 50 steps ($\ell_\infty = 4/255$), consistent with the attack budgets in Tables 1 and 2, respectively. Since PEIP employs random initialization, we also evaluate Adaptive PGD-EoT: at each attack step, the defended loss is averaged over 10 independent PEIP initializations before updating the adversarial perturbation. This variant explicitly accounts for the stochasticity of the defense.

As shown in Table 6, attacking the full PEIP output is stronger than standard PGD: under the 10-step attack, average robustness decreases from $45.0\%$ to $42.9\%$ with Adaptive PGD and to $41.2\%$ with EoT. Under the stronger 50-step attack, EoT causes only a moderate additional decrease across all

Table 6: Top-1 zero-shot classification accuracy (%) of PEIP under adaptive attacks, averaged within the four dataset groups. The 10-step and 50-step attacks use $\ell_\infty$ budgets of $1/255$ and $4/255$, respectively. "Avg." is weighted over all 16 datasets. Adaptive PGD optimizes the complete defended output, and EoT averages 10 random PEIP initializations per step.

| Steps | Attack | General | Fine-Grained | Scene | Domain | Avg. |
|---|---|---|---|---|---|---|
| 10 | PGD | 57.0 | 52.4 | 29.3 | 27.4 | 45.0 |
| | Adaptive PGD | 54.3 | 50.6 | 28.6 | 25.2 | 42.9 |
| | Adaptive PGD-EoT | 52.7 | 47.9 | 26.9 | 24.2 | 41.2 |
| 50 | PGD | 35.1 | 28.6 | 19.6 | 24.7 | 28.9 |
| | Adaptive PGD | 32.7 | 27.6 | 19.0 | 22.9 | 27.3 |
| | Adaptive PGD-EoT | 31.5 | 25.3 | 18.9 | 21.4 | 25.9 |

four dataset groups. Thus, PEIP retains non-trivial robustness and does not rely solely on random initialization; even a fully informed attacker leaves a prompt-sensitivity gap that prompt-consensus-guided purification can exploit.

### 5.5 Extension to Semantic Segmentation

Beyond image-level classification, we extend our evaluation to CLIP-based zero-shot semantic segmentation to assess the generality of PEIP. We adopt MaskCLIP [52] and ClearCLIP [23] with a ViT-B/16 backbone. Adversarial samples are generated using 10-step PGD with $\ell_\infty = 1/255$ budget to disrupt visual features. As to text embeddings, we adopt the canonical prompt "`A photo of a {class}`" for both adversarial generation and evaluation, following standard classification protocols.

Experiments are conducted on three benchmarks: Pascal VOC [13], Pascal Context [29], and COCO Stuff [4]. For each dataset, we randomly sample 2,000 images and report the average over three runs. Due to dense prediction characteristics, we solely adopt $\mathcal{L}_{var}$ here (without $\mathcal{L}_{sem}$ and Early Exit). Other settings are consistent with the default settings in classification tasks.

As shown in Table 5, PEIP consistently improves the adversarial robustness of both MaskCLIP and ClearCLIP, and significantly outperforms TTC (e.g., +7.4% and +9.4% robust mIoU on Pascal VOC). Qualitative results in Figure 4 further demonstrate that PEIP effectively restores predictions under attack, recovering accurate object boundaries aligned with benign outputs on MaskCLIP.

## 6 Conclusion

In this paper, we identified a key weakness in adversarial attacks on vision-language models: they overfit to the decision boundary of the specific canonical prompt. We formalized this as the Prompt Sensitivity Gap, showing that using semantically varied prompts shifts the decision boundary and disrupts adversarial alignment, leading to unstable predictions for malicious inputs. Based on this insight, we proposed Prompt Ensemble Image Purification (PEIP), a test-time defense strategy. PEIP jointly reduces prediction variance across prompts and restores semantic confidence, mitigating attacks while preserving natural semantic information. To enhance efficiency, we introduced a statistically calibrated Early Exit mechanism that skips benign samples, maintaining zero-shot accuracy and reducing computation. Extensive experiments across 16 benchmarks, multiple CLIP models, and zero-shot segmentation tasks demonstrate that PEIP achieves state-of-the-art robustness. Overall, PEIP offers an effective and efficient approach for securing CLIP in practical applications.

## Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 62672218, in part by the National Foreign Experts Program under Grant DL2023144001L, in part by the Fundamental Research Funds for the Central Universities under Grant JUSRP221027, National Natural Science Foundation of China (T2541014 and 62602624), and CCF-Baidu Open Fund (CCF-BAIDU OF202502).

## References


[1] Shashank Agnihotri, Steffen Jung, and Margret Keuper. Cospgd: an efficient white-box adversarial attack for pixel-wise prediction tasks. *arXiv preprint arXiv:2302.02213*, 2023.

[2] 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*, volume 36, pages 5992–6000, 2022.

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

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

[6] 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*, pages 3606–3613, 2014.

[7] 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*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

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

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Junhao Dong, Piotr Koniusz, Liaoyuan Feng, Yifei Zhang, Hao Zhu, Weiming Liu, Xinghua Qu, and Yew-Soon Ong. Robustifying zero-shot vision language models by subspaces alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21037–21047, 2025.

[11] Junhao Dong, Piotr Koniusz, Yifei Zhang, Hao Zhu, Weiming Liu, Xinghua Qu, and Yew-Soon Ong. Improving zero-shot adversarial robustness in vision-language models by closed-form alignment of adversarial path simplices. In *Forty-second International Conference on Machine Learning*, 2025.

[12] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, CAMELYON16 consortium, Meyke Hermsen, Quirine F Manson, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[14] Hao Fang, Jiawei Kong, Bin Chen, Tao Dai, Hao Wu, and Shu-Tao Xia. Clip-guided generative networks for transferable targeted adversarial attacks. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024.

[15] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[16] Xiaowei Fu and Lei Zhang. Adversarial defense in vision-language models: An overview. In *2025 4th International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, pages 1828–1831. IEEE, 2025.

[17] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset.

[18] Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip HS Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *European Conference on Computer Vision*, pages 308–325. Springer, 2022.

[19] 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.

[20] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171, 2024.

[21] 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*, pages 554–561, 2013.

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

[23] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pages 143–160. Springer, 2024.

[24] Lin Li, Haoyan Guan, Jianing Qiu, and Michael Spratling. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24408–24419, 2024.

[25] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 102–111, 2023.

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

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

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

[29] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014.

[30] Sen Nie, Jie Zhang, Zhuo Wang, Shiguang Shan, and Xilin Chen. Contrastive spectral rectification: Test-time defense towards zero-shot adversarial robustness of clip. *arXiv preprint arXiv:2601.19210*, 2026.

[31] 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 *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 93122–93143. PMLR, 06–11 Jul 2026.

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

[33] David Osowiechi, Mehrdad Noori, Gustavo A Hakim, Moslem Yazdanpanah, Ali Bahri, Milad Cheraghalikhani, Sahar Dastani, Farzad Beizaee, Ismail B Ayed, and Christian Desrosiers. Watt: Weight average test time adaptation of clip. *Advances in neural information processing systems*, 37:48015–48044, 2024.

[34] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

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

[36] 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 *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[37] 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 *International Conference on Machine Learning*, pages 43685–43704. PMLR, 2024.

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

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

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

[41] 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.

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

[43] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 15172–15182, 2025.

[44] Hunmin Yang, Jongoh Jeong, and Kuk-Jin Yoon. Prompt-driven contrastive learning for transferable adversarial attacks. In *European Conference on Computer Vision*, pages 36–53. Springer, 2024.

[45] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15922–15932, 2023.

[46] Kun Zhai, Siheng Chen, Xingjun Ma, and Yu-Gang Jiang. Fedapt: Federated adversarial prompt tuning for vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4310–4318, 2025.

[47] 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*, pages 56–72. Springer, 2024.

[48] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022.

[49] Yuxuan Zhang, Zhenbo Shi, Shuchang Wang, Wei Yang, Shaowei Wang, and Yinxing Xue. Rp-pgd: Boosting segmentation robustness with a region-and-prototype based adversarial attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10338–10347, 2025.

[50] Yuxuan Zhang, Zhenbo Shi, Wei Yang, Shuchang Wang, Shaowei Wang, and Yinxing Xue. Genseg: on generating unified adversary for segmentation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1733–1742, 2024.

[51] Yuxuan Zhang, Shuchang Wang, Xiaoman Liu, Zhenbo Shi, Zhidong Yu, Wei Song, and Wei Yang. Rethinking adversary in semantic segmentation: An out-of-distribution perspective. In *European Conference on Computer Vision*, pages 277–293. Springer, 2026.

[52] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European conference on computer vision*, pages 696–712. Springer, 2022.

[53] Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6311–6320, 2023.

# Supplementary Material

## A More Proofs

In this section, we provide detailed proofs of Proposition 1, Proposition 2, Theorem 1, and Corollary 1.

### A.1 Proof of Proposition 1

**Proposition 1** (Margin Perturbation)**.** *For any image $x$, the variation in the classification margin induced by a prompt perturbation $p_i$ is a function of the image embedding's projection onto the boundary shift $\Delta w_{y,k}^{(i)}$:*

$$m_{y,k}(x, p_i) - m_{y,k}(x, p_0) = \langle f(x), \Delta w_{y,k}^{(i)} \rangle \tag{18}$$

*Proof.* We begin by recalling the formal definition of the classification margin between the ground-truth class $y$ and the competitive class $k$. Under the perturbed prompt $p_i$, the margin is given by the inner product of the image embedding and the corresponding classification normal vector:

$$m_{y,k}(x, p_i) = \langle f(x), w_{y,k}(p_i) \rangle \tag{19}$$

According to our prompt perturbation formulation, the boundary normal vector under $p_i$ decomposes additively into the canonical normal vector and a boundary shift term, i.e., $w_{y,k}(p_i) = w_{y,k}(p_0) + \Delta w_{y,k}^{(i)}$. Substituting this decomposition into the margin equation yields:

$$m_{y,k}(x, p_i) = \langle f(x), w_{y,k}(p_0) + \Delta w_{y,k}^{(i)} \rangle \tag{20}$$

Leveraging the fundamental linearity property of the inner product in the Euclidean embedding space, we can explicitly distribute the image feature $f(x)$ across the vector addition:

$$m_{y,k}(x, p_i) = \langle f(x), w_{y,k}(p_0) \rangle + \langle f(x), \Delta w_{y,k}^{(i)} \rangle \tag{21}$$

Recognizing that the first term on the right-hand side is exactly the definition of the margin under the canonical prompt $p_0$, we substitute $m_{y,k}(x, p_0) = \langle f(x), w_{y,k}(p_0) \rangle$ back into the equation:

$$m_{y,k}(x, p_i) = m_{y,k}(x, p_0) + \langle f(x), \Delta w_{y,k}^{(i)} \rangle \tag{22}$$

This concludes the proof, mathematically demonstrating that the prompt-induced fluctuation in the decision margin is strictly determined by the geometric projection of the image representation $f(x)$ onto the boundary shift direction $\Delta w_{y,k}^{(i)}$. □

### A.2 Proof of Proposition 2

**Proposition 2** (Prompt-Induced Decision Boundary Rotation)**.** *Given a canonical prompt $p_0$ and a semantic substitution $p_i$, the decision boundary hyperplanes $\mathcal{H}_{y,k}(p_0)$ and $\mathcal{H}_{y,k}(p_i)$ intersect at a dihedral angle $\theta$ given by:*

$$\theta = \arccos\left( \frac{\langle w_{y,k}(p_0), w_{y,k}(p_i) \rangle}{\|w_{y,k}(p_0)\| \|w_{y,k}(p_i)\|} \right) \tag{23}$$

*Proof.* For a given prompt $p$, the decision boundary $\mathcal{H}_{y,k}(p)$ is a $(d-1)$-dimensional hyperplane in $\mathbb{R}^d$ defined as

$$\mathcal{H}_{y,k}(p) = \{ v \in \mathbb{R}^d \mid \langle v, w_{y,k}(p) \rangle = 0 \},$$

where $w_{y,k}(p) = t_p(y) - t_p(k) \neq 0$. Since the classifier has no bias term, all such hyperplanes pass through the origin and are uniquely determined by their normal vectors.

Therefore, the angle between $\mathcal{H}_{y,k}(p_0)$ and $\mathcal{H}_{y,k}(p_i)$ is defined as the angle between their normal vectors $w_{y,k}(p_0)$ and $w_{y,k}(p_i)$. By the definition of the angle in an inner product space,

$$\theta = \arccos\left(\frac{\langle w_{y,k}(p_0), w_{y,k}(p_i)\rangle}{\|w_{y,k}(p_0)\|\|w_{y,k}(p_i)\|}\right).$$

Since $w_{y,k}(p)$ depends on the prompt through $t_p(\cdot)$, a semantic substitution induces a change in the normal vector, resulting in a corresponding rotation of the decision boundary. □

### A.3 Proof of Theorem 1

**Theorem 1** (Adversarial Margin Amplification under Boundary Rotation)**.** *Let $x_{adv}$ be an adversarial example optimized under prompt $p_0$ with attack strength $\alpha > 0$. Under a prompt perturbation $p_i$, the margin shift decomposes as:*

$$m_{y,k}(x_{adv}, p_i) - m_{y,k}(x_{adv}, p_0) = \langle f(x), \Delta w_{y,k}^{(i)}\rangle - \alpha\langle w_{y,k}(p_0), \Delta w_{y,k}^{(i)}\rangle + \langle \zeta, \Delta w_{y,k}^{(i)}\rangle \quad (24)$$

*Proof.* According to Proposition 1, the shift in the classification margin for any arbitrary image under a prompt perturbation $p_i$ is determined by the geometric projection of its feature representation onto the boundary shift vector $\Delta w_{y,k}^{(i)}$. For an adversarial example $x_{adv}$, this margin variation is explicitly expressed as:

$$\Delta m = m_{y,k}(x_{adv}, p_i) - m_{y,k}(x_{adv}, p_0) = \langle f(x_{adv}), \Delta w_{y,k}^{(i)}\rangle \quad (25)$$

Recalling the composition of the adversarial image embedding, we have $f(x_{adv}) = f(x) + \Delta f$, where $f(x)$ represents the original clean semantic feature and $\Delta f$ denotes the adversarial perturbation in the latent space. As established, the adversary actively optimizes $\Delta f$ to cross the canonical decision boundary by aligning it against the canonical normal vector $w_{y,k}(p_0)$. Substituting the orthogonal decomposition of the adversarial shift $\Delta f = -\alpha w_{y,k}(p_0) + \zeta$ into the feature representation yields:

$$f(x_{adv}) = f(x) - \alpha w_{y,k}(p_0) + \zeta \quad (26)$$

By substituting Equation 26 into the inner product in Equation 25, the margin shift becomes:

$$\Delta m = \langle f(x) - \alpha w_{y,k}(p_0) + \zeta, \Delta w_{y,k}^{(i)}\rangle \quad (27)$$

Utilizing the fundamental distributive property of the inner product in the Euclidean space, we can rigorously decouple this overall shift into three distinct geometric components:

$$\Delta m = \langle f(x), \Delta w_{y,k}^{(i)}\rangle - \alpha\langle w_{y,k}(p_0), \Delta w_{y,k}^{(i)}\rangle + \langle \zeta, \Delta w_{y,k}^{(i)}\rangle \quad (28)$$

□

### A.4 Proof of Corollary 1

**Corollary 1** (Prompt Sensitivity Gap)**.** *Let $\mathcal{P}_{test} = \{p_0\} \cup \mathcal{P}$ be a distribution of semantic prompts, including the canonical prompt and several variants. Define the prompt sensitivity $\mathcal{S}(x)$ as the variance of the margin across $\mathcal{P}_{test}$:*

$$\mathcal{S}(x) = \mathrm{Var}_{p_i \sim \mathcal{P}_{test}}\left(m_{y,k}(x, p_i)\right) \quad (29)$$

*Assuming the prompt-induced boundary shift $\Delta w_{y,k}^{(i)}$ has isotropic variance properties, there exists a mathematical gap between the expected sensitivities of adversarial and clean examples. Because the attack strength $\alpha$ is substantial, this gap is strictly lower-bounded by the adversarial penalty:*

$$\mathbb{E}_x[\mathcal{S}(x_{adv})] - \mathbb{E}_x[\mathcal{S}(x)] \geq \alpha^2 \mathrm{Var}_{p_i}\left(\langle w_{y,k}(p_0), \Delta w_{y,k}^{(i)}\rangle\right) \quad (30)$$

*Proof.* By definition, the prompt sensitivity $\mathcal{S}(x_{adv})$ is the variance of the classification margin across the distribution of test prompts $\mathcal{P}_{test}$:

$$\mathcal{S}(x_{adv}) = \mathrm{Var}_{p_i}\left(m_{y,k}(x_{adv}, p_i)\right) \tag{31}$$

Since the margin under the canonical prompt $m_{y,k}(x_{adv}, p_0)$ is a deterministic constant with respect to the prompt perturbation, shifting the random variable by a constant does not change its variance. Thus, the variance of the absolute margin is strictly equal to the variance of the margin shift:

$$\mathcal{S}(x_{adv}) = \mathrm{Var}_{p_i}\left(m_{y,k}(x_{adv}, p_i) - m_{y,k}(x_{adv}, p_0)\right) \tag{32}$$

Crucially, Theorem 1 provides the exact geometric decomposition of this margin shift. By directly substituting the result of Theorem 1 into the variance operator, we obtain:

$$\mathcal{S}(x_{adv}) = \mathrm{Var}_{p_i}\left(\langle f(x), \Delta w_{y,k}^{(i)}\rangle - \alpha\langle w_{y,k}(p_0), \Delta w_{y,k}^{(i)}\rangle + \langle \zeta, \Delta w_{y,k}^{(i)}\rangle\right) \tag{33}$$

For notation simplicity, let the three latent components be $A = \langle f(x), \Delta w_{y,k}^{(i)}\rangle$, $B = -\alpha\langle w_{y,k}(p_0), \Delta w_{y,k}^{(i)}\rangle$, and $C = \langle \zeta, \Delta w_{y,k}^{(i)}\rangle$. The variance of their sum mathematically expands as:

$$\mathrm{Var}(A+B+C) = \mathrm{Var}(A)+\mathrm{Var}(B)+\mathrm{Var}(C)+2\mathrm{Cov}(A,B)+2\mathrm{Cov}(B,C)+2\mathrm{Cov}(A,C) \tag{34}$$

Under the assumption that the prompt-induced boundary shift $\Delta w_{y,k}^{(i)}$ has isotropic variance properties in the embedding space, its covariance matrix is proportional to the identity matrix, i.e., $\Sigma = \sigma^2\mathbf{I}$. Consequently, the covariance between the projections of any two deterministic vectors $u$ and $v$ onto the random vector $\Delta w_{y,k}^{(i)}$ is proportional to their inner product:

$$\mathrm{Cov}(\langle u, \Delta w_{y,k}^{(i)}\rangle, \langle v, \Delta w_{y,k}^{(i)}\rangle) = u^T\Sigma v = \sigma^2\langle u, v\rangle \tag{35}$$

Applying this geometric property to evaluate the covariance terms: For $\mathrm{Cov}(B, C)$, since the residual $\zeta$ is strictly defined in Theorem 1 as orthogonal to the normal vector $w_{y,k}(p_0)$ (i.e., $\langle w_{y,k}(p_0), \zeta\rangle = 0$), we have:

$$\mathrm{Cov}(B, C) = -\alpha\sigma^2\langle w_{y,k}(p_0), \zeta\rangle = 0 \tag{36}$$

Thus, the covariance between the adversarial targeted shift and the residual strictly vanishes. For the cross-terms involving the clean feature representation $f(x)$ (i.e., $\mathrm{Cov}(A, B)$ and $\mathrm{Cov}(A, C)$), we take the expectation $\mathbb{E}_x$ over the dataset distribution. In the high-dimensional embedding space, the natural variation directions of clean image features $f(x)$ are statistically independent of the instance-specific adversarial perturbation vectors. Thus, the expected inner products $\mathbb{E}_x[\langle f(x), w_{y,k}(p_0)\rangle]$ and $\mathbb{E}_x[\langle f(x), \zeta\rangle]$ average out as non-correlated baseline constants, yielding $\mathbb{E}_x[\mathrm{Cov}(A, B)] = 0$ and $\mathbb{E}_x[\mathrm{Cov}(A, C)] = 0$. Taking the expectation $\mathbb{E}_x$ on both sides of the variance expansion globally eliminates all cross-covariance terms:

$$\mathbb{E}_x[\mathcal{S}(x_{adv})] = \mathbb{E}_x[\mathrm{Var}(A)] + \mathbb{E}_x[\mathrm{Var}(B)] + \mathbb{E}_x[\mathrm{Var}(C)] \tag{37}$$

Noting that the expected variance of the clean sample margin is strictly the variance of the semantic component, $\mathbb{E}_x[\mathcal{S}(x)] = \mathbb{E}_x[\mathrm{Var}(A)]$, and substituting back the original expressions for $B$ and $C$, we obtain the exact variance decomposition:

$$\mathbb{E}_x[\mathcal{S}(x_{adv})] = \mathbb{E}_x[\mathcal{S}(x)] + \alpha^2\mathrm{Var}_{p_i}\left(\langle w_{y,k}(p_0), \Delta w_{y,k}^{(i)}\rangle\right) + \mathrm{Var}_{p_i}\left(\langle \zeta, \Delta w_{y,k}^{(i)}\rangle\right) \tag{38}$$

**Algorithm 1** Prompt Ensemble Image Purification (PEIP)

**Require:** Test image $x$, prompts $p_i \in \mathcal{P}$, iterations $K$, step size $\eta$, budget $\epsilon_d$, weight $\lambda$.
**Ensure:** Purified robust prediction $\hat{y}$

{**Phase 1: Early Exit Mechanism**}
Compute initial predictive probability $z_{i,c} = \langle f(x), t_{p_i}(c) \rangle$ for all $i \in [0, M], c \in [1, C]$
Compute initial sensitivity: $\mathcal{S}_{init} = \max_c \mathrm{Var}_i(\mathrm{softmax}(z_{i,c}))$
**if** $\mathcal{S}_{init} \leq \tau$ **then**
  **return** $\hat{y} = \arg\max_c z_{0,c}$
**end if**
{**Phase 2: Purification Loop**}
Obtain pseudo-label $\tilde{y} = \arg\max_c \left( \bar{P}_{aug}(c|x) - P_0(c|x) \right)$
Initialize $x^{(0)} \leftarrow x + \epsilon_d \cdot \mathcal{U}(-1, 1)$ {Superscript denotes the iteration number in the loop.}
**for** $k = 1, \ldots, K$ **do**
  Compute logits $z_{i,c}^{(k-1)} = \langle f(x^{(k-1)}), t_{p_i}(c) \rangle$
  Compute predictive distributions $P_i(c|x^{(k-1)}) = \mathrm{softmax}(z_{i,c}^{(k-1)})$
  $\mathcal{L}_{var} \leftarrow \sum_{c=1}^{C} \bar{P}_{aug}(c|x) \log \left( \frac{\bar{P}_{aug}(c|x)}{P_0(c|x^{(k-1)})} \right)$
  $\mathcal{L}_{sem} \leftarrow -\log P_0(\tilde{y}|x^{(k-1)})$
  $\mathcal{L}_{total} \leftarrow \mathcal{L}_{var} + \lambda \mathcal{L}_{sem}$
  $x^{(k)} \leftarrow x^{(k-1)} - \eta \cdot \mathrm{sign}(\nabla_{x^{(k-1)}} \mathcal{L}_{total})$ {Gradient Descent}
  $x^{(k)} \leftarrow \Pi_{\mathcal{B}(x, \epsilon_d)}(x^{(k)})$ {Project back to valid image bounds}
**end for**
**return** $\hat{y} = \arg\max_c \langle f(x^{(K)}), t_{p_0}(c) \rangle$

Since variance is inherently non-negative, the variance of the residual term strictly satisfies $\mathrm{Var}_{p_i}(\langle \zeta, \Delta w_{y,k}^{(i)} \rangle) \geq 0$. Discarding this non-negative term explicitly establishes the strict lower bound for the sensitivity gap:

$$\mathbb{E}_x[\mathcal{S}(x_{adv})] - \mathbb{E}_x[\mathcal{S}(x)] \geq \alpha^2 \mathrm{Var}_{p_i} \left( \langle w_{y,k}(p_0), \Delta w_{y,k}^{(i)} \rangle \right) \tag{39}$$

□

## B Overall Algorithm of PEIP

Algorithm 1 summarizes the complete inference flow of our PEIP framework. The procedure is conceptually divided into two phases. Phase 1 (Early Exit Mechanism): Given an unverified test image $x$, we first perform a single forward pass across the prompt ensemble to compute the initial sensitivity $\mathcal{S}_{init}$. If $\mathcal{S}_{init}$ falls below the statistically calibrated threshold $\tau$, the image is deemed benign, and the framework instantly yields the prediction, effectively bypassing any optimization overhead. Phase 2 (Purification Loop): If high sensitivity is detected ($\mathcal{S}_{init} > \tau$), an iterative purification process is triggered to break adversarial alignment. The process first identifies a pseudo-label $\tilde{y}$ by selecting the category with the maximum probability rebound from the canonical prediction $P_0$ to the ensemble consensus $\bar{P}_{aug}$. After initializing the image with minimal random noise to break geometric symmetry, the framework conducts projected gradient descent for $K$ steps. Each iteration jointly optimizes a Variance Suppression Objective $\mathcal{L}_{var}$, which minimizes the KL divergence between the compromised prediction and the robust consensus to neutralize prompt-specific adversarial shifts, and a Semantic Restoration Objective $\mathcal{L}_{sem}$, which employs a cross-entropy objective to pull the prediction toward the pseudo-label $\tilde{y}$. By minimizing the combined objective $\mathcal{L}_{total}$, the framework expels adversarial perturbations while remaining within the budget $\epsilon_d$, resulting in a robust, prompt-invariant final prediction $\hat{y}$.

## C Prompt Template

In this work, we align our prompt template settings with the most recent work TTC [43]. Specifically, we adopt a dataset-specific canonical text prompt to match the image features, guide the generation

of adversarial perturbations, and evaluate the performance [1]. Furthermore, we introduce $M = 4$ substitution variants for each dataset. To source these variants, we primarily refer to the official CLIP repository [2]. If the repository provides sufficient alternative templates of a specific dataset (e.g., ImageNet), we randomly sample four prompt variants and repeat this process to construct three independent sets for comprehensive evaluation. In cases where the official templates are insufficient, we leverage GPT-5 to generate the remaining variants using the query suggested by Large Language Model (LLM) below.

**LLM Instruction for Generating Prompt Variants**

```
Act as a prompt engineer specialized in Vision-Language Models. Your goal is
to generate 4 subtle variants of the following canonical prompt: '[canonical
prompt]'.

Core Objective: The variants must stay semantically near the original to
maintain the class identity, but introduce minor geometric shifts in the CLIP
embedding space.

Strict Constraints:
1. Preserve Key Semantic Anchors: Do not change the core category name
   {class} or any dataset-specific context (e.g., ', a type of pet').
2. Modify Non-Essential Elements Only: Limit changes to articles (a/an/the),
   adjectives, or subtle rephrasing of the template prefix.
3. Diversity of Perspective: Ensure the 4 variants represent slightly
   different observational nuances:
   • Variant 1 (Article): Adjust specificity (e.g., 'a' → 'the').
   • Variant 2 (Quality): Add a subtle quality adjective (e.g., 'good',
     'clear').
   • Variant 3 (Framing): Add a neutral framing descriptor (e.g.,
     'centered').
   • Variant 4 (Syntactic): Minor reordering of the template prefix.

Output Format: Provide only the 4 template strings compatible with CLIP's
tokenization.
```

## D Results on More CLIP Models

To evaluate the defense generality of PEIP across different CLIP architectures, we further conduct extensive experiments on CLIP ViT-B/32 and ViT-L/14 models. As shown in Table 7, PEIP achieves stronger robustness against 10-step PGD attacks on both models across four types of datasets, with notable improvements on General object recognition datasets (+6.7% and +5.7% on CLIP ViT-B/32 and CLIP ViT-L/14, respectively, compared to TTC). Moreover, PEIP consistently maintains competitive clean accuracy (with only a 0.2% and 1.0% average performance drop on CLIP ViT-B/32 and CLIP-L/14, respectively) while enhancing adversarial robustness.

## E Statistical Observation

We randomly sample 2,000 images on each dataset to analyze the statistical behaviors of different components.

For clean images, we find that only approximately 57% of pseudo-labels are consistent with the ground-truth, indicating that the most responsive probability across prompt variants does not necessarily correspond to the ground-truth class. As a result, $\mathcal{L}_{sem}$ may degrade clean accuracy to some extent. Meanwhile, about 98.3% of clean samples exhibit an initial sensitivity $S_{init} \leq \tau$, meaning that the Early Exit mechanism bypasses the majority of clean inputs, effectively shielding them from the adverse effects of $\mathcal{L}_{sem}$.

For adversarial settings, we produce adversarial perturbation on the images under a PGD attack with a budget of $1/255$ and 10 iterations. Approximately 93% of pseudo-labels align with the ground-truth

[1] https://github.com/Sxing2
[2] https://github.com/openai/CLIP

Table 7: Comparison of top-1 zero-shot classification accuracy (%) on CLIP ViT-B/32 and CLIP ViT-L/14 under 10-step PGD ($\ell_\infty = 1/255$).

| Method | CLIP ViT-B/32 | | | | | | | | CLIP ViT-L/14 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | General | | Fine-Grained | | Scene | | Domain | | General | | Fine-Grained | | 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 | 36.8 | 0.4 | 35.9 | 6.6 | 82.1 | 4.0 | 79.2 | 0.3 | 42.3 | 0.2 | 44.8 | 0.3 |
| CLIP-FT | 73.7 | 5.1 | 60.4 | 1.0 | 36.2 | 0.4 | 37.7 | 1.3 | 78.6 | 8.4 | 75.0 | 2.1 | 40.4 | 1.9 | 39.7 | 3.3 |
| TeCoA | 57.7 | 37.9 | 35.1 | 18.4 | 19.1 | 9.5 | 23.5 | 17.6 | 72.8 | 48.6 | 41.6 | 24.0 | 25.7 | 12.7 | 28.8 | 21.1 |
| PMG-AFT | 59.9 | 39.2 | 38.7 | 22.5 | 20.6 | 11.3 | 22.8 | 19.1 | 67.1 | 47.4 | 44.5 | 26.0 | 23.0 | 15.3 | 25.6 | 20.7 |
| FARE | 59.6 | 33.1 | 52.6 | 14.8 | 23.2 | 8.5 | 27.7 | 18.1 | 75.3 | 40.3 | 63.9 | 19.3 | 32.7 | 9.4 | 32.7 | 11.0 |
| R-TPT | 75.0 | 42.7 | 68.1 | 47.5 | 34.6 | 23.0 | 34.8 | 21.9 | 80.6 | 50.8 | 77.4 | 49.0 | 39.3 | 25.8 | 42.4 | 25.3 |
| TAPT | 75.5 | 41.8 | 66.9 | 44.1 | 35.4 | 22.8 | 34.5 | 23.0 | 79.7 | 46.7 | 78.5 | 46.2 | 40.8 | 25.6 | 43.0 | 27.4 |
| TTE | **76.3** | 47.3 | 69.8 | 27.9 | **36.5** | 9.6 | 37.7 | 19.3 | 81.4 | 54.5 | 76.6 | 42.2 | **47.8** | 19.1 | **43.7** | 15.5 |
| Anti-Adv | 75.3 | 22.0 | 63.4 | 10.6 | 32.9 | 4.1 | 35.8 | 3.7 | 80.2 | 26.4 | 75.3 | 14.5 | 39.7 | 6.2 | 40.7 | 8.1 |
| HD | 73.5 | 20.6 | 67.0 | 7.4 | 34.2 | 3.2 | **39.8** | 15.6 | 78.7 | 25.4 | 73.4 | 8.7 | 36.0 | 6.4 | 37.8 | 18.2 |
| TTC | 74.5 | 46.9 | 66.5 | 47.2 | 32.9 | 23.7 | 34.2 | 25.5 | 80.5 | 51.2 | **79.3** | 48.4 | 39.0 | 26.3 | 43.2 | 28.1 |
| PEIP | 76.1$^{\downarrow 0.6}$ | **53.6**$^{\uparrow 49.6}$ | **71.0**$^{\downarrow 0.8}$ | **50.7**$^{\uparrow 50.4}$ | 35.8$^{\downarrow 1.0}$ | **27.5**$^{\uparrow 27.1}$ | 37.7$^{\uparrow 1.8}$ | **27.1**$^{\uparrow 20.5}$ | **81.6**$^{\downarrow 0.5}$ | **57.5**$^{\uparrow 53.5}$ | 78.5$^{\downarrow 0.7}$ | **55.1**$^{\uparrow 54.8}$ | 40.8$^{\downarrow 1.5}$ | **31.6**$^{\uparrow 31.4}$ | 43.4$^{\downarrow 1.4}$ | **29.5**$^{\uparrow 29.2}$ |

Table 8: More ablation of each component on clean and adversarial performance.

| $\mathcal{L}_{var}$ | $\mathcal{L}_{sem}$ | E.E. | ImageNet | | OxfordPets | | SUN397 | | EuroSAT | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Clean | Rob. | Clean | Rob. | Clean | Rob. | Clean | Rob. |
| | | | **59.7** | 1.2 | **87.4** | 1.0 | **58.5** | 1.1 | **42.6** | 0.0 |
| ✓ | | | 55.2 | 27.4 | 83.6 | 40.5 | 56.4 | 35.3 | 42.0 | 12.7 |
| | ✓ | | 28.5 | 38.8 | 47.1 | 59.9 | 31.5 | 47.1 | 22.4 | 15.2 |
| ✓ | | ✓ | 58.3 | 25.6 | 86.8 | 38.6 | 57.0 | 33.7 | 42.5 | 11.8 |
| | ✓ | ✓ | 55.4 | 37.5 | 84.7 | 58.0 | 55.2 | 45.4 | 40.8 | 14.5 |
| ✓ | ✓ | | 51.3 | **41.2** | 80.2 | **64.8** | 52.8 | **47.9** | 37.3 | **16.3** |
| ✓ | ✓ | ✓ | 57.8 | 40.7 | 86.9 | 64.3 | 57.0 | 47.3 | 42.5 | 15.9 |

labels, making $\mathcal{L}_{sem}$ effective for optimizing and recovering benign semantic cues. Moreover, around 3.7% of adversarial samples exhibit an initial sensitivity below the statistical threshold $\tau$, leading them to be bypassed and thus slightly reducing robustness.

# F More Ablation Study

In this section, we investigate more impacts of key components and hyperparameters on both clean and robust performance.

**More Analysis of Component Combinations in PEIP.** Table 8 provides additional ablation results beyond those reported in Table 4 of the main paper. As demonstrated in the table, incorporating the Early Exit mechanism significantly preserves clean performance, especially by mitigating the negative effect of $\mathcal{L}_{sem}$ (e.g., improving clean accuracy from 28.5% to 55.4% on ImageNet). For clean images, the most responsive pseudo-label is often inconsistent with the ground-truth (43% as demonstrated in Appendix E). The Early Exit mechanism bypasses most clean samples (98.3%), preventing them from being misled by $\mathcal{L}_{sem}$. In addition, $\mathcal{L}_{var}$ further helps maintain clean accuracy by enforcing consistency with the prompt consensus.

On the other hand, Early Exit mechanism slightly reduces adversarial robustness, as a small portion of adversarial samples are also bypassed. Specifically, under a PGD attack with budget $1/255$ and 10 iterations, about 3.7% of adversarial samples exhibit initial sensitivity below the statistical threshold $\tau$, somewhat limiting the adversarial robustness.

**Impact of Iteration Number $K$.** The iteration step $K$ controls the depth of test-time purification. As shown in Figure 5(a), setting $K = 3$ already yields substantial robustness gains, while further increasing $K$ brings only marginal improvement at a much higher computational cost. Moreover, due to the Early Exit strategy, most clean samples bypass the purification loop. As a result, clean performance remains largely unaffected (Figure 6(a)). Therefore, we set $K = 3$ as a practical trade-off between robustness and efficiency.

**Sensitivity to the Balance Factor $\lambda$.** The balance factor $\lambda$ controls the trade-off between $\mathcal{L}_{var}$ and $\mathcal{L}_{sem}$. As shown in Figure 5(b), the best robustness is achieved at $\lambda = 1.0$ in most cases (Early Exit is not applied here). When $\lambda$ is too small, the optimization is dominated by $\mathcal{L}_{var}$. If the

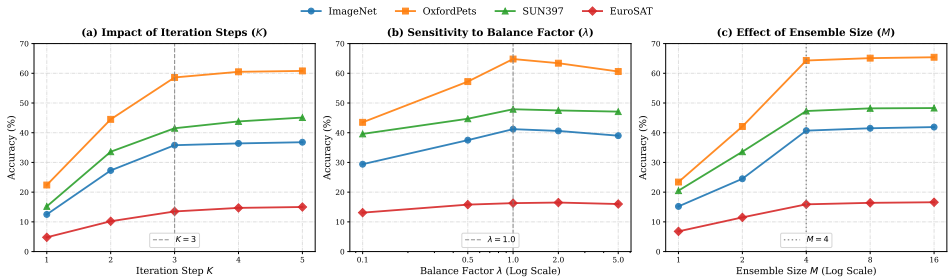



Figure 5: Ablation study of adversarial robust on iteration step $K$, balance factor $\lambda$, and ensemble size $M$.

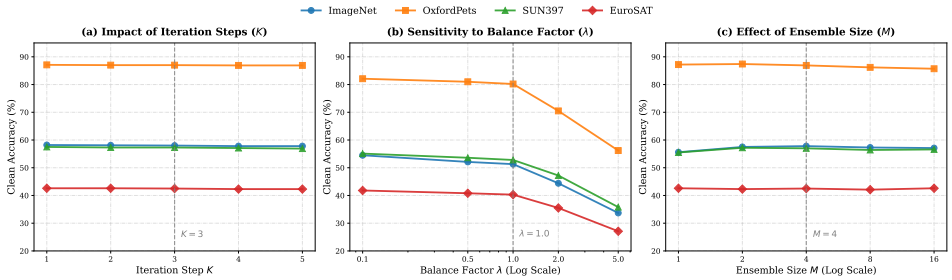



Figure 6: Ablation study of clean performance on iteration step $K$, balance factor $\lambda$, and ensemble size $M$.

ensemble consensus is also biased by the adversary, the update is insufficient to correct the canonical prediction, thereby resulting in insufficient robustness. In contrast, a large $\lambda$ degrades clean accuracy, as shown in Figure 6(b). For clean samples, prompt variants typically induce only minor changes in the ground-truth class, making the pseudo-label less reliable (only $\sim$67%). As a result, $\mathcal{L}_{sem}$ may introduce incorrect guidance. Without the regularizing effect of $\mathcal{L}_{var}$, this leads to a significant drop in clean accuracy.

**Effect of the Ensemble Size $M$.** The prompt ensemble size $M$ defines the number of distinct semantic views used to estimate the variance. As depicted in Figure 5(c), deploying PEIP with insufficient prompts ($M \leq 2$) fails to provide enough geometric constraints to accurately expose the adversarial perturbation, yielding weak defense performance. As $M$ increases to 4, both robustness and clean accuracy surge significantly, as the model successfully captures the invariant semantic core across diverse textual views. Further expanding the ensemble (e.g., $M = 8$ or 16) brings marginal robustness gains, but it increases the computational overhead. Hence, $M = 4$ is adopted as an elegant balance between multi-view semantic coverage and real-world deployment efficiency.

## G Multi-Prompt Evaluation

While prior adversarial-defense protocols for CLIP typically adopt a single canonical prompt, standard zero-shot CLIP evaluation often ensembles multiple dataset-specific templates. We therefore evaluate PEIP under a stronger multi-prompt setting in which the prompt-ensemble prediction is used for both adversarial-example generation and final evaluation. Because not all 16 benchmarks provide multiple official templates, we conduct this experiment on the 12 datasets listed in Table 9.

Let $N$ denote the number of official templates available for a dataset and used by the attacker and evaluator. Since $N$ varies across datasets, we follow the prompt-generation procedure in Appendix C to construct $M \in \{N, 2N, 4N, 8N\}$ additional variants for PEIP purification; we also include $M = 0$, which applies the attacked prompt-ensemble prediction without PEIP purification. All remaining settings follow Table 1, including the 10-step PGD attack with an $\ell_\infty$ budget of $1/255$.

Table 9: Robust zero-shot classification accuracy (%) under multi-prompt attack and evaluation. $N$ is the number of official templates used for attack generation and evaluation, and $M$ is the number of additional prompt variants used by PEIP. "Avg." is the unweighted average over the 12 evaluated datasets.

| $M/N$ | **CIFAR10** | **CIFAR100** | **STL10** | **ImageNet** | **Caltech101** | **Caltech256** |
|---|---|---|---|---|---|---|
| 0 | 0.5 | 0.2 | 8.7 | 1.1 | 11.5 | 7.3 |
| 1 | 29.1 | 16.3 | 47.8 | 24.7 | 33.5 | 37.3 |
| 2 | 41.8 | 20.9 | 65.0 | 34.3 | 58.5 | 54.7 |
| 4 | 47.4 | 24.0 | 73.7 | 36.1 | 62.5 | 59.9 |
| 8 | 48.7 | 24.5 | 74.2 | 36.1 | 63.8 | 60.0 |

| $M/N$ | **StanfordCars** | **SUN397** | **Country211** | **FGVCAircraft** | **EuroSAT** | **DTD** | **Avg.** |
|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.6 | 0.1 | 0.0 | 0.0 | 2.4 | 2.7 |
| 1 | 25.0 | 26.6 | 6.8 | 9.3 | 7.1 | 15.9 | 23.3 |
| 2 | 33.7 | 40.2 | 9.2 | 11.5 | 10.6 | 22.3 | 33.6 |
| 4 | 37.2 | 43.8 | 10.5 | 13.3 | 12.1 | 25.2 | 37.1 |
| 8 | 37.6 | 44.3 | 10.7 | 14.1 | 13.0 | 25.5 | 37.7 |

As shown in Table 9, robustness improves substantially as the purification ensemble grows. Increasing $M$ from 0 to $N$ raises the average accuracy from $2.7\%$ to $23.3\%$, while $M = 2N$ further improves it to $33.6\%$. Notably, expanding the ensemble to $M = 4N$ provides an additional $3.5\%$ gain and achieves $37.1\%$, already approaching the $37.7\%$ obtained with $M = 8N$. Thus, $M = 4N$ emerges as an effective operating point. It provides sufficient prompt diversity for strong purification while avoiding the increased computational cost of a larger ensemble, whose additional robustness gain is only marginal.

This setting is stronger and partially adaptive because the attack targets an ensemble prediction rather than a single canonical prompt. Although robustness remains slightly below that of the standard single-prompt attack in Table 1, PEIP retains meaningful robustness when supplied with a sufficiently diverse purification ensemble. In particular, the strong performance at $M = 4N$ demonstrates that PEIP can effectively withstand multi-prompt attacks without requiring an excessively large ensemble. These results show that PEIP is not limited to attacks generated from a single prompt.

## H Limitation and Future Work

First, we evaluate adaptive PGD, EoT, and different- and multi-prompt attacks to account for PEIP's optimization and randomness. Nevertheless, stronger joint attacks may exist, so our results reflect empirical robustness under the evaluated threat models rather than certified guarantees. Moreover, PEIP relies on the Prompt Sensitivity Gap, which may narrow under broader prompt ensembles. Moreover, adversarial inputs below the Early Exit threshold may bypass purification, creating a trade-off among clean accuracy, efficiency, and robustness. In addition, more potent adversarial attacks [49, 18, 51, 50, 1] in semantic segmentation should also be evaluated in the future work. And PEIP is designed primarily for fixed-label zero-shot classification. Its semantic-restoration and Early Exit components do not directly extend to dense or open-ended outputs. Generalizing the full framework to such tasks remains future work.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction have clearly stated the proposed PEIP framework, its theoretical motivation, and empirical results, strictly aligning with the actual methodological contributions and experimental findings presented in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitation of this paper in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We have given the full set of assumption and a complete proof of the proposition, theorem, and corollary in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have exhibited the algorithm of the proposed PEIP in Algorithm 1, and report all the experimental setups in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is available at `https://github.com/yuxuan357/PEIP`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings have been provided in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Each experiment is conducted three times, and the corresponding standard deviations for the main results are reported in Table 1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided the type of GPU and the execution time in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the potential positive societal impact in the abstract, introduction, and conclusion, highlighting that PEIP can help CLIP defend against various adversarial attacks while maintaining efficient inference in practical scenarios where adversarial samples are rare. However, there remains ample room for improvement in terms of clean accuracy.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly cited the used code, data, and models in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.