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

NeurIPS 2026

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

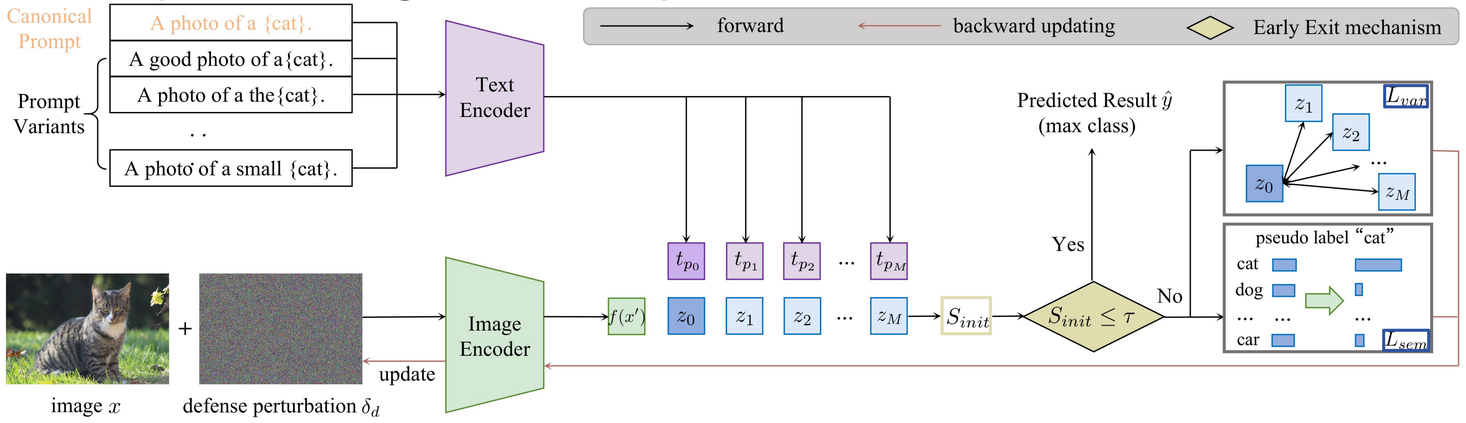
[2] School of Computer Science and Technology, University of Science and Technology of China

Vision-Language Models like CLIP exhibit remarkable zero-shot capabilities but remain vulnerable to adversarial attacks. Existing defenses often suffer from catastrophic forgetting, high computational costs, or limited robustness against strong attacks. In this work, we reveal that adversarial perturbations tend to overfit the decision boundary of a single canonical prompt, leading to a Prompt Sensitivity Gap, where adversarial examples exhibit greater prediction variance across semantic prompt variants than benign images. Building on this insight, we propose Prompt Ensemble Image Purification (PEIP), an efficient test-time defense that jointly minimizes prediction variance to disrupt adversarial alignment and reinforces the most responsive class to recover semantic consistency. A statistically calibrated Early Exit mechanism further improves efficiency by bypassing purification for benign images. Extensive experiments on 16 classification benchmarks, multiple CLIP architectures, and zero-shot semantic segmentation demonstrate that PEIP achieves state-of-the-art adversarial robustness while preserving zero-shot accuracy with high efficiency.

45.0%
robust accuracy / 16 datasets

7.0×
faster than TTC at 1% attacks

01 Prompt ensembles guide selective purification



02 Prompt Sensitivity Gap

Prompt changes expose unstable adversarial predictions.

Early Exit (τ = 0.016) bypasses 98.3% of clean inputs.

03 Dual-objective purification

Freeze the initial prompt consensus as a semantic anchor. Restore the class with the largest probability rebound.

$$\mathcal{L}_{\text{var}} = \mathrm{KL}\big(\bar{P}_{\text{aug}} \,\|\, P_0(x+\delta_d)\big)$$
$$\mathcal{L}_{\text{sem}} = -\log P_0(\tilde{y} \mid x+\delta_d)$$
$$\delta_d^* = \arg\min_{\|\delta_d\|_\infty \le \epsilon_d} \big[\mathcal{L}_{\text{var}} + \lambda \mathcal{L}_{\text{sem}}\big]$$

04 Robustness across 16 benchmarks

Method	Clean (%)	Robust (%)
CLIP	61.5	2.7
PMG-AFT	42.3	28.8
R-TPT	59.8	39.8
TTC	59.8	39.2
PEIP (ours)	**60.5**	**45.0**

+5.8 points robustness over TTC;
clean accuracy stays within 1.0 point of CLIP.

05 A fully informed adaptive adversary

Attack the full defense; EoT uses 10 random initializations.

Attack	10 steps[1]	50 steps[2]
PGD	45.0	28.9
Adaptive PGD	42.9	27.3
Adaptive PGD-EoT	**41.2**	**25.9**

[1] ℓ_∞ = 1/255 [2] ℓ_∞ = 4/255; mean accuracy (%).

1% attacks: 3.97 ms vs. TTC 27.98 ms / image.
Single RTX 5090; 5% attacks: 4.94 ms vs. 28.13 ms.

06 Zero-shot segmentation

Pascal VOC robust mIoU: 48.5 / 54.4
MaskCLIP / ClearCLIP • PGD-10, ℓ_∞ = 1/255 • Lvar only