TAPT: Test-Time Adversarial Prompt Tuning for Robust Inference in Vision-Language Models

Supplementary Material

7. Hand-Crafted Prompt Templates

We report the hand-crafted prompt templates used for zero-shot CLIP and APT methods, evaluated on 11 benchmark datasets in Table 4.

Dataset	Template
ImageNet	`"a photo of a <class>."`
Caltech101	`"a photo of a <class>."`
DTD	`"<class> texture."`
EuroSAT	`"a centered satellite photo of <class>."`
Pets	`"a photo of a <class>, a type of pet."`
Aircraft	`"a photo of a <class>, a type of aircraft."`
Food101	`"a photo of a <class>, a type of food."`
Flowers	`"a photo of a <class>, a type of flower."`
Cars	`"a photo of a <class>."`
SUN397	`"a photo of a <class>."`
UCF101	`"a photo of a person doing <class>."`

Table 4. Hand-crafted prompt templates on 11 image datasets. The placeholder `<class>` is replaced with the corresponding label.

8. Adaptive Attacks

We introduced two adaptive attacks. The first is a **defense-aware attack**, where the attacker is aware of our test-time defense that adaptively adjusts prompts based on the input. This allows the attacker to leverage input diversity to enhance the transferability of adversarial examples, such as the DI attack. The second is a **white-box attack** on our defense, where the attacker adversarially optimizes against both the CLIP cosine similarity loss and the alignment loss. As shown in Table 5, the defense-aware attack performs similarly to the baseline PGD. While the white-box attack is effective, it becomes much slower to converge. Our defense creates a strategic dilemma for attackers: ignoring it enables better adversarial robustness performance, while accounting for it diminishes the attack's effectiveness.

	ImageNet	Caltech101	DTD	EuroSAT	Pets	Aircraft	Food101	Flowers	Cars	SUN397	UCF101	**Avg.**
Defense-aware	53.8	80.5	32.3	39.6	69.4	13.3	70.6	51.1	42.5	50.3	48.2	50.1
White-box	48.9	76.4	31.9	29.5	65.9	12.9	63.5	47.9	38.4	48.2	45.5	46.3

Table 5. Zero-shot adversarial robustness (%) of TAPT from ImageNet to downstream datasets, evaluated against adaptive attacks.

9. Prompt Depth and Prompt Length

Table 6 presents the chosen prompt hyperparameter for TAPT. We further investigated the effects of prompt depth and prompt length on robust accuracy. Figure 6 shows that increasing the prompt depth from 0 to 3 significantly improved robust accuracy, from 1.4% to 43.26%. Further increases in prompt depth yielded diminishing returns. Similarly, increasing the prompt length from 0 to 2 improved the robust accuracy from 1.4% to 49.92%, and further to 53.36% with the length $= 4$. However, robust accuracy dropped to 48.11% at the length $= 8$. These findings suggest that increasing prompt depth or length can enhance adversarial robustness, but excessive values may not lead to further improvements and could even reduce effectiveness.

Prompt Design	**Prompt Depth**	**Prompt Length**	
		Visual Prompts	Textual Prompts
Visual Only	9	2	0
V-L Joint	9	2	2
V-L Independent	9	2	2

Table 6. TAPT's hyperparameter settings include three prompt designs: visual only, V-L joint, and V-L independent. The number of prompt tokens in the visual and textual branches are denoted as visual prompt and textual prompt, respectively.

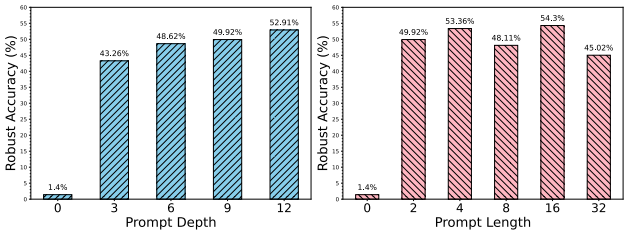

Figure 6. Effect of prompt depth and prompt length on ImageNet.

10. Effectiveness of TAPT Loss

We conducted an ablation analysis on various components of TAPT, specifically examining the impact of multi-view entropy minimization and adversarial-clean alignment. As shown in Table 7, the baseline APT method achieves a robust accuracy of 21.16%. Incorporating only the multi-view entropy minimization loss (TAPT†) significantly improves robust accuracy to 38.25%, while utilizing only the adversarial-clean alignment (TAPT‡) results in a robust accuracy of 28.74%. The best performance 48.18% is achieved when combining the multi-view entropy minimization with adversarial-clean alignment (TAPT), which suggests that these two components contribute to the im-

provement of adversarial robustness in complementary ways.

Method	Entropy Loss	Alignment Loss	Robust Accuracy
APT			21.16
TAPT†	✓		38.25
TAPT‡		✓	28.74
TAPT	✓	✓	**48.18**

Table 7. Analysis of the impact of entropy and alignment losses. The average robust accuracy (%) on ImageNet is reported. TAPT† denotes our method excluding the adversarial-clean alignment. TAPT‡ denotes our method excluding the multi-view entropy loss.