Towards Fine-Grained Robustness: Attention-Guided Test-Time Prompt Tuning for Vision-Language Models

Jia-Wei Hai[1], Yijun Wang[1], Xiu-Shen Wei[1,2*]

1. School of Computer Science and Engineering, and Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Southeast University, China.
2. School of Intelligence Science and Engineering, Southeast University, China.

1 - Motivation

Adversarial attacks corrupt image features, causing the existing Gradient Attention Rollout (GAR) fails to find discriminative fine-grained parts, making it hard to protect and use these parts during test-time adaptation.

3 - Contributions

- ✓ **We shift semantics-preserving augmentation from corrupted feature space to robust visual attention space.**
- ✓ **We refine the gradient signal of GAR with a robust term against adversarial perturbations, enabling effective identification of unperturbed semantic information under adversarial attacks.**
- ✓ **To our knowledge, A-TPT is the first method that exploits discriminative semantic regions for guiding test-time adaptation, particularly in fine-grained scenarios.**
- ✓ **A-TPT consistently outperforms existing SOTA methods on both clean and adversarial data.**

2 - The Proposed Method: A-TPT

$$\nabla_{\mathbf{T}^{(b)}(\mathbf{x})} S(x)$$

moves attribution from attention edges to token space, avoiding direct use of perturbation-sensitive edge-wise attention gradients.

$$\mathbf{W}^{(b)}(x) = \mathcal{N}\left(\left[\langle \mathbf{T}^{(b)}(x), \nabla_{\mathbf{T}^{(b)}(x)} S(x)\rangle_d\right]_+\right)$$

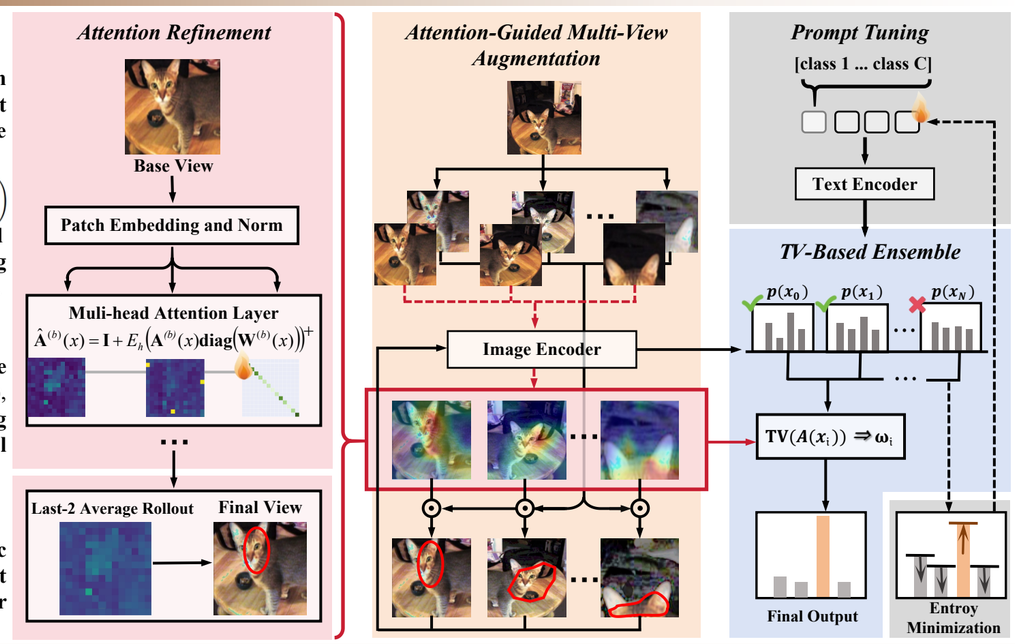
converts token gradients into bounded source-token weights by aggregating sensitivity over embedding dimension.

$$\hat{\mathbf{A}}^{(b)}(x) = \mathbf{I} + E_h\left(\mathbf{A}^{(b)}(x)\mathrm{diag}\left(\mathbf{W}^{(b)}(x)\right)\right)^+$$

scale the attention matrix along the source-token dimension using $\mathbf{W}^{(b)}(\mathrm{x})$, averaging across heads, and adding the identity matrix for residual connections.

$$\hat{\mathbf{A}}_{\mathrm{avg}}(x) = \frac{\hat{\mathbf{A}}^{(B-1)}(x) + \hat{\mathbf{A}}^{(B)}(x)}{2}$$

restricts rollout to high-level semantic transitions and averages adjacent layers to suppress shallow-layer perturbation noise.



4 - Experiments

Adversarial Robustness

Table 1. Adversarial accuracy of test-time adaptation methods on ImageNet and fine-grained classification tasks via pretrained ViT-B/16 and ViT-B/32, with the best results shown in **bold red** (ViT-B/16) and **bold blue** (ViT-B/32).

Methods		Pets	Caltech101	Cars	DTD	UCF101	EuroSAT	Flower102	Aircraft	Average
CLIP	ViT-B/16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ViT-B/32	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TTC	ViT-B/16	10.4	8.4	2.9	4.5	1.6	0.4	7.4	0.5	4.5
	ViT-B/32	11.8	22.7	2.3	4.7	6.1	3.0	3.2	1.0	6.9
TPT-Ensemble	ViT-B/16	51.2	74.7	26.0	25.1	30.6	2.2	36.3	8.7	31.9
	ViT-B/32	52.5	74.9	25.9	28.6	36.9	11.9	36.1	7.9	34.3
MTA	ViT-B/16	51.8	72.1	18.5	16.2	27.5	1.2	27.9	4.3	27.4
	ViT-B/32	53.6	76.3	26.4	28.8	39.1	11.3	36.5	8.2	35.0
R-TPT	ViT-B/16	60.2	82.0	34.7	32.8	43.2	8.5	44.6	13.2	39.9
	ViT-B/32	55.8	76.4	28.4	29.1	41.0	5.1	37.6	9.2	35.3
A-TPT (Ours)	ViT-B/16	**70.5**	**85.6**	**39.2**	**37.8**	**51.7**	**13.1**	**52.6**	**15.1**	**45.7**
	ViT-B/32	**66.4**	**79.8**	**31.8**	**31.1**	**46.9**	**12.7**	**43.2**	**10.4**	**40.3**

Table 5. The average accuracy across 8 fine-grained datasets compared A-TPT with training-time defense methods (ViT-B/32).

Methods	Stage	Clean Acc.	Adv. Acc.
PAFT	Training time	54.9	27.5
TeCoA	Training time	33.4	32.2
APT+TeCoA	Training time	40.2	38.9
FARE	Training time	42.2	40.3
A-TPT (Ours)	Test time	**58.6**	**40.3**

Table 4. Adversarial accuracies under various attacks (ViT-B/32).

Method	Flower			
	CW	DF	FGSM	Avg.
CLIP	0.8	0.4	4.8	2.0
TPT-Ensemble	50.1	52.2	46.6	49.7
MTA	34.5	35.4	36.6	35.5
R-TPT	51.6	54.7	49.2	51.8
A-TPT (Ours)	**56.2**	**59.6**	**55.1**	**57.0**

Quality of Semantic Identification & TV-Based Ensemble: Ablation Studies

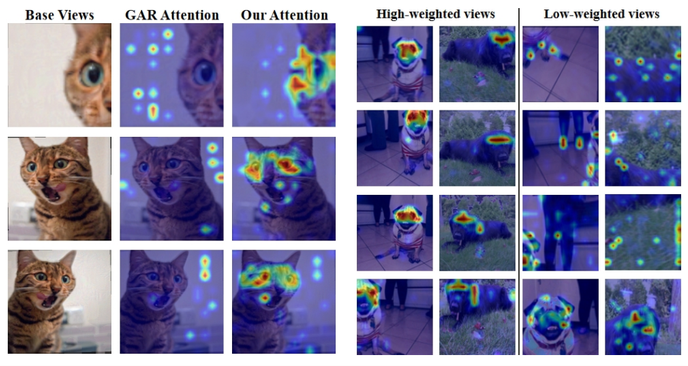


Table 7. Ablation studies: adversarial accuracy on eight fine-grained datasets (ViT-B/16).

A-Refine	A-Aug	A-TV	Adversarial Acc.
×	×	×	31.8
✓	×	×	31.8
×	✓	✓	32.3
✓	×	✓	38.1
✓	✓	×	41.6
✓	✓	✓	45.7

A-Refine: Attention Refinement

A-Aug: Attention-Guided Augmentation

A-TV: TV-Based Ensemble