

SS-TPT

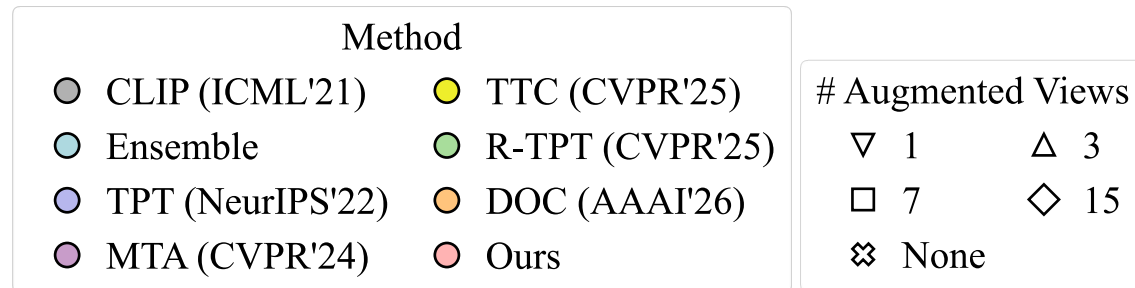
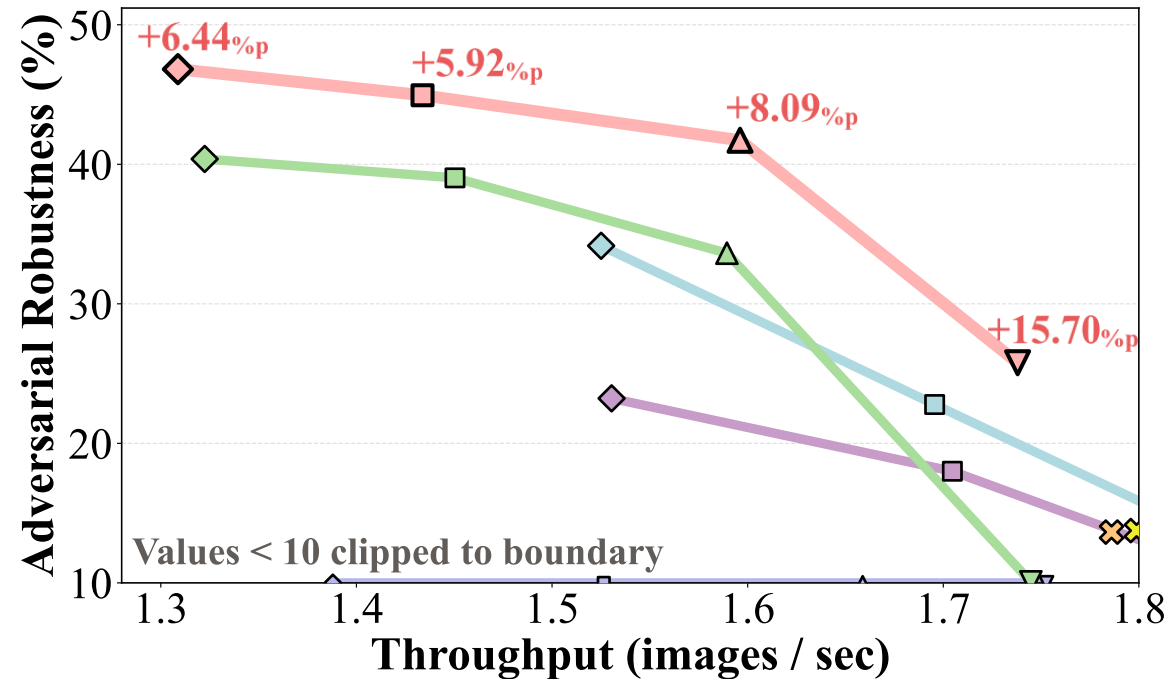
Stability and Suitability-Guided Test-Time Prompt Tuning for Adversarially Robust Vision-Language Models

Sunoh Kim¹, Daeho Um²

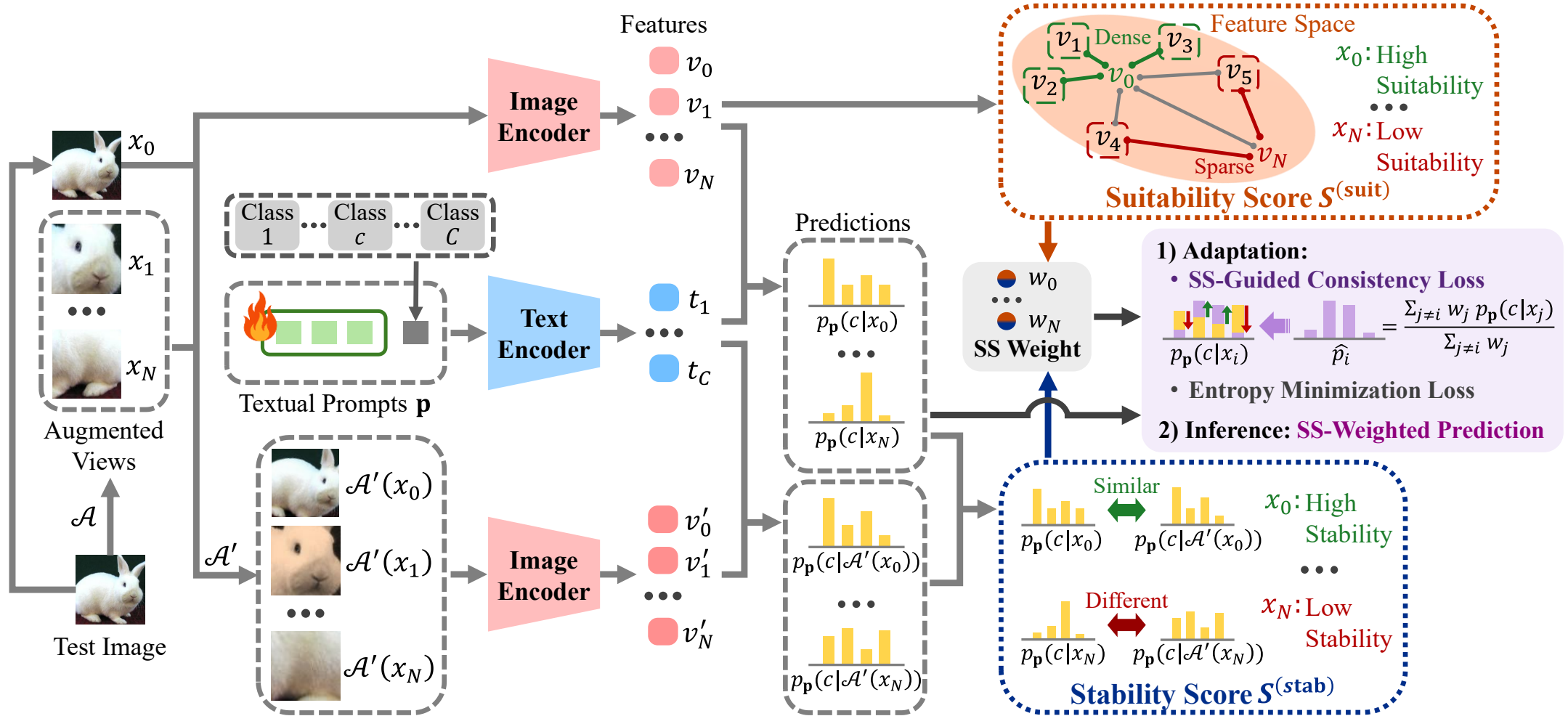
¹Dankook University, ²University of Seoul



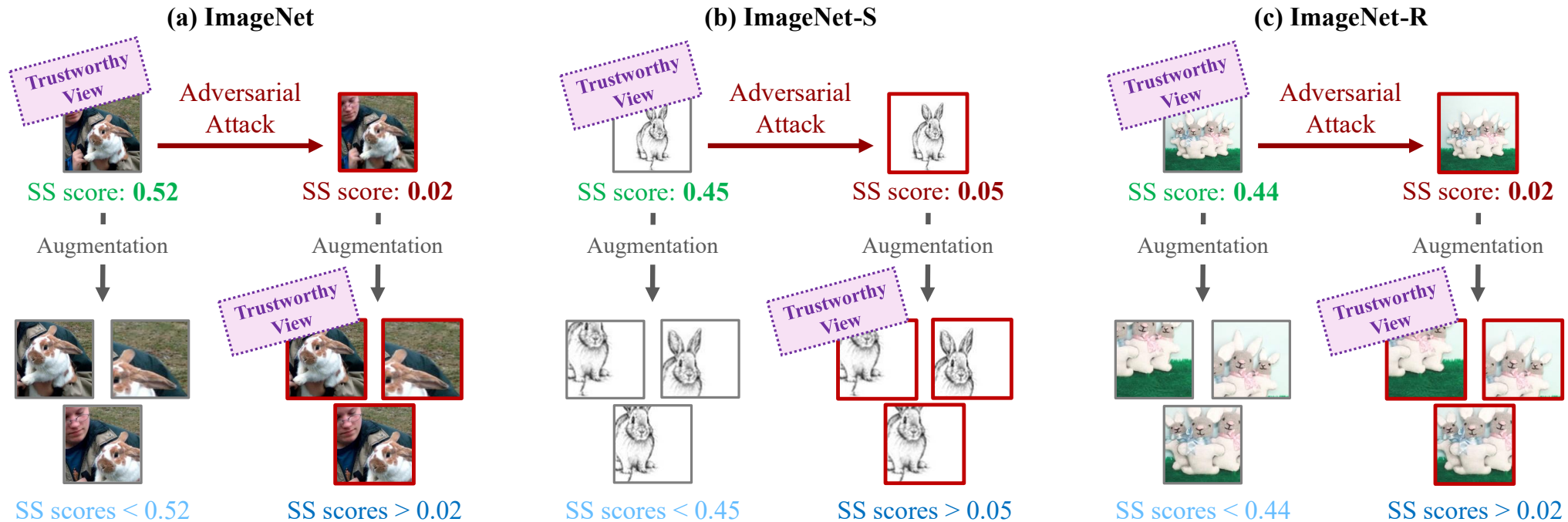
❖ Robustness–Throughput Trade-off in Test-Time Defenses



❖ SS-TPT: View Quality over View Quantity



❖ Does the SS Score Really Find Trustworthy Views?



- > SS scores dynamically identify which views should be trusted.
- > Clean original views receive high SS scores.
- > Adversarially attacked originals are strongly downweighted.
- > SS-TPT shifts trust from corrupted originals to reliable augmented views.



❖ What Actually Makes SS-TPT Robust?

- Ablation study on 10 fine-grained classification datasets

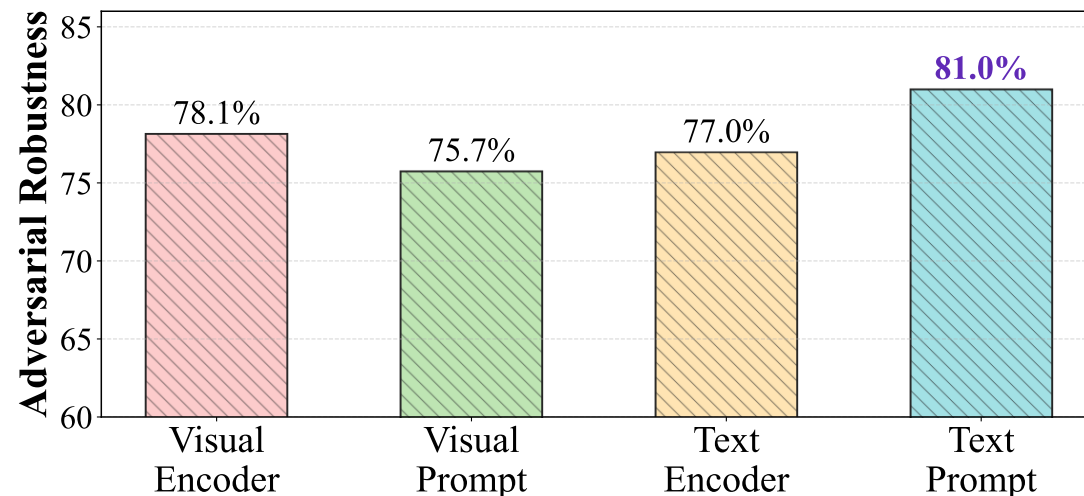
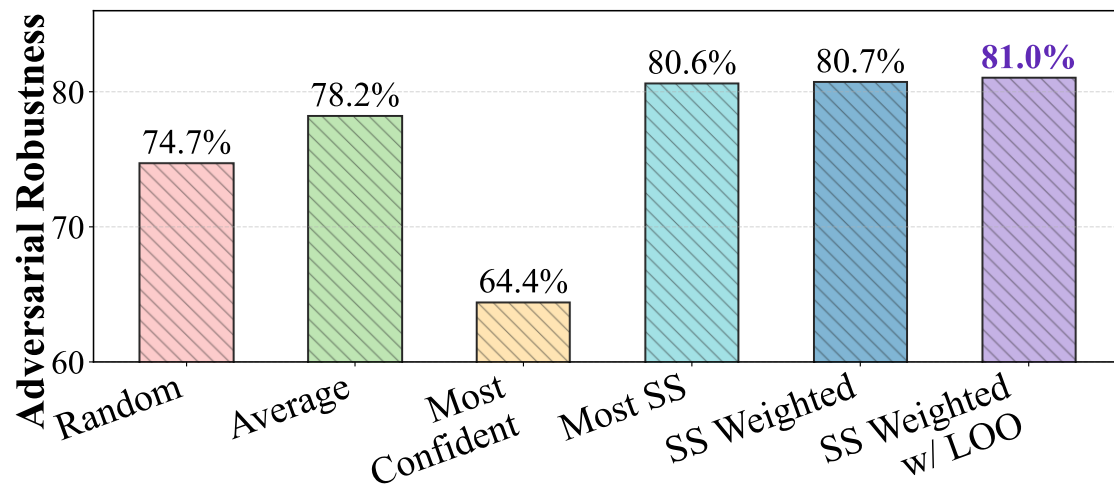
Stability	Suitability	SS-Guided Consistency	SS-Weighted Prediction	Acc.	Rob.
✗	✗	✗	✗	53.8	40.8
✓	✗	✓	✗	53.6	41.5
✓	✗	✗	✓	53.0	44.8
✓	✗	✓	✓	52.4	45.4
✗	✓	✓	✗	53.5	42.0
✗	✓	✗	✓	53.9	45.4
✗	✓	✓	✓	53.6	46.3
✓	✓	✓	✗	54.0	42.0
✓	✓	✗	✓	54.2	45.5
✓	✓	✓	✓	55.1	46.8

- > Stability and suitability provide complementary view-quality signals.
- > SS-guided consistency improves adaptation by aligning to trustworthy views.
- > SS-weighted prediction improves inference by suppressing corrupted views.
- > Using all components achieves the best performance: **55.1 Acc. / 46.8 Rob.**



❖ What Should We Trust, and What Should We Tune?

■ Experiments on the Caltech101 dataset



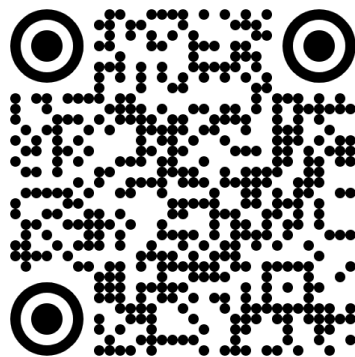
- > The choice of reference strongly affects robustness.
- > The most confident view is not always reliable under adversarial attacks.
- > SS-weighted references provide more trustworthy guidance.
- > Among different tuning targets, tuning the text prompt achieves the best robustness.



SS-TPT makes CLIP robust at test time by identifying and using trustworthy views.

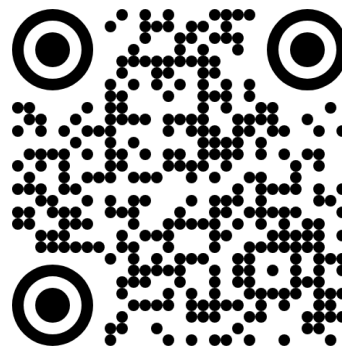
Code:

<https://github.com/sunoh-kim/SS-TPT>



Homepage:

<https://ducklab.dankook.ac.kr>



Thank You