Self-Calibrated Consistency can Fight Back for Adversarial Robustness in Vision-Language Models

Jiaxiang Liu[1] Jiawei Du[2] Xiao Liu[1] Shangyang Li[3] Songchen Ma[4]
Changshuo Wang[5] Prayag Tiwari[6] Mingkun Xu[1,*]
[1]GDIIST | [2]A*STAR | [3]BUPT | [4]HKUST | [5]UCL | [6]Halmstad University



jxliu-ai.github.io/scc-page

Motivation

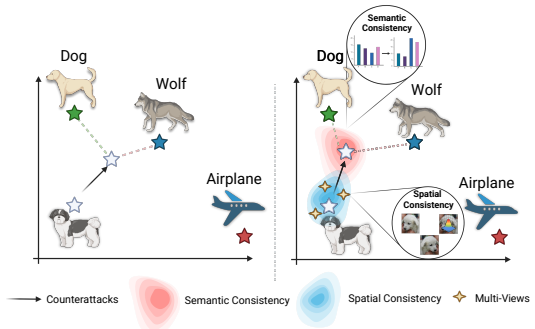
CLIP-style VLMs power photo search, content moderation, medical triage.

The threat. An imperceptible perturbation flips the prediction with high confidence:

"dog" → "wolf"
"stop sign" → "speed limit"
benign scan → positive finding

Standard fix — adversarial fine-tuning — is expensive, label-hungry, hurts clean accuracy.

We seek a retraining-free, inference-time defense.



Where existing test-time defenses fail

Counterattack methods (e.g. TTC) inject a corrective δ. Helpful, but unevenly.

Two coupled failure modes:

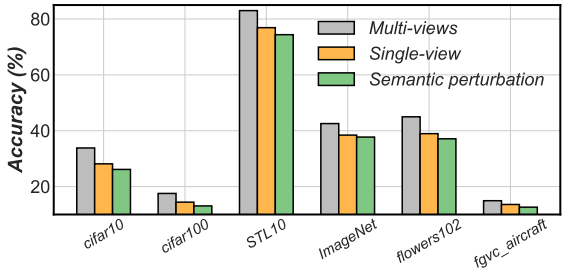
(1) Semantic fragility

- No anchor in text space
- Drifts to hard negatives

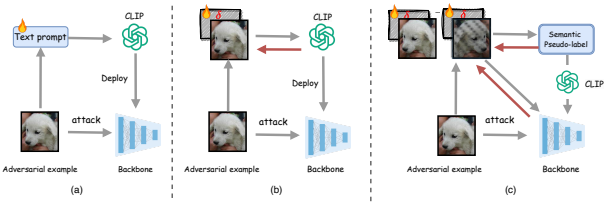
(2) Spatial fragility

- Single-view gradient is noisy
- Tiny shifts flip predictions

Fixing either alone is not enough.



Our insight: a paradigm shift



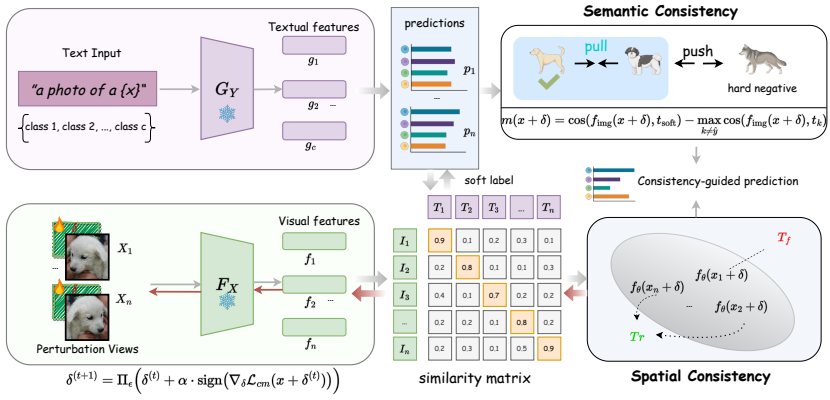
Use the model's own beliefs to constrain its own perturbation update.

(a) R-TPT — adapts prompts, still leaks.

(b) TTC — counterattack, no anchor.

(c) SCC (ours) — two self-checks:
text anchor + multi-view aggregation.

SCC pipeline at a glance



No retraining
Frozen backbone

No labels
Self-pseudo-labeled

One config for all
22 benchmarks

Module 1: Semantic consistency (text space)

Build a soft text anchor from the model's warm-up prediction:

Module 1: Semantic consistency (text space)

Build a soft text anchor from the model's warm-up prediction:

$$t_{soft} = \sum_k p_k \, t_k, \qquad p_k = \mathrm{softmax}\big(\langle f_{img}(x^w), t_k \rangle / T\big)$$

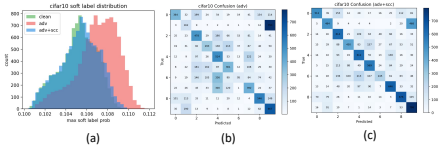
Build a soft text anchor from the model's warm-up prediction:

$$t_{soft} = \sum_k p_k \, t_k, \qquad p_k = \mathrm{softmax}\big(\langle f_{img}(x^w), t_k \rangle / T\big)$$

Two key differences from existing counterattacks:

- Anchors δ to a full distribution, not a single label.
- Separates target from confusable negatives.

Settings: w=5 warm-up steps, T=0.5, λ_{cm}=4.



SCC shifts max-prob back to clean.

Module 2: Spatial consistency (image space)

Aggregate predictions across L=2 augmented views:

Module 2: Spatial consistency (image space)

Aggregate predictions across L=2 augmented views:

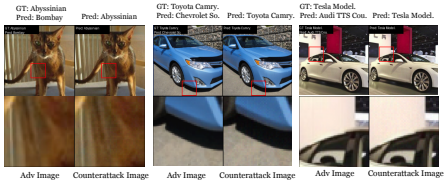
$$\bar{q}(x^w) = \frac{1}{L} \sum_{i=1}^{L} \mathrm{softmax}\big(\langle f_{img}(v_i(x^w)), t_k \rangle\big)$$

Module 2: Spatial consistency (image space)

Aggregate predictions across L=2 augmented views:

$$\bar{q}(x^w) = \frac{1}{L} \sum_{i=1}^{L} \mathrm{softmax}\big(\langle f_{img}(v_i(x^w)), t_k \rangle\big)$$

- Cheap views: flip + Gaussian noise (σ=6/255).
- Variance-reduced gradient signal.
- Stops single-view noise from misleading δ.



Imperceptible corrections reverse misclassification.

Is the soft anchor trustworthy?

Concern. Under attack, the top-1 pseudo-label is often wrong — can t_{soft} still point at the truth?

Cosine to ground-truth text embedding (PGD-10, ϵ=1/255):

Anchor	ImNet	Cal101	OxPet	Flowers	C-10	C-100
random	0.630	0.705	0.570	0.614	0.872	0.766
hard argmax	0.826	0.917	0.872	0.799	0.912	0.821
soft (ours)	0.860	0.934	0.900	0.840	0.923	0.845
clean (UB)	0.926	0.986	0.972	0.924	0.987	0.941

Soft beats hard everywhere, stays close to the clean upper bound.
The distribution carries semantic mass an argmax discards.

Headline results — 16 zero-shot benchmarks

Attack (ϵ_a)	CLIP	TTC	SCC (ours)
PGD-10 (1/255)	2.70	39.17	51.68
PGD-10 (4/255)	0.09	20.63	27.88
CW-10 (1/255)	3.54	33.46	49.42
AutoAttack	0.04	7.36	36.78
PGD-100 (1/255)	1.83	28.45	47.91
Targeted PGD-10	29.13	53.15	63.49

+12–16 pts over TTC across all 5 attack families.
Clean accuracy: −1.3 pts only vs. vanilla CLIP.

Plug-in beyond CLIP: BioMedCLIP

Same SCC, same hyperparameters, no retraining.

	BioMedCLIP	+ TTC	+ SCC
Robust (%)	0.05	10.62	34.00
Clean (%)	47.21	44.91	44.63

+23.4 pts over TTC on medical robustness.

Take-home:

The gains are not specific to a backbone. They come from the self-consistency principle itself.

soft anchor + multi-view aggregation

transfer cleanly to OpenCLIP, BioMedCLIP.

Takeaway

Diagnose. Two coupled failure modes — semantic & spatial fragility — explain where current test-time defenses break.

Fix. Two self-checks — a soft text anchor and multi-view aggregation — neutralize both using only the model's own predictions.

Free lunch. Retraining-free, label-free, plug-in; generalizes across CLIP, OpenCLIP, BioMedCLIP.

Self-consistency is enough.

jxliu-ai.github.io/scc-page

project page | code | paper