

DITHERING DEFENSE: ADVERSARIAL ROBUSTNESS OF VISION FOUNDATION MODELS VIA MULTI-LEVEL FLOYD-STEINBERG DITHERING

Yury Belousov · Brian Pulfer · Vitaliy Kinakh · Slava Voloshynovskiy
Department of Computer Science, Faculty of Science, University of Geneva, Switzerland



ONE ATTACK ON A SHARED VISION MODEL CAN BREAK MANY TASKS.
WE DEFEND THEM ALL WITH FLOYD-STEINBERG COLOR DITHERING.

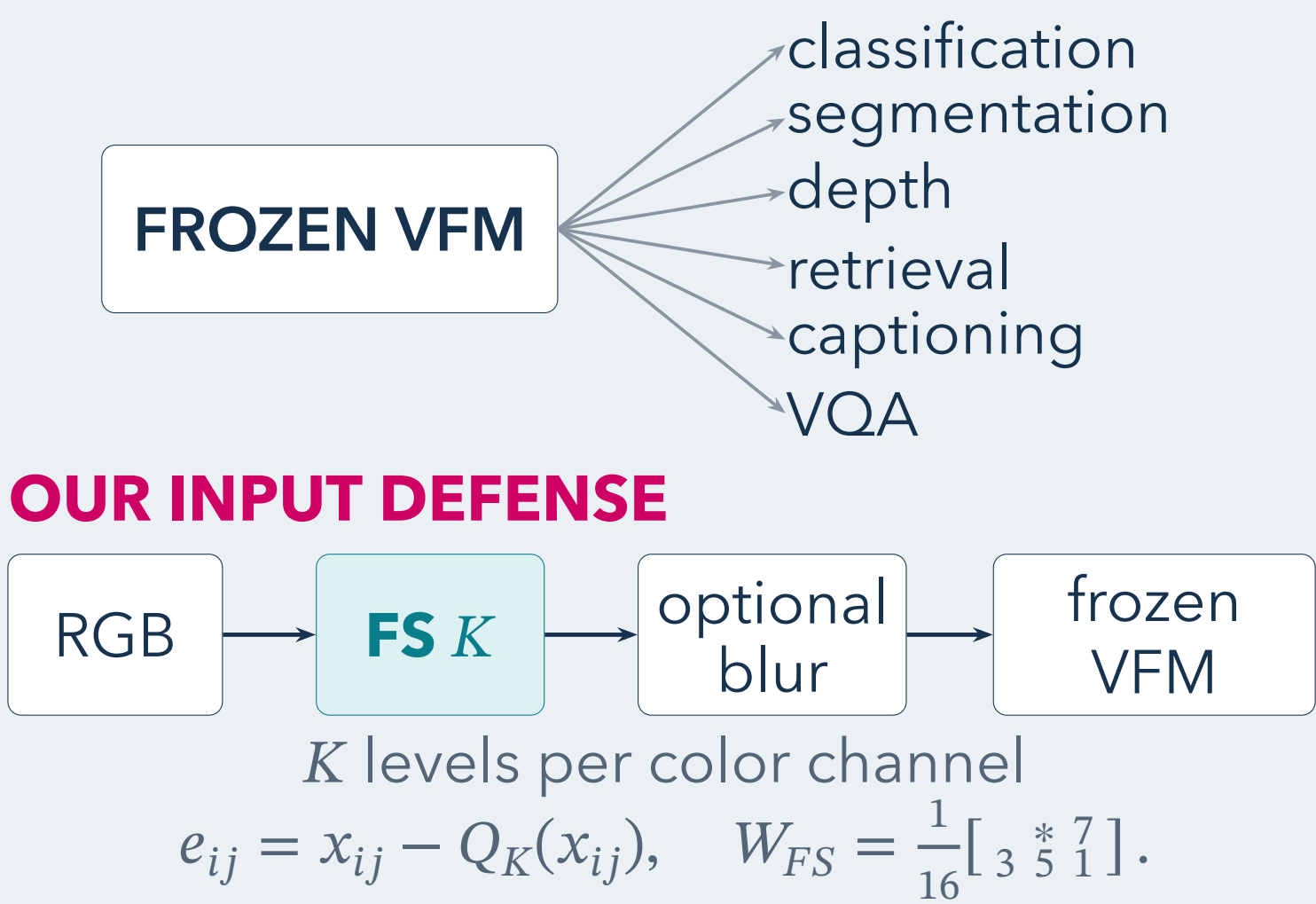
One universal, training-free input transform works across 6 tasks, 2 model families, and 3 attacks, with no model changes or retraining.

CLASSIFICATION ACCURACY UNDER SIA
0.39% → 79.55%
NO DEFENSE OUR DEFENSE

We restore accuracy, and it barely drops when the attacker knows our defense.
We beat tested baselines under adversarial attack while preserving substantially more clean performance.

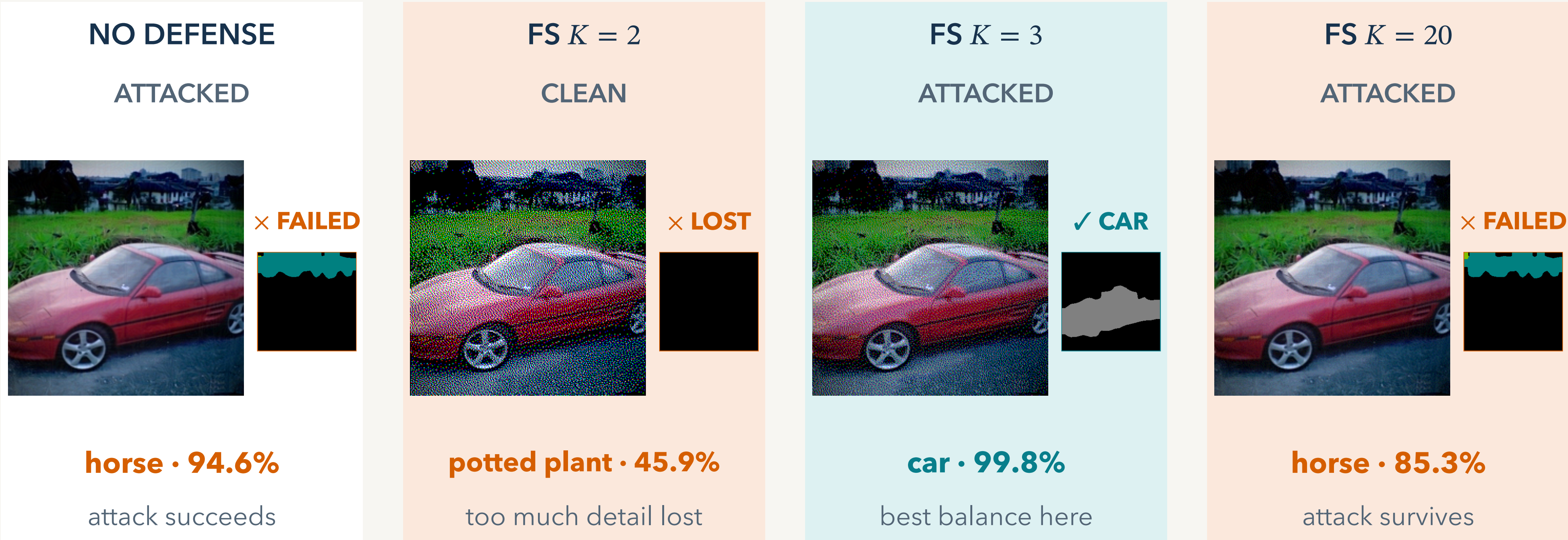
WHY DEFEND AT THE INPUT?

VISION FOUNDATION MODEL (VFM)
A frozen VFM serves many downstream tasks and provides visual input to multimodal LLMs.
THE RISK. Perturbations can break many tasks. Adversarial training across many tasks is costly; it can also harm learned image-text alignment.



K BALANCES IMAGE QUALITY AND ADVERSARIAL ROBUSTNESS

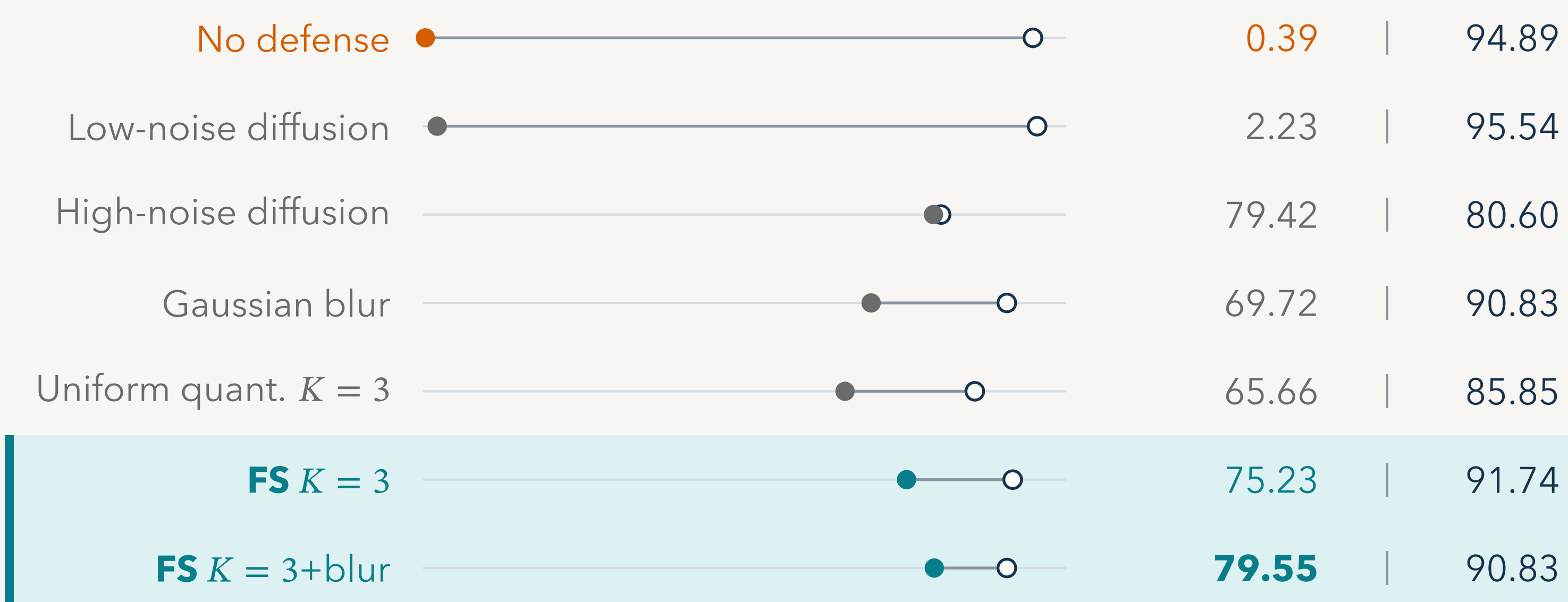
One source image, two tasks · separate attacks for classification and segmentation · shown without blur to reveal the effect of K



FS RECOVERS ACCURACY UNDER ATTACK WHILE RETAINING CLEAN ACCURACY

ATTACKED: ■ NO DEFENSE ■ BASELINES ■ OUR FS • ATTACKED (SIA) | • CLEAN

DINOv2 ViT-S/14 · PASCAL VOC · ACCURACY (%) ↑



Under attack, FS gains **9.57 pp** over uniform quantization; FS+blur gains **9.83 pp** over blur alone and matches high-noise diffusion under attack with **10.23 pp** more clean accuracy.

TASK-AGNOSTIC FS PROTECTS SEGMENTATION, DEPTH, AND RETRIEVAL

DINOv2 ViT-S/14 · TASK-SPECIFIC SIA · ATTACKED | CLEAN · BOLD = BEST SIA SHOWN

	SEGMENTATION VOC · mIoU ↑		DEPTH NYUv2 · RMSE ↓		RETRIEVAL rOxford5k · mAP ↑	
No defense	21.26	79.26	4.11	0.55	0.63	84.46
Low-noise diffusion	25.98	79.03	3.66	0.58	15.42	83.82
High-noise diffusion	61.51	61.30	0.81	0.79	58.21	62.03
Gaussian blur	57.53	71.16	0.92	0.68	75.08	82.08
Uniform quant. $K = 3$	51.96	64.08	0.97	0.74	70.47	77.65
FS $K = 3$	60.83	72.97	0.87	0.59	72.03	77.12
FS $K = 3$ +blur	64.58	72.12	0.80	0.68	71.66	77.14

FS+blur beats high-noise diffusion under attack across classification, segmentation, depth, and retrieval while preserving more clean performance.

SIA is our strongest tested transform-aware attack. Under iterative PGD and momentum-based MI-FGSM, FS variants take both top ranks in **7 of 8** settings and rank second in the eighth.

FS holds when the attacker adapts

Defense-aware SIA adapts image changes to survive our input defense.

Yet all four task metrics barely change.

CLASSIFICATION ↑	SEGMENTATION ↑
2.22 pp lower 79.55 → 77.33	1.43 points lower 64.58 → 63.15
DEPTH · RMSE ↓	RETRIEVAL · mAP ↑
no degradation 0.80 → 0.79	0.83 points lower 71.66 → 70.83

FS $K = 3$ +blur. Arrows: standard SIA → worst defense-aware SIA over all settings.

ViT scale: robustness ↑, clean cost ↓

FS robustness rises while clean cost falls.

DINOv2 · rOxford5k · SIA mAP ↑

FS $K = 3$, no blur, for every size

	ViT-S	ViT-B	ViT-L
NO DEFENSE	0.63	0.87	1.56
WITH FS	72.03	82.76	88.85

CLEAN COST 7.34 pp 2.94 pp 0.33 pp

ViT-L original clean: 88.91
With FS under SIA: just a **0.06 pp** gap.

Model-agnostic input defense

PaliGemma 3B combines a vision encoder with an LLM for diverse multimodal tasks.

SIA: no defense → FS $K = 3$, no blur

IMAGE CAPTIONING · SPICE ↑

5.79 → 16.75

COCOCaptions · FS clean: 19.26

VISUAL Q&A · ACCURACY ↑

32.70 → 62.15

VQAv2 “other” · FS clean: 69.22

Read the full paper

Explore more experiments and results in the paper and supplement.



arxiv.org/abs/2605.23065

TAKE-HOME. The key is not quantization alone: at intermediate K , Floyd-Steinberg color dithering preserves semantics while disrupting adversarial structure.

It works on both DINOv2 and PaliGemma without retraining; with FS+blur, all four DINOv2 task metrics change only slightly when the attacker adapts.

ATTACK SETUP: white-box · $\epsilon_\infty = 3/255$ (≈ 40 dB PSNR) · $T = 50$ · $\nu = 4\epsilon_\infty/T$ | DEFENSE-AWARE: STE · grid: $K_{\text{attack}} \in \{3, 6, 9, 12, 15\}$ · $p_q \in \{0.5, 1\}$ | OPTIONAL BLUR: $9 \times 9, \sigma = 3$