MECHANISTIC INTERPRETABILITY WORKSHOP · ICML 2026

SoftServe

Probing Perturbation Invariance in DINOv2

MECHANISTIC GAPS BETWEEN REAL & GENERATED IMAGE REPRESENTATIONS

Alina Labaz · Taras Rumezhak · Volodymyr Karpiv – SoftServe Inc.

TL;DR

Cross **two out-of-distribution axes** — an untrained-for **perturbation** (Gaussian noise) applied to OOD **inputs** (AI-generated images) — and a static distributional difference becomes a **dynamic stability gap** in DINOv2 patch space. A training-free detector with no generator-specific fitting.

01 THE TWO OOD AXES

DINOv2 learns invariance to Gaussian **blur** but never **noise**. Each axis alone is detectable; combining them mechanistically is the contribution.

PERTURBATION AXIS

Gaussian noise lies **outside** DINOv2's trained invariance regime → an OOD **stressor**.

INPUT AXIS

Generated images are OOD relative to DINOv2's natural-photo pretraining → an OOD **input**.

OOD PERTURBATION × OOD INPUT → STABILITY GAP

REAL PHOTO in-dist. | **AI-GENERATED** OOD

↓

+ GAUSSIAN NOISE σ=100 · OOD perturbation, never trained

↓

DINOv2 ViT-L/14 · mean cosine shift, 256 patches

↓

REAL

GENERATED

→ GENERATED IMAGES MOVE FARTHER UNDER NOISE

02 METHOD · TRAINING-FREE PROBE

STEP 1

Extract 257 final-layer tokens (1 CLS + **256 patch**) from a frozen ViT-L/14.

STEP 2

Add Gaussian noise (σ=100), re-encode to perturbed features.

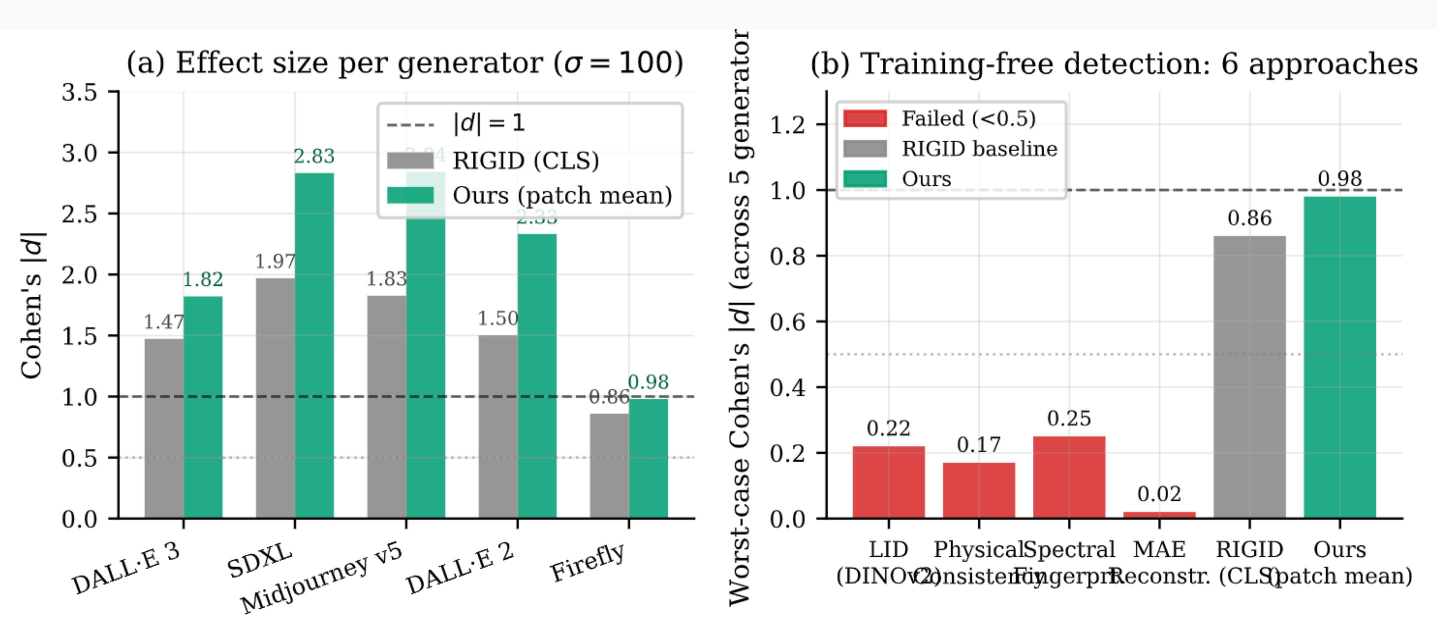
STEP 3

Score = **mean per-patch cosine distance** (— not the CLS token).

STEP 4

Compare via Cohen's |d| & AUROC. No training, no per-generator fitting.

03 KEY RESULT · PATCH BEATS CLS



Patch-mean (teal) beats CLS-only RIGID (grey) on all five generators; four of five non-overlapping 95% CIs.

0.86 → 0.98

WORST-CASE |d| · CLS → PATCH MEAN

+14% WORST-CASE GAIN, FREE

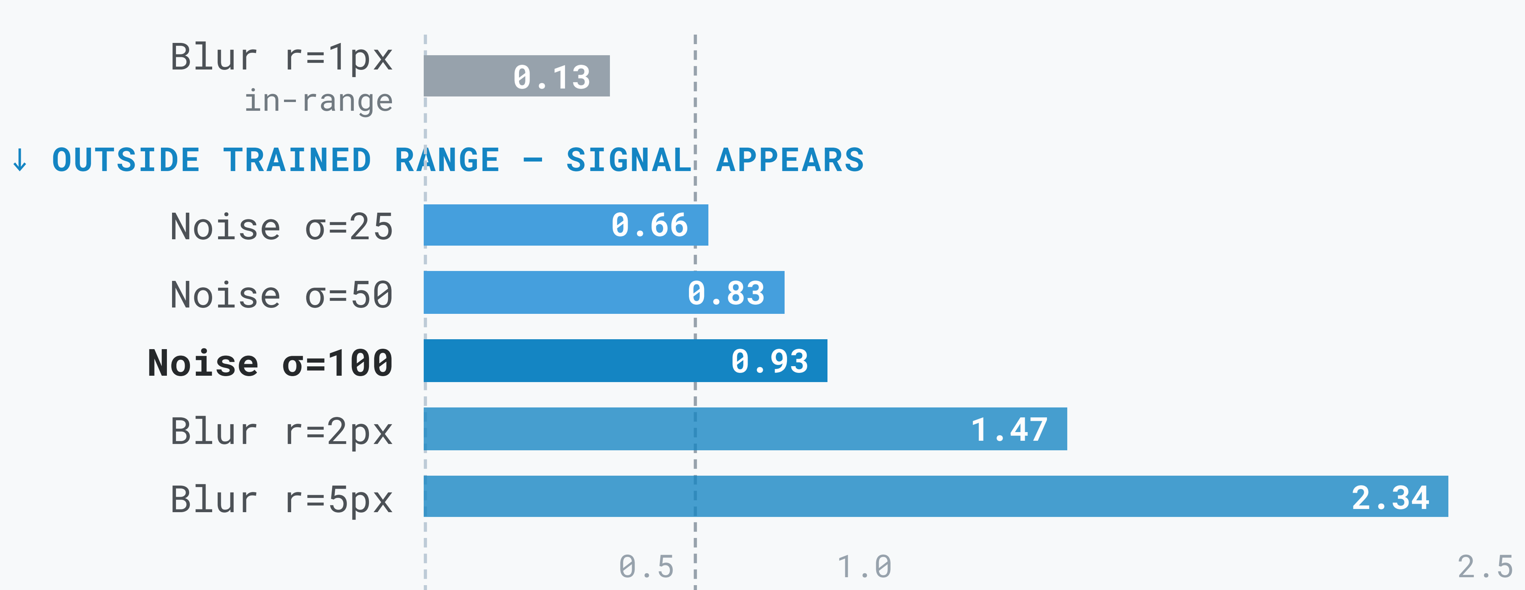
5/5 GENERATORS, SAME DIRECTION

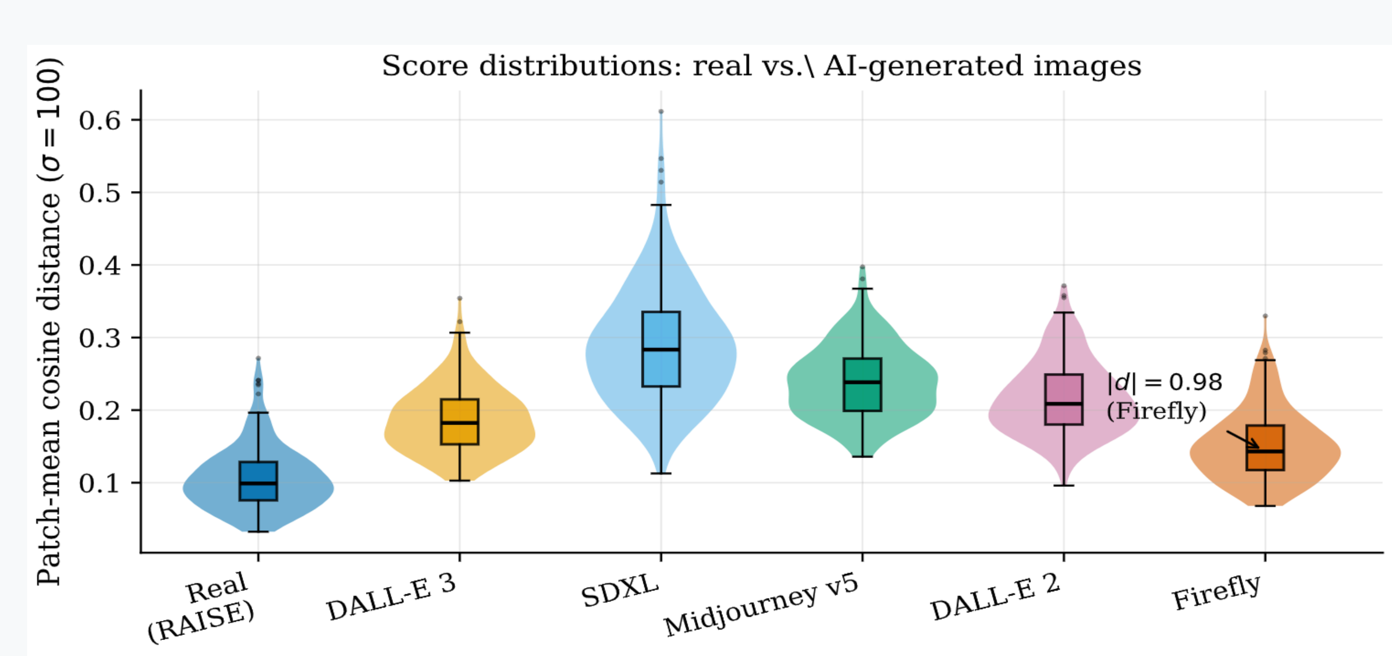
Averaging 256 patch tokens recovers local perturbation signal the CLS summary compresses away — at **no extra cost**.

04 MECHANISM · IT'S THE OOD STATUS

PERTURBATION-TYPE NATURAL EXPERIMENT

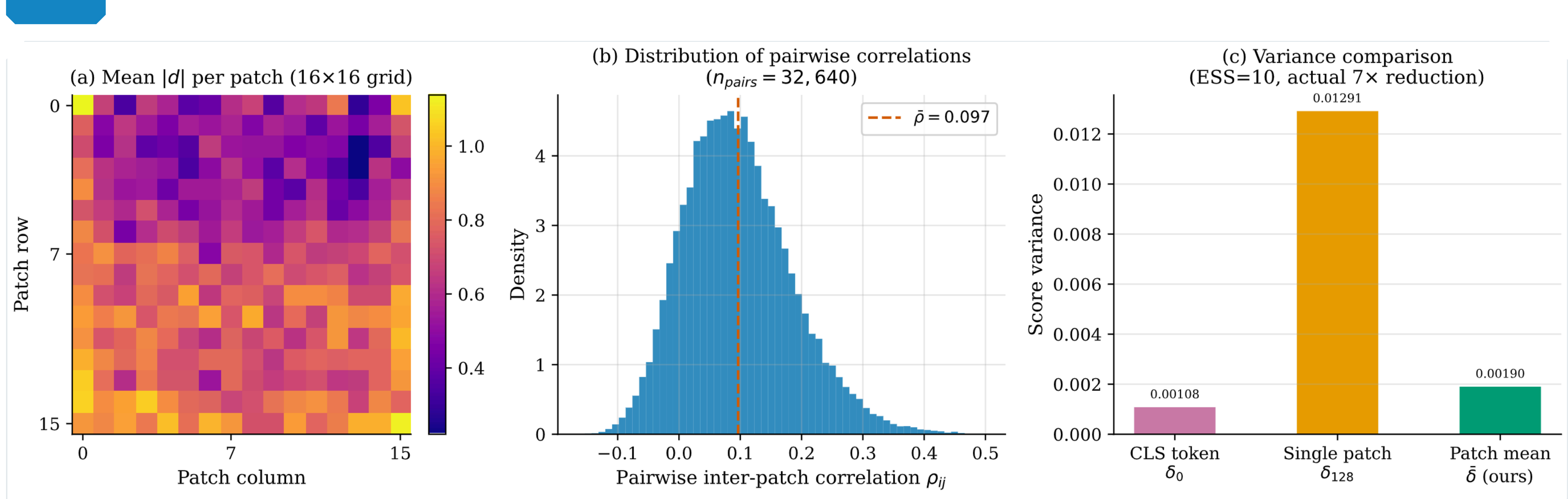
Worst-case |d| across 5 generators





All five generators shift to higher cosine distance (direction-consistent) – the property absent from five competing training-free signals, which flip sign per generator.

05 WHY PATCH TOKENS WIN



Distributed signal. The gap spreads across all patches — no corner/edge artefact.

Lossy bottleneck. CLS pooling = 11.9× lower variance but discards local responses.

Read at the right granularity. Mean patch keeps the signal (ESS ≈ 10).

TAKEAWAY — PROBE SPATIALLY-DISTRIBUTED PROPERTIES AT THE PATCH LEVEL

Diagnostic probe, not a deployed detector · uneven across generators (TPR@5%FPR 0.93–0.33), |d|=0.74 on JPEG web images.

CODE

github.com/alinamuliak/probing-perturbation-invariance-dinov2