Probing Perturbation Invariance in DINOv2: Mechanistic Gaps Between Real and Generated Image Representations

Alina Labaz [1] **Taras Rumezhak** [1] **Volodymyr Karpiv** [1]

Abstract

A representation can encode whether an image is real or AI-generated yet not reveal it under static probing. We show the distinction becomes legible when two out-of-distribution (OOD) conditions are crossed: DINOv2 is trained with Gaussian blur but never Gaussian noise, so noise is an OOD *perturbation*, while generated images are OOD *inputs*. Applying the former to the latter turns a static distributional difference into a *dynamic* stability gap — generated images move farther in DINOv2 patch space under noise than real photographs — giving a training-free detector with no generator-specific fitting. The signal lives in the spatial patch tokens: averaging the 256 patch tokens, rather than the CLS summary used by the RIGID baseline, recovers local perturbation responses and raises worst-case Cohen's $|d|$ across five Synthbuster generators from 0.86 to 0.98 at no extra cost. A perturbation-type natural experiment isolates the mechanism — blur inside DINOv2's trained range gives a near-null gap ($|d|$=0.13) while noise (0.93) and out-of-range blur (≥ 1.47) do not — and cross-validation plus non-perturbative baselines rule out cherry-picked noise levels and static-shift explanations. We frame this as a diagnostic probe, not a deployable detector: detection is uneven (TPR@5%FPR 0.93–0.33) and weakens on JPEG web images. The broader lesson for ViT interpretability is that probes for spatially distributed properties should read patch tokens, since CLS pooling is a lossy bottleneck. Code: alinamuliak/probing-perturbation-invariance-dinov2-EB58.

[1]SoftServe Inc.. Correspondence to: Alina Labaz <almuli@softserveinc.com>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

DINOv2 (Oquab et al., 2024) is trained with the iBOT objective (Zhou et al., 2022) on natural photographs under augmentations that include Gaussian *blur* but not Gaussian noise. The probe in this paper sits at the intersection of two out-of-distribution (OOD) axes that we keep separate throughout: (1) the *perturbation* axis — Gaussian noise lies outside DINOv2's trained invariance regime, in the spirit of Hendrycks & Dietterich (2019); and (2) the *input* axis — AI-generated images are OOD relative to DINOv2's pretraining corpus, with the fine-grained texture and spectral statistics that supervised forensic detectors have long exploited (Wang et al., 2020; Corvi et al., 2023; Ojha et al., 2023). Each axis alone is detectable; our contribution is to combine them mechanistically — an OOD perturbation makes the OOD input status visible as a *dynamic* stability gap in DINOv2 patch space rather than a static distributional difference. We measure the gap via patch-token cosine distance on the Synthbuster benchmark (Bammey, 2024), and additionally uncover a CLS-vs.-patch information asymmetry: averaging the 256 patch tokens recovers locally distributed perturbation signals that the CLS suppresses, raising worst-case Cohen's $|d|$ by 14% over the CLS-only RIGID baseline (He et al., 2024) at no additional computational cost.

Contributions.

1. **OOD instability as a mechanistic probe**: Blur within DINOv2's training range (r=1 px, $|d|$=0.134) is 7× weaker than Gaussian noise ($|d|$=0.93); blur outside the training range ($r\geq$2 px, $|d|\geq$1.47) is even stronger, confirming OOD status drives the signal.
2. **CLS vs. patch-token asymmetry**: Averaging 256 patch tokens improves worst-case $|d|$ from 0.86 to 0.98 (95% CI: $[0.78, 1.20]$) over CLS-only RIGID; non-perturbative baselines reach only $|d|\leq$0.071 (13.8× gap), confirming perturbation is essential.
3. **Validated σ selection**: LOGO cross-validation ($|d|$=0.949, Δ=0.031) confirms σ=100 is not cherry-picked.
4. **Direction-inconsistency diagnosis**: Systematic study of five alternative training-free signals with per-generator failure analyses.

2. Related Work

Mechanistic interpretability of vision transformers. Prior work probes DINO/DINOv2 for semantic segmentation, depth, surface normals, and texture (Oquab et al., 2024), finding that CLS tokens encode global scene semantics while patch tokens preserve fine-grained spatial structure. Perturbation-based probing is a causal complement to linear probing: it reveals what a network is *sensitive to*, not merely what its representations can decode, and is related to adversarial robustness analysis (Goodfellow et al., 2015). We apply it to characterise OOD-induced instability of generated-image representations and the CLS-vs.-patch information asymmetry.

Supervised detection. Wang et al. (2020) showed that a ResNet-50 trained on ProGAN outputs generalises to other GAN families via data augmentation. Subsequent work used frequency-domain features (Frank et al., 2020) and inter-pixel noise correlations (Xi et al., 2023). DIRE (Wang et al., 2023) reconstructs an image through a diffusion model and uses the reconstruction error to detect generated content, but requires a diffusion model at test time. Supervised methods achieve high accuracy within their training distribution but degrade on novel generators; our work specifically targets the training-free regime to avoid this limitation.

Training-free detection. Ojha et al. (2023) showed that nearest-neighbour distances in CLIP feature space separate real from generated images without fine-tuning. AEROBLADE (Wang et al., 2024) uses reconstruction error of a frozen VAE encoder-decoder; this signal is strong for latent-diffusion models but weaker for generators not based on a shared VAE (one of the direction-inconsistency failures we analyse). Corvi et al. (2023) demonstrated that diffusion models leave detectable spectral artefacts, though these weaken as generators improve. Most relevant to our work, RIGID (He et al., 2024) applies Gaussian perturbations, computes cosine similarity of the perturbed *CLS* token of a DINOv2 model, and uses the distance as a detection score; we extend it by replacing the single CLS token with the mean of all 256 spatial patch tokens and providing a systematic mechanistic analysis of why the improvement occurs. A concurrent training-free method, High-Frequency Influence (HFI), detects AI images via frequency-domain statistics and has been reported to improve over AEROBLADE and RIGID-like methods on Synthbuster and GenImage (Bammey, 2024); we were unable to obtain its implementation for direct comparison. HFI and our method operate in fundamentally different signal spaces (frequency domain vs. feature-space perturbation response) and are likely complementary.

3. Method

Feature extraction. We centre-crop image I to a square and pass it through DINOv2 ViT-L/14 (Oquab et al., 2024) at 224×224, extracting all final-layer tokens $\mathbf{F}=[\mathbf{f}_0,\ldots,\mathbf{f}_P]\in\mathbb{R}^{257\times1024}$, where $\mathbf{f}_0$ is the CLS token and $\mathbf{f}_{i\geq1}$ are the $P=256$ spatial patch tokens.

Perturbation. We apply isotropic Gaussian noise with standard deviation σ to the pixel values before DINOv2 processing, clipping to $[0,255]$:

$$I' = \mathrm{clip}(I+\boldsymbol{\varepsilon}), \quad \boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0},\,\sigma^2\mathbf{I}). \tag{1}$$

This yields a perturbed feature matrix $\mathbf{F}'$.

Detection score. The per-patch cosine distance between original and perturbed features is

$$\delta_i = 1-\frac{\mathbf{f}_i\cdot\mathbf{f}_i'}{\|\mathbf{f}_i\|\,\|\mathbf{f}_i'\|}, \quad i=1,\ldots,P. \tag{2}$$

Our detection score is the mean over all $P=256$ patch tokens:

$$s_{\mathrm{ours}} = \frac{1}{P}\sum_{i=1}^{P}\delta_i. \tag{3}$$

Higher values indicate a larger feature shift under perturbation. For comparison, the CLS-token baseline (RIGID) uses only δ_0.

CLS vs. patch tokens: an information asymmetry. The CLS token aggregates all spatial patches through self-attention into a single scene-level summary, suppressing local perturbation responses. Averaging the $P=256$ patch distances accesses local signals the CLS compresses away, yielding strictly higher $|d|$ despite moderate inter-patch correlation ($\bar{\rho}=0.097$, ESS≈10); the decomposition is quantified in Section 4.5.

Mechanistic basis. Natural photographs are in-distribution for DINOv2 and produce smooth feature trajectories under noise; generated images are OOD inputs and occupy less-smooth regions of feature space, yielding larger shifts under the same OOD perturbation. Section 4.6 tests this prediction directly with a perturbation-type natural experiment, and Section 4.10 contrasts the dynamic stability gap with the static local intrinsic dimension (LID) signal.

Noise level selection. We pick σ by maximising worst-case $|d|$ over the grid $\{10,25,50,75,100,150\}$; the optimum is $\sigma=100$ for the patch mean and $\sigma=150$ for CLS (RIGID). Robustness of this choice is established by leave-one-generator-out cross-validation (Section 4.4); a generator-agnostic multi-σ ensemble (Section 4.4) provides

a fallback when no per-deployment σ tuning is possible. Results report threshold-free Cohen's $|d|$ and AUROC, defined formally in Section 4; operational TPR at 1%/5%FPR is in Section 4.3.

4. Experiments

Datasets. **Real images**: RAISE-1k (Dang-Nguyen et al., 2015), a dataset of uncompressed RAW photographs. We use 200 images (centre-cropped to square, resized to 224×224 by the DINOv2 preprocessor). **Generated images**: Synthbuster (Bammey, 2024), providing 200 PNG images from each of five generators: DALL-E 3 (Betker et al., 2023), Stable Diffusion XL (SDXL) (Podell et al., 2023), Midjourney v5, DALL-E 2, and Adobe Firefly. Generators cover four distinct architectures: latent VAE (SDXL, Firefly), unCLIP (DALL-E 2), cascaded diffusion (DALL-E 3), and proprietary (Midjourney v5). RAISE-1k is uncompressed RAW photography and is atypical of deployment distributions (which are dominated by web-scraped JPEGs). We therefore treat the RAISE-1k numbers in Sections 4.1–4.7 as an *upper bound* on detection performance and report the deployment-aligned counterpart in Section 4.9 using COCO val2017 (worst-case $|d|$=0.74). The gap between the two settings is itself diagnostic and is discussed in Section 5.

Metrics. Let $\{s_i^{\mathrm{r}}\}_{i=1}^{N_r}$ and $\{s_j^{\mathrm{g}}\}_{j=1}^{N_g}$ denote scores s_{ours} computed on the real and generated splits. Cohen's $|d|$ (Cohen, 1988) is the standardised mean difference,

$$|d| \;=\; \frac{\left|\,\mu_g - \mu_r\,\right|}{s_p}, \qquad s_p \;=\; \sqrt{\frac{(N_r{-}1)\,\hat{\sigma}_r^2 + (N_g{-}1)\,\hat{\sigma}_g^2}{N_r + N_g - 2}}, \tag{4}$$

with $\mu_\cdot$, $\hat{\sigma}_\cdot^2$ the per-split sample mean and unbiased variance. Area under the ROC curve (AUROC) measures the rank separability

$$\mathrm{AUROC} \;=\; \frac{1}{N_r N_g} \sum_{i,j} \mathbf{1}\big[s_j^{\mathrm{g}} > s_i^{\mathrm{r}}\big], \tag{5}$$

and is invariant to monotonic rescaling of s_{ours}. Standard thresholds: $|d|\geq$1.0 (strong), $|d|\geq$0.5 (marginal). For statistical reliability, all point estimates are accompanied by 95% bootstrap CIs (10 000 resamples, percentile method). All results use N=200 per split and `facebook/dinov2-large` (ViT-L/14, 256 patch tokens, hidden dim 1024) unless stated otherwise. Noise-sample reproducibility is addressed in Section 4.8: results are stable across seeds 42–45 (CV=1.8%).

4.1. Main Results

Table 1 reports Cohen's $|d|$ and AUROC at σ=100 for our method and RIGID, with 95% bootstrap CIs. Our method

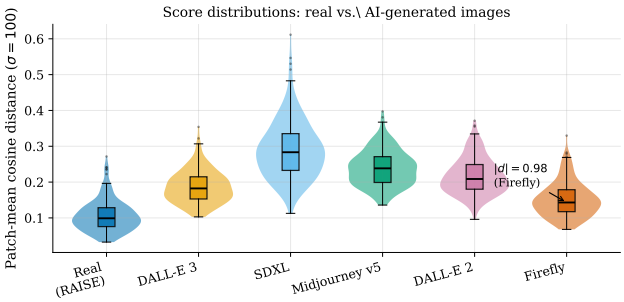

Figure 1. Patch-mean cosine distance distributions (σ=100) for RAISE-1k (real, blue) and five generators. All generators shift to higher values (direction-consistent). Firefly has the smallest shift ($|d|$=0.98); SDXL and Midjourney have the largest ($|d|$>2.8).

consistently outperforms RIGID in Cohen's $|d|$ across all five generators, with improvements from +14% on Firefly to +55% on Midjourney v5. Bootstrap CIs are non-overlapping between methods for four of five generators; even on Firefly the lower bound of our method (0.78) exceeds the lower bound of RIGID (0.66).

On Firefly, $|d|$ rises (0.98 vs. 0.86) but AUROC dips marginally (0.773 vs. 0.786): the Firefly distribution has higher variance than real (ratio 1.11), so upper-tail overlap depresses the rank statistic more than the mean-shift statistic (Section 4.3).

Score distributions are visualised in Figure 1. Generated images shift to higher cosine distances under perturbation for all five generators, confirming direction-consistent separation — the key property absent from the five failed approaches (Section 4.10).

4.2. Noise Level Ablation

Figure 2 shows worst-case $|d|$ as a function of σ. For the patch mean, the signal grows from $|d|$=0.28 at σ=10 to a peak of $|d|$=0.98 at σ=100, then decreases to 0.95 at σ=150. For RIGID (CLS token), the signal is already marginal at σ=10 and plateaus around σ=100–150. Patch mean strictly outperforms CLS at every $\sigma \geq 50$.

4.3. Operational Thresholds and Score Analysis

Table 2 reports TPR at 1% and 5% FPR — the thresholds most relevant to deployment, where false accusations carry real costs.

TPR@5%FPR ranges from 0.69 to 0.93 on the four non-Firefly generators; Firefly remains hard (0.325). Figure 3 shows ROC curves and the Firefly score distributions: the $|d|$/AUROC divergence reflects a real mean shift ($\Delta\mu$=0.046) masked by upper-tail variance overlap (gen/real std ratio 1.11). The CLS baseline shows the same divergence in the opposite direction (AUROC 0.786, $|d|$

Table 1. Detection performance at σ=100, N=200, with 95% bootstrap CIs (10 000 resamples) shown in brackets. Worst-case generator (Firefly) is in **bold**. $|d|\geq 1.0$ = strong; $|d|\geq 0.5$ = marginal.

	CLS baseline (RIGID) (He et al., 2024)		Ours (patch mean, P=256)	
Generator	$\lvert d\rvert$ [95% CI]	AUROC [95% CI]	$\lvert d\rvert$ [95% CI]	AUROC [95% CI]
DALL-E 3	1.473 [1.260, 1.712]	0.898 [0.865, 0.928]	1.820 [1.577, 2.103]	0.908 [0.876, 0.936]
SDXL	1.969 [1.792, 2.196]	0.974 [0.960, 0.986]	2.831 [2.570, 3.159]	0.985 [0.975, 0.993]
Midjourney v5	1.826 [1.603, 2.100]	0.943 [0.918, 0.964]	2.841 [2.550, 3.186]	0.974 [0.958, 0.986]
DALL-E 2	1.503 [1.320, 1.722]	0.924 [0.896, 0.950]	2.332 [2.077, 2.624]	0.951 [0.928, 0.971]
Firefly	**0.860** [0.660, 1.078]	**0.786** [0.740, 0.828]	**0.980** [0.776, 1.200]	**0.773** [0.726, 0.816]
Worst-case $\lvert d\rvert$	0.860	—	**0.980**	—
Generator-agnostic multi-σ ensemble (Section 4.4)				
Worst-case $\lvert d\rvert$	—	—	**0.826**	—

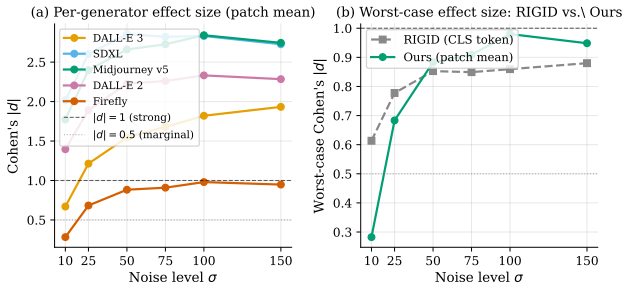

Figure 2. **(a)** Per-generator Cohen's $|d|$ vs. σ for the patch-mean method. **(b)** Worst-case $|d|$ for RIGID (CLS, dashed) vs. ours (patch mean, solid). Our method peaks at σ=100; CLS at σ=150. Patch mean outperforms CLS at every $\sigma \geq 50$.

Table 2. TPR at 1% and 5% FPR for our method (σ=100). Real images are the positive class.

Generator	TPR@1%FPR	TPR@5%FPR
DALL-E 3	0.580	0.690
SDXL	0.755	0.925
Midjourney v5	0.845	0.900
DALL-E 2	0.685	0.865
Firefly	0.195	0.325

0.860), so this is a structural property of the Firefly distribution rather than an artifact of patch aggregation.

4.4. Leave-One-Generator-Out Sigma Validation

To rule out σ=100 being overfit to the five evaluation generators, for each held-out generator g we select $\sigma^* = \arg\max_\sigma \min_{g' \neq g} |d(g', \sigma)|$ on the remaining four and evaluate on g (Table 3). With Firefly held out, σ^*=150 yields $|d|$=0.949; LOGO worst-case $|d|$=0.949 vs. full-data 0.980 (Δ=0.031) confirms σ=100 is not a post-hoc cherry-pick.

Multi-sigma ensemble. A generator-agnostic ensemble averages standardised patch-mean scores across all six sigmas $\{10, 25, 50, 75, 100, 150\}$, with per-sigma standardisation fit on real images only. Worst-case $|d|$ is 0.826 vs. 0.980 for σ=100: marginal detection is maintained on all five generators even with no sigma access to generator-specific information.

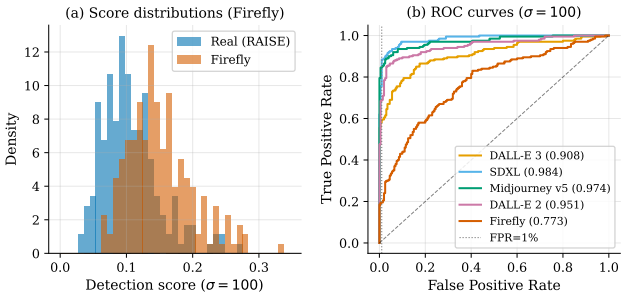

Figure 3. **(a)** Score distributions (σ=100) for real (RAISE-1k, blue) and Firefly (orange), illustrating distributional overlap. **(b)** ROC curves for all five generators; dashed vertical line at FPR = 1%.

Table 3. Leave-one-generator-out (LOGO) sigma validation. For each held-out generator, σ^* is selected using remaining four. LOGO worst-case —d— = 0.949 (full-data: 0.980, Δ=0.031).

Held-out generator	Selected σ^*	Train WC $\lvert d\rvert$	LOGO $\lvert d\rvert$
DALL-E 3	100	0.980	1.820
SDXL	100	0.980	2.831
Midjourney v5	100	0.980	2.841
DALL-E 2	100	0.980	2.332
Firefly	150	1.933	**0.949**
LOGO worst-case	—	—	**0.949**

4.5. Mechanistic Analysis: CLS vs. Patch Token Information Content

We extract all P=256 per-patch probe responses δ_i for the N=200 real and generated images (Figure 4).

Spatial distribution of the invariance gap. Mean Cohen's $|d|$ per spatial patch (16×16 grid, averaged over generators; Figure 4a) is distributed across the image with no

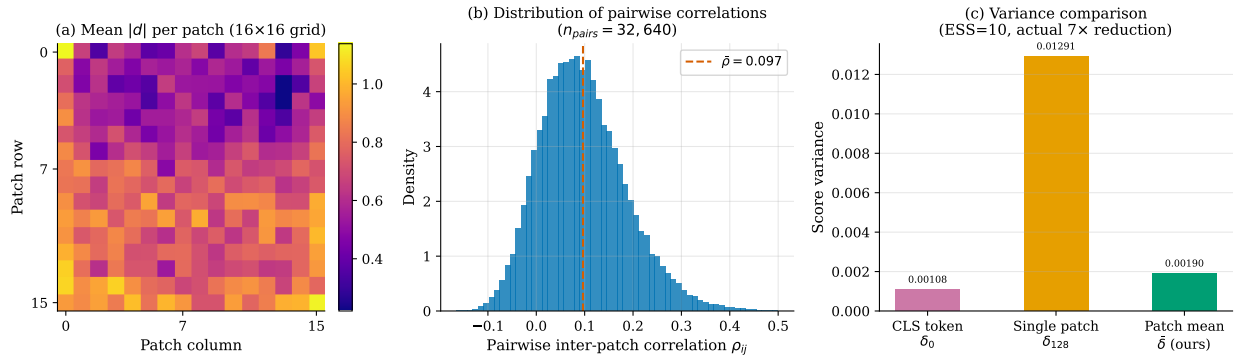

Figure 4. **(a)** Spatial heat map of the invariance gap: mean Cohen's $|d|$ per patch position (16×16 grid, averaged over five generators). The OOD-induced representation instability gap is spatially distributed, not localised to any region. **(b)** Inter-patch correlation structure: histogram of 32,640 pairwise correlations ($\bar{\rho}$=0.097), following a smooth spatial decay; most distant pairs are nearly uncorrelated. **(c)** Variance decomposition: the CLS token achieves 11.9× lower variance through global attention pooling but discards local signal; the patch mean achieves 6.8× lower variance than a single patch while retaining it.

strong regional concentration, confirming the signal is a global property of patch-token representations rather than a corner/edge artefact.

Inter-patch correlation structure. Mean pairwise correlation across the $\binom{256}{2}$=32,640 patch pairs is $\bar{\rho}$=0.097, giving an effective sample size $\mathrm{ESS} = P/(1+(P-1)\bar{\rho}) \approx 10$, rather than P=256 under independence. The structure is spatially smooth (Figure 4b): adjacent patches have $\rho\approx$0.20–0.30 via overlapping attention windows, while distant pairs are nearly uncorrelated.

Why patch mean beats CLS despite CLS having lower variance. The CLS token has lower variance than the patch mean ($\mathrm{Var}(\delta_0)$=0.00108 vs. 0.00190; Table 4), because global attention pooling produces an extremely smooth, compressed representation. The patch mean nevertheless yields larger $|d|$ on all five generators because $\Delta\mu$ grows proportionally faster than the score noise: the local spatial discriminative signal the CLS compresses away is recovered at no additional computational cost.

4.6. Perturbation Type Comparison: Training-Aug vs. Out-of-Training

The OOD-perturbation mechanism predicts an asymmetry between perturbations *within* DINOv2's training augmentation range and those *outside* it. DINOv2 inherits iBOT's (Zhou et al., 2022) blur augmentation $\sigma_{\mathrm{blur}} \sim \mathrm{Uniform}[0.1, 2.0]$ px; Gaussian noise is absent. Pillow's `ImageFilter.GaussianBlur(radius=r)` uses r as the Gaussian kernel standard deviation in pixels, so $r=\sigma_{\mathrm{blur}}$: r=1 lies inside the trained range, r=2 at its upper boundary, and $r\in\{3,5\}$ outside. We test the prediction with blur $r\in\{1,2,3,5\}$ px and noise $\sigma\in\{25,50,100\}$.

Table 5 reports worst-case $|d|$ (fixed seed 42; multi-seed stability in Section 4.8). Blur within the trained range (r=1 px) gives $|d|$=0.134; all noise variants give $|d|\geq$0.65 ($|d|$=0.93 at σ=100) — a ≈7× gap. Out-of-range blur ($r\in\{2,3,5\}$) gives $|d|\in\{1.47, 2.04, 2.34\}$, stronger than any noise variant. The ordering supports the mechanism: any perturbation outside DINOv2's trained invariance regime — noise or large-radius blur alike — reveals the OOD-induced stability gap.

4.7. Non-Perturbative Patch-Mean Baselines

We compare our perturbative probe against three non-perturbative patch-mean statistics computed from clean (unperturbed) features, to isolate the added value of perturbation vs. static distributional shifts:

1. **Intra-patch cosine distance** (s_{intra}): mean pairwise cosine distance between clean patch features within each image — a direct clean analog of our probe (which measures cosine distance under perturbation).
2. **Patch-norm std** (s_{norm}): standard deviation of patch-feature L2 norms within each image.
3. **CLS norm** (s_{cls}): L2 norm of the CLS token.

Table 6 reports the worst-case $|d|$ for each statistic. All three clean baselines are substantially weaker than the perturbative probe (worst-case $|d|$=0.98), confirming that the detection signal is not simply a static property of clean feature distributions but is *revealed* by perturbation. This rules out the hypothesis that patch-mean features trivially separate real from generated images without the perturbative probe.

4.8. Multi-Seed Stability

We assess the sensitivity of our results to the specific Gaussian noise samples used for perturbation by rerunning the

Table 4. Variance components of the detection score for real images (N=200, σ=100, RAISE-1k). ESS = effective sample size; reduction is relative to a single patch.

Score	Variance	$\sqrt{\mathrm{Var}}$	Reduction vs. single patch
Single patch δ_{128}	0.01291	0.1136	1×
Patch mean $\bar{\delta}$ (ours)	0.00190	0.0436	6.8×
CLS token δ_0 (RIGID)	0.00108	0.0329	11.9×

Table 5. Perturbation type comparison: worst-case $|d|$ across 5 generators (N=200, DINOv2-L, RAISE-1k). Only blur *within* DINOv2's training augmentation range (r=1 px) gives weak signal ($|d|$=0.134); blur *exceeding* the training range ($r\geq$2 px) and Gaussian noise both give strong signals, consistent with OOD status w.r.t. training augmentation driving probe sensitivity.

Type	Parameters	WC $\lvert d\rvert$
Gaussian noise (not in DINOv2 training aug)		
Noise	σ=25	0.656
Noise	σ=50	0.831
Noise	σ=100	**0.931**
Gaussian blur (r=1: within training aug; $r\geq$2: outside)		
Blur	r=1 px	0.134
Blur	r=2 px	1.466
Blur	r=3 px	2.043
Blur	r=5 px	2.342

Table 6. Non-perturbative vs. perturbative baselines (σ=100, N=200, DINOv2-L, RAISE-1k). Worst-case (WC) $|d|$ and AUROC across five generators. All clean statistics are substantially weaker than the perturbative probe.

Statistic	WC $\lvert d\rvert$	WC AUROC
Intra-patch cosine dist. (s_{intra})	0.071	0.520
Patch-norm std (s_{norm})	0.028	0.498
CLS norm (s_{cls})	0.028	0.506
Perturbative patch mean (ours)	**0.980**	**0.773**

main experiment with four independent noise seeds (42–45), image ordering fixed (Table 7). Cross-seed mean is 0.956±0.017 (CV 1.8%); the reported $|d|$=0.980 sits within 1.4σ of this mean.

4.9. Second Real Dataset: COCO val2017 (JPEG)

To probe the RAISE-1k upper-bound caveat, we rerun the detection experiment using MS-COCO 2017 validation images (Lin et al., 2014) (N_{real}=500, standard web JPEGs) as the real class against the same Synthbuster splits (Table 8).

Key findings. The signal persists on JPEG web images (worst-case $|d|$=0.74, marginal range), but the worst-case generator *switches* from Firefly to SDXL. This reveals a **generator-dataset alignment effect**: SDXL was trained on LAION (web data similar to COCO) and becomes harder to distinguish, whereas Firefly (Adobe Stock) is now easy against COCO ($|d|$=2.57) but was hardest against the also-professional RAISE. Practical takeaways: evaluate training-free detectors on multiple real datasets, and report worst-case generator relative to the deployment distribution rather than a fixed benchmark.

Table 7. Multi-seed stability (σ=100, N=200 per split). Worst-case (WC) $|d|$ across five generators for four independent noise seeds.

Seed	42	43	44	45	Mean±Std
WC $\lvert d\rvert$	0.931	0.951	0.964	0.977	0.956 ± 0.017

Table 8. Detection on COCO val2017 (JPEG real images), σ=100, N_{real}=500, N_{gen}=200. $\Delta|d|$: difference from RAISE-1k result. 95% bootstrap CIs in brackets. Worst case (SDXL) is bold.

Generator	$\lvert d\rvert$ [95% CI]	AUROC	$\Delta\lvert d\rvert$ vs RAISE
DALL-E 3	2.115 [1.956, 2.294]	0.967	+0.295
SDXL	**0.743 [0.579, 0.910]**	**0.714**	−2.088
Midjourney v5	1.464 [1.326, 1.611]	0.887	−1.377
DALL-E 2	1.720 [1.573, 1.881]	0.924	−0.612
Firefly	2.570 [2.388, 2.778]	0.986	+1.590
Worst-case $\lvert d\rvert$	0.743	—	—

4.10. Why Five Other Approaches Fail

Table 9 summarises five additional training-free signals, all of which fail to achieve worst-case $|d|\geq$0.5. The failures share a common structure we call *direction-inconsistency*: a signal that separates real from generated images for some generators flips sign or collapses to near-zero on at least one other, so no generator-independent threshold exists. Our perturbation score, in contrast, separates real from generated in the same direction (real < gen) for all five generators (Figure 1).

Local Intrinsic Dimension (LID). We compute Levina–Bickel LID estimates on DINOv2 ViT-L/14, CLIP ViT-L/14, and SigLIP patch embeddings. The signal is meaningful for DALL-E 3 ($|d|$=0.58) and SDXL ($|d|$=0.43) but collapses on DALL-E 2 ($|d|$=0.22, direction inconsistent) because DALL-E 2's unCLIP architecture explicitly conditions on CLIP embeddings during generation, placing its outputs close to the CLIP manifold and eliminating the LID gap. Crucially, *LID and perturbation sensitivity measure differ-*

Table 9. Training-free approaches evaluated on Synthbuster (N=100 pilot). Worst-case $|d|$ = minimum across five generators. "Killer" = the generator that causes failure (sign reversal or near-zero effect).

Approach	WC $\|d\|$	Best $\|d\|$	Killer (reason)
LID (DINOv2/CLIP/SigLIP)	0.22	0.58	DALL-E 2 (unCLIP targets CLIP space)
Physical consistency	0.17	0.44	Firefly (more consistent than real)
Spectral fingerprinting	0.25	0.71	Midjourney (matches camera noise)
MAE recon. error	0.02	0.31	SDXL/MJ (lower recon. loss than real)
Gradient field PCA	—	—	Degenerate (equivalent to spectral)
CLS baseline (RIGID)	0.860	1.969	—
Ours (patch mean)	**0.980**	**2.841**	—

ent things: LID estimates the effective dimensionality of the static feature distribution, while perturbation sensitivity measures dynamic response to a stressor. DALL-E 2 can match real images in LID (structural position) while still exhibiting the OOD-induced representation instability that our perturbative probe captures (functional robustness).

Physical consistency. We apply Depth Anything V2 to estimate depth maps, then measure four consistency metrics (depth–edge alignment, illumination coherence, normal roughness, gradient direction alignment). The signal is measurable for SDXL ($|d|$=0.44) and DALL-E 3 ($|d|$=0.38) but inverts for Firefly ($|d|$=0.17, generated images score as *more* physically consistent than real photographs), preventing a universal threshold.

Spectral fingerprinting. We compute residual variance, kurtosis, spectral slope, and high-frequency flatness on 512×512 native-resolution centre crops. The signal reaches $|d|$=0.71 on DALL-E 3 but collapses on Midjourney v5 ($|d|$=0.25): Midjourney's high-quality rendering matches camera noise statistics closely enough to collapse the residual variance signal.

MAE reconstruction error. We measure reconstruction loss of a frozen MAE ViT-L (`facebook/vit-mae-large`). Real RAISE photographs are *harder* to reconstruct than DALL-E 3/2 outputs ($|d|$=0.31), but SDXL and Midjourney v5 reconstruct with *lower* loss than real photographs ($|d|$=0.02), a direction reversal. The cause is training-distribution bias: MAE was pre-trained on ImageNet, whose aesthetic (object-centric, curated) is closer to SDXL/Midjourney than to RAISE's geographically diverse RAW photography.

Why Gaussian perturbation sensitivity avoids these failures. Unlike LID (structural), physical consistency (scene-level), spectral features (frequency domain), or MAE reconstruction (absolute loss), our score measures a *relative change* under a controlled stressor applied to the same image. This ratio-like structure is robust to distributional differences across generators: all five are OOD inputs to DINOv2, so their representations are consistently less stable under OOD perturbations than those of real photographs, regardless of generator architecture, training data, or output resolution.

5. Discussion

Implications for ViT probing. The CLS-vs.-patch asymmetry generalises beyond detection: global attention pooling is a lossy bottleneck for any spatially distributed property, so probes targeting such properties should be applied to patch tokens, and the inter-patch correlation structure (ESS$\approx P/10$ here) should be accounted for when estimating statistical power.

Robustness, generality, and JPEG. The signal transfers across backbones (DINOv2-B $|d|$=0.887, CLIP ViT-L/14 $|d|$=0.846; Appendix A) and survives realistic post-processing (JPEG 90 $\Delta|d|$=−0.019, resize +0.059, combined $|d|$=0.972; Appendix B). JPEG Q=30 as the perturbation gives near-zero signal ($|d|$<0.25 worst-case): DINOv2 was pre-trained with aggressive JPEG augmentation, so JPEG sits inside its trained invariance regime, mirroring the in-range-blur result in Section 4.6.

Comparison to HFI. High-Frequency Influence (HFI) is a concurrent training-free method based on frequency-domain statistics, reported to surpass AEROBLADE and RIGID on Synthbuster and GenImage; its implementation was not available to us. HFI and our method operate in different signal spaces (frequency domain vs. feature-space perturbation response) and are likely complementary.

Threat model and limitations. We frame the method as a passive forensic probe rather than an adversarial-robust detector. The operating point σ=100 corresponds to visually significant noise (roughly 39% of the pixel range), so the probe is conspicuous if applied in an interactive pipeline; this is acceptable for offline analysis but precludes deployment scenarios where covert perturbation is required. The same property also means the perturbation is trivially detectable by a preprocessing filter, and a determined adversary could fine-tune a generator to minimise DINOv2 patch shift under Gaussian noise — we have not run this adver-

sarial stress test and treat it as informative future work. At the operational thresholds reported in Section 4.3, Firefly remains hard (TPR@5%FPR=0.33), and the COCO results (Section 4.9) caution that the worst-case generator depends on the real-image distribution, so a single deployment threshold cannot be derived from RAISE-1k alone. The Synthbuster benchmark covers generators from 2022–2023; more recent systems (FLUX, Midjourney v6, newer Firefly, Ideogram 2) are not included, and we cannot claim the signal persists for all future architectures. Although the spatial heat map (Figure 4a) shows the invariance gap is distributed across patch positions, we did not perform spatial ablation experiments (e.g., restricting the probe to a subset of patches or masking image regions), which could further illuminate which parts of the image drive the signal. These limitations reinforce treating our method as a diagnostic baseline rather than a deployed system.

6. Conclusion

Crossing two OOD axes — an OOD perturbation applied to OOD inputs — turns a static distributional gap between real and AI-generated images into a *dynamic* stability gap in DINOv2 patch space. The gap is consistent across five architecturally diverse generators (worst-case $|d|=0.98$ on RAW, $|d|=0.74$ on JPEG web images), supported by a perturbation-type natural experiment ($r=1$ blur vs. noise vs. $r\geq 2$ blur), and recovered by aggregating patch tokens rather than the CLS — a transferable lesson for ViT probes targeting spatially distributed properties. Beyond the detection result, the study illustrates a more general methodological point: perturbing a frozen representation along an axis it was never trained to be invariant to exposes structure that static probing leaves hidden, and reading that response at the token granularity where it is encoded matters as much as the probe itself. The same recipe — pair an untrained-for perturbation with an OOD input class, then measure the response in patch space — may transfer to other frozen encoders and other notions of distributional shift, which we leave as future work alongside backbones beyond DINOv2.

The method is a diagnostic baseline, not a deployed detector. Operational thresholds remain uneven across generators and real-image distributions, and the conspicuous noise level makes it unsuitable where covert probing is required. The clearest open questions are therefore adversarial robustness (a generator fine-tuned to minimise DINOv2 patch shift under Gaussian noise) and generalisation to post-2023 models such as FLUX and Midjourney v6; both are tractable extensions of the present setup, and we view the dynamic stability gap reported here as a starting point rather than a finished detector.

References

Bammey, Q. Synthbuster: Towards detection of diffusion model generated images. volume 5, pp. 1–9, 2024.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. *Computer Science. https://cdn.openai.com/papers/dall-e-3.pdf*, 2(3):8, 2023.

Cohen, J. Statistical power analysis for the behavioral sciences. 1988.

Corvi, R., Cozzolino, D., Zingarini, G., Gragnaniello, D., Nagano, K., and Verdoliva, L. On the detection of synthetic images generated by diffusion models. 2023.

Dang-Nguyen, D.-T., Pasquini, C., Conotter, V., and Boato, G. RAISE: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM Multimedia Systems Conference (MMSys)*, pp. 219–224, 2015.

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., and Holz, T. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 3247–3258, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

He, Y. et al. Towards universal fake image detection exploiting foundation models. 2024.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Ojha, U., Li, Y., and Lee, Y. J. Towards universal fake image detection exploiting vision foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489, 2023.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. DINOv2: Learning robust visual features without supervision. In *Transactions on Machine Learning Research*, 2024.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8695–8704, 2020.

Wang, T., Gragnaniello, D., Mandelli, S., Verdoliva, L., and Cozzolino, D. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., and Li, H. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22445–22455, 2023.

Xi, Z., Huang, W., Wei, K., Huang, W., Zheng, H., Zhang, B., and Chen, B. PatchCraft: Exploring texture patch for efficient AI-generated image detection. 2023.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*, 2022.

A. Backbone Comparison

Table 10 compares patch-mean perturbation sensitivity at σ=100 across three backbone families. The signal is present in all three, confirming it is not specific to DINOv2-L's training objective or scale.

Table 10. Worst-case Cohen's $|d|$ by backbone (σ=100, N=200, 256 patches). Worst case is always Firefly in all three runs.

Backbone	WC $\|d\|$	FF AUROC	Verdict
DINOv2-L (ours)	**0.954**	0.765	MARGINAL
DINOv2-B	0.887	0.756	MARGINAL
CLIP ViT-L/14	0.846	0.738	MARGINAL

All three backbones show marginal-to-strong effect sizes for four of the five generators (DALL-E 3, SDXL, Midjourney v5, DALL-E 2 all achieve $|d|>1.7$ across all three backbones). DINOv2-L outperforms DINOv2-B, consistent with the known scaling behaviour of iBOT/DINO representations. CLIP-L/14 achieves $|d|=0.846$ despite a fundamentally different training objective (contrastive rather than self-supervised augmentation-based), which has two mechanistic implications. First, the OOD instability signal is not unique to DINOv2: CLIP's contrastive training on web-crawled natural images also produces real-image representations that are more stable under Gaussian noise than those of generated images, suggesting the effect generalises across training objectives. Second, DINOv2's augmentation-based objective yields a larger gap (higher worst-case $|d|$), consistent with the hypothesis that aggressive self-supervised augmentation creates more tightly anchored in-distribution representations, amplifying the contrast with OOD generated images.

B. Post-Processing Robustness

Table 11 reports worst-case $|d|$ after four post-processing operations applied to generated images before scoring. Real images are not post-processed (consistent with the deployment scenario where the adversary controls the generated image).

Table 11. Worst-case $|d|$ under post-processing (σ=100, N=200, DINOv2-L). Worst case is Firefly in all conditions. $\Delta|d|$: change from unprocessed; positive = detection improves.

Post-processing	FF $\|d\|$	WC $\|d\|$	$\Delta\|d\|$
None (baseline)	0.954	0.954	—
JPEG quality 90	0.935	0.935	-0.019
Resize 50% $\rightarrow$ 100% (bilinear)	1.012	1.012	$+0.059$
Gaussian blur σ=1.0	0.974	0.974	$+0.020$
Combined (JPEG 90 + blur)	0.972	0.972	$+0.018$

JPEG compression at quality 90 degrades worst-case $|d|$ by only 0.019. Bilinear resizing to 50% and back slightly improves detection ($+0.059$): interpolation appears to exaggerate the feature fragility of generated images. Mild Gaussian blur (σ=1) and the combined chain preserve $|d|>0.97$. These results confirm that the signal is robust to realistic benign post-processing without any adaptation.