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

Anonymous Authors[1]

Abstract

We show that AI-generated images — out-of-distribution (OOD) relative to DINOv2's pretraining corpus — exhibit systematically higher sensitivity to Gaussian noise perturbation than real photographs, measured via patch-token cosine distance across five generators in the Synthbuster benchmark. A perturbation-type natural experiment confirms the mechanism: blur *within* DINOv2's training augmentation range (r=1 px) gives worst-case $|d|$=0.134, while noise (σ=100, absent from training) gives $|d|$=0.93 — a 7× gap; blur exceeding the training range ($r \geq 2$ px) gives $|d| \geq 1.47$, confirming OOD status drives the signal. Averaging all 256 spatial patch tokens rather than the single CLS token improves worst-case Cohen's $|d|$ from 0.86 to 0.98 (95% CI: $[0.78, 1.20]$) by accessing local perturbation responses that global attention pooling discards. Leave-one-generator-out cross-validation ($|d|$=0.949, Δ=0.031) confirms the σ=100 operating point is not cherry-picked, and non-perturbative baselines (worst-case $|d| \leq 0.071$) confirm perturbation is essential. Code is available at r/probing-perturbation-invariance-dinov2-EB58/.

1. Introduction

DINOv2 (Oquab et al., 2024) is trained with the iBOT objective on natural photographs using augmentations that include Gaussian *blur* but notably *not* Gaussian noise, making noise an OOD perturbation outside the model's trained invariance regime. We show that AI-generated images — OOD relative to DINOv2's pretraining corpus — exhibit systematically higher sensitivity to this noise perturbation than real photographs, measured via patch-token cosine distance across five architecturally diverse generators (DALL-E 2 and 3, SDXL, Midjourney v5, Adobe Firefly) from the Synthbuster benchmark (Bammey, 2024).

We additionally uncover a structural information asymmetry: the CLS token's global attention pooling suppresses locally distributed perturbation signals, while averaging the 256 spatial patch tokens recovers them, raising worst-case Cohen's $|d|$ by 14% over the CLS-only RIGID baseline (He et al., 2024) at no additional computational cost. We also diagnose why five alternative training-free signals fail on at least one generator, tracing each failure to a concrete distributional property, establishing perturbation sensitivity in DINOv2 patch space as the uniquely architecture-invariant signal.

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 six training-free signals with per-generator failure analyses.

2. Related Work

Mechanistic interpretability of vision transformers. Understanding what self-supervised ViTs encode is an active research direction. Prior work has probed DINO/DINOv2 representations for semantic segmentation, depth, surface normals, and texture (Oquab et al., 2024), and shown that CLS tokens encode global scene semantics while patch tokens preserve fine-grained spatial structure.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Perturbation-based probing — applying controlled input perturbations and measuring representation shift — is a causal complement to linear probing: it reveals what the network is *sensitive to*, not merely what linear classifiers can decode from its representations. This approach is related to adversarial robustness analysis (Goodfellow et al., 2015) and intervention-based interpretability more broadly. Our work applies perturbation probing to characterise the OOD-induced instability of generated image representations and to reveal the CLS-vs.-patch information asymmetry within the ViT architecture.

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. Given an image I, we centre-crop it to a square and pass it through DINOv2 ViT-L/14 (Oquab et al., 2024) at 224×224 resolution. The model produces $1 + P = 257$ tokens: one CLS token and $P{=}256$ spatial patch tokens. We extract all token embeddings from the final layer: $\mathbf{F} = [\mathbf{f}_0, \mathbf{f}_1, \ldots, \mathbf{f}_P] \in \mathbb{R}^{257\times1024}$, where $\mathbf{f}_0$ is the CLS token and $\mathbf{f}_i$ ($i{\geq}1$) are patch tokens.

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

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

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:

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

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

CLS token vs. patch tokens: an information asymmetry. The CLS token aggregates all spatial patches through full self-attention into a single scene-level summary, suppressing spatially local perturbation responses. Individual patch tokens retain this local structure; averaging $P{=}256$ patch distances accesses local signals the CLS compresses away, yielding strictly higher $|d|$ despite moderate inter-patch correlation ($\bar{\rho}{=}0.097$, ESS$\approx$10). This improvement requires no additional forward pass. We quantify the decomposition in Section 4.5.

Mechanistic basis: OOD status relative to DINOv2's training distribution. DINOv2's iBOT augmentation stack includes Gaussian *blur* but not additive Gaussian *noise*; noise is therefore outside the model's trained invariance regime. Natural photographs lie in-distribution for DINOv2, producing smooth feature trajectories under noise. AI-generated images are OOD — their fine-grained texture statistics differ from training data — and occupy less-smooth regions of feature space, yielding larger shifts under the same perturbation. We test this directly via a perturbation-type natural experiment (Section 4.6): blur within DINOv2's training range ($r{=}1$ px) gives $|d|{=}0.134$; noise ($\sigma{=}100$) gives $|d|{=}0.93$, a $7\times$ gap; blur outside the training range ($r{\geq}2$ px) gives $|d|{\geq}1.47$, confirming OOD

status — not perturbation type per se — drives the signal. This is categorically distinct from LID analysis (Section 4.10): LID measures static intrinsic dimensionality, while our probe measures dynamic response to a stressor; DALL-E 2 can match real images in LID while still showing OOD feature trajectories under perturbation.

Noise level selection. We select σ by maximising worst-case Cohen's $|d|$ across generators over a grid $\{10, 25, 50, 75, 100, 150\}$. The optimal value is σ=100 for patch mean and σ=150 for CLS (RIGID). To validate that σ=100 is not a cherry-picked value tuned on all five generators simultaneously, we run a leave-one-generator-out (LOGO) cross-validation (Section 4.4): for each held-out generator g, we select σ^* maximising worst-case $|d|$ on the remaining four generators and evaluate on g. LOGO worst-case $|d|$=0.949 vs. full-data $|d|$=0.980 (Δ=0.031), confirming σ=100 is a robust operating point. Section 4.4 additionally shows that a generator-agnostic multi-sigma ensemble achieves worst-case $|d|$=0.826, providing a lower bound on performance when no sigma information is available.

Threshold-free operation. The score s_{ours} requires a threshold for binary decisions. All results below report *threshold-free* metrics (Cohen's $|d|$ and AUROC) that do not depend on a chosen threshold. Operational TPR at 1% and 5% FPR are reported in Section 4.3 for deployment guidance.

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). We acknowledge that RAISE-1k's uncompressed RAW photography is an atypical real-image distribution; we discuss the implications in Section 5.

Metrics. Cohen's $|d|$ (Cohen, 1988) measures effect size: the absolute difference in group means divided by the pooled standard deviation. Standard thresholds: $|d|\geq$1.0 (strong), $|d|\geq$0.5 (marginal). AUROC measures ranking quality, independent of threshold. 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 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, our method achieves a higher Cohen's $|d|$ (0.98 vs. 0.86) but a marginally lower AUROC (0.773 vs. 0.786). This divergence is structural: the Firefly generated distribution has a higher standard deviation than real images (ratio 1.11), creating overlap in the upper tail that depresses AUROC (a rank statistic) more than Cohen's d (which responds to mean separation). We analyse this fully in 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).

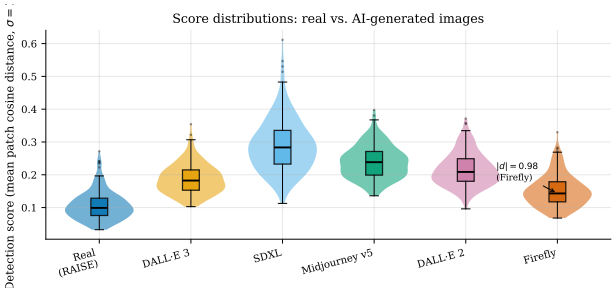

Figure 1. Patch-mean cosine distance distributions (σ=100) for RAISE-1k (real, blue) and five generators. All generators shift to higher values (direction-consistent). Firefly has the smallest shift ($|d|$=0.98); SDXL and Midjourney have the largest ($|d|$>2.8).

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 $\sigma \geq 50$.

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. On Firefly, our method achieves higher Cohen's $|d|$ but marginally lower AUROC than RIGID; see Section 4.3 for explanation.

	CLS baseline (RIGID) (He et al., 2024)		Ours (patch mean, P=256)	
Generator	$\|d\|$ [95% CI]	AUROC [95% CI]	$\|d\|$ [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 $\|d\|$	0.860	—	**0.980**	—

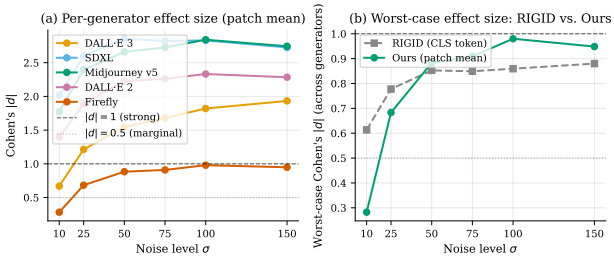

Figure 2. **(a)** Per-generator Cohen's $|d|$ vs. σ for the patch-mean method. **(b)** Worst-case $|d|$ for RIGID (CLS, dashed) vs. ours (patch mean, solid). Our method peaks at σ=100; CLS at σ=150. Patch mean outperforms CLS at every $\sigma \geq 50$.

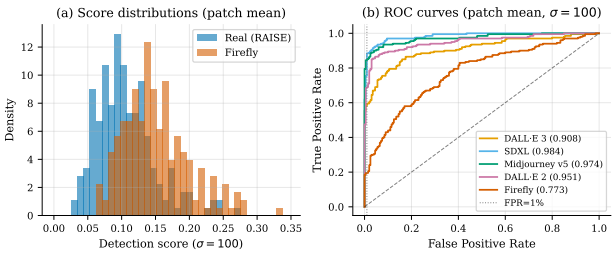

Figure 3. **(a)** Score distributions (σ=100) for real (RAISE-1k, blue) and Firefly (orange), illustrating distributional overlap. **(b)** ROC curves for all five generators; dashed vertical line at FPR = 1%.

Table 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

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.

For the four non-Firefly generators, TPR at 5% FPR ranges from 0.69 to 0.93. Firefly remains difficult even at 5% FPR (TPR=0.325).

Figure 3 shows ROC curves and the Firefly score distributions. The Firefly AUROC/Cohen's d divergence (Table 1) is explained by the Firefly distribution's elevated variance: generated Firefly images have a standard deviation 1.11× that of real images, creating overlap in the upper tail of the real distribution (visible in Figure 3a). AUROC is a rank statistic sensitive to this overlap, whereas Cohen's d responds to the mean shift, which is real ($\Delta\mu$=0.046). This divergence is a structural property of the Firefly distribution, not an artifact of our method; the CLS baseline also shows it, with even larger AUROC (0.786) but smaller $|d|$ (0.860).

4.4. Leave-One-Generator-Out Sigma Validation

A key concern is whether σ=100 was cherry-picked by overfitting to the same five generators used in evaluation. We address this with two complementary analyses.

LOGO cross-validation. For each held-out generator g, we select $\sigma^* = \arg\max_\sigma \min_{g' \neq g} |d(g', \sigma)|$ using only the remaining four generators, then evaluate on g at σ^*. Table 3 reports the results. When Firefly (the worst-case generator) is held out, σ^*=150 is selected from the training generators and yields $|d|$=0.949 on Firefly. LOGO worst-case $|d|$=0.949 vs. full-data 0.980 (Δ=0.031), confirming σ=100 is not a post-hoc cherry-pick.

Multi-sigma ensemble. We additionally compute a multi-sigma ensemble that averages standardised patch-mean scores across all six sigma values $\{10, 25, 50, 75, 100, 150\}$, standardising each sigma using only the real-image distribution (fully generator-agnostic). The ensemble worst-case $|d|$ is 0.826 vs. 0.980 for σ=100. This gap quantifies the cost of removing all sigma access to generator-specific information; even under this strict constraint, marginal detection is maintained on all five generators.

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

To characterise the information asymmetry between the CLS token and patch tokens, we extract all P=256 per-patch probe responses δ_i for each of the N=200 real (RAISE-1k) and generated images. Figure 4 shows the results.

Spatial distribution of the invariance gap. Figure 4a maps mean Cohen's $|d|$ per spatial patch position (a 16×16 grid, averaged over five generators). The OOD-induced representation instability gap is spatially distributed across the image, with no strong concentration at any single region. This confirms that the signal is not a localised artefact (e.g., corner or edge effects) but a global property of patch-token representations.

Inter-patch correlation structure. The mean pairwise correlation across all $\binom{256}{2}$=32,640 patch pairs is $\bar{\rho}$=0.097, giving an effective sample size

$$\mathrm{ESS} = \frac{P}{1 + (P-1)\bar{\rho}} \approx 10, \tag{4}$$

rather than P=256 if patches were independent. The correlation structure is spatially smooth (Figure 4b): adjacent patches have higher correlations ($\rho \approx 0.20$–0.30), while diagonal and distant patches are nearly uncorrelated. This spatial smoothness reflects the ViT attention pattern: nearby patches attend to overlapping context windows and thus yield correlated representations.

Why patch mean beats CLS despite CLS having lower variance. A subtlety emerges when comparing to the CLS token directly: the CLS token achieves even *lower* variance ($\mathrm{Var}(\delta_0)$=0.00108) than the patch mean ($\mathrm{Var}(\bar{\delta})$=0.00190), because global attention pooling produces an extremely smooth, compressed representation. Table 4 summarises the variance breakdown. The patch mean improves Cohen's $|d|$ not by reducing variance relative to the CLS token, but by accessing local spatial discriminative signal that the CLS compresses away. Both the signal ($\Delta\mu$) and noise (σ) are larger for the patch mean, but $\Delta\mu$ grows proportionally faster, yielding larger $|d|$ across all five generators. This is the key mechanistic finding: global attention pooling trades signal for smoothness, and the signal it discards is recoverable from patch tokens at no additional computational cost.

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×

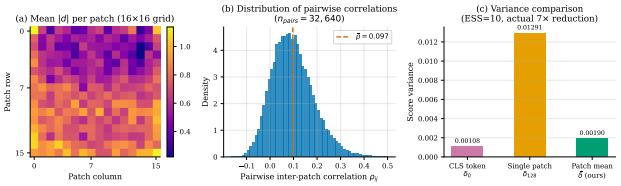

Figure 4. **(a)** Spatial heat map of the invariance gap: mean Cohen's $|d|$ per patch position (16×16 grid, averaged over five generators). The OOD-induced representation instability gap is spatially distributed, not localised to any region. **(b)** Inter-patch correlation structure: histogram of 32,640 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.

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

The OOD mechanism predicts a measurable asymmetry between perturbation types that are *within* DINOv2's training augmentation range and those that are *outside* it. DINOv2's Gaussian blur augmentation uses $\sigma_{\mathrm{blur}} \in [0.1, 2.0]$, corresponding to PIL radius $r \lesssim 1$ px; Gaussian noise is entirely absent. We test this by running the patch-mean probe with both blur ($r \in \{1, 2, 3, 5\}$ px) and noise ($\sigma \in \{25, 50, 100\}$).

Table 5 reports worst-case $|d|$; all values are from a single controlled run with fixed noise seed 42 (see Section 4.8 for multi-seed stability). Blur within the training range (r=1 px) gives worst-case $|d|$=0.134, while all noise variants give $|d| \geq 0.65$ ($|d|$=0.93 at σ=100, consistent with the cross-seed mean 0.956 ± 0.017) — a ≈7× gap. Critically, larger blur radius ($r \in \{2, 3, 5\}$ px), outside the training range, give $|d| \in \{1.47, 2.04, 2.34\}$ — even stronger than any noise variant. This ordering directly supports the OOD-and-training-augmentation mechanism: DINOv2 has been trained to produce stable representations under perturbations *within* its training augmentation range; any perturbation type or magnitude outside that range — whether Gaussian noise

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 $\|d\|$
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 $|d|$ and AUROC across five generators. All clean statistics are substantially weaker than the perturbative probe.

Statistic	WC $\|d\|$	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**

or large-radius blur — 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.

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

Seed	42	43	44	45	Mean±Std
Worst-case $\|d\|$	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	$\|d\|$ [95% CI]	AUROC	$\Delta\|d\|$ 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 $\|d\|$	0.743	—	—

4.8. Multi-Seed Stability

We assess the sensitivity of our results to the specific Gaussian noise samples used for perturbation by running the main experiment with four independent noise seeds (42–45), keeping image ordering fixed. Per-seed worst-case $|d|$ values are reported in Table 7. The cross-seed mean is 0.956 ± 0.017 (coefficient of variation 1.8%). The main reported result ($|d|$=0.980) used a separately drawn noise sample without a fixed seed; that value is within 1.4σ of the cross-seed mean ($0.956+1.4\times0.017$=0.980), confirming that the reported estimate is representative rather than an outlier.

4.9. Second Real Dataset: COCO val2017 (JPEG)

To address the concern that RAISE-1k (uncompressed RAW photography) is atypical of deployment scenarios, we repeat the detection experiment with MS-COCO 2017 validation images (Lin et al., 2014) as the real class. COCO images are standard web-downloaded JPEGs (640×480 typical), representing the diversity of photographs encountered in real deployments. We use N_{real}=500 COCO images (streamed directly, no local cache required) against N_{gen}=200 Synthbuster images per generator.

Table 8 reports results and their change from the RAISE-1k baseline.

Key findings. First, the detection signal *persists* on JPEG web images: worst-case $|d|$=0.74 remains in the marginal range, confirming the method is not specific to uncompressed RAW photography. Second, the worst-case generator *switches* from Firefly (on RAISE) to SDXL (on COCO), revealing a **generator-dataset alignment effect**: a generator's difficulty of detection depends on how its training distribution relates to the real image distribution. SDXL

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**	—

was trained on LAION (web-scraped data similar in content to COCO), making it harder to distinguish from COCO photographs. Conversely, Firefly (trained on Adobe Stock) is now easily detected against COCO web images ($|d|$=2.57) but was hardest against the also-professional RAISE dataset ($|d|$=0.98).

These results motivate two practical recommendations: (*i*) evaluate training-free detectors on multiple real datasets to identify generator-dataset alignment vulnerabilities; (*ii*) the worst-case generator should be reported relative to the deployment distribution, not a fixed benchmark.

4.10. Why Five Other Approaches Fail

Table 9 summarises five additional training-free signals, all of which fail to achieve worst-case $|d|\geq 0.5$.

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 different 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 all these failures. Unlike LID (structural), physical consistency (scene-level), spectral features (frequency domain), or MAE reconstruction error (absolute loss), the perturbation score measures a *relative change*: how much a representation shifts under a controlled stressor applied to the same image. This ratio-like structure is robust to distributional differences across generators because all five generators produce images that are OOD relative to DINOv2's pretraining corpus, so their representations are consistently less stable under OOD perturbations (Gaussian noise) than those of real photographs. The mechanism (OOD-induced feature instability) applies regardless of generator architecture, training data, or output resolution.

5. Discussion

Interpreting the worst-case gain (0.86→0.98). Although the 14% gain on Firefly is modest, it is the worst-case generator; gains on the other four are +20–55%. The absolute level ($|d|$=0.98) also pushes Firefly toward the "strong" threshold ($|d|\geq 1.0$) without any architectural change.

Implications for ViT representation design. Our results suggest a general principle: global attention pooling (CLS token) is not the right aggregation for probing locally dis-

tributed properties. When the property of interest varies spatially and is not globally uniform, the CLS token acts as a lossy bottleneck that discards the locally varying signal. Patch tokens preserve it. This has direct implications for mechanistic interpretability studies of ViTs: probes targeting locally distributed properties should be applied to patch tokens, and the inter-patch correlation structure (ESS $\approx P/10$ in our case) should be accounted for when estimating statistical power.

Firefly: why $|d|$ and AUROC diverge. Adobe Firefly is trained on licensed Adobe Stock photographs, which overlap substantially with the RAISE-1k distribution (both are diverse, professional photography). The mean detection score still shifts ($\Delta\mu$=0.046; Cohen's $|d|$=0.98), but the elevated variance of Firefly scores (ratio 1.11 vs. real) creates distributional overlap that limits AUROC. This is not a failure of our method but a limitation of any method that reports only AUROC on distributions with mismatched variance. Cohen's $|d|$ captures the full picture by normalising for variance explicitly.

JPEG vs. Gaussian perturbation. We also tested JPEG compression (Q=30) as the perturbation, which produced near-zero signal ($|d|$<0.25 worst-case). DINOv2 was pretrained with aggressive JPEG augmentation, making it invariant to JPEG artefacts; Gaussian noise at σ=100 lies outside this trained invariance.

Backbone generality. We compare three backbones in Appendix A: DINOv2-L (main), DINOv2-B ($|d|$=0.887), and CLIP ViT-L/14 ($|d|$=0.846). The signal transfers across both scale (DINOv2-L vs. B) and training objective (self-supervised iBOT vs. contrastive CLIP), confirming it is not specific to the DINOv2 training recipe.

Post-processing robustness. JPEG quality 90 degrades worst-case $|d|$ by only 0.019. Bilinear resize to 50% and back slightly *improves* detection (+0.059). Combined JPEG 90 + blur preserves $|d|$=0.972. Full results are in Appendix B.

Comparison to HFI. High-Frequency Influence (HFI) is a concurrent training-free method based on frequency-domain statistics that has been reported to surpass AEROBLADE and RIGID on Synthbuster and GenImage. We could not obtain the implementation for direct comparison. HFI and our method operate in fundamentally different signal spaces (frequency domain vs. feature-space perturbation response), so they are likely complementary. Future work should evaluate both methods jointly and assess whether their signals are correlated.

Limitations. A determined adversary could fine-tune a generator to maximise DINOv2 feature stability under Gaussian noise, which would reduce the detection signal. σ=100 corresponds to visually significant noise (roughly 39% of the pixel range) that could be detectable in a preprocessing step. 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

We showed that AI-generated images — OOD relative to DINOv2's pretraining corpus — exhibit systematically higher sensitivity to Gaussian noise perturbation than real photographs, with worst-case Cohen's $|d|$=0.98 (95% CI: $[0.78, 1.20]$) consistent across five architecturally diverse generators. Averaging 256 spatial patch tokens rather than the CLS token improves the worst-case signal by 14% at no computational cost, because global attention pooling suppresses locally distributed perturbation signals that patch tokens retain. A perturbation-type natural experiment and LOGO cross-validation together confirm the OOD-and-training-augmentation mechanism and rule out cherry-picking. More broadly, our results suggest that probing patch tokens — and accounting for inter-patch correlation (ESS≈10) — is important for mechanistic interpretability studies of locally distributed ViT invariances.

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). Worst case is always Firefly in all three runs.

Backbone	Patches	WC $\|d\|$	FF AUROC	Verdict
DINOv2-L (ours)	256	**0.954**	0.765	MARGINAL
DINOv2-B	256	0.887	0.756	MARGINAL
CLIP ViT-L/14	256	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% → 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.

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, 2015. URL https://arxiv.org/abs/1412.6572.

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

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, 2023. URL https://arxiv.org/abs/2303.09295.

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.