PIXEL ECHO: MULTISCALE PHASE DECOMPOSITION OF GENERATOR FINGERPRINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generated images carry recurring traces locked to the pixel grid. Where these traces originate decides what an image can be attributed to and which operations erase them. We introduce *Pixel Echo*, a diagnostic measurement that resolves these traces by spatial phase and repetition period. Pixel Echo pools local pixel responses at matching grid positions and exactly decomposes the resulting periodic structure into count-weighted orthogonal increments at periods of 2, 4, 8, and 16 pixels. Their strengths define a *scale profile* for examining how fingerprints vary across decoders and input routes. Controlled comparisons across two decoder families show that reconstructions and generations sharing a decoder exhibit closely aligned patterns, while pixel-head replay links one decoder's pattern to phase-specific projection weights. Fixed-latent decoder swaps reveal a period-dependent distinction: decoder-associated patterns concentrate at periods up to the upsampling factor, whereas longer-period structure can pass through from the input. Specifically, six $\times 8$ decoders differ at periods 2–8 but retain a 16-pixel component shared with each other and with real photographs, whereas $\times 16$ decoders exhibit decoder-specific patterns at this period. On held-out splits of our class-to-image and text-to-image benchmarks, linear classifiers frozen beforehand achieve 95.1% four-way and 90.9% seven-way source attribution accuracy, increasing to 98.6% and 94.9% when combined with frequency features, at 0–1.4% per-source false-positive rates on real images. These results support closed-set attribution of decoder and export pipelines rather than unique identification of upstream generators, with attribution requiring preservation of the pixel grid. Code, benchmark manifests, frozen experiment definitions and an interactive demo are available at `https://pixel-echo.github.io/`.

1 INTRODUCTION

Generated images contain model fingerprints, checkerboard patterns and spectral discrepancies (Marra et al., 2019; Yu et al., 2019; Odena et al., 2016; Durall et al., 2020; Corvi et al., 2023; Tan et al., 2024). Such traces help distinguish image sources, but a classifier that uses them does not say which stage of image formation wrote them, and modern generators share stages: many diffusion and autoregressive models decode their outputs through the same or closely related latent decoders (Figure 1). Where these patterns originate has not been examined in depth, yet it decides what a trace can attribute an image to, which operations erase it, and whether the pipeline can be read back from the pixels. This raises a natural question: **what does a recurring pixel pattern reveal about the image-formation pipeline?**

To answer this question, we first propose *Pixel Echo*, a diagnostic measurement that resolves the repeating patterns a decoder writes into every image it produces, by where they sit on the pixel grid and by how many pixels apart they repeat. It measures how each pixel differs from its neighbors and averages the responses at matching positions within a repeating grid (Figure 2), so that a pattern tied to fixed positions adds up while image content averages out. Specifically, a decoder enlarges its input by repeated doubling, and a pattern written at one doubling repeats every 2, 4, 8 or 16 pixels in the output depending on how many doublings follow. Accordingly, we split the pooled map into four parts, one per period, each holding only the pattern that first appears at that spacing; the parts add back to the original exactly and do not overlap. Moreover, the *scale profile*, the strength of each increment, records at which periods a decoder leaves its imprint.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

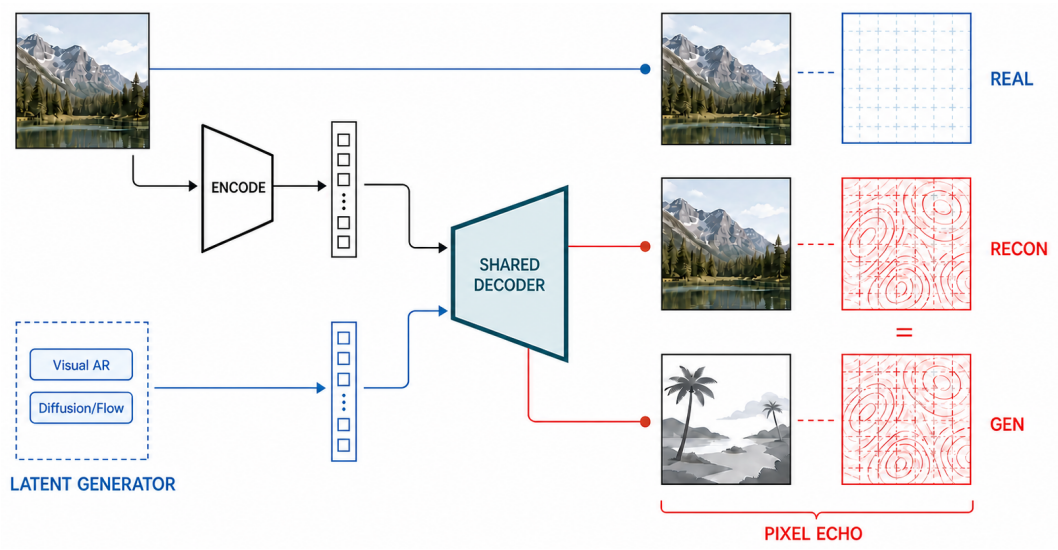


Figure 1: **Illustration of Pixel Echo, the recurring pixel pattern a decoder writes into its outputs.** A real image bypasses the decoder and carries no such pattern; an encoded image and a generated latent reach the same decoder by different routes and come out with the same pattern ("=" denotes matching cohort-level patterns, not equal images or amplitudes; schematic, not measured heatmaps).

With this measurement in hand, we then trace the pattern to its source by controlled substitution: one stage of the pipeline is exchanged at a time, all others held fixed, and the scale profile records what survives and what changes. First, holding the decoder fixed and changing only its input route, a real image encoded and reconstructed versus a generated latent, leaves the pattern unchanged (Figure 4): the same decoder writes the same fingerprint in two decoder families ($r = 0.996$ and 0.846 against 0.06 across different decoders), a single image already shows it, and pixel-head replay locates it in phase-specific weights (Sections 4.1–4.2). Second, holding the latents fixed and changing the decoder does the opposite, the fingerprint changes, but not at every period: a decoder's own pattern is concentrated at periods up to its upsampling factor and the input's structure can pass through above them, so six $\times 8$ decoders applied to the same latents differ at $P2$–$P8$ yet share $P16$ with each other and with real images, whereas $\times 16$ decoders write $P16$ themselves (Section 4.3). Two of those decodings agree at only $r = 0.12$ in the full field yet at 0.82 in $P16$: this is what the decomposition is for.

Having traced the pattern to the decoder, we use the measurement as it stands for source attribution and detection, on a benchmark built for the purpose: a class-to-image set with matched Real controls and six generators, and a text-to-image set with seven, each with public Train/Validation splits and a private Test (construction and protocol in Appendix H). Permutation tests show that real images and the four class-to-image sources differ pairwise in the measurement (Holm-adjusted $\mathbf{p \leq .0012}$; Section 5.2). On the held-out Test, heads and thresholds frozen on Validation attribute $\mathbf{95.1}\%$ of class-to-image and $\mathbf{90.9}\%$ of text-to-image images to the correct source ($\mathbf{98.6}\%$ and $\mathbf{94.9}\%$ with frequency features appended), with per-source Real false-positive rates of $\mathbf{0}$–$\mathbf{1.4}\%$ (Section 5). Attribution is closed-set and identifies the decoder and export pipeline rather than the upstream generator, and the pixel grid must be preserved.

Our contributions are:

- **Exact multiscale measurement.** A phase-pooled estimator and an exact count-weighted decomposition whose increment strengths form a decoder's scale profile (Section 3).
- **Shared-decoder evidence and the period rule.** Same-decoder alignment in two families, single-image fit, pixel-head replay, and the rule that a decoder's own pattern is concentrated at periods up to its upsampling factor (Section 4).
- **Benchmark, code and tools.** C2I/T2I datasets, source-level tests, held-out detection and attribution with frozen heads (Section 5; Appendix H), code, website and demo.

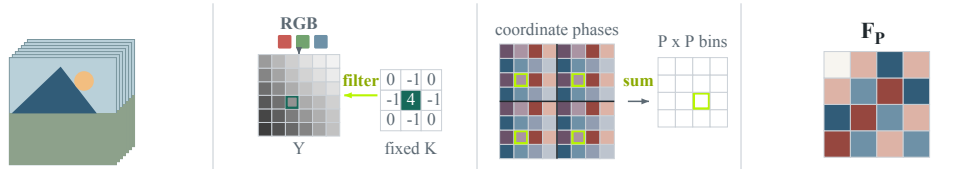


Figure 2: Pixel Echo estimator (schematic; $P = 4$). RGB images are converted to luminance Y and filtered with fixed K to obtain $E = |K * Y|$ using replicate padding. Responses at matching modulo-P coordinates are pooled into F_P by one sum/count division, without per-image normalization. Green boxes mark repeated phase classes, not patches.

2 Related Work and Terminology

We use *fingerprint* for an inherent statistical trace of generation (Marra et al., 2019; Yu et al., 2019), as opposed to a watermark deliberately embedded in outputs (Yu et al., 2021); *Pixel Echo* names the structured subclass studied here, recurring local responses tied to output-pixel coordinate classes, grouped by $(i \bmod P, j \bmod P)$ after the multirate *polyphase* convention (Vaidyanathan, 1993). Its ingredients have close precedents. Uneven upsampling overlap and phase-dependent filters explain checkerboard artifacts (Odena et al., 2016; Aitken et al., 2017; Sugawara et al., 2019); spectral discrepancies support detection (Frank et al., 2020; Durall et al., 2020; Dzanic et al., 2020), and Fourier, autocorrelation and neighbouring-pixel traces extend across modern generators (Corvi et al., 2023; Tan et al., 2024). Chandrasegaran et al. (2021) analyze architecture-dependent Fourier discrepancies, while processing or explicit mitigation weakens spectral cues and the detectors built on them (Dong et al., 2022; Tang et al., 2023); SPAI learns a masked spectral representation for any-resolution detection (Karageorgiou et al., 2025). Detection itself has been built on generic CNN features and diffusion reconstruction errors (Wang et al., 2020; 2023), and FingerprintNet synthesizes fingerprint cues from autoencoders with several upsampling configurations (Jeong et al., 2022). Reconstruction is an equally direct precedent: Vesnin et al. (2024) transfer a real-versus-autoencoder-reconstruction detector to latent-diffusion outputs, and CO-SPY combines VAE reconstruction residuals with semantic evidence (Cheng et al., 2025). Attribution, asking which model or family produced an image, is usually posed in a closed set: Deep Image Fingerprint follows few-example residual patterns across fine-tuned models (Sinitsa & Fried, 2024), ManiFPT formalizes artifacts relative to a data manifold (Song et al., 2024; Song & Itti, 2025), and causal representations decouple source traces from content (Xu et al., 2025).

Pixel Echo keeps what these methods integrate away. Spectral statistics discard where a response sits on the grid, and one fingerprint correlation mixes every repeat spacing into a single number; we retain the spatial phase and split it into nested periods, so structure that first appears at one spacing is separated from repetitions of a shorter one. A controlled substitution can then be read period by period instead of as one aggregate change, which is also how our reconstruction comparison differs in purpose: we ask which periods survive two routes through a shared decoder, not whether reconstruction residuals separate real from generated images. Pixel-space generators with no latent decoder, such as PixelDiT (Yu et al., 2026), serve as a counterfactual for decoder-induced structure; our attribution stays closed-set and decoder-level, and a classifier score is never read as a source probability. Appendix B gives the fuller detection and attribution literature, causal terminology, and the distinction between pixel-space generation and the proposed Independent Pixel Generation (IPG) counterfactual.

3 Pixel Echo Measurement

Pixel Echo is a recurring pattern in local pixel variations at regularly spaced image positions. To measure it, we first compute how each pixel differs from its immediate neighbors, then average these responses at matching positions within a repeating grid. The resulting map summarizes the periodic response in an image or an image collection (Figure 2). Specifically, Section 3.1 defines how to compute this map, and Section 3.2 separates its contributions at different spatial scales.

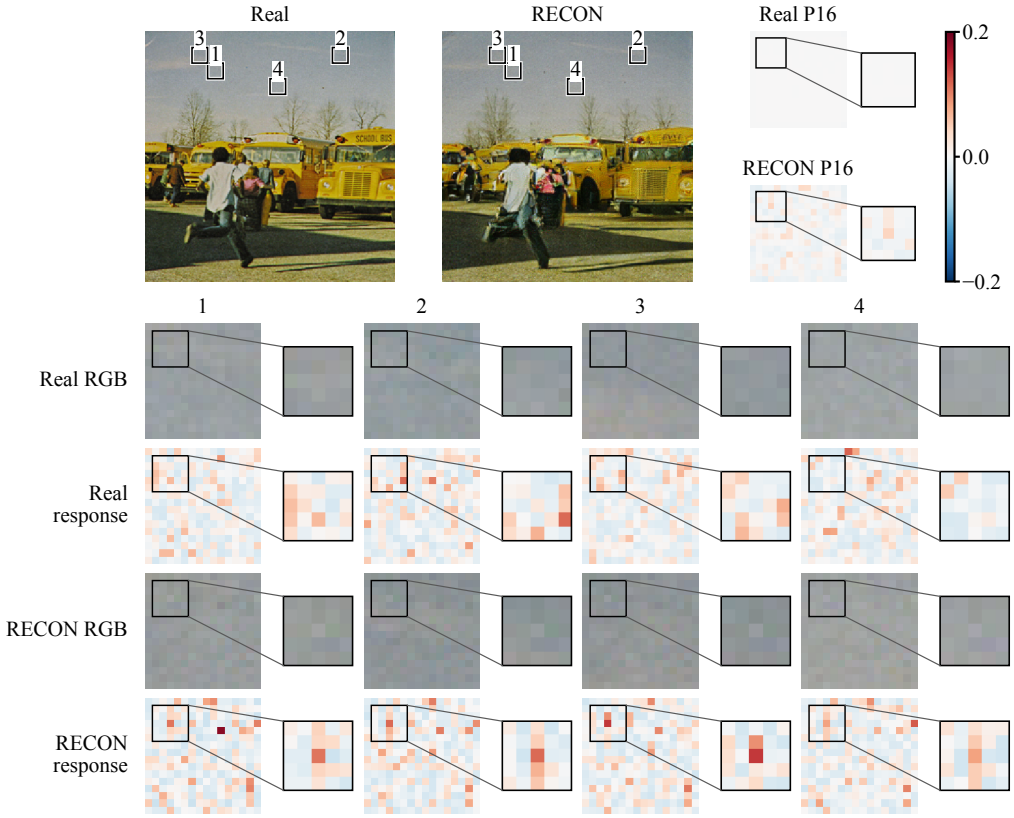


Figure 3: **Local Real–RECON responses and cohort-level Pixel Echo maps:** Real responses vary with local scene content; the selected RECON patches partly resemble the cohort pattern, with higher responses at some matching phase positions. Magnifiers show matching 5×5 regions. This pair is the best case of the search; typical pairs and the full candidate distribution are in Appendix E.2.

3.1 Phase-conditioned response

For RGB image $x \in [0,1]^{H\times W\times 3}$, luminance Y_x uses BT.709 channel weights, and $\widetilde{Y}_x$ replicates boundary values:

$$\begin{aligned} Y_x(i,j) &= 0.2126x_R(i,j) + 0.7152x_G(i,j) + 0.0722x_B(i,j), \\ \widetilde{Y}_x(i,j) &= Y_x(\mathrm{clip}(i,0,H-1), \mathrm{clip}(j,0,W-1))\,. \end{aligned} \tag{1}$$

At each original coordinate $(i,j) \in \Omega_x = \{0,\dots,H-1\} \times \{0,\dots,W-1\}$, the response is

$$E_x(i,j) = \left|4\widetilde{Y}_x(i,j) - \widetilde{Y}_x(i-1,j) - \widetilde{Y}_x(i+1,j) - \widetilde{Y}_x(i,j-1) - \widetilde{Y}_x(i,j+1)\right|. \tag{2}$$

This is the absolute four-neighbor Laplacian response. The luminance weights, filter, absolute value and replicate padding are design choices, not unique or optimal requirements. For cohort $\mathcal{D}$ and phase $(a,b) \in \Omega_{16} = \{0,\dots,15\}^2$, pool once at maximum period 16:

$$\begin{aligned} S(a,b) &= \sum_{x\in\mathcal{D}} \sum_{(i,j)\in\Omega_x} \mathbf{1}[i \equiv a,\ j \equiv b \pmod{16}] E_x(i,j), \\ C(a,b) &= \sum_{x\in\mathcal{D}} \sum_{(i,j)\in\Omega_x} \mathbf{1}[i \equiv a,\ j \equiv b \pmod{16}], \qquad F(a,b) = F_{16}(a,b) = \frac{S(a,b)}{C(a,b)}. \end{aligned} \tag{3}$$

Phase labels retain the raster origin and change under crops or shifts. Raw pooling is pixel-weighted for mixed resolutions, without per-image or cohort normalization. Display centering subtracts the unweighted phase mean; decomposition and norms use count-weighted projections.

From one block to one image to a cohort. Figure 3 shows the phenomenon on one pair; Table 13 (Appendix E.3) quantifies it against the ORT GEN Train template. A single 16×16 block barely shows Pixel Echo (RECON median $r = 0.165$), but one image's full-raster field (Equation 3 over its 256 phase-aligned blocks) already does (0.690 for RECON, 0.686 for ORT GEN, 0.009 for Real); this per-image field is exactly the classifier input, so single-image attribution (Section 5) rests on it, and averaging many images only sharpens it toward the cohort template (0.995 versus 0.160 at $n = 1{,}000$).

3.2 Exact multiscale decomposition

A latent decoder enlarges its input by a sequence of doublings; the total enlargement is its *upsampling factor* ($\times 8$ for the SD VAEs, $\times 16$ for AliTok). Every doubling turns one input position into a 2×2 block of pixels; if the network treats the four pixels differently it leaves a pattern that repeats every two pixels, and each later doubling stretches that pattern, so a pattern written early repeats every four, eight or sixteen pixels in the final image: each doubling is one period (Section 4.4). The response map of Section 3.1 shows all these patterns on top of each other; to see each period separately we tile the smaller-period pattern over the larger grid and subtract it from the larger-period pattern.

For each period $q \in \{1, 2, 4, 8, 16\}$ we average the responses of positions with the same modulo-q coordinates, weighted by the pooled counts C, and tile the resulting $q \times q$ template over the 16×16 grid:

$$(\Pi_q^C F)(a, b) = \frac{\displaystyle\sum_{\substack{u \equiv a \pmod q \\ v \equiv b \pmod q}} C(u, v) F(u, v)}{\displaystyle\sum_{\substack{u \equiv a \pmod q \\ v \equiv b \pmod q}} C(u, v)}. \tag{4}$$

Here $\Pi_1^C F$ is the overall mean. Subtracting successive templates gives the additional contribution at each scale, which we call a primitive increment:

$$\Delta F_q = \Pi_q^C F - \Pi_{q/2}^C F, \qquad q \in \{2, 4, 8, 16\}. \tag{5}$$

For example, ΔF_4 is the 4×4 template minus the tiled 2×2 template, and vanishes if the former only repeats period-two structure.

When all images share the phase origin and every phase has a positive count, the four increments are orthogonal under the count weighting, and adding them to the overall mean gives back F exactly (Appendix C.1). How strong each increment is, measured by its count-weighted RMS $e_q = \|\Delta F_q\|_C$ or by its share of the total energy $e_q^2 / \sum_{q'} e_{q'}^2$, is the *scale profile* of a cohort: four numbers that say at which periods a decoder writes, which Sections 4 and 5 compare across decoders. Numerical checks of the decomposition and boundary sensitivity are in Appendix C.2.

4 Shared-Decoder Evidence and Mechanism Hypothesis

Section 3 measures Pixel Echo; this section asks where it comes from. We compare input routes through a shared decoder (Section 4.1), probe the pixel head with fixed hidden states (Section 4.2), relate the observed periods to the decoder's upsampling factor (Section 4.3), and propose how a stage writes them (Section 4.4).

Table 1: **Shared-decoder alignment:** Pearson r for full $P16$ maps and concatenated primitive increments (340D).

Pair	$P16$	340D
Real–RECON	0.1568	0.1517
Real–GEN	0.1529	0.1557
RECON–GEN	0.9959	0.9958

4.1 Matched Real, RECON, and GEN

Do the overall patterns match? We compare 1,000 images per cohort using the estimator in Section 3. Real images supply encoder inputs and direct controls; their AliTok (Wu et al., 2025) reconstructions and ORT generations share the output decoder. Figure 4 shows aligned RECON/GEN patterns and a nearly phase-flat Real control. Table 1 quantifies this contrast: RECON and GEN align closely in both the full map and the primitive representation, whereas Real aligns weakly with either.

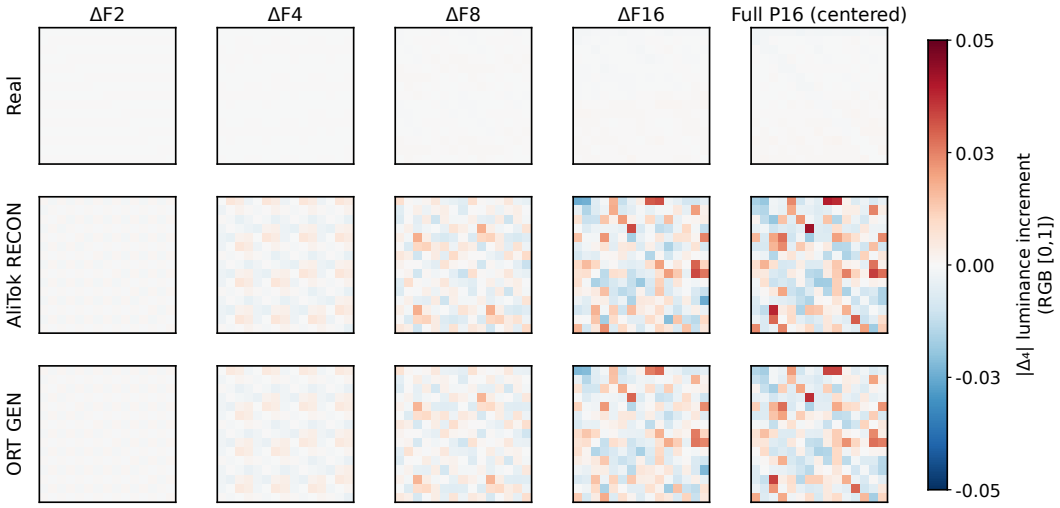


Figure 4: Shared-decoder phase geometry. Columns show the four count-weighted primitive increments ΔF_2–ΔF_{16}, each tiled to the common 16×16 grid, and the display-centered full $P16$ field, which equals their sum. Real, AliTok RECON, and ORT GEN share one linear scale. RECON/GEN align at every scale while Real is nearly phase-flat.

Table 2: **Shape vs strength:** RECON–GEN full map, 95% bootstrap CI.

Quantity	Value	95% CI
Shape (Pearson r)	0.996	[0.992, 0.997]
Strength (RMS ratio)	0.883	[0.819, 0.949]
RMS distance	0.0017	[0.0012, 0.0025]

Does similar shape mean equal strength? No: the GEN/RECON centered-RMS ratio is 0.88, and its bootstrap interval excludes 1 (Table 2; Appendix A.1). Shape and strength answer different questions: the correlation says that GEN and RECON place their highs and lows at the same positions; the RMS ratio says that GEN writes them about 12% more faintly.

Is the similarity confined to one scale? No. At every period the RECON and GEN patterns have the same shape (cosine 0.995–0.9998), and at every period GEN is about 10% weaker (Table 3). In one fixed 500/500 split, RECON/GEN increment cosines range from 0.9966 to 0.9998. A single full-map number could not separate shape from strength. Real has almost no structure at $P2$ and $P16$ and only weak structure at $P4$ and $P8$. Appendix C.3 gives energies, split diagnostics and pointwise intervals.

Table 3: **Scale alignment:** Primitive RECON–GEN weighted cosine and GEN/RECON RMS ratio.

Scale	Cosine	RMS ratio
$P2$	0.9998	0.902
$P4$	0.9946	0.894
$P8$	0.9970	0.885
$P16$	0.9957	0.881

4.2 Empirical pixel-head replay

Which parts of the pixel head retain the pattern? We fix 1,000 post-LayerNorm reconstruction states and vary only the AliTok token-to-pixel projection (full, weights-only, bias-only, phase-averaged), with the rest of the head unchanged. Weights-only retains the reference pattern ($r = 0.996$); bias-only (0.26), which keeps a fixed offset per position but no input-dependent term, and phase-averaged (-0.18), which gives every position the same weight vector, do not (Figure 5). The projection gives each of the 256 positions in a token's 16×16 block its own weight vector, so the pattern lies in how these weights differ. Phase averaging also alters texture, so the result localizes the pattern to the readout path under fixed hidden states; an output-matched edit attenuates all scales, not $P16$ alone (Appendix I).

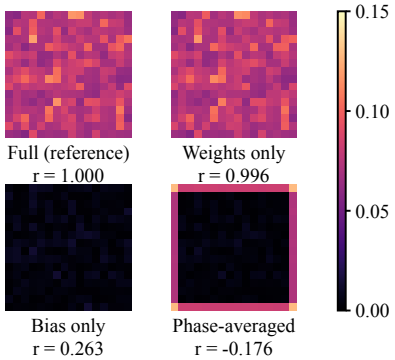


Figure 5: **Pixel-head replay:** shared raw scale; r is relative to the full head.

4.3 WHICH PERIODS A DECODER WRITES

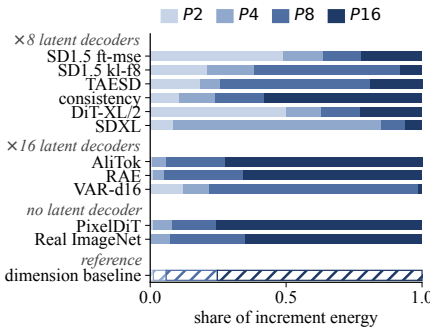


Figure 6: **Scale profile by decoder**: share of each increment in the cohort template's energy (1,000 images per cohort). Hatched: the split from degrees of freedom alone ($3/12/48/192$ of 255). All cohorts and pair statistics: Appendix E.6.

A decoder's own pattern is concentrated at periods up to its upsampling factor; above them, the input's pattern can pass through. *First, swap the decoder and keep the input: periods the decoder writes change, periods from the input do not.* Six $\times 8$ decoders decoding one latent batch differ at $P2$–$P8$ (15 pairs: median $r = 0.21/0.00/0.01$), while $P16$ keeps one shape for all six (0.72–0.93; only the amplitude varies) and matches real photographs (0.71–0.85); for the $\times 16$ decoders AliTok and RAE, $P16$ itself changes with the decoder (0.09) and matches no photographs (0.19, 0.11); to period 32, 97% of AliTok's energy up to $P32$ lies at ≤ 16, with a reproducible 2.7% residual at $P32$ (Appendix E.6). *Second, swap the input and keep the decoder: periods the decoder writes stay, periods from the input change.* Rolling sd-vae-ft-mse's latents by one latent pixel carries $P16$ along (r after the predicted 8-pixel shift 0.52–0.97 vs -0.26–0.40 unshifted); feeding the same decoder real-encoded or DiT-generated latents leaves $P2$–$P8$ intact ($0.92/0.94/0.89$) and moves $P16$ most (0.74; Appendix E.6). Below the factor, how the decoder enlarges sets which periods carry most (Figure 6): VAR-d16 is $\times 16$ but, enlarging like the SD VAEs (nearest-neighbour copying plus convolution), writes mostly at $P8$.

4.4 A POSSIBLE EXPLANATION FOR MULTISCALE PIXEL ECHO

The preceding comparisons show, period by period, what changes with the decoder and what changes with the input, but not where inside the decoder a period is written. One candidate is spatial expansion: an expansion by s_ℓ maps an input at u to output positions $s_\ell u + \phi$, $\phi \in \{0, \ldots, s_\ell - 1\}^2$, through position-dependent maps $g_{\ell,\phi}$ (Equation 11, Appendix C.3); systematic differences among the ϕ could repeat across blocks and be carried to larger periods by later expansions, or be altered by mixing. Inputs also matter: for fixed-size outputs the mean phase response is

$$F_P(D_\theta, p_z) = \mathbb{E}_{z \sim p_z}[\Phi_P(D_\theta(z))]\,, \tag{6}$$

where $\Phi_P(x)$ is Equation 3 applied to the single image $x = D_\theta(z)$. Decoder and input enter jointly; the two substitutions of Section 4.3 reveal period-dependent sensitivity to each, not a unique split of their contributions, and localization inside a decoder remains open.

5 APPLICATIONS: SOURCE COMPARISON, DETECTION AND ATTRIBUTION

Does Pixel Echo separate images from different sources? Source differences are tested on public Validation data; detection and attribution are evaluated on the held-out Test with heads and thresholds frozen beforehand (Appendix H).

5.1 FEATURE CONTRACT

Classifiers use the four increments together as one 340-dimensional feature, the centered $P16$ field in other coordinates (identical performance, Appendix H.3); this keeps one feature contract for the scale profile and the classifiers, and single increments are weaker (Table 10).

5.2 HYPOTHESIS TESTS FOR SOURCE-LEVEL DIFFERENCES

With $f \in \mathbb{R}^{340}$ the per-image feature of Section 5.1, $d = \bar{f}_A - \bar{f}_B$ compares two cohorts' arithmetic image means without standardization. With a Ledoit–Wolf covariance $\widehat{\Sigma}_{\text{Real}}$ fitted only on 1,000 Train Real images, we use

$$T_M = \sqrt{d^\top \widehat{\Sigma}_{\text{Real}}^{-1} d}, \qquad p = \frac{1 + \sum_{b=1}^{B} \mathbf{1}[T_M^{(b)} \geq T_M]}{B + 1}. \tag{7}$$

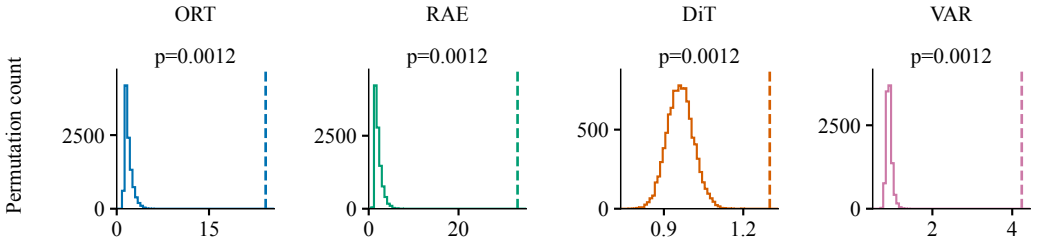


Figure 7: Real–GEN mean-distance tests. Solid histograms show 9,999 class-block swaps; dashed lines mark observations. All labels retain Holm12-adjusted $p = .0012$ after post-hoc distance selection. Axes vary independently: visual separation and histogram height do not compare effect sizes.

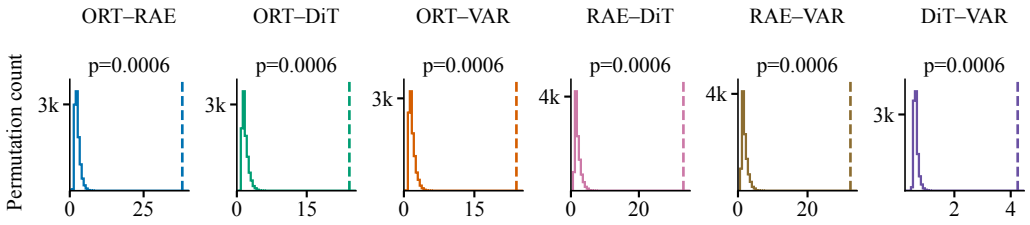


Figure 8: GEN–GEN mean-distance tests (500 images per source). Colors identify pairs; solid histograms count 9,999 swaps (k = 1,000), with observed distances dashed. All Holm6-adjusted $p = .0006$ values are resolution-limited, not equal effects. Axes vary independently.

The covariance stays fixed during $B = 9{,}999$ whole-image swaps within class blocks (exchangeability assumed within class).

Real versus generated. For ORT (ort, 2026), RAE (Zheng et al., 2025), DiT and VAR (Tian et al., 2024), each test compares 500 Real and 500 GEN images in 500 class blocks. Every observed distance exceeds all swap distances (Figure 7): raw $p = .0001$, Holm12-adjusted $p = .0012$ (three distances inspected; Appendix A). The distances are not equal: 33 for RAE and 24 for ORT, whose $\times 16$ templates are 14–22$\times$ Real's in amplitude, but 4.2 for VAR and 1.3 for DiT, whose $\times 8$ decoder shares its $P16$ with Real (Section 4.3). The tests establish mean differences, not single-image detection; their ordering anticipates which sources Section 5.3 finds hard.

Between generated sources. The six GEN pairs use the same frozen covariance and 500 class blocks (Figure 8). No swap exceeds an observation; each Holm6-adjusted $p = .0006$ is resolution-limited and does not rank source similarity. The distances do carry structure (Table 6): pairs involving ORT or RAE lie at 24–38, whereas DiT–VAR lies at 4.2, as far as VAR from Real; a linear head still separates them by direction (Section 5.3).

5.3 Detection and Source Attribution

Tables 4 and 5 report the held-out Test (500 images per cohort, evaluated once with heads and thresholds frozen beforehand). At thresholds frozen on 500 public Validation Real images the four Pixel Echo heads keep the Real false-positive rate at $1.0/0.0/0.6/1.4\%$ and detect $91.4/99.0/1.6/30.0\%$ of ORT/RAE/DiT/VAR (Max: $99.8/100/26.6/95.4\%$); four-way attribution is 95.1% for Pixel Echo and 98.6% for Max. Baselines are the pretrained detectors and frozen-representation probes of UnivFD (Ojha et al., 2023), RINE (Koutlis & Papadopoulos, 2024), SPAI (Karageorgiou et al., 2025) and NPR (Tan et al., 2024). The tables use heads run to the stopping criterion or a 50k-update budget; at an L-BFGS stationary point the NPR probe overtakes Pixel Echo alone on C2I attribution (97.5 vs 95.6%), while Pixel Echo Max leads under all three optimizers and the T2I ordering holds (Appendix H.3): the advantage lies in the combined feature, not in one representation. Attribution is closed-set; sources sharing one decoder are separated only 9–11 pp above chance (Appendix H.4). The detection columns are per-source heads, not one detector: applied jointly they flag 2.6% of

Method (Venue)	Feat. dim.	Per-source real/fake discrimination (C2I)							Source attribution		
		ORT TPR_τ	RAE TPR_τ	DiT TPR_τ	VAR TPR_τ	Real FPR	Avg. TPR_τ	Min. TPR_τ	Acc_4	Macro F1	Min Recall
Pretrained detectors											
UnivFD (CVPR 2023)	768	31.4	13.6	8.0	25.2	1.8	19.6	8.0	88.7	88.7	80.6
NPR (CVPR 2024)	512	–	–	–	–	–	–	–	**94.4**	**94.4**	**91.0**
RINE (ECCV 2024)	1024	**62.8**	17.2	24.6	**55.2**	1.6	**40.0**	**17.2**	67.9	67.5	54.4
SPAI (CVPR 2025)	1096	16.0	**19.8**	**50.2**	26.8	**0.6**	28.2	16.0	58.7	58.2	47.6
Statistical methods											
Frequency-64D[d]	64	90.2	99.4	9.6	17.8	1.6	54.3	9.6	95.3	95.3	92.4
Dzanic (NeurIPS 2020)[a]	6	5.2	0.6	15.4	23.2	1.3	11.1	0.6	55.4	53.5	27.4
Frank (ICML 2020)[a]	64	3.0	0.6	1.2	2.2	1.2	1.8	0.6	46.0	44.2	24.0
NPR-stat (CVPR 2024)[a]	111	27.0	56.8	7.2	94.0	1.4	46.2	7.2	76.3	75.8	60.2
Pixel Echo (Ours)	340	91.4	99.0	1.6	30.0	**0.8**	55.5	1.6	95.1	95.1	91.2
Pixel Echo Max (Ours)	404	**99.8**	**100.0**	**26.6**	**95.4**	1.0	**80.5**	**26.6**	**98.6**	**98.6**	**97.0**

Table 4: **C2I per-source discrimination and attribution** (held-out Test; 500 images per cohort). TPR_τ: recall of each Real-versus-generator head on its held-out cohort at its threshold, frozen on 500 public Validation Real images ($\leq$ 5 false positives); official detectors use one score and threshold. Real FPR: false-positive rate on the held-out Real 500, mean over a method's four per-source heads (the joint any-head rate is in Appendix H.6). Avg./Min.: mean/minimum over generators. Acc_4/Macro-F1/Min recall: four-way source attribution on the same held-out cohorts. Percent. Primary heads (stopping criterion or 50k budget). [a]Custom adaptation; [d]Frequency/Durall counted once; – unavailable. Bold/underline mark groupwise best/second values, including ties.

Method (Venue)	Feat. dim.	Text-to-Image GenAI Recall							Source attribution		
		GPT Image 1	GPT Image 2	Imagen 3	Nano Banana	SD1.5	SDXL	Seed. 3	Acc_7	Macro F1	Min Recall
Frozen representations + trained probes											
UnivFD (CVPR 2023)	768	**94.8**	**91.0**	**78.2**	**82.4**	**92.4**	**89.4**	**99.4**	**89.7**	**89.7**	**78.2**
NPR (CVPR 2024)	512	69.6	34.8	49.2	62.6	86.0	71.2	91.6	66.4	65.8	34.8
RINE (ECCV 2024)	1024	86.0	55.2	51.6	43.2	61.2	43.4	81.4	60.3	60.3	43.2
Statistical methods											
Frequency-64D[d]	64	83.8	**99.0**	56.0	81.2	79.0	83.8	99.0	83.1	83.0	56.0
Dzanic (NeurIPS 2020)[a]	6	69.2	68.0	24.0	22.8	60.0	13.6	61.8	45.6	43.2	13.6
Frank (ICML 2020)[a]	64	46.8	16.8	17.8	45.0	24.2	16.4	25.0	27.4	26.4	16.4
NPR-stat (CVPR 2024)[a]	111	86.8	77.0	71.0	71.2	72.6	66.0	97.6	77.5	77.4	66.0
Pixel Echo (Ours)	340	87.6	92.8	91.6	98.4	80.4	85.4	**99.8**	90.9	90.9	80.4
Pixel Echo Max (Ours)	404	**96.0**	96.8	**93.2**	**98.6**	**85.6**	**94.4**	99.6	**94.9**	**94.9**	**85.6**

Table 5: **T2I source attribution** (held-out Test; 500 images per generator). Acc_7/Macro F1/Min Recall: mean recall, macro-F1 and minimum recall over the seven generators (%); Seed.: Seedream. No matched Real controls. Superscripts and highlighting follow Table 4. Primary heads (stopping criterion or 50k budget). SPAI deferred pending preprocessing verification.

held-out Real images for Pixel Echo and 3.6% for Max (Appendix H.6). All of this assumes a preserved raster: one bicubic resampling brings attribution to chance, JPEG q90 to $81.5/40.4\%$ (Appendix I).

6 Conclusion

Pixel Echo measures phase-locked raster structure, and its exact multiscale decomposition shows which period increments a controlled comparison shares or alters. Reconstruction and generation through one decoder share aligned patterns of unequal amplitude; a decoder's own structure is concentrated at periods up to its upsampling factor, and the input's can pass through above them. On the held-out benchmark the measurement attributes images to their source and, with frequency features appended, does so more accurately than either alone; attribution is closed-set, identifies the decoder and export pipeline rather than the upstream generator, and requires a preserved pixel grid.

AI USE STATEMENT

In this work we used generative AI tools to propose and refine hypotheses, to give feedback on research methodology and experimental design, to implement the measurement, analysis and website code, to clean and reformat cohort records into the analysis tables, to interpret preliminary results, to summarise and identify relevant literature, to create and edit figures, to format references, and to draft and edit parts of the manuscript. The benchmark itself consists of images produced by generative models: they are the object of study, and the generators, checkpoints, prompts, seeds and export settings behind every cohort are listed in Appendix H. We did not use generative AI tools to develop the theoretical framework, to formulate the mathematical claims or to write their proofs, or for translation; qualitative and thematic data analysis and interview or survey design are not applicable to this work.

We have reviewed all AI-assisted work. Reported measurements are accepted only from executable artifacts and registered evidence; AI-assisted code is tested before it supports a claim; and new citation metadata is checked against Google Search, Google Scholar, arXiv, and an authoritative venue or publisher record. The authors take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Image-source forensics can support provenance analysis, but erroneous or overconfident attribution can also harm creators and users. We therefore separate passive statistical fingerprints from intentional watermarks, treat the identification demo as closed-set and sensitive to processing (source rejection with an explicit unknown outcome is not implemented), and do not present similarity as a source probability without open-set calibration. Any public release of model outputs, derived templates, or datasets will undergo a rights and privacy audit. We will also document robustness limits and potential evasion risks rather than framing non-detection as proof that an image is real.

REPRODUCIBILITY STATEMENT

The anonymous artifact links every paper-ready claim to a machine-readable evidence record. The appendix specifies the phase measurement and evaluation protocols; the supplementary repository contains tested extraction code, frozen experiment definitions, seeds, checksums, and an independently deployable project website. Results that are not backed by a registered artifact remain explicitly marked as proposed or missing evidence.

REFERENCES

ORT: Revisiting random generation order—ordinal-biased random training for efficient visual autoregressive models. `https://github.com/Xudangliatiger/ORT`, 2026.

Andrew Aitken, Christian Ledig, Lucas Theis, Jose Caballero, Zehan Wang, and Wenzhe Shi. Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize. *arXiv preprint arXiv:1707.02937*, 2017.

Keshigeyan Chandrasegaran, Ngoc-Trung Tran, and Ngai-Man Cheung. A closer look at Fourier spectrum discrepancies for CNN-generated images detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. CO-SPY: Combining semantic and pixel features to detect synthetic images by AI. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025.

Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: From generative adversarial networks to diffusion models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2023.

Chengdong Dong, Ajay Kumar, and Eryun Liu. Think twice before detecting GAN-generated fake images from their spectral domain imprints. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.

Tarik Dzanic, Karan Shah, and Freddie D. Witherden. Fourier spectrum discrepancies in deep network generated images. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020.

Yonghyun Jeong, Doyeon Kim, Youngmin Ro, Pyounggeon Kim, and Jongwon Choi. FingerprintNet: Synthesized fingerprints for generated image detection. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution AI-generated image detection by spectral learning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs leave artificial fingerprints? In *IEEE Conf. Multimedia Inf. Process. Retrieval (MIPR)*, 2019.

Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023.

Sergey Sinitsa and Ohad Fried. Deep image fingerprint: Towards low budget synthetic image detection and model lineage analysis. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2024.

Hae Jin Song and Laurent Itti. Riemannian-geometric fingerprints of generative models. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025.

Hae Jin Song, Mahyar Khayatkhoei, and Wael AbdAlmageed. ManiFPT: Defining and analyzing fingerprints of generative models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

Yusuke Sugawara, Sayaka Shiota, and Hitoshi Kiya. Checkerboard artifacts free convolutional neural networks. *APSIPA Trans. Signal Inf. Process.*, 2019.

Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

Ling Tang, Wen Shen, Zhanpeng Zhou, Yuefeng Chen, and Quanshi Zhang. Defects of convolutional decoder networks in frequency representation. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024.

P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, 1993.

Dmitry Vesnin, Dmitry Levshun, and Andrey Chechulin. Detecting autoencoder is enough to catch LDM-generated images. *arXiv preprint arXiv:2411.06441*, 2024.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023.

Pingyu Wu, Kai Zhu, Yu Liu, Longxiang Tang, Jian Yang, Yansong Peng, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Towards sequence modeling alignment between tokenizer and autoregressive model. *arXiv preprint arXiv:2506.05289*, 2025.

Hui Xu, Chi Liu, Congcong Zhu, Minghao Wang, Youyang Qu, and Longxiang Gao. Causal fingerprints of AI generative models. *arXiv preprint arXiv:2509.15406*, 2025.

Ning Yu, Larry S. Davis, and Mario Fritz. Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. PixelDiT: Pixel diffusion transformers for image generation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2026.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

Table 6: Saved permutation results. Holm adjusts the upper block over 12 Real–GEN tests and the lower block over six GEN pairs. Distances have different units; raw magnitudes and floor-valued p-values do not rank methods.

Comparison	Distance	Observed	Raw p	Holm p
Real–ORT	L_2	0.159072	0.0001	0.0012
Real–RAE	L_2	0.222815	0.0001	0.0012
Real–DiT	L_2	0.010080	0.0114	0.0142
Real–VAR	L_2	0.021617	0.0001	0.0012
Real–ORT	Cosine	0.894639	0.0001	0.0012
Real–RAE	Cosine	0.977952	0.0001	0.0012
Real–DiT	Cosine	0.550940	0.0071	0.0142
Real–VAR	Cosine	0.871431	0.0001	0.0012
Real–ORT	Mahalanobis	24.204919	0.0001	0.0012
Real–RAE	Mahalanobis	33.141342	0.0001	0.0012
Real–DiT	Mahalanobis	1.300184	0.0001	0.0012
Real–VAR	Mahalanobis	4.238770	0.0001	0.0012
ORT–RAE	Mahalanobis	38.102332	0.0001	0.0006
ORT–DiT	Mahalanobis	24.153782	0.0001	0.0006
ORT–VAR	Mahalanobis	23.951430	0.0001	0.0006
RAE–DiT	Mahalanobis	33.104553	0.0001	0.0006
RAE–VAR	Mahalanobis	32.334446	0.0001	0.0006
DiT–VAR	Mahalanobis	4.247059	0.0001	0.0006

A Exploratory Statistical Comparisons

The source-level tests of Section 5.2 use saved canonical Full340 features on the same 500 public Validation Real images and 500 images per GEN source, with equal-image arithmetic means; class matching supplies exchange blocks, not reconstruction pairs. The covariance is a Ledoit–Wolf identity-target fit on the 1,000 Train Real vectors (shrinkage 0.1274, rank 340, no extra ridge), held fixed for every comparison and permutation; it is not a mean standard error or Hotelling statistic. Each of 9,999 permutations swaps whole images independently within each class (one saved swap matrix, PCG64 seed 20260910); the plus-one correction makes the minimum raw p equal to $.0001$. Mahalanobis was chosen for display after viewing the Validation results for three distances (raw L_2, cosine, Mahalanobis); Holm adjustment is kept across all $3 \times 4 = 12$ Real–GEN tests and across the six GEN pairs, and Table 6 retains every result. Floor-valued p-values cannot rank effects or distances, and this mean-sensitive statistic need not detect every distributional difference.

A.1 AliTok shape and amplitude uncertainty

For the historical AliTok RECON and ORT GEN cohorts (1,000 images each), 2,000 independent image-bootstrap replicates per cohort (seed 20260910) pool raw phase sums and counts before count-weighted centering; the cohorts are not bootstrap-paired. The centered-$P16$ RMS distance is 0.00173 with 95% percentile interval $[0.00123, 0.00254]$; shape correlation $r = 0.99595$ $[0.99153, 0.99679]$ and GEN/RECON centered-RMS ratio 0.88284 $[0.81932, 0.94907]$ are descriptive pointwise intervals, not simultaneous bounds or cross-checkpoint replication. No scientifically justified equivalence margin was available independently of these cohorts, so equivalence is undecided. A cohort-level same-decoder RECON control now exists for DiT (Appendix E.5, template $r = 0.846$) but not for RAE or VAR.

B Extended Related Work and Terminology

Inherent fingerprints and active watermarks. We reserve *fingerprint* for an inherent statistical trace of a generative process, as in learned GAN fingerprints for discrimination and attribution (Marra et al., 2019; Yu et al., 2019), Deep Image Fingerprint's few-example residual patterns (Sinitsa & Fried, 2024) and ManiFPT's manifold-relative artifacts (Song et al., 2024; Song & Itti, 2025); these

differ from active watermarks deliberately embedded in data or outputs (Yu et al., 2021). Our traces are passive.

Periodic, spectral, and neighboring-pixel artifacts. Uneven overlap in transposed convolution and phase-dependent sub-pixel filters are established sources of checkerboard structure, and an avoidance condition exists for nonlinear CNNs (Odena et al., 2016; Aitken et al., 2017; Sugawara et al., 2019); phase imbalance is prior art. Spectral discrepancies support strong classification, sometimes from very few examples (Frank et al., 2020; Durall et al., 2020; Dzanic et al., 2020), are architecture-dependent, and can be weakened by processing or explicit mitigation (Chandrasegaran et al., 2021; Dong et al., 2022; Tang et al., 2023); SPAI learns a masked spectral representation for any-resolution detection (Karageorgiou et al., 2025). Fourier, autocorrelation and neighboring-pixel traces have been measured across modern generator families (Corvi et al., 2023; Tan et al., 2024). We inherit these observations; our distinction is to retain spatial phase, nested scale structure and decoder GEN–RECON correspondence. Earlier AliTok shared-decoder and controlled-input probes used legacy estimators and are hypothesis-generating. The canonical matched comparison is scoped as post-hoc association.

Detection, reconstruction, and attribution. Detection has exploited generic CNN features, frequency cues and diffusion reconstruction errors (Wang et al., 2020; Frank et al., 2020; Wang et al., 2023); FingerprintNet synthesizes fingerprint cues from autoencoders with several upsampling configurations (Jeong et al., 2022); Vesnin et al. (2024) transfer a real-versus-reconstruction detector to latent-diffusion outputs, and CO-SPY uses VAE reconstruction residuals with semantic evidence (Cheng et al., 2025). Attribution asks which model or family produced an image, usually in a closed set (Yu et al., 2019; Sinitsa & Fried, 2024; Song et al., 2024); causal representations have been used to decouple source traces from content (Xu et al., 2025). Our attribution is closed-set and decoder-level (Appendix H.4); a library softmax is not treated as a source probability.

Terminology. Grouping samples by coordinates modulo a sampling factor is the standard polyphase view in multirate signal processing (Vaidyanathan, 1993); we use *polyphase* descriptively for the partition by $(i \bmod P, j \bmod P)$. Pixel Echo names the structured reappearance studied here, not all generator fingerprints. Pixel-space generators without a VAE decoder, such as PixelDiT (Yu et al., 2026), are candidate counterfactuals for decoder-induced structure; the matched-control test is not reported here.

C Supplementary Measurement and Mechanism Derivation

C.1 Exact decomposition and scale strength

For common phase origin and positive $P16$ counts, define the count-weighted inner product

$$\langle A, B\rangle_C = \sum_{(a,b)\in\Omega_{16}} C(a,b)A(a,b)B(a,b). \tag{8}$$

Let $\mathcal{V}_q$ be the space of $q \times q$ templates repeated across the common 16×16 grid. The averaging operation in Equation 4 is the orthogonal projection onto $\mathcal{V}_q$ under this inner product.

Proposition 1 (exact nested decomposition). The primitive increments in Equation 5 satisfy

$$F = \Pi_1^C F + \sum_{q\in\{2,4,8,16\}} \Delta F_q, \qquad \langle \Delta F_q, \Delta F_{q'}\rangle_C = 0 \quad (q \neq q'). \tag{9}$$

The reconstruction sum telescopes. Orthogonality follows from $\mathcal{V}_1 \subset \mathcal{V}_2 \subset \mathcal{V}_4 \subset \mathcal{V}_8 \subset \mathcal{V}_{16}$ and $\Delta F_q \in \mathcal{V}_q \cap \mathcal{V}_{q/2}^{\perp_C}$. Thus the increments preserve the full response map without duplicating scale energy; they do not add information, imply statistical independence, or identify a generating mechanism. Summing S, C over modulo-q classes gives $S_q/C_q = \Pi_q^C F$ at each class, even for unequal phase counts and image dimensions not divisible by 16.

We measure each increment's strength by its count-weighted root-mean-square response:

$$\mathrm{RMS}_C(\Delta F_q) = \left(\frac{\langle \Delta F_q, \Delta F_q\rangle_C}{\sum_{a,b} C(a,b)}\right)^{1/2}. \tag{10}$$

Cohort	Boundary/ratio	$P2$	$P4$	$P8$	$P16$
ImageNet Real	valid	1.27e-05	1.56e-04	1.82e-04	4.58e-04
ImageNet Real	repl.	4.45e-05	1.68e-04	3.17e-04	4.72e-04
ImageNet Real	ratio	3.5128	1.0766	1.7392	1.0311
PixelDiT	valid	1.13e-05	2.09e-04	3.38e-04	6.15e-04
PixelDiT	repl.	5.16e-05	1.13e-04	1.72e-04	4.38e-04
PixelDiT	ratio	4.5837	0.5414	0.50855	0.71215
GPT Image 1	valid	3.46e-05	2.41e-04	3.44e-04	3.62e-04
GPT Image 1	repl.	4.03e-05	1.35e-04	2.43e-04	2.76e-04
GPT Image 1	ratio	1.1658	0.5584	0.70806	0.76182
Seedream 3.0	valid	4.14e-03	3.61e-03	7.61e-03	4.53e-04
Seedream 3.0	repl.	4.15e-03	3.59e-03	7.57e-03	4.43e-04
Seedream 3.0	ratio	1.0024	0.99434	0.99438	0.97837

Table 7: Four-cohort primitive count-weighted RMS under valid and replicate boundaries, with replicate/valid ratios. No per-image normalization; all four cohorts and periods are retained. These are full-cohort sensitivity measurements, not uncertainty estimates or canonical split reproducibility.

This reports the magnitude of the increment; it does not normalize the response map by that magnitude.

C.2 Boundary records, shape summaries, and a local approximation

Table 7 exports the four measured cohorts from the registered four-cohort audit under both boundaries with the same count-weighted nested decomposition; the valid side is reconstructed from archived $P16$ means and counts, no normalization is applied, and the cohorts are not content- or resolution-matched. Per-period reconstruction errors are at floating-point level and the weighted inner products between increments vanish where recorded (valid-boundary orthogonality was not recorded; NR is not an inferred zero); the machine-readable export and source-to-field mappings are in `REPRODUCTION_D.md` and `tables/c14_reporting_data.json`.

C.3 Canonical increment structure and scale ablation

Shared-decoder energy estimands. The shared-decoder analysis uses all 1,000 Real, 1,000 paired AliTok RECON and 1,000 ORT GEN per-image sums/counts with identical phase support. With $w = C/\sum C$ and $f = F - \Pi_1^C F$, $N_q = \|\Pi_q^C f\|_w^2$, $e_q = \|\Delta F_q\|_w^2$ and $N_q = N_{q/2} + e_q$; Table 8 reports cumulative and increment energies and the share e_q/N_q (the inherited share is $1 - e_q/N_q$). $P2$'s unit share and the monotone cumulative energy are algebraic properties, not discoveries. Intervals are 1,000 pointwise percentile image-bootstrap replicates (Real/RECON paired by source image, GEN independent) that pool raw sums/counts before division; a frozen heuristic flags increment energy at most four times the bootstrap noise as low-signal. These are exploratory, pointwise and not coverage-guaranteed; nonnegative energy intervals excluding zero are not evidence of population fingerprint existence. Table 9 separates route shape, absolute RMS and amplitude ratios at all four scales.

C2I effective-coordinate sanity and retrained ablations. On the separate K4 cohort (four source classes, 4,000 Train and 2,000 Validation images) the fixed uniform phase geometry makes centered $P16$ and the complete compact340 the same 255-dimensional subspace: deterministic within-parent contrasts with signs $(+,-,+,-)$, $(+,+,-,-)$, $(+,-,-,+)$ and amplitude $q/2$, lifted to $P16$ and orthonormal under $w = 1/256$, reconstruct the field exactly, and the two mappings agree numerically: Across 6000 cached C2I rows, the independently implemented mappings give maximum absolute differences feature 2.43×10^{-17}, logit 3.20×10^{-14}, probability 8.44×10^{-15}, margin 4.49×10^{-14}; label disagreements: 0. This replays one common-coordinate head, not two competing fits or a fresh pixel audit. All full/only/drop heads are fitted separately with the common objective, scaler and Train-only CV of Appendix H.2; removing a scale retrains on the remaining coordinates (intrinsic dimensions 3/12/48/192, leaving 252/243/207/63), so different dimensionalities preclude per-dimension specificity claims. Table 11 gives the 2,000-update fits with 1,000 paired class-stratified Validation image-bootstrap replicates on fixed predictions; drop-$P4$'s interval includes

Cohort	q	N_q	e_q	New share (%)	Flag
		1.98e-09	1.98e-09	100.000	
Real	2	[3.48e-10, 7.10e-09]	[3.48e-10, 7.10e-09]	[100.000, 100.000]	low
		3.03e-08	2.83e-08	93.468	
Real	4	[1.79e-08, 5.66e-08]	[1.45e-08, 5.33e-08]	[76.832, 98.952]	–
		1.31e-07	1.00e-07	76.851	
Real	8	[1.00e-07, 2.19e-07]	[6.86e-08, 1.84e-07]	[62.068, 88.521]	–
		3.54e-07	2.23e-07	63.066	
Real	16	[3.13e-07, 5.73e-07]	[1.84e-07, 4.13e-07]	[51.610, 76.020]	–
		9.30e-07	9.30e-07	100.000	
RECON	2	[8.13e-07, 1.06e-06]	[8.13e-07, 1.06e-06]	[100.000, 100.000]	–
		8.72e-06	7.79e-06	89.337	
RECON	4	[7.82e-06, 9.68e-06]	[6.98e-06, 8.66e-06]	[88.271, 90.279]	–
		3.94e-05	3.07e-05	77.858	
RECON	8	[3.54e-05, 4.37e-05]	[2.76e-05, 3.41e-05]	[77.351, 78.357]	–
		1.43e-04	1.04e-04	72.499	
RECON	16	[1.30e-04, 1.57e-04]	[9.45e-05, 1.14e-04]	[71.937, 73.092]	–
		7.56e-07	7.56e-07	100.000	
GEN	2	[6.65e-07, 8.51e-07]	[6.65e-07, 8.51e-07]	[100.000, 100.000]	–
		6.98e-06	6.23e-06	89.168	
GEN	4	[6.21e-06, 7.84e-06]	[5.52e-06, 7.00e-06]	[88.217, 90.040]	–
		3.10e-05	2.40e-05	77.462	
GEN	8	[2.76e-05, 3.48e-05]	[2.14e-05, 2.70e-05]	[76.956, 77.915]	–
		1.12e-04	8.06e-05	72.237	
GEN	16	[1.01e-04, 1.24e-04]	[7.28e-05, 8.93e-05]	[71.694, 72.838]	–

Table 8: Shared-decoder energy and new-subspace share, all 12 points. Entries show estimates and pointwise 95% image-bootstrap intervals; low flags unstable cumulative-energy denominators. Energy units are squared residual luminance. At $P2$, the 100% share is algebraic; positive energy intervals alone do not establish a fingerprint.

q	RECON–GEN cosine	RECON RMS	GEN RMS	GEN/RECON
	1.000	9.64e-04	8.70e-04	0.902
2	[0.994, 1.000]	[9.02e-04, 1.03e-03]	[8.16e-04, 9.22e-04]	[0.827, 0.989]
	0.995	2.79e-03	2.50e-03	0.894
4	[0.987, 0.998]	[2.64e-03, 2.94e-03]	[2.35e-03, 2.65e-03]	[0.830, 0.971]
	0.997	5.54e-03	4.90e-03	0.885
8	[0.994, 0.998]	[5.25e-03, 5.84e-03]	[4.62e-03, 5.19e-03]	[0.818, 0.958]
	0.996	1.02e-02	8.98e-03	0.881
16	[0.991, 0.997]	[9.72e-03, 1.07e-02]	[8.53e-03, 9.45e-03]	[0.823, 0.945]

Table 9: Per-increment shared-decoder alignment and absolute amplitudes, with pointwise image-bootstrap intervals. Shared shape does not imply amplitude equivalence; no route increment triggers the recorded low-signal rule.

zero, only $P2$ meets the 10^{-6} gradient diagnostic, and the same ablation under the stopping rule keeps its direction (Appendix H.3). Independently standardizing the centered256 and primitive340 storage coordinates changes the optimization geometry; any resulting difference is recipe sensitivity, not information created by the decomposition (Table 28). The T2I common-geometry comparison remains blocked because normalized phase weights vary across images. Source and review-scope map: `REPRODUCTION_A1_A2.md`.

The full representation outperforms each single-scale head, consistent with complementary cues in this K4 task; dropping $P4$ has a paired macro-F1 interval of $[-0.79, 0.10]$ percentage points, which includes zero.

Shape, signed amplitude, and residual. For centered template T and response W, $r(W,T) = \mathrm{corr}(W,T)$, $\alpha(W,T) = \langle W,T\rangle/\langle T,T\rangle$ and $R = W - \alpha T$, with $\|R\|_2/\|W\|_2 = \sqrt{1-r^2}$ for nonzero centered fields, so the residual fraction is not independent evidence; the signed amplitude α and the GEN/RECON RMS ratio are different quantities, and neither high correlation nor a small residual establishes equivalence.

Table 10: **Scale ablation:** C2I K4 Validation macro-F1 (%). Only retains one increment; Drop removes it. Heads are fitted separately; intervals are in Table 11.

Full common255: 255 dimensions, F1 95.43%				
Scale	Only dim.	Only F1	Drop dim.	Drop F1
$P2$	3	67.09	252	93.02
$P4$	12	84.17	243	95.08
$P8$	48	90.22	207	94.05
$P16$	192	80.04	63	94.45

Branch	Dim.	C	F1 (%)	ΔF1 (pp)	$\|\nabla J\|_\infty$
Full	255	10.0	95.431 [94.557, 96.360]	0.000 [0.000, 0.000]	6.59e-03
Only P2	3	0.1	67.086 [65.136, 69.176]	-28.345 [-30.439, -26.225]	1.95e-16
Only P4	12	1.0	84.165 [82.747, 85.562]	-11.266 [-12.798, -9.924]	3.13e-04
Only P8	48	10.0	90.218 [89.001, 91.514]	-5.214 [-6.377, -4.131]	2.66e-03
Only P16	192	10.0	80.035 [78.463, 81.500]	-15.396 [-17.036, -13.828]	2.60e-03
Drop P2	252	10.0	93.017 [91.930, 94.043]	-2.414 [-3.311, -1.630]	3.79e-03
Drop P4	243	1.0	95.085 [94.212, 96.068]	-0.347 [-0.792, 0.100]	6.44e-03
Drop P8	207	10.0	94.045 [93.123, 94.987]	-1.386 [-2.220, -0.643]	6.83e-03
Drop P16	63	10.0	94.454 [93.527, 95.436]	-0.978 [-1.679, -0.280]	2.92e-03

Table 11: C2I common255 full, only and retrained-drop branches. Pointwise 95% intervals use 1,000 paired class-stratified image resamples with fixed predictions, not training or selection uncertainty. Only P2 passes the 10^{-6} gradient diagnostic; the other eight do not. Full/drop fits are reused unchanged by the single-scale extension.

A local decoder–input approximation. The illustrative expansion of Section 4.4 is

$$h_{\ell+1}(s_\ell u + \phi) = g_{\ell,\phi}(h_\ell(u)), \qquad \phi \in \{0, \dots, s_\ell - 1\}^2, \tag{11}$$

which distinguishes output positions without modelling contextual decoders or asserting unequal responses; for fixed-size outputs the mean phase response is

$$F_P(D_\theta, p_z) = \mathbb{E}_{z\sim p_z}[\Phi_P(D_\theta(z))]\,, \tag{12}$$

with $\Phi_P(x)$ Equation 3 applied to $x = D_\theta(z)$. Linearizing $D_\theta(z) \approx D_\theta(\mu_z) + J_D(\mu_z)(z - \mu_z)$ and writing k_p for the row of the replicate-padded filter at output position p, the squared high-pass response has mean $\mathbb{E}[(KLD_\theta(z))_p^2] \approx (KLD_\theta(\mu_z))_p^2 + k_p L J_D(\mu_z)\Sigma_z J_D(\mu_z)^\top L^\top k_p^\top$: the Jacobian defines locally available output directions and the latent covariance weights their excitation. This assumes local linearity and finite second moments and is a motivation, not a derivation of the measured phase map or evidence that decoder weights alone identify it.

D Artifact and Reproduction Checklist

Table 12 lists every registered claim and ledger item with its evidence status; it is generated from the claim registry by `export_evidence_status_table.py`.

The public release includes immutable cohort manifests (source identifiers, hashes, provenance, model version, acquisition date, media pipeline), environment locks, checkpoint hashes, exact commands and seeds, scripts that produce every metric and figure from manifests, split files, a machine-readable claim registry, unit tests for phase indexing, crop-origin behavior, nested-period projection, template leakage, paired folds and bootstrap units, the anonymous project site, and a de-anonymization procedure that changes identity metadata without touching experimental assets.

ID	Status	Claim (registry statement, truncated)	Scope / measurement state
C14	Canonical	A registered one-pixel replicate-padding audit shows that much of the…	full cohort sensitivity only
E27	Canonical	canonical-post-hoc shared-decoder rerun, not in claims registry	ledger only (post-hoc rerun)
E29	Canonical	Under the canonical estimator, matched Real ImageNet, AliTok reconstru…	post hoc matched cohort
E30	Canonical	Replaying empirical hidden states through one AliTok pixel head preser…	same checkpoint mechanism discovery
E50	Canonical	The Figure 3 local Real–RECON example is a top-0.1% best case of a 19…	post hoc candidate distribution
E51	Canonical	Under the canonical count-weighted estimator, disjoint random halves o…	within cohort repartition stability
E52	Canonical	Across twenty canonical 1,000-image cohorts, cohorts that share a deco…	cross cohort template alignment
E53	Canonical	On the matched E29 pairs, phase-aligned averaging reveals a coherent A…	descriptive convergence illustration
E54	Canonical	The count-weighted P16 field of a single image already carries the coh…	single image template similarity
C01	Legacy est.	Under a migrated legacy normalized absolute-Laplacian estimator, indep…	migrated method development
C02	Legacy est.	Against a migrated legacy normalized P16 target, the exact AliTok outp…	migrated method development
C03	Legacy est.	For one AliTok checkpoint, model-only synthetic probes recover a stabl…	migrated method development
C04	Legacy est.	The migrated legacy AliTok analyses motivate modeling their numerical…	migrated method development
C05	Legacy est.	Against a migrated legacy normalized P16 target and with AliTok decode…	migrated method development
C06	Legacy est.	Under the migrated legacy multiscale analysis, decoder families exhibi…	migrated method development
C09	Legacy est.	In a method-development reanalysis of legacy normalized per-image feat…	method development reanalysis
C10	Legacy est.	In a method-development reanalysis of legacy normalized per-image feat…	method development reanalysis
C11	Legacy est.	Under the legacy valid-boundary, per-image-normalized estimator frozen…	registered test split
C12	Legacy est.	Under the legacy valid-boundary, per-image-normalized estimator, the r…	registered training split
C13	Legacy est.	Under the legacy valid-boundary, per-image-normalized estimator frozen…	registered test split
A01	Deferred	Detection application (not yet in the canonical claim registry	plan / proposal
C07	Deferred	An Independent Pixel Generation (IPG) architecture, in which each outp…	unknown
C08	Deferred	Under the proposed Multiscale Phase-Selective Decoding mechanism, prim…	proposal
N01	Deferred	Novelty boundary (authoring rule, not an empirical claim	authoring rule, not evidence
P01	Deferred	Localized phase-balancing intervention family (design proposal, not an	plan / proposal

Table 12: Evidence status of every registered claim and ledger item: Canonical = completed under the canonical estimator with the stated scope (post-hoc or development unless marked otherwise); Legacy est. = legacy or method-development estimator, exploratory or migration evidence only; Deferred = proposal, plan, or unknown. Generated from the claim registry; statements are truncated.

Statistical protocol. Splits are made by source image, prompt, generation seed and latent before feature extraction, and paired reconstructions stay in the same fold as their source. Image-level bootstrap intervals quantify conditional uncertainty only and do not inflate the number of checkpoints or decoder families; with few families, results are reported as replicated cases, not population generalization. Multiple period, edit or family hypotheses use the prespecified correction, and any endpoint-informed choice on a target family removes that family from confirmatory analysis. Every table row points to a versioned manifest, command, environment lock, checkpoint hash and machine-readable metric file.

E Deferred Failure Conditions and Phase-Origin Controls

The decoder-geometry hypothesis would be weakened if within-decoder GEN–RECON transfer is no stronger than matched between-decoder transfer; if parameter edits produce inconsistent directions across latent sources; or if random matched edits change the response as much as mechanism-selected edits. The IPG counterfactual would be weakened if primitive responses remain unchanged after

verifying the dedicated output unit and removing all candidate phase-selective paths. These are future tests, not completed or already preregistered results.

Three different phase-origin conditions. Let $E'_x(i,j) = E_x(i - \delta_{xi}, j - \delta_{xj})$ on a common controlled support. With one offset δ shared by every image, the phase map is relabeled by that offset; compare $F'(a,b)$ with $F(a - \delta_i, b - \delta_j)$ modulo P, not with unshifted cells. With independently randomized offsets and no registration, differently shifted templates are pooled together and can cancel. Template disappearance in this second condition is expected for a phase-locked response and is not, by itself, evidence against the mechanism. With known per-image offsets, register each contribution using $(i - \delta_{xi}, j - \delta_{xj}) \bmod P$ before pooling both sums and counts. Recovery then tests coordinate registration. These identities concern shifted response fields on controlled support; recomputing residuals after an image crop, wrap, or shift can change edge responses. A future image-level test must report its border/support policy and separate those changes from relabeling.

E.1 Matched Real–RECON–GEN atlas

The shared-decoder atlas reports the canonical primitive increments $\Delta F_2/\Delta F_4/\Delta F_8/\Delta F_{16}$ and the full $P16$ field for the matched Real ImageNet, AliTok RECON, and ORT GEN cohorts; the coarsened phase means $\Pi_q^C F$ are recorded in the same registered arrays. Scalar phase means are subtracted for display only; the underlying measurement uses raw pooled sums/counts and one final division. This comparison is post-hoc and does not provide held-out, causal, universal, or generator-specific confirmation.

The atlas is shown in the mechanism discussion as Figure 4. Its first four columns are the primitive increments tiled to 16×16, and the final column is the display-centered full-period map, which they reconstruct exactly (maximum absolute error 1.4×10^{-17}). The shared display range permits visual comparison without per-image normalization or rescaling each cohort independently. This display does not change the post-hoc evidence boundary above.

E.2 Local Real–RECON illustration and its candidate distribution

The candidates are the first 100 existing matched Real–RECON pairs; each pair's score is the mean of its four highest block correlations to the historical 1,000-image ORT GEN selection template, which excludes the displayed pairs but does not imply independence from model training. Responses use the full-raster replicate-padded absolute Laplacian before block extraction; heatmaps subtract their means without amplitude rescaling, on a common $[-0.20, 0.20]$ range in RGB$[0,1]$ luminance-response units; RGB is unenhanced. The displayed cohort maps are the matched Real and RECON pooled $P16$ fields from 1,000 images including the displayed pair, i.e. descriptive references, not independent matching targets. Reconstruction error, texture, blur and color changes remain alternative contributions to any single local example.

E.3 Block-to-cohort convergence

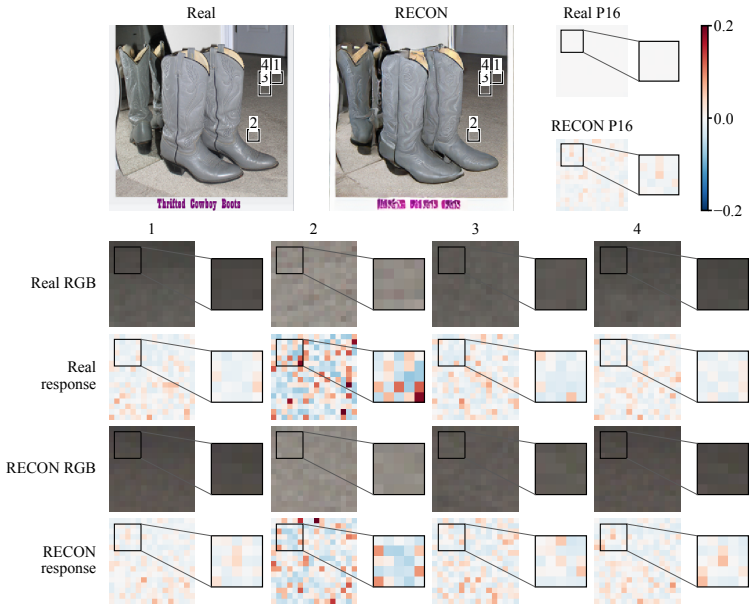
Figure 10 traces the representative pair of Table 13 from one block through the mean of its 196 interior phase-aligned blocks (a display variant; Table 13 uses the full-raster 256-block field that the classifier consumes) to nested random cohort subsets; Figure 11 gives the corresponding split-half agreement curve.

E.4 Canonical sample-size stability on the matched cohorts

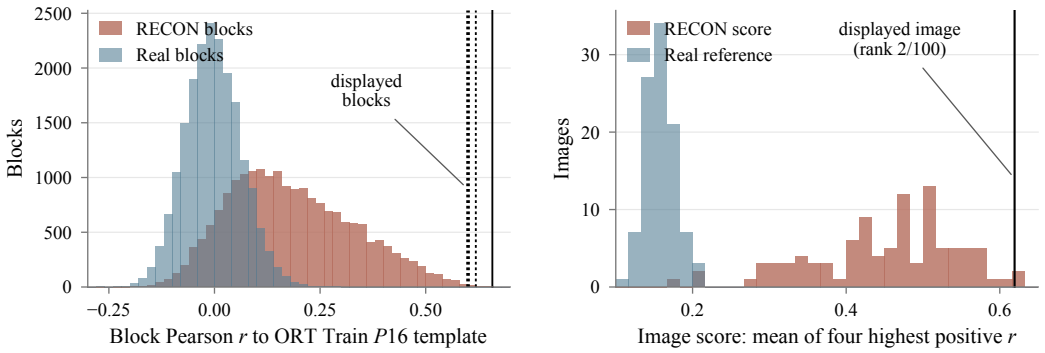
Under the canonical estimator, disjoint halves of the AliTok RECON and ORT GEN cohorts agree at median $r \geq 0.95$ from $n = 32$ images (0.964/0.961) and at $r \geq 0.99$ from $n = 128$; the corresponding 5% quantiles cross these thresholds at $n = 64$ and $n = 256$. The Real ImageNet centered field remains below $r = 0.95$ for every $n \leq 500$. Figure 11 and Table 14 give the full curves and repartition ranges.

E.5 Cross-decoder alignment matrix

Table 15 summarizes the controls of this section.



(a) Representative pair 000038, image rank 50 of 100: image score 0.469; displayed blocks $r = 0.47$; block median 0.113. Shared ± 0.20 scale (three cells of block 2 beyond it).



(a) All 19,600 interior blocks

(b) Image-level selection score, 100 pairs

(b) Candidate distribution behind Figure 3: left, Pearson r between each centered 16×16 block response and the centered ORT GEN Train $P16$ template for all 19,600 interior blocks of the first 100 matched development pairs (RECON median 0.165; Real median -0.005, maximum 0.280), with the four displayed best-case blocks of pair 000096 marked ($r = 0.66$–0.60, ranks 1–19 of 19,559, top 0.1%); right, the image-level selection score with the ranks 2, 25, 50, 75 and 95 marked.

Figure 9: **Representative local Real–RECON example and the selection distribution.** The main-text Figure 3 is the best case of this distribution; the typical RECON block resembles the cohort template only weakly (93% of the 100 pairs have image score above 0.3, 76% above 0.4), and Real blocks are centered at zero. Magnifiers show matching 5×5 regions.

With a second same-decoder control, the picture is consistent across two decoder families: reconstructions and generations that share a decoder align at or near their split-half reproducibility (AliTok $r = 0.996$, sd-vae-ft-mse $r = 0.846$), a generator pair that shares only the decoder aligns at $r = 0.864$, a pair that shares the generator but not the decoder does not ($r = 0.20$), and decoding the very same latents with six decoders gives r between -0.18 and 0.68 (median 0.21) against split-half reproducibility of 0.71–0.98, while the $P16$ increment alone stays 0.72–0.93 over the 15 decoder pairs (0.78–0.93 against ft-mse) and is shared with Real ImageNet, so no single increment is a decoder-specific band in this family; cohorts with different decoders do not align (median $r = 0.06$,

Level	Cohort	N	Mean	Median	[q25, q75]	$r > 0.5$
One 16×16 block	RECON	19,559	0.183	0.165	[0.071, 0.284]	2.1%
One 16×16 block	Real	19,538	-0.004	-0.005	[-0.049, 0.041]	0.0%
One image (its own $P16$ field)	RECON	1,000	0.653	0.690	[0.568, 0.783]	82.4%
One image (its own $P16$ field)	ORT GEN	1,000	0.630	0.686	[0.536, 0.777]	78.4%
One image (its own $P16$ field)	Real	1,000	0.011	0.009	[-0.036, 0.055]	0.0%
Cohort ($n = 1{,}000$)	RECON	1	0.995	–	–	–
Cohort ($n = 1{,}000$)	ORT GEN	1	0.999	–	–	–
Cohort ($n = 1{,}000$)	Real	1	0.160	–	–	–

Table 13: Pearson r of centered $P16$ fields to the ORT GEN Train template at three aggregation levels. Block: all interior blocks of the 100 searched development pairs. Image: each image of the matched E29 cohorts by its own full-raster count-weighted $P16$ field (Equation 3 over all 256 phase-aligned blocks, boundary rows included), which is exactly the Pixel Echo classifier input, so this row is the direct single-image evidence; the figures' 196-block interior mean is a display variant that agrees closely for RECON (e.g. $r = 0.58$ vs 0.62 for pair 000038) and moves Real by up to ± 0.2 around zero; ORT GEN images are disjoint from the Train template. Cohort: the count-weighted mean over all 1,000 images. descriptive correlations to a selection template, not significance tests.

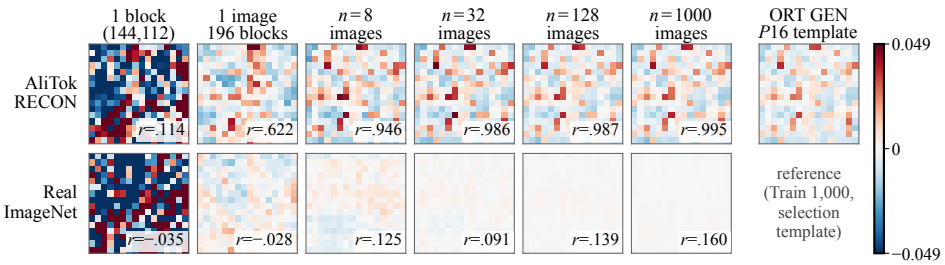
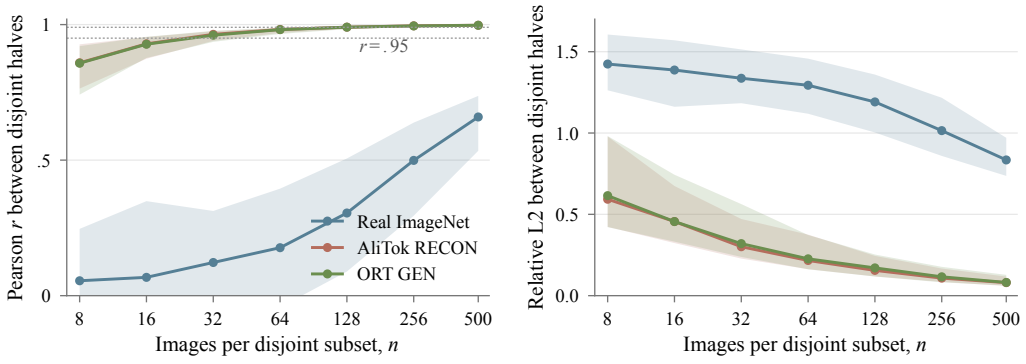


Figure 10: From one block to a cohort. Top: AliTok reconstruction; bottom: the paired Real ImageNet input. Left to right: the centered response of one 16×16 block of a representative pair (the block with the median template correlation of its image; image rank 50 of 100 in Appendix E.2), the mean of that image's 196 interior phase-aligned blocks (display variant; the full-raster 256-block field used by the classifier gives $r = 0.58$ for this RECON image), and the canonical count-weighted cohort field for nested random subsets of $n = 8$, 32, 128 and all 1,000 matched images. Right: the ORT GEN Train template used for selection. All panels are centered $P16$ fields in the same physical units with one color scale (the single-block panels exceed it and are clipped for display; values are unclipped in the data). The number in each cell is the Pearson r to the reference template. For RECON the pattern is faint in one block ($r = 0.11$), visible after one image ($r = 0.62$) and stable from a few dozen images ($r \geq 0.99$ at $n \geq 32$) with unchanged amplitude; for Real the same averaging only shrinks the field ($r \leq 0.16$ at every n). This is a descriptive illustration; the r values are correlations to a selection template, not held-out or significance evidence.

maximum 0.55 over 84 pairs). On the full raster the sd-vae-ft-mse cohorts also correlate 0.5–0.6 with Real ImageNet; restricting every field to the raster interior (margin 16, Figure 13) removes most of that shared component ($0.46 \rightarrow 0.07$) while the same-decoder alignment persists ($0.85 \rightarrow 0.78$), so the decoder-conditioned part survives the boundary control. The alignment is condition dependent: for the 512^2 COCO SD1.5 triplet the same decoder yields only $r = 0.49$ (unchanged in the interior, 0.50), where the Real inputs were Lanczos-resampled to 512^2 and the 354 rows at the exact 16/15 factor carry a phase-locked $P16$ interpolation artefact (split-half $r = 1.00$) that the reconstruction inherits; on the 572 rows without such a period the same decoder aligns RECON and GEN at $r = 0.66$. The weak COCO pair is therefore attributed primarily to a media-pipeline confound; a decoder contribution is not excluded (see the paragraph below). Table 16 groups the pairs; Figure 12 shows the full matrix and Figure 13 its interior-only counterpart.

Table 17 separates the same comparisons by primitive scale: matched ImageNet same-decoder pairs align at every scale (the COCO 512^2 pairs only at $P2$/$P4$); cohorts with different decoders show



(a) Shape agreement

(b) Relative L2 distance

Figure 11: Canonical sample-size stability for the matched E29 cohorts (1,000 images each). For each n, $B = 200$ random pairs of disjoint n-image subsets are pooled with the canonical count-weighted estimator and compared on the centered full $P16$ field; lines are medians and bands the 5–95% repartition range (not confidence intervals). Median agreement reaches $r \geq 0.99$ at $n = 128$ for AliTok RECON ($r = 0.9905$) and ORT GEN ($r = 0.9905$) but not at $n = 64$ (0.982/0.981), so the threshold lies in $(64, 128]$; the 5% quantile exceeds 0.99 from $n = 256$. The Real ImageNet field does not reach $r \geq 0.95$ for any $n \leq 500$ (median 0.659 at $n = 500$), which bounds the reproducibility of its weak centered structure and does not by itself establish that Real images carry no response. All subsets are drawn from one pool; this is within-cohort repartition evidence, not held-out confirmation, and is not comparable to the legacy per-image-normalized GPT-image-1 curve of Appendix F.

	Pearson r			Relative L_2		
n	Real	RECON	GEN	Real	RECON	GEN
8	0.0550 [-0.2253, 0.2464]	0.8590 [0.7639, 0.9272]	0.8575 [0.7416, 0.9204]	1.4241 [1.2626, 1.6057]	0.5929 [0.4230, 0.9789]	0.6150 [0.4213, 0.9816]
16	0.0676 [-0.1944, 0.3487]	0.9289 [0.8740, 0.9540]	0.9276 [0.8769, 0.9551]	1.3873 [1.1614, 1.5701]	0.4559 [0.3234, 0.6743]	0.4557 [0.3322, 0.7436]
32	0.1224 [-0.1261, 0.3124]	0.9640 [0.9401, 0.9758]	0.9609 [0.9363, 0.9770]	1.3367 [1.1834, 1.5141]	0.3009 [0.2278, 0.4714]	0.3200 [0.2383, 0.5631]
64	0.1769 [-0.0429, 0.3948]	0.9821 [0.9693, 0.9880]	0.9814 [0.9655, 0.9884]	1.2936 [1.1187, 1.4572]	0.2163 [0.1628, 0.3755]	0.2265 [0.1618, 0.3730]
128	0.3049 [0.0891, 0.5061]	0.9905 [0.9841, 0.9939]	0.9905 [0.9833, 0.9936]	1.1912 [1.0034, 1.3594]	0.1555 [0.1182, 0.2447]	0.1707 [0.1177, 0.2529]
256	0.4990 [0.2949, 0.6384]	0.9956 [0.9931, 0.9969]	0.9951 [0.9917, 0.9969]	1.0148 [0.8593, 1.2168]	0.1079 [0.0847, 0.1669]	0.1159 [0.0818, 0.1773]
500	0.6592 [0.5344, 0.7370]	0.9976 [0.9959, 0.9984]	0.9975 [0.9958, 0.9983]	0.8344 [0.7364, 0.9706]	0.0802 [0.0593, 0.1173]	0.0813 [0.0620, 0.1279]

Table 14: Canonical sample-size stability on the matched E29 cohorts. Each cell is the median and [5%, 95%] range over $B = 200$ random pairs of disjoint n-image subsets drawn from the same 1,000-image pool; halves are pooled with the canonical count-weighted estimator and compared on the centered full $P16$ field. Ranges are repartition ranges, not confidence intervals; they are not comparable to the legacy GPT-image-1 curve of Appendix F.

only weak median alignment at every scale ($|r| \leq 0.12$ on the full raster; the interior $P2$ median is -0.27 on three degrees of freedom), and Real-versus-generated pairs show only weak positive $P8/P16$ values consistent with shared content or pipeline.

Fixed latents, six decoders (three pre-registered rounds). One sealed batch of 1,000 SD1.5 latents was decoded by six decoders of the same kl-f8 latent space and measured with the canonical estimator; each round's predictions were written and committed before the weights were downloaded (`dataset_benchmark/w4_fixed_latent_decoder_20260918/`). Round 1 (original kl-f8 vs ft-mse) showed $P2$ and $P16$ aligned (0.90, 0.78) and $P4/P8$ not (-0.37, -0.12), and we predicted

Control (what is shared)	full $P16$	$P2$	$P4$	$P8$	$P16$	interior
AliTok RECON vs ORT GEN (same decoder, matched ImageNet)	0.996	–	–	–	–	–
SD-VAE RECON vs DiT GEN (same decoder, matched ImageNet)	0.846	–	–	–	–	0.784
SD1.5 same latents, vs ft-mse: original kl-f8 (same lineage)	0.302	0.899	-0.366	-0.121	0.779	0.331
ft-ema (same lineage)	0.531	0.602	0.002	0.284	0.927	0.409
kl-f8-anime2 (same lineage, other data)	0.206	-0.310	0.553	0.473	0.911	0.118
TAESD (different architecture)	0.120	-0.024	0.031	-0.160	0.820	-0.027
consistency decoder (different architecture)	0.677	0.824	0.493	0.627	0.891	0.553
TAESD 384^2 vs TAESD 512^2 (output size)	0.972	0.990	0.967	0.984	0.929	0.968
SD1.5 COCO 512^2 RECON vs GEN (same decoder, resampled Real)	0.49	–	–	–	–	0.50
rows without a phase-locked resampling period	0.66	–	–	–	–	–
Different decoders (84 pairs, median)	0.06	-0.06	0.06	0.02	0.12	-0.03

Table 15: **Decoder controls at a glance** (Pearson r of centered cohort fields; scale columns are the nested increments of Table 17, "interior" excludes a 16-px margin). Sharing a decoder aligns reconstruction and generation on matched ImageNet inputs; across six decoders of the same kl-f8 latent space the $P16$ increment is shared (0.78–0.93 against ft-mse; 0.72–0.93 over all 15 pairs) and, for one decoder, stable under a latent-crop change of output size, while $P2$/$P4$/$P8$ change with the decoder in a way that is not predictable from lineage or architecture (three pre-registered rounds, Appendix E.5); the COCO 512^2 exception is attributed primarily to a Real-side resampling artefact; different decoders show only weak median alignment of the full field.

that $P16$ is architecture-shared and $P4$/$P8$ weight-specific. Round 2 tested this on ft-ema (same lineage; prediction $P2$, $P16 > 0.6$, $P4$/$P8 < 0.3$: confirmed, $P2 = 0.60$ at the threshold) and TAESD (different architecture; prediction all four < 0.3: *failed*, $P16 = 0.82$). Round 3 tested the consistency decoder (different architecture; prediction $P16 > 0.6$, $P2$/$P4$/$P8 < 0.3$: $P16 = 0.89$ confirmed, $P2$/$P4$/$P8 = 0.82/0.49/0.63$ *failed*), kl-f8-anime2 (same lineage, other training data; prediction $P16 > 0.6$, $P2 > 0.5$, $P4$/$P8 < 0.3$: $P16 = 0.91$ confirmed, $P2 = -0.31$ and $P4$/$P8 = 0.55/0.47$ *failed*) and the output-size geometry (TAESD on the central 48×48 latent crop, 384^2, vs its 512^2 decode; prediction $P16 > 0.6$: confirmed, 0.93; $P2$–$P8 \geq 0.97$). Table 18 lists each prediction and outcome; only the geometry prediction and the $P16$ predictions held. Reading: within this latent space the $P16$ increment template is shared by all six decoders (0.78–0.93 against ft-mse, 0.72–0.93 over all 15 pairs, including both heterogeneous architectures) and, for TAESD, is stable under a latent-crop change of output size (0.93), and it also correlates with Real ImageNet (0.72–0.75 for ft-ema and TAESD), so it is not a decoder signature; $P2$/$P4$/$P8$ change with the decoder, but the direction of change is not predictable from lineage or architecture (a same-lineage fine-tune inverts $P2$, -0.31, while a heterogeneous decoder keeps it, 0.82); the full centered field still separates decoder identity consistently (same decoder 0.996 vs different 0.06). The AliTok family behaves differently: its $P16$ increment is decoder-specific (0.996 within vs ≤ 0.13 against every non-AliTok cohort, including ft-ema 0.12 and TAESD 0.09). Which single scale carries decoder specificity is therefore family-dependent; the earlier statement that $P16$ is inherited across architectures and $P4$/$P8$ are weight-specific is withdrawn.

Media-pipeline confound (COCO 512^2). The estimator responds to any phase-locked periodic structure in the raster, not only to decoder-induced structure. The E32 COCO Real inputs were Lanczos-resampled to 512^2 by rational factors; for the 354 source images at the exact $16/15$ factor the interpolation-kernel phase repeats every 16 output pixels and produces a reproducible $P16$ field in the Real cohort (split-half $r = 1.00$ for those rows versus 0.53 for the 572 rows with long or irrational phase periods) that the reconstruction inherits and that is absent from the ImageNet tracks, which crop without resizing. Restricted to the unresampled-like rows the same decoder aligns reconstruction and generation at $r = 0.66$, so the weak COCO 512^2 pair in Table 16 is attributed primarily to this Real-side resampling artefact; a decoder contribution is not excluded, because resolution, export pipeline and resampling history are not separated experimentally. Cohorts must therefore be compared under one resize policy, with resampling recorded as part of the measurement contract (Appendix I).

Boundary ring. Refitting the main-table heads on interior-only features shows that the replicate-padding boundary is not the source of the attribution signal: dropping the kernel-affected 1-px ring leaves every paired Validation interval containing zero, and dropping a 16-px margin (23% of a 256^2 raster) lowers only Real-vs-VAR, by 2.5 pp (`tables/interior_only_heads.tex` in the artifact).



Figure 12: Pairwise Pearson correlation between centered full $P16$ cohort templates for twenty-seven canonical 1,000-image cohorts (Real, AliTok-decoder, other latent decoders, pixel-native, sd-vae-ft-mse family incl. the new ImageNet reconstruction and the COCO triplets, the original-VAE SD1.5 cohort, SDXL, closed T2I). Same-decoder RECON–GEN alignment on matched inputs reaches $r = 0.996$ for the AliTok decoder and $r = 0.846$ for sd-vae-ft-mse (split-half reproducibility 0.998 and 0.86–0.92), and DiT-XL/2 GEN aligns with SD1.5 GEN through the shared sd-vae-ft-mse decoder at $r = 0.864$ across resolution and content domain, whereas the 84 pairs of cohorts with different decoders have median $r = 0.057$ (maximum 0.549). Real ImageNet correlates 0.5–0.6 with the sd-vae-ft-mse cohorts (vs 0.16 with AliTok cohorts), so part of that family's similarity may be natural-image structure preserved by a high-fidelity VAE. Two SD1.5 cohorts that share the UNet but were decoded by the original SD1.5 VAE and by sd-vae-ft-mse agree only at $r = 0.20$, and the same 1,000 latents decoded by six decoders agree with sd-vae-ft-mse at $r = 0.677$ (consistency decoder), 0.531 (EMA sibling), 0.302 (original weights), 0.206 (anime fine-tune) and 0.120 (TAESD), the mirror image of the DiT–SD1.5 pair.

E.6 Scale profile by upsampling factor

Table 19 lists the scale profile of every cohort behind Figure 6, Table 20 the per-period pair statistics, and Table 21 the upsampling operators of each decoder read from its code, configuration and weight shapes. Templates are cohort means of the canonical 340-D primitive vectors (1,000 images each;

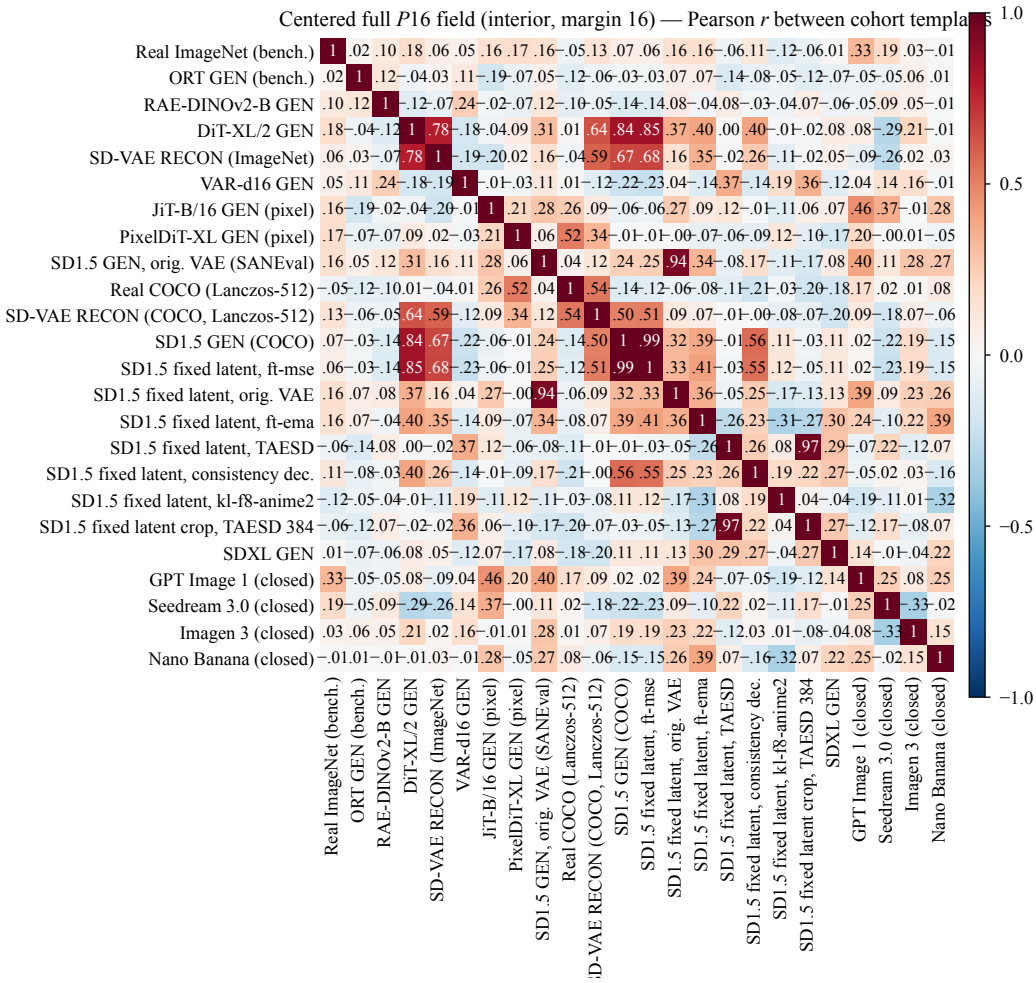


Figure 13: Interior-only variant of Figure 12: the same cohort templates computed from the interior window (margin 16) of each raster, which equals the valid-boundary response and removes the replicate-padding rows; the three E29 cohorts have no interior cache. Same-decoder alignment persists (SD-VAE RECON vs DiT-XL/2 $r = 0.784$, DiT vs SD1.5 COCO 0.838), different-decoder pairs have median -0.03 (maximum 0.40), and the Real-ImageNet correlations of the sd-vae family fall to 0.06 (from 0.46), while the weak COCO 512^2 RECON–GEN pair is unchanged ($0.49 \rightarrow 0.50$).

public and mechanism cohorts); per-cohort values and the correlation of every increment with Real ImageNet are in `experiments/scale_profile_20260920/`.

Stage-to-period rule. A $\times 2$ stage that takes the grid from a spacing of $2d$ to d pixels creates output period $2d$; a $\times s$ decoder built from such stages therefore creates period s at its first stage and period 2 at its last, and a linear projection of a token to a $p \times p$ pixel block creates every period dividing p. Its own stages therefore create no period above s; whether structure above s appears depends on the input and on any position-dependent component of the decoder (see the period-32 measurement below). In the $\times 8$ family the increment above s carries the structure of the decoder's input: for the $\times 8$ kl-f8 decoders the 16-pixel period is two latent pixels, so all six decoders of one latent batch keep the same $P16$ (0.72–0.93), it survives a change of output size (TAESD 384^2 crop vs 512^2: 0.93) and it correlates with Real ImageNet (0.71–0.85); the consistency decoder, whose only latent upsampling is a weight-free nearest-neighbour $\times 8$, has the smallest template (RMS $= 0.0003$) and the largest $P16$ share (58%); the U-Net that follows (`ConsistencyDecoderVAE`: four `ResnetDownsampleBlock2D` and four `ResnetUpsampleBlock2D`) runs at full resolution but internally downsamples by 8 and

Pair group	n	Full raster	Interior	$P16$ incr.
Same decoder, RECON–GEN, matched	5	0.846 [0.492,0.996]	0.514 [0.504,0.784]	0.736 [-0.226,0.996]
AliTok RECON–ORT GEN (ImageNet 256)	1	0.996	n/a	0.996
SD-VAE RECON–DiT GEN (ImageNet 256)	1	0.846	0.784	0.736
SD-VAE RECON–SD1.5 GEN (COCO 512)	1	0.492	0.504	-0.226
Same decoder, diff. generator (DiT–SD1.5)	1	0.864	0.838	0.776
Same latents, diff. decoder	15	0.206 [-0.182,0.677]	0.188 [-0.311,0.553]	0.839 [0.719,0.927]
ft-mse–ft-ema (siblings)	1	0.531	0.409	0.927
ft-mse–kl-f8-anime2 (same lineage)	1	0.206	0.118	0.911
ft-mse–original kl-f8	1	0.302	0.331	0.779
ft-mse–consistency dec. (other arch.)	1	0.677	0.553	0.891
ft-mse–TAESD (other architecture)	1	0.120	-0.027	0.820
Same decoder, latent crop (TAESD 384 vs 512)	1	0.972	0.968	0.929
Same generator, diff. decoder	13	0.201 [-0.191,0.689]	0.166 [-0.166,0.563]	0.460 [0.350,0.922]
Same kl-f8 arch., diff. weights	21	0.100 [-0.238,0.497]	0.107 [-0.201,0.405]	0.532 [-0.621,0.797]
Different decoder	84	0.057 [-0.162,0.549]	-0.025 [-0.230,0.404]	0.118 [-0.190,0.768]
Real vs generated (ref.)	72	0.052 [-0.602,0.721]	0.028 [-0.208,0.537]	0.153 [-0.882,0.852]
Real ImageNet–SD-VAE RECON (IN)	1	0.456	0.065	0.852
Real COCO (Lanczos-512)–its RECON	1	0.433	0.537	0.810
Split-half reproducibility (median r, 500/500, $B = 200$): full, interior, $P16$ incr. full				
AliTok RECON	1,000	0.998	n/a	0.997
ORT GEN	1,000	0.997	0.997	0.997
SD-VAE RECON (IN)	1,000	0.923	0.845	0.777
DiT-XL/2	1,000	0.855	0.802	0.626
SD1.5 GEN (COCO)	1,000	0.885	0.820	0.751
Real COCO	1,000	0.991	0.992	0.994
Real ImageNet	1,000	0.658	0.397	0.666

Table 16: Cross-decoder alignment of canonical cohort templates (Pearson r; median [min, max] over pairs; 27 cohorts of 1,000 images), on the full replicate-padded raster and on the interior window (margin 16, equal to the valid-boundary response; E29 cohorts have no interior cache). Matched-domain same-decoder pairs exist for two decoder families (AliTok; sd-vae-ft-mse via DiT-XL/2 GEN and an sd-vae-ft-mse reconstruction of the same 1,000 ImageNet inputs). The same-latents rows decode the same 1,000 SD1.5 latents with six decoders (sd-vae-ft-mse, its EMA sibling, a same-lineage anime fine-tune, the original SD1.5 VAE, and the architecturally unrelated consistency decoder and TAESD) under identical settings: every decoder swap leaves the full field far below the split-half reproducibility (0.71–0.98), the $P16$ increment is shared by all six decodings (0.74–0.93) and by Real ImageNet, and which lower bands two decoders share is pair-specific (no increment is a consistent decoder-identity band within this family; the AliTok family is decoder-specific at every scale). Decoding the centre 48×48 latent crop reproduces the TAESD 512^2 field (0.97); the same-generator rows compare SD1.5 cohorts that additionally differ in prompts and steps. The interior column shows that the weak COCO 512^2 RECON–GEN pair is not a boundary artefact, whereas the Real-ImageNet correlations of the sd-vae family and Real ImageNet's own reproducibility largely are. The COCO Real inputs were Lanczos-resampled to 512^2; 354 rows at the exact 16/15 factor carry a phase-locked $P16$ resampling artefact (split-half $r = 1.00$) that the reconstruction inherits, and the weak COCO RECON–GEN pair is attributed to this Real-side artefact (restricted to the 572 unresampled-like rows the same decoder gives RECON–GEN $r = 0.66$). Templates are cohort means; reproducibility values are repartition medians, not confidence intervals.

upsamples back in three $\times 2$ stages, which create periods 8, 4 and 2 again; its own stages create no period above 8, the entry factor.

Operator sets the periods, weights set the amplitude. AliTok's pixel head is one linear map per token reshaped into a $16 \times 16 \times 3$ block, followed only by a pointwise 1×1 convolution and a sigmoid; nothing mixes across block borders, so the cohort-mean block is a fixed pattern on the 16-pixel lattice, exactly the support of $P16$ and its divisors. Its shares ($0.6/5.4/21.4/72.5\%$) match those of a white 16×16 template ($1.2/4.7/18.8/75.3\%$, the dimension counts $3/12/48/192$; verified with the repository decomposition over 2,000 draws), its RMS is $20\times$ Real's, and ORT, which uses the same decoder, agrees at $r = 0.996$. RAE (Zheng et al., 2025), JiT (Li & He, 2025) and PixelDiT (Yu et al., 2026) use the same operator class (linear token-to-block, no spatial convolution) and have $P16$ shares of $66/56/76\%$ with RMS differing by $17\times$. VAR-d16 (Tian et al., 2024) is $\times 16$ but built

Pair group	n	$P2$	$P4$	$P8$	$P16$	Full $P16$
Same decoder, RECON–GEN, ImageNet	3/1	0.999/0.913	0.994/0.839	0.996/0.743	0.995/0.506	0.995/0.784
Same decoder, RECON–GEN, COCO 512	2/2	0.994/0.990	0.667/0.514	0.333/0.409	-0.220/-0.246	0.496/0.509
Same decoder, diff. generator	2/2	0.986/0.983	0.589/0.435	0.877/0.866	0.781/0.543	0.868/0.844
Same latents, decoder only changed	15/15	0.206/0.175	0.002/0.079	0.006/0.013	0.839/0.769	0.206/0.188
Same decoder, latent crop (TAESD)	1/1	0.990/0.987	0.967/0.993	0.984/0.980	0.929/0.816	0.972/0.968
Same generator, diff. decoder	13/13	0.415/0.417	-0.132/-0.073	-0.119/-0.014	0.460/0.052	0.201/0.166
Same kl-f8 arch., diff. weights	21/21	0.677/0.670	-0.241/-0.209	0.062/0.017	0.532/0.250	0.100/0.107
Different decoder	84/54	-0.064/-0.265	0.060/0.038	0.022/-0.012	0.118/0.025	0.057/-0.025
Real vs generated	72/44	-0.113/-0.065	-0.049/-0.007	0.136/0.096	0.153/0.184	0.052/0.028

Table 17: Median Pearson r per nested increment ($P2$, $P4$, $P8$, $P16$: 4, 16, 64, 256 values) and for the centered full $P16$ field, by pair group, full raster/interior (margin 16; E29 pairs absent from the interior). Shared-decoder pairs on ImageNet 256 align at every scale. Decoding the same latents with six decoders keeps $P16$ (0.72–0.93 over 15 pairs, also shared with Real ImageNet) while the lower increments and the full field vary by decoder pair without a consistent band: within the SD-latent family no increment is decoder-identity-specific, whereas in the AliTok family $P16$ is (0.996 within, ≤ 0.13 across). Decoding the centre latent crop reproduces the TAESD field at every scale. Different-decoder pairs show only weak median alignment at every scale ($|r| \leq 0.12$; the interior $P2$ median is -0.27 on three degrees of freedom); Real-vs-generated pairs show only a weak positive $P8$/$P16$ median. The COCO 512^2 RECON–GEN pairs align at $P2$/$P4$ only (Lanczos artefact on the Real side). $P2$ has three degrees of freedom.

Decoder (vs ft-mse)	Prediction	$P2$	$P4$	$P8$	$P16$ [split-half]	Outcome
R1 original kl-f8 (same lineage)	none (discovery)	0.90	-0.37	-0.12	0.78 [0.48, 0.80]	motivated the rule
R2 ft-ema (same lineage)	$P2$, $P16 > .6$; $P4$, $P8 < .3$	0.60	0.00	0.28	0.93 [0.66, 0.86]	held ($P2$ at threshold)
R2 TAESD (other arch.)	all four $< .3$	-0.02	0.03	-0.16	0.82 [0.57, 0.82]	**failed** $P16$
R3 consistency dec. (other arch.)	$P16 > .6$; $P2$, $P4$, $P8 < .3$	0.82	0.49	0.63	0.89 [0.61, 0.83]	**failed** $P2$, $P4$, $P8$
R3 kl-f8-anime2 (same lineage, other data)	$P16 > .6$; $P2 > .5$; $P4$, $P8 < .3$	-0.31	0.55	0.47	0.91 [0.64, 0.86]	**failed** $P2$, $P4$, $P8$
R3 TAESD 384^2 crop vs its 512^2 decode	$P16 > .6$	0.99	0.97	0.98	0.93 [0.61, 0.86]	held

Table 18: **Fixed-latent series: predictions and outcomes.** One sealed batch of 1,000 SD1.5 latents; each round's predictions were written and committed before the weights were downloaded. Cells are cohort-template Pearson r per increment against the ft-mse decode of the same latents (row 6: against the 512^2 TAESD decode); the $P16$ bracket is the 5–95% range over $B = 200$ split-half repartitions of the two cohorts, a paired-uncertainty summary rather than a confidence interval. The rule that $P16$ is architecture-shared and $P2$–$P8$ weight-specific, formed after round 1, held once and failed three times; only the statements "$P16$ is shared across decoders of this latent space" and "the template is stable under the latent-crop size change" survived. The 15 decoder pairs share inputs and are not independent replications.

from four nearest-neighbour-$\times 2$-plus-3×3-convolution stages followed by ResNet blocks, the same operator type and the same full-resolution depth (seven 3×3 convolutions) as the SD VAEs: its $P16$ is created only at the first stage and by the token lattice, then passes through twelve ResNet blocks and three further stages, and its energy sits at $P8$ (77%), the residual of the second stage; which stage's residual dominates is set by the weights (original kl-f8 $P8$, ft-mse and ft-ema $P2$, SDXL $P4$).

Input phase shift with the decoder fixed. The second swap of Section 4.3 is also performed directly on the input's phase: the sealed SD1.5 latent batch is rolled circularly by one latent pixel, (1,0), (0,1) or (1,1), and decoded by the same decoder (`dataset_benchmark/w4_fixed_latent_decoder_20260918`, extension v4; predictions frozen in `PREREGISTERED_PREDICTION_v4.md`). One latent pixel is 8 output pixels, half a $P16$ period and a whole number of periods 2, 4 and 8. Table 22 reports the interior window (margin 16; the wrapped latent row lies in the boundary ring). $P2$–$P8$ remain aligned (all four cases ≥ 0.98), but an 8-pixel shift is a symmetry of these increments, so this test does not distinguish their origin. For sd-vae-ft-mse, $P16$ moves with the input: in all three shifts the shifted template agrees better with the unshifted template rolled by the predicted 8 pixels than with it unrolled ($0.52/0.97/0.53$ against $-0.14/0.40/-0.26$), and the two values near 0.5 exceed the split-half reproducibility of the interior $P16$ template itself (0.38 at $n = 1{,}000$, odd versus even images), the reference level for correlations read in this variant. For TAESD both comparisons are

Cohort	$P2$	$P4$	$P8$	$P16$	RMS $\times 10^3$	r_{P16}(Real)
Latent decoders, $\times 8$ (SD kl-f8 latent space)						
SD1.5 ft-mse decoder	0.49	0.15	0.14	0.22	0.5	0.75
SD1.5 original kl-f8 (same latents)	0.21	0.17	0.54	0.08	0.8	0.56
TAESD (same latents)	0.18	0.07	0.55	0.19	0.6	0.72
consistency decoder (same latents)	0.11	0.13	0.18	0.58	0.3	0.73
DiT-XL/2 (sd-vae-ft-mse)	0.50	0.13	0.15	0.23	0.8	0.71
SDXL	0.08	0.76	0.09	0.06	0.6	0.65
Latent decoders, $\times 16$						
AliTok RECON	0.01	0.05	0.21	0.72	12.0	0.19
RAE-DINOv2-B	0.01	0.04	0.29	0.66	16.6	0.11
VAR-d16	0.12	0.09	0.77	0.01	2.7	0.55
No latent decoder						
Real ImageNet	0.00	0.07	0.28	0.65	0.7	–
PixelDiT-XL (pixel-space)	0.01	0.07	0.17	0.76	3.9	-0.26

Table 19: **Scale profile by upsampling factor.** Share of each increment in the cohort template's energy, total RMS, and Pearson r between the cohort's $P16$ increment and Real ImageNet's (1,000 images per cohort). All 27 cohorts and the pair statistics are in Appendix E.6.

Pair group	n	$P2$	$P4$	$P8$	$P16$	full field
Same decoder, AliTok RECON vs ORT GEN	1	1.00	0.99	1.00	1.00	1.00
Same decoder, SD-VAE RECON vs DiT GEN	1	0.92	0.94	0.89	0.74	0.85
Same latents, six $\times 8$ decoders (15 pairs)	15	0.21	0.00	0.01	0.84	0.21
AliTok RECON vs $\times 8$ decoders (7 pairs)	7	0.30	-0.09	-0.05	0.10	0.06
Different $\times 16$ decoders (AliTok, RAE, VAR; 3 pairs)	3	-0.26	0.22	0.24	0.09	0.11

Table 20: **Per-period template correlation by pair group** (median Pearson r between cohort templates, per primitive increment and for the centered full field). Same-decoder pairs agree at every period; the six $\times 8$ decoders of one latent batch agree only at $P16$; AliTok ($\times 16$) agrees with $\times 8$ decoders at no period, and different $\times 16$ decoders do not share $P16$. Source: `experiments/scale_profile_20260920`.

near the noise level (0.15, 0.14; reliability 0.37), so its $P16$ is undetermined here. On the full raster the unrolled comparison is higher ($0.59/0.75/0.47/0.82$), as expected when the boundary ring stays fixed on the output grid.

Period 32 on the $\times 16$ decoder. For $\times 16$ decoders the main-text estimator reaches only $P16$; we therefore extend it to $P32$ to test for structure beyond the upsampling factor. Running the same estimator with maximum period 32 on the matched E29 cohorts (Table 23; `experiments/p32_x16_20260921`, predictions frozen before the run) reproduces the registered $P2$–$P16$ values exactly and adds one row: AliTok writes a small residual at $P32$. It carries 2.7% of the RECON template's energy and 2.9% of GEN's (RMS 0.0020 against 0.0102 at $P16$; a structureless template would put 75% there by count), it is reproducible (split-half 0.93), it is the same for real-encoded and generated inputs (0.95), and it is not Real's (0.30, 0.32; Real's own $P32$ is only weakly reproducible, 0.60, and mostly boundary: 0.28 on the interior window). An interior window of margin 32 gives the same shares and correlations, so the residual is not a boundary artefact. The bound is therefore where the decoder's writing concentrates (97% of AliTok's energy up to $P32$ at periods ≤ 16), not a hard cutoff. Under shift-compatible boundary conditions, equivariance of the decoder to shifts of its input grid and a shift-stationary input distribution imply an s-periodic population response, so every larger increment vanishes in expectation; the measurement's replicate padding is not shift-compatible at the image border, which is why the interior window is reported. Position-dependent components break the equivariance; AliTok's transformer decoder carries positional embeddings over its 16×16 token grid, and a two-token (32-pixel) periodicity in them would appear exactly as this residual. Where it is written is not identified.

F Legacy Estimator Analyses and Migration Details

Legacy cross-decoder primitive spectrum. Before the canonical estimator was registered (2026-08-18), an exploratory cross-decoder primitive spectrum was computed with a valid-boundary, per-image-normalized estimator on $N = 1{,}000$ cohorts of ORT GEN, AliTok RECON, RAE, SD-

Table 21: Decoder upsampling operators and the measured scale spectrum. s is the total upsampling factor; "own periods" are the output periods created by the decoder's own stages (a $\times 2$ stage from spacing $2d$ to d px creates period $2d$; a linear patch→pixel projection of patch p creates every period dividing p). Shares are the count-weighted energy fractions of the P2/P4/P8/P16 increments of the cohort template; RMS is the template amplitude. A structureless template has shares 1.2/4.7/18.8/75.3 by dimension count.

Decoder	Upsampling operators (stages)	s	Own periods	Full-res convs	Cohort	Share $P2/P4/P8/P16$ (%)	RMS
AliTok pixel head (ORT, AliTok)	linear token→16×16 block (1×1 conv 1024→768 + rearrange), 1×1 conv 3→3, sigmoid	16	2,4,8,16	0	AliTok RECON / ORT GEN	0.6/5.4/21.4/72.5	0.0120
RAE decoder (ViT-XL)	linear token→16×16 block (Linear 1152→768 + unpatchify)	16	2,4,8,16	0	RAE GEN	1.0/4.0/29.4/65.6	0.0166
VAR VQGAN decoder	4×[3 ResnetBlock → nearest ×2 + conv3×3]; latent f_{hat} = bicubic sum of 10 scales	16	2,4,8,16	7	VAR GEN	12.2/9.2/77.2/1.4	0.0027
JiT-B/16 (pixel-native)	linear token→16×16 block (Linear 768→768 + unpatchify), every step	patch 16	2,4,8,16	0	JiT GEN	0.0/16.2/28.1/55.7	0.00095
PixelDiT-XL (pixel-native)	4 residual linear patch projections (1152→4096 per 16×16 patch) + per-pixel Linear 16→3	patch 16	2,4,8,16	0	PixelDiT GEN	1.0/6.9/16.5/75.5	0.0039
SD kl-f8 VAE (SD1.5 orig., ft-mse, ft-ema, anime2; SDXL same arch.)	3×[3 ResnetBlock2D → nearest ×2 + conv3×3]	8	2,4,8	7	fixed-latent ft-mse	48.9/14.6/14.3/22.2	0.00052
					fixed-latent orig. kl-f8	21.2/17.0/53.8/8.0	0.00078
					fixed-latent ft-ema	47.8/7.0/27.6/17.5	0.00058
					fixed-latent anime2	21.3/11.9/62.2/4.6	0.00115
					SDXL GEN	8.3/76.5/9.1/6.1	0.00058
TAESD	3×[3 tiny blocks → nearest ×2 → conv3×3]	8	2,4,8	4	fixed-latent TAESD	18.4/7.4/55.1/19.2	0.00062
Consistency decoder	nearest ×8 of the latent (no weights) + UNet at full resolution with 3 internal ×2 down/up stages, 2 steps	8	2,4,8	9	fixed-latent consistency	10.7/13.1/18.2/58.1	0.00030
Real ImageNet (reference)	–	–	–	–	E29 / E46	0.6/8.0/28.4/63.1	0.00059

Decoder	Latent shift	$P2$	$P4$	$P8$	$P16$ direct	$P16$ moved with input
sd-vae-ft-mse	(1,0)	1.00	1.00	0.99	-0.14	0.52
sd-vae-ft-mse	(0,1)	1.00	1.00	0.98	0.40	0.97
sd-vae-ft-mse	(1,1)	1.00	0.99	0.98	-0.26	0.53
TAESD	(1,1)	1.00	1.00	1.00	0.15	0.14
sd-vae-ft-mse	split-half reliability	1.00	0.92	0.70	0.38	–
TAESD	split-half reliability	0.99	0.96	0.95	0.37	–

Table 22: **Input phase shift, decoder fixed.** The sealed SD1.5 latent batch is rolled by one latent pixel (8 output pixels, half a $P16$ period) and decoded by the same decoder; Pearson r between the shifted and unshifted cohort templates per increment (interior window, margin 16; 1,000 images). $P2$–$P8$ are invariant to an 8-pixel shift. For $P16$, "direct" compares the templates as measured and "moved with input" compares against the unshifted template rolled by the shift: a $P16$ that enters with the input follows the second, one written on the output lattice follows the first (`PREREGISTERED_PREDICTION_v4.md`). The split-half rows give the odd/even-image reproducibility of each unshifted template per increment, the reference level for correlations read in the same variant.

VAE and Cosmos outputs (`figures/primitive_response_spectrum.pdf` in the artifact). It is not evidence for the canonical replicate-padding, raw-pooling estimator; it is retained in the artifact as a Legacy-estimator exploratory reanalysis, not a canonical-estimator figure, and canonical templates now exist for every cohort except Cosmos (Figure 4, Appendix E.5).

Legacy within-source split analyses. Two T2I cohorts were acquired under the legacy contract: 2,000 hash-verified original GPT Image 1 PNGs from the SANEval snapshot, split 1,000/500/500 by deterministic prompt group with duplicate prompts kept in one fold, and 2,000 prompt-matched

	$P2$	$P4$	$P8$	$P16$	$P32$
Energy share (%)					
Real ImageNet	0.2	2.4	8.7	19.2	69.5
AliTok RECON	0.6	5.3	20.8	70.6	2.7
ORT GEN	0.7	5.4	20.9	70.2	2.9
Degrees-of-freedom split	0.3	1.2	4.7	18.7	75.1
Split-half reliability (median r, 500/500)					
Real ImageNet	0.66	0.85	0.76	0.62	0.60
AliTok RECON	1.00	1.00	1.00	1.00	0.93
ORT GEN	1.00	1.00	1.00	1.00	0.93
Template correlation (r)					
AliTok RECON–ORT GEN	1.00	0.99	1.00	1.00	0.95
Real ImageNet–AliTok RECON	0.23	-0.06	0.25	0.14	0.30
Real ImageNet–ORT GEN	0.24	-0.13	0.24	0.15	0.32

Table 23: **Period 32 on the $\times 16$ decoder cohorts.** Canonical estimator with maximum period 32 on the matched E29 cohorts (1,000 images each, 256^2): energy share of each nested increment, split-half reliability, and template correlation per period. AliTok RECON and ORT GEN share the $\times 16$ decoder; RECON's inputs are the Real images. Predictions in `experiments/p32_x16_20260921/PREREGISTERED_PREDICTION.md`.

Seedream 3.0 rows (1,995 exact matches; five deterministic replacements marked in the manifest; all 1024×1024 PNG). In the legacy GPT analysis, median agreement between disjoint groups rises from $r = .9037$ at $n = 98$ to .9775 at $n = 500$ over 100 repartitions of one 1,000-image pool (`figures/subset_size_stability_labeled.pdf`); this curve does not establish canonical-estimator reliability or a minimum sample count. A Post-hoc migration view using the legacy valid-boundary, per-image-normalized estimator (`figures/seedream_gpt_prompt_matched_mpf.pdf`) shows GPT Image 1 peaking at $P8$ (.0073) against a stronger Seedream $P8$ response (.2276) under matched prompts. Under the legacy contract, the GPT test fold passed both legacy gates with all settings frozen beforehand (norm ratio .9941; $r = .9716$ and relative $\ell_2 = .2383$ against the training reference). The canonical held-out evaluation is the single-pass Test of Appendix H.6.

G Empirical Pixel-Head Replay

The pixel-head replay uses an already exposed, same-instance cohort. Layerwise, parameter-path and causal-intervention diagnostics are deferred; the one completed output-matched intervention is summarized in Appendix I.

G.1 Empirical hidden-state replay

Figure 5 reports empirical mechanism discovery from one AliTok checkpoint and 1,000 fixed post-LayerNorm reconstruction states. The full-head panel is an identity check. The modified projection maps 1,024 hidden channels to 16×16 RGB output positions per token. Weights-only zeroes its bias; bias-only zeroes its weights; phase averaging replaces the weights across the two spatial phase axes by their mean while retaining the original bias. The subsequent output convolution and sigmoid are unchanged in all conditions.

Pearson correlations with the empirical full-head map are 1, 0.995970, 0.263264, and -0.175752 for full, weights-only, bias-only, and phase-averaged replay, respectively. The figure uses the existing raw pooled $P16$ fields with no display centering or amplitude normalization, one $[0, 0.15]$ color range, and raster row zero at the top. These are not comparisons to an independently predicted map. The result is not held-out, cross-checkpoint, quality-preserving causal, or generator-specific evidence.

H Benchmark Protocol and Head Implementation

This appendix is the formal protocol for the benchmarks in Section 5.

Result	Estimator	Data	Heads / optimizer	Threshold	Evidence status
Shared-decoder alignment (Fig. 4, Tables 1–3)	canonical (2026-08-18)	E29 matched Real/RECON/ORT GEN cohorts, 1,000 each	none (cohort templates)	–	cohort templates; descriptive
Pixel-head replay (Fig. 5)	canonical	1,000 AliTok reconstruction states, same checkpoint	none	–	same-instance; descriptive
Scale ablation (Table 10)	canonical	public Train/Validation, C2I K4	only/drop heads, 2,000 updates	–	development
Source-level tests (Figs. 7–8)	canonical	public Validation, 500 per source	none (permutation, Ledoit–Wolf on Train Real)	–	development; Holm-adjusted
Held-out detection and attribution (Tables 4–5, Appendix H.6)	canonical	private Test, 500 per cohort, scored once	primary GD heads (stopping rule or 50k), frozen 2026-09-18	500 public Validation Real, ≤ 5 FP, frozen	held-out; single pass
Convergence and L-BFGS (Appendix H.3)	canonical	public Validation	GD to stopping rule or 50k; L-BFGS-B to 10^{-8}	–	development
Attribution level (Appendix H.4)	canonical	public/mechanism rows	primary recipe (2,000-update contrast in caption)	–	development
Benign transforms (Appendix I)	canonical	public Validation, transformed	frozen primary heads	frozen Validation thresholds	development
Legacy analyses (Appendix F)	pre-canonical	historical folds	historical	historical	superseded

Table 24: **Protocol summary.** Estimator version, data, head set and optimizer, threshold source and evidence status of each reported result. "Development" results use public Validation data; only the held-out row was scored once on the private Test with everything frozen beforehand.

		Class-to-Image GenAI Detection							Source attribution		
Method (Venue)	Feat. dim.	ORT TPR_1	RAE TPR_1	DiT TPR_1	VAR TPR_1	Avg. AUC	Avg. TPR_1	Min. TPR_1	Acc_K	Macro F1	Min Recall
Pretrained detectors											
UnivFD (CVPR 2023)	768	28.8	12.6	6.8	25.2	0.839	18.4	6.8	**89.1**	**89.0**	**79.0**
NPR (CVPR 2024)	512	–	–	–	–	–	–	–	87.2	87.0	78.8
RINE (ECCV 2024)	1024	**62.0**	17.0	21.2	**58.0**	**0.920**	**39.5**	**17.0**	57.4	56.9	41.8
SPAI (CVPR 2025)	1096	14.0	**22.2**	**48.2**	31.2	0.887	28.9	14.0	51.9	51.1	39.0
Statistical methods											
Frequency-64D[d]	64	68.2	98.6	13.4	31.6	0.946	53.0	13.4	91.7	91.7	85.6
Dzanic (NeurIPS 2020)[a]	6	3.6	1.4	16.6	20.2	0.825	10.5	1.4	55.2	53.4	27.2
Frank (ICML 2020)[a]	64	2.2	1.6	4.6	1.8	0.615	2.6	1.6	47.0	45.3	28.0
NPR-stat (CVPR 2024)[a]	111	32.4	42.0	4.6	79.2	0.920	39.6	4.6	74.3	73.6	56.2
Pixel Echo (Ours)	340	87.4	99.2	1.6	30.8	0.912	54.8	1.6	95.2	95.2	91.0
Pixel Echo Max (Ours)	404	**99.6**	**99.8**	**21.6**	**94.0**	**0.971**	**78.8**	**21.6**	**98.4**	**98.4**	**97.4**

Table 25: **C2I detection and attribution** (public Validation, $K = 4$). TPR_1: per-generator ROC TPR at Real FPR $\leq 1\%$, not one frozen threshold. Avg./Min.: source mean/minimum; AUC: AUROC; Acc_K: mean source recall. Scores are percentages except AUC. [a]Custom adaptation; Attribution probes use 2,000 updates; official binary scores are unchanged. [d]Frequency/Durall counted once; – unavailable. Bold/underline mark groupwise best/second values, including ties.

H.1 Datasets, splits, and tasks

Table 24 lists, for every reported result, the estimator version, data, head set and optimizer, threshold source and evidence status.

We assemble the C2I and T2I evaluation datasets and define their splits and tasks for this study. Within each training regime, prior methods are ordered chronologically and our methods appear last. Bold and underline identify the best and second distinct values at displayed precision, with ties sharing emphasis; feature dimensions are not ranked. The current C2I and T2I comparisons use 1,000 Train and 500 Validation images per generated source. C2I has Real controls; the current T2I export is GEN-only, and its K7 task includes GPT Image 2 (7,000 Train/3,500 Validation total). Table 26 replaces the former K6 main comparison; six-source diagnostics and ablations below remain historical

Method (Venue)	Feat. dim.	Per-source recall (%)							Source attribution		
		GPT Image 1	GPT Image 2	Imagen 3	Nano Banana	SD1.5	SDXL	Seed. 3	Acc_7	$F1_7$	Min Recall
Frozen representations + trained probes (2,000 updates)											
UnivFD (CVPR 2023)	768	**95.4**	**91.0**	**79.6**	**84.8**	**94.4**	**90.2**	**99.0**	**90.6**	**90.6**	**79.6**
NPR (CVPR 2024)	512	63.6	25.6	38.2	42.0	77.0	54.6	90.6	55.9	55.2	25.6
RINE (ECCV 2024)	1024	71.4	47.8	33.0	34.8	62.2	12.8	53.6	45.1	43.8	12.8
Statistical methods (2,000 updates)											
Frequency-64D[d]	64	73.4	95.6	38.0	48.2	71.2	66.8	92.4	69.4	68.3	38.0
Dzanic (NeurIPS 2020)[a]	6	72.6	49.8	20.2	17.6	58.4	13.6	63.0	42.2	39.5	13.6
Frank (ICML 2020)[a]	64	39.2	16.6	15.6	43.4	23.8	16.4	24.2	25.6	24.8	15.6
NPR-stat (CVPR 2024)[a]	111	80.8	61.8	46.6	58.4	72.4	51.8	98.2	67.1	66.6	46.6
Pixel Echo (Ours)	340	87.8	92.0	92.4	98.0	77.6	83.4	**99.8**	90.1	90.2	77.6
Pixel Echo Max (Ours)	404	**90.8**	**96.0**	**93.4**	**98.6**	**84.8**	**90.4**	99.6	**93.4**	**93.4**	**84.8**

Table 26: **T2I source attribution** (public Validation, $K = 7$). Acc_7/$F1_7$/MinRecall: mean recall/macro-F1/minimum recall (%); Seed.: Seedream. No matched Real controls. Superscripts and highlighting follow Table 25. SPAI deferred pending preprocessing verification.

and are not K7 results. A suitable Real-control design remains under study. The private splits are used only for the one-pass held-out Test of Appendix H.6. Train selects the declared probe. Where source rejection is defined, its single family-wise threshold is selected using Train Real; achieved Validation FPR must be reported separately. Binary ROC summaries use the descriptive Validation definition in the main-table caption, not that source-rejection threshold. Native and center-crop tracks are separate, retired `matched_256` is excluded, and missing fields are not completed results.

H.2 Head implementation and operating points

All fitted C2I/K7 linear heads (statistical features, Max, frozen-feature probes) share one implementation. For each fit with n training rows, features are standardized with that fit's mean and population standard deviation (deviations $\leq 10^{-12}$ replaced by 1), weights and bias are initialized to zero in float64, and full-batch gradient descent minimizes

$$J(W, b) = \frac{1}{n} \sum_{i=1}^{n} - \log \operatorname{softmax}(z_i W + b)_{y_i} + \frac{\|W\|_F^2}{2Cn} \tag{13}$$

with the constant step $\eta = [\frac{1}{2}\lambda_{\max}(A^\top A/n) + 1/(Cn) + 10^{-12}]^{-1}$, $A = [Z, \mathbf{1}]$, no class weights, an unpenalized intercept, and no minibatches, momentum or schedule. $C \in \{0.1, 1, 10\}$ is chosen by the mean held-fold macro-F1 over three class-stratified index-modulo-three folds (ties to the smaller C); each fold refits its scaler, and the final fit uses all Train rows. Two head sets exist: the *secondary* heads take exactly 2,000 updates (38 of the 44 comparable heads, including every Pixel Echo head, do not meet the gradient-$\infty \leq 10^{-6}$ / relative-change $\leq 10^{-6}$ diagnostic at that budget); the *primary* heads, which produce the main-text tables, run the same solver until that stopping criterion is met or a 50,000-update budget is reached, with C re-selected by the same CV (Appendix H.3). Pixel Echo uses Full340; Max concatenates it with Frequency64 into 404 features under the same protocol (no score fusion; Frequency/Durall counted once). C2I pretrained rows keep official binary scores and separately fitted attribution probes; T2I frozen-feature probes are trained, not zero-shot; statistical adapters are not official reproductions. NPR111 bins were fitted on all Train images before the folds, so its CV is fixed-cache classifier CV. Settings, traces and provenance: `REPRODUCTION_D.md`, `tables/head_protocol_records.json`, `tables/unified_head_review_data.json`.

Operating points. Each C2I detector is a Real-versus-one-generator head with its own threshold, frozen on the 500 public Validation Real images at ≤ 5 false positives and applied unchanged on the Test (Appendix H.6); binary ROC summaries on Validation use the descriptive definition in the table captions. No rule combines the four heads for an image of unknown source: the family-wise false-positive rate of applying them jointly is reported in Table 32, and their joint recall is not available from the single Test pass. Source rejection with an explicit unknown outcome is not implemented.

H.3 CONVERGENCE SENSITIVITY AND CONTROLLED REPRESENTATION COMPARISON

All 67 heads were refitted with the same objective, solver, step rule and initialization as the frozen 2,000-update heads, re-selecting C by the same Train CV, until the stopping criterion of Appendix H.2 was met or a 50,000-update budget was reached (*primary* heads; SC/B in Table 27); only 18 of 67 meet the criterion, so "primary" does not mean converged, and a second solver (L-BFGS-B to gradient-$\infty \leq 10^{-8}$) reaches stationarity on 66 of 67. The budget is not neutral: from the 2,000-update to the primary heads Validation macro-F1 moves by a median of $+0.8$ pp but up to $+17$ pp for the frozen-feature probes and Frequency-64D, and at the L-BFGS stationary points the probes gain up to a further $+13$ pp (C2I RINE $66.8 \rightarrow 79.6$, NPR $94.2 \rightarrow 97.5$; T2I RINE $60.9 \rightarrow 77.5$), whereas Pixel Echo Full340 and Max move by at most 0.3 pp; the single-head C2I source order at stationarity is Max, NPR probe, Frequency-64D, Full340, while Max stays first everywhere and the T2I order is unchanged. Table 28 shows that Full340, centered F_{16} and common255 differ by less than 0.05 pp under equalized optimization, as the linear equivalence of Appendix C.3 predicts, and that centered F_{16} concatenated with Frequency64 reproduces the Pixel Echo Max gain, so the fusion gain is not an artifact of the primitive coordinates; the Only/Drop scale ablation keeps its direction under the stopping rule (dropping $P2$ remains the largest loss, magnitudes shift by at most 1 pp; `tables/scale_ablation_converged.tex` in the artifact). The Test heads (Appendix H.6) are the primary gradient-descent heads frozen before the Test and are not refit.

H.4 ATTRIBUTION LEVEL: GENERATOR OR DECODER FAMILY

Reviewers asked whether Pixel Echo separates generators that share one decoder, i.e. where its signal sits. Table 29 answers this with the primary head recipe (stopping criterion or 50k budget) on the accepted 340-D caches (no Test, no new pixels); the 2,000-update recipe gives the same picture (caption). Within one decoder the head is barely above baseline: plain accuracy exceeds the majority-class baseline by at most 3 pp on the imbalanced tasks (9 pp on the balanced AliTok pair) (DiT GEN versus an sd-vae-ft-mse reconstruction of the same ImageNet inputs 62.9% against a 60.0% majority baseline; AliTok RECON versus ORT GEN 59.0% against 50.0%; all five heads meet the stopping criterion) and balanced accuracy exceeds $1/K$ by 9–11 pp, and adding more same-decoder cohorts collapses toward baseline, whereas swapping only the decoder of one generator (SD1.5 with its original kl-f8 VAE versus ft-mse) is separated at 86.6%. On the same rows a 13-class generator head reaches 54.9% plain accuracy (balanced 51.0%; majority baseline 8.8%) while a 7-class decoder-family head reaches 77.4% (balanced 77.7%; majority 26.5%; 82.6% on 256^2 cohorts only, which removes the resolution confound); folding the generator predictions to families recovers 79.0%, and half of the generator-level errors stay inside the family. Across 78 pairwise heads, pairs with high template correlation are the least separable (same-family pairs 61.8% versus 91.6%; Spearman -0.38 among latent-decoder pairs). The signal therefore sits predominantly at the decoder level, consistent with the alignment matrix of Appendix E.5: Pixel Echo supports decoder-family attribution more than exact generator attribution. Some upstream information remains (the same-decoder heads are not at chance), and the COCO 512^2 cohorts carry the resampling confound of Appendix E.5. PixelDiT remains nearly inseparable from the ft-mse cohorts under both recipes (13-class recall 2.6%), so this is not a budget artifact.

H.5 T2I METADATA-ONLY SHORTCUT AUDIT

In T2I the image size and export pipeline coincide with the source, so we measure how far source attribution gets without pixels. Two predictors are fixed before evaluation (`dataset_benchmark/audits/t2i_metadata_shortcut_20260921/metadata_rules.json`): the majority source in Train for each pixel size (H, W) (ties by source order, unseen sizes to the global majority), and the same with every file-visible field added (format, bit depth, colour mode, EXIF/ICC/text chunks). Table 30 reports them on the same public Validation rows as the Pixel Echo heads. Size alone reaches 43.4% Acc_7: it exploits source-associated dimensions, above all SD1.5 (512^2) and many GPT Image 2 images (variable sizes), and within the shared 1024^2 group it reduces to a tie-broken constant prediction; the file fields add nothing (every file is an 8-bit RGB PNG without EXIF, ICC or text chunks). Five sources (GPT Image 1, Imagen 3, Nano Banana, SDXL, Seedream) share 1024^2; within that block size gives 20.4% (the majority-class baseline: 500 of 2,452) and the Pixel Echo head 93.6% (94.5% primary). A metadata predictor this high shows that the benchmark contains a source shortcut for two sources; it does not show that the Pixel Echo heads use it, and a low metadata

Domain / task	Representation	C (2000/prim.)	GD@2000	GD rule/50k	Δ (pp)	L-BFGS	Updates	Status
C2I / source	Pixel Echo Full340	10/0.1	95.23	95.71	+0.48	95.61	50,000	B/L
C2I / source	Pixel Echo Max (404)	1/1	98.41	98.75	+0.35	98.60	50,000	B/L
C2I / source	Frequency-64D	10/10	91.70	95.21	+3.51	96.20	50,000	B/L
C2I / source	NPR-stat (111)	10/10	73.62	77.06	+3.44	77.95	50,000	B/L
C2I / source	Dzanic (6)	10/10	53.42	54.35	+0.93	54.24	50,000	B/L
C2I / source	Frank DCT (64)	10/10	45.30	45.30	+0.00	45.30	1,600	SC/L
C2I / source	UnivFD probe (768)	0.1/0.1	89.04	88.61	-0.42	88.61	50,000	B/L
C2I / source	NPR probe (512)	1/10	87.01	94.24	+7.23	97.45	50,000	B/L
C2I / source	RINE probe (1024)	10/10	56.86	66.84	+9.98	79.62	50,000	B/L
C2I / source	SPAI probe (1096)	1/10	51.09	57.70	+6.61	70.44	50,000	B/L*
C2I / Real vs ORT	Pixel Echo Full340	0.1/0.1	96.10	96.50	+0.40	96.60	50,000	B/L
C2I / Real vs ORT	Pixel Echo Max (404)	0.1/1	99.60	99.50	-0.10	99.50	50,000	B/L
C2I / Real vs ORT	Frequency-64D	10/10	93.48	96.00	+2.52	96.80	50,000	B/L
C2I / Real vs ORT	NPR-stat (111)	10/10	83.18	84.19	+1.01	84.59	50,000	B/L
C2I / Real vs ORT	Dzanic (6)	0.1/10	70.38	69.79	-0.60	69.79	31,700	SC/L
C2I / Real vs ORT	Frank DCT (64)	1/1	50.90	50.90	+0.00	50.90	200	SC/L
C2I / Real vs RAE	Pixel Echo Full340	0.1/0.1	98.90	99.20	+0.30	99.40	50,000	B/L
C2I / Real vs RAE	Pixel Echo Max (404)	1/10	99.80	99.90	+0.10	100.00	50,000	B/L
C2I / Real vs RAE	Frequency-64D	10/1	95.09	99.60	+4.51	99.70	50,000	B/L
C2I / Real vs RAE	NPR-stat (111)	1/10	86.60	88.90	+2.30	88.20	50,000	B/L
C2I / Real vs RAE	Dzanic (6)	10/10	72.64	73.40	+0.76	73.40	50,000	B/L
C2I / Real vs RAE	Frank DCT (64)	1/1	48.20	48.20	+0.00	48.20	200	SC/L
C2I / Real vs DiT	Pixel Echo Full340	0.1/0.1	65.84	65.84	+0.00	65.84	2,500	SC/L
C2I / Real vs DiT	Pixel Echo Max (404)	0.1/0.1	80.26	80.66	+0.40	80.66	27,500	SC/L
C2I / Real vs DiT	Frequency-64D	10/1	74.62	75.79	+1.17	76.28	50,000	B/L
C2I / Real vs DiT	NPR-stat (111)	10/10	77.13	78.35	+1.22	77.55	50,000	B/L
C2I / Real vs DiT	Dzanic (6)	10/10	72.43	72.23	-0.19	72.23	18,400	SC/L
C2I / Real vs DiT	Frank DCT (64)	1/1	51.27	51.27	+0.00	51.27	200	SC/L
C2I / Real vs VAR	Pixel Echo Full340	0.1/0.1	84.25	83.94	-0.31	83.94	8,200	SC/L
C2I / Real vs VAR	Pixel Echo Max (404)	10/0.1	97.60	97.40	-0.20	97.40	50,000	B/L
C2I / Real vs VAR	Frequency-64D	10/10	93.69	96.00	+2.31	97.20	50,000	B/L
C2I / Real vs VAR	NPR-stat (111)	10/1	95.20	97.40	+2.20	97.60	50,000	B/L
C2I / Real vs VAR	Dzanic (6)	10/10	91.48	91.89	+0.41	91.89	50,000	B/L
C2I / Real vs VAR	Frank DCT (64)	0.1/0.1	90.98	90.98	+0.00	90.98	500	SC/L
T2I / source	Pixel Echo Full340	10/0.1	90.23	91.27	+1.03	91.27	50,000	B/L
T2I / source	Pixel Echo Max (404)	1/1	93.40	94.82	+1.42	95.05	50,000	B/L
T2I / source	Frequency-64D	10/10	68.29	82.82	+14.53	85.59	50,000	B/L
T2I / source	NPR-stat (111)	10/10	66.63	77.60	+10.97	78.73	50,000	B/L
T2I / source	Dzanic (6)	10/10	39.50	42.49	+2.99	43.35	50,000	B/L
T2I / source	Frank DCT (64)	1/1	24.76	24.76	+0.00	24.76	400	SC/L
T2I / source	UnivFD probe (768)	0.1/0.1	90.61	90.23	-0.38	90.20	50,000	B/L
T2I / source	NPR probe (512)	10/10	55.17	65.19	+10.02	76.19	50,000	B/L
T2I / source	RINE probe (1024)	10/10	43.78	60.87	+17.09	77.53	50,000	B/L
T2I / source	SPAI probe (1096)	1/10	53.59	69.85	+16.26	79.69	50,000	B/L

Table 27: **Convergence sensitivity of the comparable heads** (public Validation macro-F1, %). Same objective, solver, step rule and initialization as the frozen heads; C re-selected by the same Train CV; gradient descent run until the stopping criterion (gradient-$\infty \leq 10^{-6}$ and last-100-update relative objective change $\leq 10^{-6}$) was met or the 50,000-update budget was reached (*primary* heads). F1 GD@2000 is the frozen 2,000-update head. F1 L-BFGS: the same objective, scaler, C and initialization optimized with L-BFGS-B to gradient-$\infty \leq 10^{-8}$ as a second solver. Status: SC = stopping criterion met, B = 50k budget reached; L = L-BFGS reached 10^{-8}, L* = L-BFGS stopped by zero relative progress. 18 of 67 final GD heads (comparable, scale-ablation and controlled) meet the stopping criterion before the budget. Owner-reported development analysis on opened Validation; not held-out evidence.

score does not rule out pipeline traces that have already entered the pixels (resampling, compression; Appendix I).

Representation (C2I source, $K = 4$)	dim.	F1 @2000	F1 rule/50k	Updates
Pixel Echo Full340	340	95.13	95.71	50,000
Centered F16 (256)	256	95.39	95.71	50,000
Common255	255	95.33	95.67	50,000
Centered F16 + Frequency64 (320)	320	98.31	98.65	50,000
Pixel Echo Max (404)	404	98.41	98.75	50,000

Table 28: **Controlled representation comparison under the stopping rule or 50k budget** (C2I source attribution, public Validation macro-F1, %). Full340 (the primitive representation), centered F_{16} (256) and common255 are linear re-coordinatizations of the same information; under the stopping rule or 50k budget they agree within 0.05 pp, and centered F_{16} concatenated with Frequency64 reproduces the Pixel Echo Max gain. F1 @2000 uses the same C at the 2,000-update snapshot.

(a) Same decoder, different upstream: separability of the Pixel Echo head

Task	n_{val}/class	Acc [95%]	Bal.	Macro-F1	Maj.	$1/K$
DiT GEN vs SD-VAE RECON (ft-mse, both ImageNet 256^2) (SC)	500/333	62.9 [59.7, 65.8]	59.4	59.4	60.0	50.0
DiT GEN / SD-VAE RECON / SD1.5 GEN (ft-mse; 256^2/256^2/512^2) (SC)	500/333/333	44.5 [41.7, 47.3]	42.5	42.8	42.9	33.3
+ SD1.5 RECON (four ft-mse cohorts) (SC)	500/333/333/333	36.8 [34.3, 39.2]	35.9	36.2	33.4	25.0
AliTok RECON vs ORT GEN (E29, same decoder, 256^2) (SC)	333/333	59.0 [55.4, 62.6]	59.0	58.9	50.0	50.0
SD1.5 GEN: original kl-f8 VAE vs ft-mse (decoder swap) (SC)	500/333	86.6 [84.3, 89.0]	85.4	85.8	60.0	50.0

(b) Generator-level versus decoder-family-level labels on the same rows

Row set	n_{val}	generator head: Acc [95%] / Bal.	K; Maj.	family head: Acc [95%] / Bal.	K; Maj.
all 13 known-decoder cohorts (256^2–1024^2) (B/B)	5,665	54.8 [54.0, 55.8] / 51.0	13; 8.8	77.4 [76.4, 78.4] / 77.7; folded 79.0	7; 26.5
256^2 cohorts only (B/B)	3,999	57.6 [56.4, 58.8] / 54.5	9; 12.5	82.6 [81.5, 83.6] / 84.6; folded –	5; 29.2

Table 29: **Attribution level.** Pixel Echo Full340 heads with the primary recipe (Train-only standardization, C by within-class Train CV, gradient descent to the stopping criterion or the 50k budget; SC/B marks which) on opened public/mechanism rows. Cohorts contribute 500 (official split) or 333 (mod-3 split) Validation rows, so plain accuracy (Acc, all rows; 95% class-stratified image-bootstrap interval, $B = 2{,}000$) and balanced accuracy (Bal., mean per-class recall) differ; Maj. is the majority-class baseline for plain accuracy and $1/K$ the balanced-chance level. (a) Cohorts that share one decoder are separated only 9–11 pp above $1/K$ in balanced accuracy (at most 3 pp above the majority baseline on the imbalanced tasks), whereas swapping the decoder of one generator is easily separable. (b) Acc [95%] / Bal. with K and the majority baseline; decoder-family labels are recovered far better than generator labels, and folding the generator predictions to families (*folded*) recovers most of the family-head accuracy. (c) Over 78 pairwise heads, same-family pairs average 61.8% accuracy versus 91.6% for different-family pairs; among latent-decoder pairs, Spearman correlation between the W4 template correlation and separability is -0.38. Under the 2,000-update recipe the same heads give 59.0/62.9/86.4% for the AliTok, DiT and decoder-swap rows of (a) and 53.9/74.2% for the generator/family heads of (b); the conclusions do not change.

H.6 Held-out Test

The estimator, features and both head sets were frozen before any Test feature was computed (2,000-update heads on 2026-09-16; primary heads and their thresholds, chosen on the 500 public Validation Real images at ≤ 5 false positives, before the run), and the Test was executed once (2026-09-18) on the private splits: Real 500, 500 images each of ORT, RAE, DiT-XL/2, VAR-d16, JiT and PixelDiT, an additional Real cohort of 500, and seven T2I generators of 500. Table 31 lists every detector's Real false-positive rate and per-cohort recall with Wilson intervals at the frozen thresholds: the four Pixel Echo detectors produce 5/0/3/7 false positives on the Real 500 ($1.0/0.0/0.6/1.4\%$) and 2/0/1/4 on the additional cohort, while NPR-stat, Pixel Echo Max and Frequency-64D exceed 1% on at least one head (up to 4.6, 3.0 and 2.6%); recall is $91.4/99.0/1.6/30.0\%$ for Pixel Echo and $99.8/100.0/26.6/95.4\%$ for Max on ORT/RAE/DiT/VAR, and DiT remains the hard source. Under an any-head-fires rule (Table 32) Pixel Echo flags 2.6% of the Real 500 (1.2% of the additional cohort) and Max 3.6% (6.4%), against 4.2–5.6% (5.2–7.8%) for the statistical baselines; joint recall is not derivable because each detector scored only its own target cohort. Four-way C2I attribution (main-text Table 4) is 98.6% for Max, 95.3% for Frequency-64D, 95.1% for Pixel Echo and 94.4%

Predictor	Acc_7 (%)	Macro-F1 (%)	Min recall (%)	unseen key (%)
Size only (H, W)	43.4	35.4	0.0	0.6
Size + file metadata	43.4	35.4	0.0	0.6
Pixel Echo (secondary, 2,000 updates)	90.1	90.2	77.6	–
Pixel Echo Max (secondary, 2,000 updates)	93.4	93.4	84.8	–
Pixel Echo (primary, stopping rule / 50k)	91.2	91.3	79.4	–
Pixel Echo Max (primary, stopping rule / 50k)	94.8	94.8	87.2	–
Within the 1024×1024 block (2,452 rows, 5 sources): size only 20.4; Pixel Echo 93.6 (secondary) / 94.5 (primary)				

(H, W)	GPT-1	GPT-2	Imagen-3	Nano	SD1.5	SDXL	Seedream
(1024,1024)	919/452	0/0	1000/500	1000/500	0/0	1000/500	1000/500
(512,512)	0/0	0/0	0/0	0/0	1000/500	0/0	0/0
(1122,1402)	0/0	236/135	0/0	0/0	0/0	0/0	0/0
(1254,1254)	0/0	147/69	0/0	0/0	0/0	0/0	0/0
(1181,1331)	0/0	91/45	0/0	0/0	0/0	0/0	0/0
(1024,1536)	63/40	17/6	0/0	0/0	0/0	0/0	0/0
other sizes (< 50 Train rows each)	527 Train rows over 127 sizes (GPT-1 / GPT-2 only)						

Table 30: **T2I metadata-only shortcut audit** (public Validation, $K = 7$, 500 rows per source). Top: attribution from file-visible metadata alone against the Pixel Echo heads on the same rows; "unseen key" is the share of Validation rows whose size did not occur in Train. Bottom: source by pixel size (Train/Validation counts).

Method	Heads	FPR Real	FPR New-Real	TPR ORT	TPR RAE	TPR DiT	TPR VAR
Frequency-64D[d]	primary	1.8/1.6/2.4/0.4	2.2/1.8/2.6/0.2	90.2 [87.3,92.5]	99.4 [98.3,99.8]	9.6 [7.3,12.5]	17.8 [14.7,21.4]
	secondary	1.8/2.2/3.0/1.2	1.6/1.6/2.2/1.4	69.2 [65.0,73.1]	95.4 [93.2,96.9]	12.4 [9.8,15.6]	31.2 [27.3,35.4]
Dzanic	primary	1.0/1.6/1.2/1.2	2.2/0.4/2.2/1.8	5.2 [3.6,7.5]	0.6 [0.2,1.7]	15.4 [12.5,18.8]	23.2 [19.7,27.1]
	secondary	1.6/1.8/1.4/1.4	1.8/0.6/2.2/1.8	4.4 [2.9,6.6]	1.2 [0.6,2.6]	15.4 [12.5,18.8]	23.8 [20.3,27.7]
Frank	primary	2.6/0.8/1.4/0.0	2.4/2.0/2.0/0.2	3.0 [1.8,4.9]	0.6 [0.2,1.7]	1.2 [0.6,2.6]	2.2 [1.2,3.9]
	secondary	2.6/0.8/1.4/0.0	2.4/2.0/2.0/0.2	3.0 [1.8,4.9]	0.6 [0.2,1.7]	1.2 [0.6,2.6]	2.2 [1.2,3.9]
NPR-stat	primary	0.0/1.0/0.4/4.2	0.6/2.6/0.2/4.6	27.0 [23.3,31.1]	56.8 [52.4,61.1]	7.2 [5.2,9.8]	94.0 [91.6,95.8]
	secondary	0.0/0.8/0.2/2.4	0.6/1.2/0.2/2.8	32.0 [28.1,36.2]	38.6 [34.4,42.9]	3.2 [2.0,5.1]	78.8 [75.0,82.2]
Pixel Echo	primary	1.0/0.0/0.6/1.4	0.4/0.0/0.2/0.8	91.4 [88.6,93.6]	99.0 [97.7,99.6]	1.6 [0.8,3.1]	30.0 [26.1,34.2]
	secondary	0.8/0.2/0.6/1.4	0.6/0.0/0.2/0.8	89.4 [86.4,91.8]	98.2 [96.6,99.1]	1.6 [0.8,3.1]	29.6 [25.8,33.7]
Pixel Echo Max	primary	0.6/1.0/1.4/1.2	0.6/3.0/1.6/2.2	99.8 [98.9,100.0]	100.0 [99.2,100.0]	26.6 [22.9,30.6]	95.4 [93.2,96.9]
	secondary	0.6/0.8/0.8/0.8	0.6/1.8/1.2/2.0	99.8 [98.9,100.0]	100.0 [99.2,100.0]	22.2 [18.8,26.0]	95.0 [92.7,96.6]

Official detectors (one score, one threshold); TPR on all six held-out cohorts

Method		Real FPR	New-Real FPR	ORT / RAE	DiT / VAR	JiT / PixelDiT
UnivFD	official	1.8	1.2	31.4 / 13.6	8.0 / 25.2	50.0 / 11.8
RINE	official	1.6	1.8	62.8 / 17.2	24.6 / 55.2	67.6 / 38.6
SPAI	official	0.6	0.4	16.0 / 19.8	50.2 / 26.8	29.0 / 5.0

Table 31: **Held-out Test (C2I).** Primary = heads run to the stopping criterion or the 50k budget, secondary = 2,000-update heads. Each fitted head is Real versus one generator; its threshold was frozen on 500 public Validation Real images (≤ 5 false positives) and applied unchanged. FPR Real: false-positive rates of the four heads (ORT/RAE/DiT/VAR heads) on the private Real 500; New-Real FPR: on an additional Real cohort of 500 images materialized after freezing. TPR: recall of each head on its own held-out generated cohort (500 images) at that threshold, with Wilson 95% intervals in brackets; official detectors report one FPR and TPR on all six cohorts. Percent.

for the NPR probe, the last three within each other's Wilson intervals; T2I seven-way attribution (Table 33) is 90.9% for Pixel Echo (minimum recall 80.4% on SD1.5) and 94.9% for Max, against 89.7% for the UnivFD probe, 83.1% for Frequency-64D and 70.3% for the SPAI probe. Primary and 2,000-update heads differ marginally for Pixel Echo and more for the baselines; empirical calibration on 500 images is not a population guarantee.

I Limitations and Scope

This section collects every boundary of the evidence in one place.

Representation	Heads	Real 500: joint FPR [95%] (per-head ORT/RAE/DiT/VAR)		second Real 500: joint FPR [95%] (per-head ORT/RAE/DiT/VAR)	
Pixel Echo	primary	13/500 = 2.6 [1.5,4.4]	(1.0/0.0/0.6/1.4)	6/500 = 1.2 [0.6,2.6]	(0.4/0.0/0.2/0.8)
	secondary	14/500 = 2.8 [1.7,4.6]	(0.8/0.2/0.6/1.4)	8/500 = 1.6 [0.8,3.1]	(0.6/0.0/0.2/0.8)
Pixel Echo Max	primary	18/500 = 3.6 [2.3,5.6]	(0.6/1.0/1.4/1.2)	32/500 = 6.4 [4.6,8.9]	(0.6/3.0/1.6/2.2)
	secondary	12/500 = 2.4 [1.4,4.1]	(0.6/0.8/0.8/0.8)	25/500 = 5.0 [3.4,7.3]	(0.6/1.8/1.2/2.0)
Frequency-64D^{d}	primary	23/500 = 4.6 [3.1,6.8]	(1.8/1.6/2.4/0.4)	26/500 = 5.2 [3.6,7.5]	(2.2/1.8/2.6/0.2)
	secondary	29/500 = 5.8 [4.1,8.2]	(1.8/2.2/3.0/1.2)	26/500 = 5.2 [3.6,7.5]	(1.6/1.6/2.2/1.4)
NPR-stat	primary	28/500 = 5.6 [3.9,8.0]	(0.0/1.0/0.4/4.2)	39/500 = 7.8 [5.8,10.5]	(0.6/2.6/0.2/4.6)
	secondary	17/500 = 3.4 [2.1,5.4]	(0.0/0.8/0.2/2.4)	24/500 = 4.8 [3.2,7.0]	(0.6/1.2/0.2/2.8)
Dzanic	primary	21/500 = 4.2 [2.8,6.3]	(1.0/1.6/1.2/1.2)	27/500 = 5.4 [3.7,7.7]	(2.2/0.4/2.2/1.8)
	secondary	21/500 = 4.2 [2.8,6.3]	(1.6/1.8/1.4/1.4)	27/500 = 5.4 [3.7,7.7]	(1.8/0.6/2.2/1.8)
Frank	primary	22/500 = 4.4 [2.9,6.6]	(2.6/0.8/1.4/0.0)	28/500 = 5.6 [3.9,8.0]	(2.4/2.0/2.0/0.2)
	secondary	22/500 = 4.4 [2.9,6.6]	(2.6/0.8/1.4/0.0)	28/500 = 5.6 [3.9,8.0]	(2.4/2.0/2.0/0.2)

Table 32: **Family-wise Real false-positive rate on the held-out Test.** Fraction of Real images on which at least one of the four Real-versus-{ORT, RAE, DiT, VAR} detectors fires at its own frozen threshold, computed from the per-image decisions of the single Test pass (no new scores); Wilson 95% intervals; per-head rates in parentheses. The union rate is bounded by the sum of the four per-head rates. Joint recall under the same any-head rule is not available: in the frozen protocol each detector was applied only to its own target cohort and the two Real cohorts, so the other three detectors have no decisions on that generated cohort.

Method	Heads	Acc$_7$ [95%]	Macro-F1	Min recall	GPT 1	GPT 2	Imagen 3	Nano B.	SD1.5	SDXL	Seed. 3
Frozen representations + trained probes											
UnivFD probe	primary	89.7 [88.6,90.6]	89.7	78.2	94.8	91.0	78.2	82.4	92.4	89.4	99.4
	secondary	90.2 [89.2,91.2]	90.2	78.4	96.2	91.4	78.4	83.0	93.0	90.4	99.2
NPR probe	primary	66.4 [64.8,68.0]	65.8	34.8	69.6	34.8	49.2	62.6	86.0	71.2	91.6
	secondary	56.1 [54.5,57.8]	55.0	24.4	69.8	24.4	39.8	36.4	82.4	52.2	87.8
RINE probe	primary	60.3 [58.7,61.9]	60.3	43.2	86.0	55.2	51.6	43.2	61.2	43.4	81.4
	secondary	47.2 [45.5,48.8]	45.8	14.4	78.4	46.6	37.6	35.2	59.4	14.4	58.6
SPAI probe	primary	70.3 [68.8,71.8]	69.8	41.4	85.4	41.4	54.2	62.8	88.2	73.2	86.8
	secondary	57.6 [56.0,59.3]	55.7	11.4	76.2	11.4	43.4	62.0	85.4	57.6	67.4
Statistical methods											
Frequency-64D^{d}	primary	83.1 [81.8,84.3]	83.0	56.0	83.8	99.0	56.0	81.2	79.0	83.8	99.0
	secondary	71.0 [69.5,72.5]	70.1	39.2	77.2	97.2	39.2	52.8	70.0	67.4	93.2
Dzanic	primary	45.6 [44.0,47.3]	43.2	13.6	69.2	68.0	24.0	22.8	60.0	13.6	61.8
	secondary	42.9 [41.3,44.6]	40.3	13.2	73.4	50.4	21.2	17.8	62.2	13.2	62.2
Frank	primary	27.4 [26.0,28.9]	26.4	16.4	46.8	16.8	17.8	45.0	24.2	16.4	25.0
	secondary	27.4 [26.0,28.9]	26.4	16.4	46.8	16.8	17.8	45.0	24.2	16.4	25.0
NPR-stat	primary	77.5 [76.0,78.8]	77.4	66.0	86.8	77.0	71.0	71.2	72.6	66.0	97.6
	secondary	69.1 [67.5,70.6]	68.5	50.6	84.0	60.6	50.6	63.0	75.6	50.8	98.8
Pixel Echo	primary	90.9 [89.9,91.8]	90.9	80.4	87.6	92.8	91.6	98.4	80.4	85.4	99.8
	secondary	89.4 [88.3,90.3]	89.5	76.2	85.0	93.6	90.0	98.0	76.2	83.0	99.8
Pixel Echo Max	primary	94.9 [94.1,95.6]	94.9	85.6	96.0	96.8	93.2	98.6	85.6	94.4	99.6
	secondary	93.2 [92.3,94.0]	93.3	85.0	91.8	97.2	90.2	98.4	85.0	90.2	99.8

Table 33: **Held-out Test (T2I source attribution).** Seven-generator attribution on the held-out private cohorts (500 images per generator, GEN-only: no Real cohort and no detection threshold) with the frozen heads (primary = stopping criterion or 50k budget, secondary = 2,000-update). Acc$_7$: mean per-generator recall with Wilson 95% interval; Macro-F1 and minimum recall over the seven generators; per-generator recall in percent. Frozen-feature probes use the official released representations with heads trained on public Train; SPAI's preprocessing remains unverified and its row is exploratory.

The statistic is tied to the raster phase. Pixel Echo measures phase-locked structure on the output raster and is, by design, not invariant to operations that move or resample that raster. Table 34 applies seven benign transforms to every public Validation image and scores them with the frozen primary heads and thresholds, without refitting. One bicubic resampling by $\times 0.9$ or $\times 1.1$ brings C2I four-way attribution from 95.7 to 23% (chance) and T2I seven-way from 91.2 to 17%; JPEG at quality 90 gives $81.5/40.4\%$ and quality 75 gives $57.0/15.9\%$; an unknown integer translation gives $50.0/42.1\%$, while re-registering the phase origin to a known offset restores $95.0/91.2\%$, and a $\sigma{=}0.5$ Gaussian blur leaves $94.0/89.5\%$. Detection at the frozen thresholds keeps the Real false-positive rate at 0–1.8% under every transform, but recall falls in step (Real-vs-ORT $91.4 \rightarrow 64.6\%$ at q90, 0.0% after resampling, 29.6% after a random shift, 89.4% after re-registration). Frequency-64D

Method	Ref.	JPEG q90	JPEG q75	Bicubic $\times 0.9$	Bicubic $\times 1.1$	Gauss σ=0.5	Random shift	Shift + re-reg.
C2I source attribution, $K = 4$ (Acc_4, %)								
UnivFD probe	88.7	57.4	52.1	75.6	79.0	86.1	89.1	89.1
NPR probe	94.3	27.1	27.3	79.0	91.2	91.5	89.8	89.8
RINE probe	67.0	43.4	30.3	60.3	65.4	66.8	67.5	67.5
SPAI probe	58.2	49.4	44.8	51.8	53.3	54.9	58.1	58.1
Frequency-64D[d]	95.2	82.1	70.3	46.2	47.1	91.2	95.3	95.3
Dzanic[a]	55.8	56.3	55.1	52.6	53.1	53.0	55.7	55.7
Frank[a]	47.0	47.1	47.1	47.1	45.9	47.1	46.3	46.3
NPR-stat[a]	77.5	59.6	56.5	53.2	56.3	64.7	58.3	58.3
Pixel Echo (Ours)	95.7	81.5	57.1	23.1	23.4	94.0	50.1	95.1
Pixel Echo Max (Ours)	98.8	85.0	51.4	37.7	37.2	97.7	56.6	98.6
T2I source attribution, $K = 7$ (Acc_7, %)								
UnivFD probe	90.2	31.7	38.6	80.8	76.3	84.8	90.2	90.2
NPR probe	65.7	18.6	18.1	66.1	65.1	65.9	66.4	66.4
RINE probe	60.7	29.8	27.1	55.1	56.7	55.9	60.9	60.9
SPAI probe	70.3	40.6	32.1	65.1	65.2	63.0	70.4	70.4
Frequency-64D[d]	82.7	71.0	42.5	46.1	35.9	67.0	82.1	82.1
Dzanic[a]	45.1	44.0	41.2	39.2	32.3	31.1	45.1	45.1
Frank[a]	25.6	25.6	25.6	25.6	23.8	25.6	25.4	25.4
NPR-stat[a]	77.6	37.5	20.9	42.4	39.9	68.0	62.7	62.7
Pixel Echo (Ours)	91.2	40.4	15.9	17.2	17.6	89.5	42.1	91.2
Pixel Echo Max (Ours)	94.8	43.2	22.0	26.9	25.4	92.5	53.8	94.5

Table 34: **Attribution under benign transforms (public Validation).** Each column applies one transform to every public Validation image (500 per source) and scores it with the frozen primary heads; no refit or recalibration. Ref.: untransformed Validation. Random shift: an unknown integer translation (the raster phase origin moves); Shift + re-reg.: the same images after re-registering the phase origin to the known offset. [a]Custom adaptation; [d]Frequency/Durall counted once. Frozen-feature probes recompute the backbone feature on the transformed image.

is translation-invariant by construction and degrades more slowly under JPEG and resampling; the frozen-backbone probes are robust to resampling, shift and blur but fail under JPEG. Pixel Echo is therefore a registered-grid, lossless-pipeline statistic: it applies when the raster is preserved or the offset is known, and it does not apply to re-encoded or resampled media. The media-pipeline confound of Appendix E.5 (a rational-factor Lanczos resampling that creates a phase-locked $P16$ pattern in Real images) is the same property seen from the other side.

What is attributed. The source attribution that the held-out Test supports is closed-set among sources whose decoder and export pipeline differ. Sources that share one decoder are separated only 9–11 pp above chance in balanced accuracy (Appendix H.4), decoder-family labels are recovered far better than generator labels, and in T2I the image size and processing pipeline coincide with generator identity, so the tables do not separate model traces from pipeline traces; size alone attributes 43.4% of T2I Validation images (two sources), and within the five sources that share 1024^2 it is at the majority-class baseline (20.4%) while Pixel Echo reaches 93.6% (Appendix H.5). Unseen generators, open-set rejection and cross-checkpoint transfer are not evaluated.

Single-representation ranking is not established. With heads run to an L-BFGS stationary point, the NPR probe overtakes Pixel Echo alone on C2I attribution and the frozen-feature probes gain up to 17 pp (Appendix H.3); only 18 of 67 primary gradient-descent heads meet the stopping criterion. The tables support complementarity between phase and frequency statistics (Pixel Echo Max), not superiority of Pixel Echo as a single representation. The comparison also mixes training regimes (official scores, frozen features with trained probes, custom adaptations) and no unified real/fake detector is built: the four per-source heads applied jointly flag 2.6% of held-out Real images for Pixel Echo and 3.6% for Max, and their joint recall is not available from the single Test pass (Appendix H.6). Pixel Echo detection of DiT (1.6% recall) and VAR (30.0%) at the 1%-FPR operating point is weak, and empirical calibration on 500 images is not a population guarantee.

Mechanism evidence is associational. The shared-decoder comparison, the pixel-head replay and the scale analyses establish decoder-conditioned association, not a localized causal mechanism: input distributions, architecture lineage, decoder weights and training data are not jointly controlled, and the fixed-latent series (Appendix E.5) withdrew our pre-registered scale-lineage prediction: across

six decoders of one latent space (one batch of 1,000 latents, 15 non-independent pairs) only the $P16$ sharing is stable, and the $P2$–$P8$ changes are decoder-specific but not predictable from lineage or architecture. The main-text estimator stops at period 16; Appendix E.6 extends the AliTok cohorts to period 32 and finds a reproducible 2.7% residual there, so the factor marks where a decoder's own pattern concentrates, not a hard cutoff. The one completed same-checkpoint intervention (800 reconstruction states; a targeted edit against random directions matched to the same output change) passed its 20 output-matching and 23 quality checks but reduced $P2/P4/P8$ RMS by $5.1/4.4/4.5\%$ against 4.3% for $P16$, a specificity of -0.36 pp (95% interval $[-0.46, -0.27]$), and shifted the high-frequency contrast by -6.7%; it therefore does not isolate a scale-selective mechanism, and an earlier equal-norm attempt failed matching. Layerwise, parameter-path and cross-checkpoint intervention diagnostics are deferred. Image-bootstrap intervals omit training and selection variation; permutation calibration requires exchangeability; compatible cross-decoder RECON controls and amplitude-equivalence criteria remain missing.

Data. The T2I cohorts have no matched Real controls, so T2I supports attribution only. T2I common255 ablations remain blocked by nonuniform phase geometry. Calibration beyond the evaluated cohorts and independent replication on other checkpoints remain open.