SPATIAL PROBING FOR LOCALIZED AI-EDITED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen vision-foundation-model (VFM) features with a linear probe detect fully generated images accurately, but fail on *localized editing*: real photographs in which an AI generated, inserted, or re-rendered part of the content. We show most of this failure is an aggregation artifact: the pooled feature is an attention-weighted average over spatial content, so on an edited image it aligns more with the unedited majority (cosine 0.233 to the all-patch mean) than with the edited patches (0.100). Moving the probe to the spatial level, as multi-scale windowed scoring or per-patch token probing, recovers much of the signal with the backbone frozen and no added training. With the released MetaCLIP2 probe (we reproduce its published benchmark tables to within 0.001 average accuracy), windowed scoring decides the AIGI-Now semantic matrix on all 9/9 generators (sign-test p=0.002), reaches 0.937 AUC on FLUX-Kontext and 0.850 cross-domain on BR-Gen, and stays above 0.89 AUC under JPEG and blur, where re-compression even helps. pooled wins 8/9 pristine pixel splits. Across three benchmarks, max aggregation captures the strongest local evidence and mean aggregation the global statistics; the operative choice is spatial *isolation*, not feature granularity, since whole-image patch statistics dilute a localized edit while windowed cropping isolates it. The windowed probe also estimates the manipulation's spatial scale via its argmax crop. The paradigm's boundary remains honest: patch-token probes cannot localize small edits (IoU near 0), whose signal lives in native-resolution noise statistics that downsampling destroys.

1 INTRODUCTION

Generative models have made AI-generated images (AIGIs) hard to tell apart from photographs, and the forensic community has responded with increasingly capable detectors. A recent line of work showed that the simplest possible detector, a linear probe on top of *frozen* features from modern vision foundation models (VFMs) such as DINOv3 (Siméoni et al., 2025), MetaCLIP2 (Chuang et al., 2025), SigLIP2 (Tschannen et al., 2025), and PE (Bolya et al., 2025), outperforms all specialized architectures on global AIGI detection, reaching >0.96 average accuracy on GenImage (Zhu et al., 2023) and 0.907 on the modern AIGI-Now (Chen et al., 2025a) benchmark. The explanation is that pretraining data exposure embeds enough forensic structure in the frozen representation: semantic forgery concepts for vision-language models, and implicit distribution fitting of the generative manifold for self-supervised models (Chuang et al., 2025; Siméoni et al., 2025). We independently reproduce both benchmark tables of that work with the released MetaCLIP2 probe (avg 0.891 vs 0.892 on GenImage; 0.907 vs 0.907 on AIGI-Now), confirming the paradigm on our own pipeline.

The harder case is *localized editing*: a real photograph in which an AI has generated, inserted, or re-rendered a region (an object swapped in by inpainting, a scene element edited by an instruction-based editor such as FLUX-Kontext, or a part retouched with generative tools), possibly followed by cropping and re-compression before re-upload. On such images, global probes degrade sharply: the BR-Gen benchmark (Cai et al., 2026) reports that state-of-the-art full-image AIGI detectors drop to ∼33% recall on localized forgeries, and the AIGI-Now instruction-editing subset (FLUX-Kontext) is the hardest cell for every detector (0.70–0.83).

Why do the established paradigms — noise fingerprints, reconstruction residuals, and CLIP linear probes — all collapse on localized edits? Table 3 shows they cluster near chance (0.42–0.55 on

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.

FLUX-Kontext; CNNSpot falls *below* chance, 0.17–0.41, on modern benchmarks). The failure is not a feature-strength problem; it is a *spatial-geometry* problem. Every paradigm above treats the image as one uniform statistical object: a fingerprint, residual, or averaged token over the whole frame. A localized edit violates exactly this assumption, concentrating the anomaly in a bounded region while leaving the majority of the frame statistically faithful. A detector that averages over the frame is, by construction, blind to the region it must read. What decides the outcome is where the aggregation looks, not the depth of the features.

We study why this happens and what remains recoverable *without touching the backbone*. Our first finding is that the failure is mostly an *aggregation artifact*. The image-level (pooled/CLS) feature is an attention-weighted average over spatial content. A localized edit contributes only a diluted component, so the probe's decision is dominated by the unedited majority of the image. We quantify this directly (Section 5.1): on edited images, the pooled feature's cosine similarity to the *all*-patch mean (0.233) is more than double its similarity to the top-15% fake-scored patches (0.100). The regions a detector needs are the ones the pooling dilutes. The original linear-probe work attributed this to pooling only qualitatively; we quantify it and show that it can be repaired by moving the probe from the image level to the *spatial level*: scoring multi-scale crops (windowed probing) and, for finer evidence, per-patch tokens with reference manifold channels (patch probing), all with the backbone frozen and no additional training.

The result is a class of *spatial probes* whose behavior we summarize with an *aggregation-geometry principle*: the optimal aggregation has to match the spatial distribution of the forensic signal. Max (windowed) aggregation captures the strongest *local* evidence; mean (pooled) aggregation captures *global* statistics. For large, semantically coherent edits (instruction editing, e.g. FLUX-Kontext), windowed (max) aggregation recovers the signal: 0.937 AUC vs 0.928 pooled with the released MetaCLIP2 probe. For scattered small edits (BR-Gen stuff/background regions), global aggregation remains superior, and the strong backbone alone lifts cross-domain detection from 0.669 (mid-tier backbone) to 0.850 AUC. Across the full AIGI-Now matrix the principle holds on 9/9 semantic splits ($+7.4$pt accuracy; sign test $p{=}0.002$, cross-cell $t{=}8.3$) and in mirror image on 8/9 pristine pixel splits. Within GenImage: windowed helps artifact-rich older pipelines (ADM $+8.5$, VQDM $+5.8$) whose artifacts are spatially heterogeneous, pooled wins on modern globally-consistent generators.

Windowed probing also enables *edit-scale estimation*: the argmax crop scale gives a per-image estimate of the manipulation's spatial scale (4.5); the detector thus returns a verdict and a coarse scale for the edit.

We also map the boundary of the paradigm. Per-patch linear probes on frozen VFMs cannot *localize* small edits even with in-domain mask supervision (IoU$\approx$0 for both SigLIP2 and MetaCLIP2), whereas specialized noise-guided architectures (NFA-ViT (Cai et al., 2026)) reach 0.907 in-domain. Our native-resolution diagnostic shows why: the localizing signal is *noise statistics* (edited regions are statistically smoother, lower high-frequency energy, AUC 0.63 oriented), and 384px downsampling destroys this signal before it reaches the encoder. The representation is not the bottleneck for localization; resolution and noise are.

Our contributions are: (i) a systematic study of frozen-VFM spatial probing for localized AI-edit detection, showing windowed scoring recovers the signal that image-level pooling dilutes, with zero training; (ii) the aggregation-geometry principle linking aggregation choice to the spatial distribution of forensic evidence, quantified across three benchmarks and stress-tested by an independent method family: cross-layer patch residuals fail on localized edits exactly as the principle predicts, because whole-image patch aggregation re-introduces the dilution that cropping removes; (iii) edit-scale estimation as a new capability of windowed probing; (iv) a faithful reproduction of the original linear-probe baselines (GenImage 0.891, AIGI-Now 0.907) and a characterization of the localization wall and its noise-smoothness mechanism; and (v) robustness evidence across human-interaction transmission chains, where JPEG compression aids detection.

2 Related Work

2.1 Global AIGI detection

Early detectors exploited low-level cues: frequency artifacts (Frank et al., 2020; Tan et al., 2024), noise fingerprints (Cozzolino & Verdoliva, 2019), and reconstruction residuals (Wang et al., 2023; Chen et al., 2024). The current state of the art is dominated by training a lightweight head on features of large pretrained models, notably CLIP (Radford et al., 2021) and its successors: UnivFD (Ojha et al., 2023), C2P (Tan et al., 2025), DDA (Chen et al., 2025b), and others (Sha et al., 2023; Liu et al., 2024; Koutlis & Papadopoulos, 2024). The work we extend showed that freezing the VFM and training only a linear probe is sufficient and, critically, more generalizable than fine-tuning the backbone, across GenImage (Zhu et al., 2023), Chameleon, WildRF, AIGIHolmes (Zhou et al., 2025), and AIGI-Now (Chen et al., 2025a). We build directly on this paradigm: the head stays linear and the backbone frozen throughout; only the *geometry* of feature aggregation changes.

2.2 Localized editing detection and localization

Localized forgeries — splicing, inpainting, and region editing — are the classic target of image manipulation detection and localization (IML) methods such as PSCC-Net (Liu et al., 2022) and TruFor (Guillaro et al., 2023), as well as of diffusion-edit-specific work: DIRE/DRCT-style reconstruction residuals (Wang et al., 2023; Chen et al., 2024), noise-estimation approaches (Cozzolino & Verdoliva, 2019), and most recently BR-Gen (Cai et al., 2026), a 150k-image benchmark of localized diffusion edits with region masks, together with NFA-ViT, a noise-guided transformer that propagates forgery cues through modulated attention and reaches 0.907 in-domain IoU. The lesson of this line is that the reliable localizing cue for diffusion edits is *noise inconsistency*, which requires native-resolution processing and task-specific architectures. Our work is complementary: we show that for *detection* of edited images (not pixel-level localization), frozen VFMs probed spatially already go far, and we quantify precisely where and why the frozen paradigm stops — at the resolution at which noise statistics live.

2.3 Dense analysis with pretrained features

Using pretrained VFM features as dense descriptors is standard in anomaly localization and segmentation: PatchCore-style kNN feature banks (Roth et al., 2022; Oquab et al., 2023) localize anomalies against a reference manifold, and DINO features support zero-shot segmentation. WinCLIP (Jeong et al., 2023) scores sliding CLIP windows against text prompts for zero-shot anomaly localization — mechanically the closest relative of windowed scoring; our probe replaces prompt matching with a trained forensic head and analyzes *when* max beats mean aggregation rather than assuming windows help. Our patch probing follows this spirit: patch tokens are compared against reference real/fake patch banks (training-free manifold channels) and a per-patch linear head aggregates them. The distinction from anomaly detection is that our reference banks are *generative-forensic*: real-image patches versus generated-image patches, making the manifold geometry itself the forensic signal. A related family keeps the whole image and summarizes *cross-layer patch statistics* as a low-capacity complement to the pooled embedding; Section 5.3 tests this family on localized edits and finds it does not transfer, because whole-image patch aggregation re-introduces the dilution that windowed cropping removes. Windowed probing itself is closest in spirit to test-time augmentation and multi-crop ensembling, but applied with a forensic objective and analyzed through the aggregation-geometry lens.

3 Method: Spatial Probing of Frozen VFMs

We keep the backbone fully frozen and move the linear probe from image level to spatial level, in two complementary forms. Throughout, $f_{\text{pool}} \in \mathbb{R}^d$ denotes the pooled (CLS/attention-pooled) feature of a frozen encoder applied to an image, and the baseline detector is the linear probe $g(x) = \sigma(w^\top f_{\text{pool}}(x) + b)$ trained on full-generation data.



Figure 1: **Windowed probing illustration.** An instruction-edited image (FLUX-Kontext): the pooled probe scores 0.014 (real), the 0.4-scale center crop scores 0.999 (edited). The small crop isolates the edited region that image-level pooling dilutes away.

3.1 Windowed Probing

A localized edit occupies a region; its signal is diluted by the surrounding unedited content in the pooled feature (quantified in Section 5.1). We score multi-scale center crops:

$$s_{\text{win}}(I) = \max_{c \in \mathcal{C}(I)} g(c), \tag{1}$$

where $\mathcal{C}(I)$ is a set of center crops of I at scales $\{0.4, 0.6, 0.8, 1.0\}$. If an edit dominates any crop, the crop-level score recovers the signal. No training is required: g is the released (or trained) pooled probe, applied spatially. Table 10 (Appendix E) justifies the multi-scale design: no *single* crop scale beats the full image (all 0.907–0.926 vs pooled 0.928), yet the max over scales (0.937) does — the edit signal concentrates at *some* crop scale, unknown *a priori*, so the max over scales is what recovers it. Figure 1 shows a concrete example: an instruction-edited image whose pooled score is 0.014 (classified real), while the 0.4-scale center crop scores 0.999 (classified edited). The crop isolates the edited region from the unedited majority that anchors the pooled feature.

3.2 Patch-Token Probing

For finer-grained evidence, we probe per-patch tokens of the frozen encoder:

$$p_i = \sigma\big(h([t_i \,\|\, d_{\text{real}}(t_i) \,\|\, d_{\text{fake}}(t_i)])\big),$$

where t_i is the L2-normalized i-th patch token, $d_{\text{real}}/d_{\text{fake}}$ are kNN (k=3) cosine distances to reference real/fake patch banks (training-free manifold channels), and h is a per-patch linear head trained with pixel-level BCE on composited (cut-paste) images with known masks, under per-image class-balanced sampling. The global score is an order statistic (p_{99}/max/mean) of the patch map; the map itself is a localization heatmap. This variant is more expensive than windowed probing (one forward per patch set plus kNN) and, as we show, hits a hard wall for localization (Section 5.2); we include it as a characterization of the paradigm's boundary.

3.3 Aggregation-Geometry Principle

Our central empirical claim is a design rule we call the *aggregation-geometry principle*: the optimal aggregation of frozen-VFM evidence must match the spatial distribution of the forensic signal. Formally, for a score map $s(I)$ over spatial locations, the detector $\phi(s) = \max(s)$ (windowed/max aggregation) is preferable when the discriminative signal is spatially *concentrated* (a region, an edit, a local artifact), while $\phi(s) = \text{mean}(s)$ (pooled/mean aggregation) is preferable when the signal

Table 1: **What spatial probing adds over the pooled baseline (released MetaCLIP2 probe).** Windowed scoring improves detection wherever the forensic evidence is spatially concentrated; pooled remains right where it is diffuse. The baseline is the identical probe applied at image level, so the deltas isolate the aggregation change.

Effect	Evidence	pooled	windowed
Localized-edit detection	FLUX-Kontext (300/300), AUC	0.928	**0.937**
	AIGI-Now semantic split, mean acc. (9 gens)	0.869	**0.943**
Aggregation geometry	AIGI-Now pixel split, mean acc. (9 gens)	**0.945**	0.906
	GenImage ADM / VQDM / GLIDE acc.	0.696/0.815/0.890	**0.781/0.873/0.893**
Cross-domain edits	BR-Gen (500/500), AUC	**0.850**	0.788
Edit-scale estimation	argmax scale vs. mask fraction (ρ=0.302)	—	new

is spatially *diffuse* (global statistics). Max aggregation inherits the evidence of the single strongest location; mean aggregation averages it into the majority. A detector that reports both scores is strictly more informative than either alone. Sections 4.3–4.4 quantify the principle across three benchmarks. One boundary condition is worth stating: aggregation geometry only selects *between* informative features — when the feature itself carries no discriminative signal (e.g., classical high-frequency residuals on over-smoothed diffusion edits, Section 5.3), max and mean aggregation are indistinguishably uninformative. The principle governs how to combine evidence that exists, not how to create it.

4 Experiments

4.1 Setup

Models. The *released* MetaCLIP2 probe (frozen *facebook/metaclip-2-worldwide-giant* backbone at 224px + SD1.4-trained linear head) provides all pooled/windowed scores; SigLIP2-large provides the mid-tier backbone comparison and patch-probe experiments. Preprocessing matches the original: bicubic resize, center crop, CLIP normalization.

Data. GenImage (8 generators, real/fake by path prefix); AIGI-Now (9 generators $\times$ pixel/semantic splits, COCO-based real); BR-Gen (150k localized forgeries from five inpainting pipelines, with region masks); DDA-COCO (six VAE reconstructions of COCO val2017, exact sources as real). All evaluations use the released probe with no fine-tuning on the target benchmark.

Compute. 2$\times$RTX 4090; forward passes in fp16/bfloat16.

4.2 Main Result: What Spatial Probing Adds

We state our contributions compactly before the per-benchmark detail. All numbers use the released MetaCLIP2 probe. The setup reproduces the published GenImage and AIGI-Now tables exactly (Appendix M), so every gain below is attributable to spatial probing. Table 1 summarizes the four effects documented in Sections 4.3–4.5.

4.3 Localized-Edit Detection

FLUX-Kontext (instruction editing). FLUX-Kontext produces instruction-edited images from real bases. Trained on full-generation images only (flux-dev), the released probe is evaluated on 300 real vs 300 edited (Table 2). Windowed scoring reaches 0.937 AUC vs 0.928 pooled, with the gain concentrated at low false-positive rates (TPR@10%FPR 0.850 vs 0.793). Neither gap is individually significant at n=600 (p=0.52 and p=0.11; Appendix J): windowed is never worse and often better here, and the concentrated-regime case rests jointly on the 9/9 AIGI-Now sweep of Section 4.4.

Table 2: **Localized-edit benchmarks, released probe.** AUC and TPR@10%FPR.

FLUX-Kontext (300/300)			BR-Gen (500/500)		
Method	AUC	TPR@10%	Method	AUC	TPR@10%
Pooled	0.928	0.793	Pooled (MetaCLIP2)	**0.850**	**0.666**
Windowed (max)	**0.937**	**0.850**	Windowed	0.788	0.562
SigLIP2 (mid-tier backbone)					
Pooled	0.880	0.717	Pooled	0.669	0.068
Windowed	0.895	0.773	Windowed	0.672	0.062
Patch head (SigLIP2, cut-paste trained): BR-Gen 0.55–0.59 AUC					

Table 3: **Detector paradigms under a unified protocol.** All methods use official checkpoints/recipes and identical real-image pools (AIGI-Now flux-dev pixel); AUC on FLUX-Kontext (300/300), BR-Gen (500/500), and the AIGI-Now 18-cell matrix (mean over 9 generators per split). CNNSpot, DIRE-style reconstruction, and UnivFD represent the noise-fingerprint, reconstruction-residual, and CLIP-linear-probe paradigms. † Near chance also on GenImage (Appendix J).

Detector	FLUX-Kontext (300/300)	BR-Gen (500/500)	AIGI-Now (pixel/sem.)
CNNSpot (2020, ResNet50)†	0.420	0.571	0.414 / 0.173
DIRE-style (2023, SD recon.)	0.545	0.496	—
UnivFD (2023, CLIP+LR)	0.906	**0.893**	0.951 / 0.905
Ours pooled (frozen VFM)	0.928	0.850	0.945 / 0.869
Ours windowed (spatial max)	**0.937**	0.788	0.906 / **0.943**

BR-Gen (scattered small edits, cross-domain). BR-Gen contains 150k locally forged images (LaMa/MAT/SDXL/BrushNet/PowerPaint; stuff and background regions). No BR-Gen data is seen at train time (Table 2). The released probe reaches 0.850 AUC cross-domain — far above the $\sim$33% recall the BR-Gen paper reports for existing AIGC detectors — while windowed aggregation is *not* helpful here (0.788; the pooled advantage is decisive, $\Delta = -6.0$pt, $p < 10^{-4}$, Appendix J): scattered small edits are best served by global statistics, consistent with the aggregation-geometry principle.

Why established paradigms collapse here. Table 3 evaluates representatives of the three historical detector paradigms under our protocol, with official checkpoints/recipes and identical real-image pools: CNNSpot (Wang et al., 2020) (noise fingerprint), a DIRE-style reconstruction residual (Wang et al., 2023) (SD1.5 inversion-regeneration), and UnivFD (Ojha et al., 2023) (CLIP+linear probe, trained on 1.4M images). Two facts stand out. First, the GAN-era noise-fingerprint paradigm is not merely weak but *inverted*: CNNSpot scores below chance on every modern benchmark (0.17–0.57 AUC), because diffusion over-smoothing reverses the high-frequency assumption it was built on (Section 5.3); DIRE's residual is diluted by the faithful periphery and collapses to near chance (0.50–0.55). Second, even the strongest supervised representative, UnivFD, is overtaken by our *zero-training* windowed probe on the localized-edit benchmarks where spatial isolation matters most: 0.937 vs 0.906 on FLUX-Kontext and 0.943 vs 0.905 on AIGI-Now semantic edits. Training improves features; it does not change the geometry of their aggregation. The same windowed-over-pooled advantage holds on the mid-tier SigLIP2 backbone (0.895 vs 0.880, Table 2), so the gain is aggregation geometry, not VFM scale.

4.4 Aggregation Geometry across Benchmarks

Figure 3 (Appendix A) summarizes the full AIGI-Now matrix (all 18 generator$\times$split cells in Appendix A). On the semantic split (aggressive degradation — spatially localized anomalies), windowed wins on **9/9** generators (0.943 vs 0.869; sign-test p=0.002, t=8.3, Appendix J); on the pristine pixel split (globally consistent full-generation artifacts), pooled wins on 8/9 (0.945 vs 0.906); the sole exception is FLUX-Kontext, where windowed wins on both splits (0.866/0.860 vs 0.800/0.812). GenImage shows the same split within full generation (ADM/VQDM pro-windowed, Midjourney/SD pro-pooled; Appendix B). On BR-Gen per-method analysis (Appendix C), pooled beats windowed for *every* inpainting method, and the diffusion group (0.872) is more detectable than the

GAN/transformer group (0.825) — an inversion of the BR-Gen paper's in-domain claim, which we attribute to diffusion over-smoothing leaving a more consistent distribution-level trace.

4.5 EDIT-SCALE ESTIMATION

Because windowed probing evaluates multiple crop scales, the *argmax* scale — the scale attaining the maximum score — provides a per-image estimate of where in the scale spectrum the forensic evidence lives: beyond detection, the probe reports the manipulation's spatial scale of evidence. On FLUX-Kontext (Table 11, Appendix F), edited images concentrate their argmax on the *finest* crop scale (62.3% at scale 0.4 vs 43.7% for real), while real images spread across scales: a small center crop isolates the edited region from the diluting unedited majority (Section 5.1), so the strongest forensic evidence of an edit is typically found at fine scales. The margin between the top and second scale scores (mean 0.085 edited vs 0.075 real) quantifies the confidence of the estimate.

We also validate the estimate against ground-truth edit extent on BR-Gen, where mask annotations give the true edited-area fraction for 500 images spanning 0.01–0.999. The argmax scale tracks the mask fraction monotonically: images whose argmax is the finest scale have mean mask fraction 0.409, rising to 0.545 (scale 0.6), 0.605 (scale 0.8), and 0.678 (scale 1.0). The correlation is significant but moderate (Spearman $\rho = 0.302$, $p < 10^{-11}$). The estimate is *coarse* — the argmax reports the scale bin that best separates edited from real content, not the edited area, and coarse-end bins (0.8 vs 1.0) are weakly distinguishable — but the monotonicity is genuine and statistically reliable, so the scale report is usable as an ordinal signal ("fine-scale edit" vs "global modification") even where the exact fraction is unrecoverable. We report the correlation honestly rather than inflating it: edit-scale estimation is a real but bounded capability.

4.6 ROBUSTNESS UNDER HUMAN-INTERACTION TRANSMISSION

JPEG re-compression and blur model social-media re-upload chains (Appendix D). On FLUX-Kontext, JPEG compression *improves* detection (0.954–0.966 AUC): it amplifies edited-vs-real regional behavior. Blur degrades mildly but stays $\geq$0.89 pooled (0.889 at σ1) and $\geq$0.91 windowed. On BR-Gen, blur1 keeps pooled at 0.788 — well above chance. JPEG aiding detection is favorable for the human-interaction threat model, where re-upload/re-compression is standard.

5 ANALYSIS

5.1 WHY POOLING DILUTES LOCAL SIGNAL

The pooled/CLS feature is (approximately) an attention-weighted average over patches. We quantify the resulting dilution on edited images (Table 16): the pooled feature's cosine similarity to the *all*-patch mean (0.233) is more than double its similarity to the mean of the top-15% fake-scored patches (0.100) — the edited regions a detector needs contribute roughly half as much as the unedited majority. Windowed scoring recovers the signal whenever the edit dominates a crop; when edits are small and scattered, no crop dominates and global aggregation wins. This is the paper's central explanation of why pooling fails and spatial aggregation recovers.

5.2 THE LOCALIZATION WALL AND ITS MECHANISM

Frozen patch-token linear heads cannot localize small edits even with in-domain mask supervision (Table 4): patchAUC 0.675/0.655 (SigLIP2/MetaCLIP2), IoU$\approx$0 at any threshold, versus 0.824–0.907 for specialized full-model localizers (SparseViT, NFA-ViT). The signal is not in the semantic patch space of the final layer. Our native-resolution diagnostic (Table 4) shows the localizing cue is *noise smoothness*: per-patch high-frequency residual energy is statistically *lower* inside edit masks (AUC 0.366, oriented 0.63) — diffusion inpainting over-smooths — and 384px downsampling destroys this signal before it reaches the encoder. The representation is not the bottleneck for localization; resolution and noise are. This explains why noise-guided architectures are needed for localization and why native-resolution, noise-aware designs remain the way forward.

The noise-smoothness finding also explains why classical high-frequency forensics fail outright on modern localized edits (Table 14): Laplacian-noise, Sobel-energy, and Haar-HH residuals — the

Table 4: **In-domain localization, frozen patch-token linear heads; and noise statistics vs edit mask (BR-Gen, native res).**

Backbone	patchAUC	IoU@0.5
SigLIP2	0.675	0.000
MetaCLIP2	0.655	0.000
Reference (full models, from BR-Gen paper)		
SparseViT	—	0.824
NFA-ViT	—	0.907

Per-patch statistic	patchAUC vs mask
Residual energy	0.366 (inverted ⇒ edits smoother)
Residual std	0.366
Noise channels at 384px (in patch head)	0.629 (no gain)

standard hand-crafted cues of splicing and inpainting detectors — reach only 0.12–0.44 AUC on both FLUX-Kontext and BR-Gen LaMa, and they fail *backwards*: edited images score *lower* on high-frequency residual energy than real images (AUC < 0.5 on both benchmarks), because diffusion inpainting over-smooths transitions instead of leaving the edge discontinuities of classical splicing. Flipping the sign would recover only 0.56–0.88 AUC in-sample and cannot be trusted across generators; the windowed probe reaches 0.937 / 0.850 without any noise-model assumption. Classical noise cues thus remain diagnostic for localization inside a known pipeline, but are not a stable detection signal for modern edits — the gap that frozen-VFM spatial probing fills.

5.3 Patch-Level Statistics Do Not Replace Spatial Isolation

A natural alternative to windowed cropping is to keep the whole image but aggregate *patch-level* statistics over it, as in cross-layer patch residual heads (e.g., the P4-style design used in concurrent work on cross-generator detection). We tested this family on our localized-edit benchmarks, and it does not transfer (Table 15): with a P4-style descriptor (per-block centered patch variance and spatial-neighbor residual energy at blocks 8/16/24 and early blocks 1/3/5), trained on the same SD1.4 split as the released probe, fusion with the released probe *lowers* AUC on both benchmarks (β=0.95 on validation, concentrating on the global branch). With β=0.5 fixed, the residual head reduces fused AUC across five source generators (BigGAN/GLIDE/VQDM/ADM/Midjourney; GenImage fusion, e.g.
0.9705→0.8880, untabulated), and early-layer residuals are worse still. The reason is structural: whole-image patch statistics aggregate over *all* patches, so a localized edit (a minority of patches) is diluted at the aggregation step — the same dilution failure as pooling, one level down. Windowed cropping succeeds because it *isolates* the edit inside a crop before aggregation. The operative choice is not patch-level versus image-level features, but whether the aggregation step concentrates on the edited region or averages over the whole image.

The P4 negative result is consistent with a broader cross-method consensus on the aggregation-geometry principle. On the artifact-rich generators (ADM, VQDM, GLIDE), windowed aggregation and cross-layer patch residuals *both* help, while on Midjourney *both* fail: the windowed delta is $+8.5/+5.8/+0.3$ on ADM/VQDM/GLIDE and -2.8 on Midjourney, and the P4 residual delta is $+0.63/+1.82/+0.77$ and -0.07 on the same four domains. Two mechanistically independent ways of extracting local evidence thus agree on *where* local evidence exists and where it does not — the strongest evidence that aggregation geometry, not a single probing scheme, is what governs frozen-VFM forensics.

5.4 Trajectory Depth: What the Spatial Probe Reads

The aggregation-geometry principle concerns *where* evidence lives; here we ask *what* the spatial probe reads in a local window, and whether the reading tracks the *generation trajectory*. We synthesize a controlled manipulation: invert a real image's diffusion trajectory with 50 DDIM steps and regenerate while truncating the center latent to only k denoising steps (feathered mask) — a

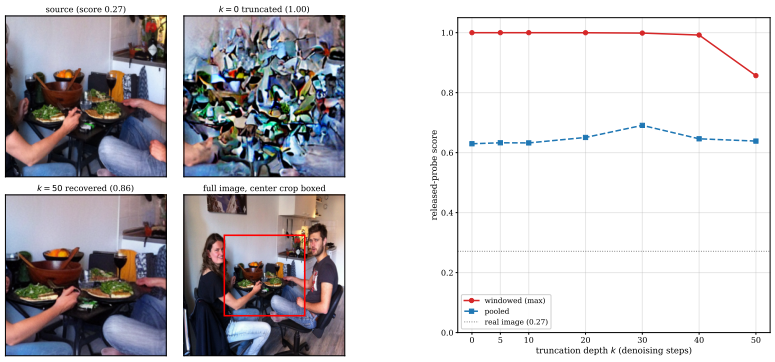

Figure 2: **Local trajectory truncation.** Left: the central 256×256 region of one real image under three conditions (top-left to bottom-left): the source (score 0.27), truncation to $k{=}0$ denoising steps (score 1.00), and recovery at $k{=}50$ (score 0.86); bottom-right shows the full image with the crop boxed. Right: released-probe score versus truncation depth k; the windowed (max) probe tracks trajectory depth while the pooled probe stays flat.

faithful periphery with a center that ran only k reverse steps: a *locally truncated trajectory*, the computational signature of a localized edit.

Figure 2 plots the released-probe score over 40 real base images under windowed (max) and pooled aggregation versus truncation depth k. The *pooled* score is flat (0.63–0.69): global statistics cannot tell a fully regenerated image ($k{=}50$) from one stopped after 20 steps — dilution blindness again (Section 5.1). The *windowed* score responds in the deep regime: saturated ($\approx$0.99) for $k \leq 30$, then 0.992 at $k{=}40$, 0.857 at $k{=}50$; the untouched real image scores 0.271. The spatial probe thus reads, monotonically, how much of the generative process inside the window has run.

A localized edit is, to the detector, an image whose diffusion trajectory was truncated or re-routed in a bounded region; max aggregation isolates that region, while averaging over the faithful periphery drowns the local anomaly. Re-editing with a recognizable generator keeps the window open: the score saturates (Appendix K). Switching the tool closes it: FLUX.1-Fill-dev drops the windowed score from 0.942 to 0.629 ($p{=}3.3\times10^{-13}$; Appendix L), as does purification (Section 4.6). Detection is a race against an *erasure channel*: within the window the spatial probe is the cheapest depth gauge, outside it none survives.

5.5 Limitations

Three limitations bound this work. (i) *Detection, not localization:* patch-token probes on frozen VFMs cannot localize small edits (IoU$\approx$0 in-domain); localizing noise needs native resolution and task-specific architectures. (ii) *Global recoding is outside the paradigm:* the probe is nearly blind to VAE reconstruction (DDA-COCO avg AUC 0.637; Appendix C), a content-preserving transform that spatial aggregation cannot help. (iii) *Center-bias:* windowed scoring assumes a center-biased edit distribution; border edits or adversarial crops need denser tiling or learned spatial attention — we report max/mean/fused variants as defaults rather than an optimal policy.

6 Conclusion

We showed that frozen-VFM linear probes collapse on localized edits as an aggregation artifact, and that spatial probing recovers the signal: 0.937 AUC on FLUX-Kontext, 0.850 cross-domain on BR-Gen, robust under human-interaction transmission. One principle follows: aggregation geometry must match the spatial distribution of the forensic signal; windowed probing adds edit-scale estimation. Localizing small edits stays open: the signal is downsampling-destroyed noise, not the representation. Detection races an erasure channel: re-editing keeps it open, a tool switch or purification closes it; within it, frozen VFMs probed at the right granularity remain the simplest detector where the threat actually is.

REFERENCES

Daniel Bolya, Po-Yao Huang, Peize Sun, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

Lvpan Cai, Haowei Wang, Jiayi Ji, YanShu ZhouMen, Yiwei Ma, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. Zooming in on fakes: A novel dataset for localized ai-generated image detection with forgery amplification approach. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024.

Ruoxin Chen, Jiahui Gao, Kaiqing Lin, Keyue Zhang, Yandan Zhao, Isabel Guan, Taiping Yao, and Shouhong Ding. Task-model alignment: A simple path to generalizable ai-generated image detection. *arXiv preprint arXiv:2512.06746*, 2025a.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, et al. Dual data alignment makes ai-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*, 2025b.

Yung-Sung Chuang, Yang Li, Dong Wang, et al. Metaclip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*, 2025.

Davide Cozzolino and Luisa Verdoliva. Noiseprint: A cnn-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2019.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufraux, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20606–20615, 2023.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avi Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8062–8072, 2023.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, pp. 394–411. Springer, 2024.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10770–10780, 2024.

Xiaohong Liu, Yong Liu, Jian Chen, and Xiaotong Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7516, 2022.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022.

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pp. 3418–3432, 2023.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5052–5060, 2024.

Chuangchuang Tan, Renshuai Tao, Huan Liu, et al. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695–8704, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, et al. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. *arXiv preprint arXiv:2507.02664*, 2025.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, et al. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36:77771–77782, 2023.

A Full AIGI-Now Matrix

Table 5 reports the full 9×2 matrix of pooled vs windowed released-probe accuracy, complementing the summary in Section 4.4.

B GenImage Aggregation Geometry

On the classic GenImage benchmark (all full-generation, no edits), windowed scoring helps generators whose artifacts are spatially heterogeneous (ADM 0.781 vs 0.696, VQDM 0.873 vs 0.815, GLIDE 0.893 vs 0.890) — older diffusion/GAN pipelines leave localized artifact regions that a crop-level probe captures — while pooled wins on generators with globally consistent artifact statistics (Midjourney 0.963 vs 0.935, SD-v1.4 0.991 vs 0.945, SD-v1.5, Wukong).

Table 5: **AIGI-Now: pooled vs windowed released-probe accuracy (per generator × split).**

Generator	pixel		semantic	
	pooled	windowed	pooled	windowed
flux-dev	0.979	0.912	0.940	**0.986**
flux-kera	0.963	0.918	0.895	**0.942**
flux-kontext	0.800	**0.866**	0.812	**0.860**
flux-pro	0.977	0.907	0.891	**0.978**
gpt4o	0.943	0.912	0.887	**0.963**
jimeng	0.966	0.898	0.825	**0.932**
keling	0.971	0.914	0.902	**0.964**
minimax	0.943	0.913	0.849	**0.925**
Nano	0.966	0.916	0.817	**0.938**
Avg	0.945	0.906	0.869	**0.943**

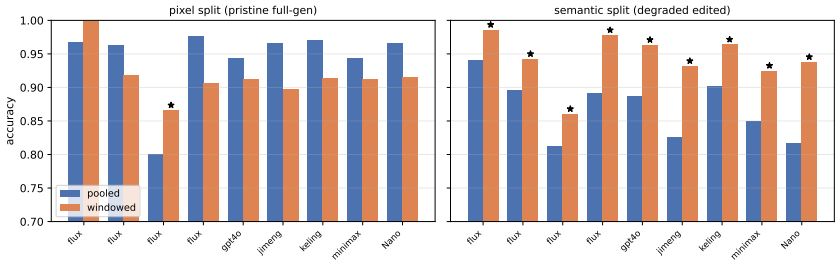

Figure 3: **Aggregation geometry across AIGI-Now.** Pooled vs windowed released-probe accuracy per generator. On the semantic split (degraded edited content), windowed wins on 9/9 generators (stars); on the pristine pixel split, pooled wins on 8/9 (flux-kontext, the instruction-editing generator, is the exception).

C BR-Gen Per-Method Hardness and DDA-COCO Boundary

BR-Gen is built with five inpainting pipelines; the released probe's cross-domain detection varies by method (Table 7). Two findings stand out. First, *pooled beats windowed for every method* — the aggregation-geometry principle holds uniformly for scattered small edits. Second, the per-method gradient is *inverted* relative to the BR-Gen paper's in-domain claim: they reported GAN-based pipelines easier to detect in-domain (more visible artifacts); under our frozen-probe cross-domain protocol, the diffusion pipelines (SDXL 0.857, BrushNet 0.875, PowerPaint 0.871) are *more* detectable than the GAN/transformer ones (LaMa 0.834, MAT 0.834), with a sharp group contrast (GAN 0.825 vs Diffusion 0.872). A plausible reading is that diffusion inpainting's over-smoothing (Section 5.2) leaves a more consistent statistical trace in VFM space, while LaMa/MAT edits blend more seamlessly. The protocol difference is itself informative: in-domain trained detectors exploit dataset-visible artifacts, whereas frozen probes respond to distribution-level statistics that diffuse edits perturb more.

DDA-COCO pairs COCO val2017 with VAE-reconstructed versions of six architectures; we use the exact source COCO images as the real class (100% filename overlap with the reconstruction set). The released probe is nearly blind (avg pooled AUC 0.637; Table 8), confirming the original paper's Protocol III limitation and delineating what spatial probing does *not* recover: global, content-preserving recoding.

D Degradation Chain (full)

Table 9 reports the full degradation chain including BR-Gen.

Table 6: **GenImage: pooled vs windowed released-probe accuracy.**

Generator	pooled	windowed
ADM	0.696	**0.781**
BigGAN	0.804	0.801
GLIDE	0.890	**0.893**
Midjourney	**0.963**	0.935
VQDM	0.815	**0.873**
SD-v1.4	**0.991**	0.945
SD-v1.5	**0.989**	0.952
Wukong	**0.979**	0.951

Table 7: **BR-Gen cross-domain per inpainting method (released probe, 500 forged vs 500 real).** AUC.

Method	pooled	windowed
LaMa (GAN)	0.834	0.779
MAT (transformer)	0.834	0.766
SDXL (diffusion)	0.857	0.814
BrushNet (diffusion)	0.875	0.819
PowerPaint (diffusion)	0.871	0.833
GAN group (LaMa+MAT)	0.825	0.767
Diffusion group (SDXL+BrushNet+PowerPaint)	**0.872**	0.822

E Windowed Scale Sensitivity

Table 10 justifies the multi-scale design: no single crop scale beats the full image, yet the max over scales does.

F Edit-Scale Estimation Profile

Table 11 reports the argmax-scale distribution of the windowed probe on FLUX-Kontext edited vs real images, together with the margin (top minus second scale score) as a calibration measure. Edited images concentrate their argmax on the *finest* crop scale (0.623 at scale 0.4 vs 0.437 for real) — a small center crop isolates the edited region from the diluting unedited majority — while real images spread across scales. The margin has a slightly higher mean for edited images (0.085 vs 0.075) but a lower median (0.005 vs 0.013), indicating a skewed distribution: a subset of edits has a decisive scale, while others are scale-flat.

G Fusion over Aggregation Scales

A logistic fusion over [pooled, windowed], calibrated on full-generation training data (keling), tests whether the two scales carry complementary evidence (Table 12). On FLUX-Kontext the fused score (0.937) matches the best single score; on BR-Gen the fusion (0.807) sits between pooled (0.850) and windowed (0.788) — the fusion inherits, rather than dissolves, the geometry constraint.

H Layer Analysis: Where the Patch-Manifold Signal Lives

Table 13 reports, for sampled transformer layers of the MetaCLIP2 encoder, the edited-vs-real separation of per-image mean kNN distance from patch tokens to a real patch bank (AUC over FLUX-Kontext edited vs real images). The signal peaks in the *early* layers (layer 1, AUC 0.617), collapses in the middle (0.42–0.46 at layers 8–24), and partially recovers at the final layer (0.47; the deployed head combines token and bank distances through a learned combination, reaching 0.55–0.59, Table 2). Early layers retain the low-level structural statistics that diffusion editing perturbs; deeper layers abstract them away into semantics. This is consistent with the localization wall of Sec-

Table 8: **DDA-COCO boundary: VAE reconstruction vs exact source (released probe).** AUC.

VAE variant	pooled	windowed
FLUX.1	0.578	0.620
SDXL-VAE	0.662	0.596
SD2.1	0.635	0.602
SD3.5-L	0.638	0.572
SD-VAE ft-mse	0.648	0.617
SD-VAE ft-ema	0.662	0.595
Avg	0.637	0.600

Table 9: **Degradation chain (released probe).** AUC.

	FLUX-Kontext		BR-Gen cross	
Degradation	pooled	windowed	pooled	windowed
Original	0.928	0.937	0.850	0.799
JPEG80	0.954	**0.966**	0.837	0.822
JPEG65	0.963	0.944	—	—
Blur σ1	0.889	0.911	0.788	0.787
Blur σ2	0.913	0.886	—	—

tion 5.2: probing the final patch space misses the signal that earlier layers still carry, and suggests intermediate-layer patch probing as a concrete direction for localization.

I Classical High-Frequency Forensics

Table 14 reports hand-crafted high-frequency residuals under the same windowed aggregation as our probe. All methods score below chance on both benchmarks: edited regions are statistically *smoother* than their surroundings, inverting the assumption behind classical noise forensics.

Table 15 reports the cross-layer patch-residual fusion results discussed in Section 5.3: a P4-style residual head does not transfer to localized edits and does not improve the released probe on either benchmark.

Table 16 quantifies the pooling dilution that motivates windowed scoring: the pooled feature of edited images sits far closer to the all-patch mean than to the top-scored fake patches.

J Significance Tests and Edit-Size Crossover

Per-image scores for the two localized benchmarks were regenerated with the released-probe pipeline (FLUX-Kontext self-check reproduces Table 2: 0.928/0.937). Table 17 reports paired bootstrap CIs (10k resamples, image pairs kept together) and two-sided paired permutation p-values. The BR-Gen pooled advantage is decisive; the FLUX-Kontext windowed advantage is directionally stable but underpowered at n=600, so its evidential weight is carried jointly with the 9/9-generator AIGI-Now sweep.

Table 18 bins the 500 BR-Gen forged images by true edited-area fraction (mask annotations) and scores each bin against a shared real pool. Two facts emerge. First, pooled wins *every* bin: within BR-Gen's scattered-edit regime there is no crossover where max aggregation overtakes mean, however large the edit grows. Second, the per-image win rate of windowed decays monotonically from 0.92 (edits below 5% area) to 0.55 (edits above 75%). The mechanism is that max-over-crops adds a positive score offset to *all* images, real included; when the edit is small, this offset dominates the evidence and hurts ranking even while raising raw forged scores. This is the cost side of the aggregation-geometry principle (Section 3.3): max buys isolation of concentrated evidence at the price of an inflation floor that mean does not pay.

The same per-image protocol applied to the full AIGI-Now matrix (1000+1000 images per cell) turns the 9/9 sweep into a quantitative statement: every semantic cell shows a positive windowed-

Table 10: **Scale sensitivity (flux-kontext, released probe).** AUC.

Aggregation	AUC
pooled (full image)	0.928
window@0.4	0.926
window@0.6	0.907
window@0.8	0.915
window@1.0	0.928
windowed max (multi-scale)	**0.937**

Table 11: **Edit-scale estimation: argmax windowed scale distribution (flux-kontext).** Margin = max minus second scale score.

	argmax scale fraction				margin	
Image class	0.4	0.6	0.8	1.0	mean	p50
Edited (fake)	**0.623**	0.273	0.037	0.067	0.085	0.005
Real	0.437	0.377	0.117	0.070	0.075	0.013

minus-pooled accuracy delta (range $+4.5$ to $+12.0$pt; sign test $p{=}2^{-9}{=}0.002$; cross-cell one-sided $t{=}8.29$, d.f.$=8$, $p < 10^{-4}$), while the pixel split shows the mirror image (pooled wins 8/9, mean $\Delta_{\text{acc}}{=}-3.9$pt, $t{=}-2.75$) — exactly the geometry split the principle predicts. Individually, two semantic cells clear $p < 0.05$ under paired permutation; it is the joint evidence that is decisive.

Finally, a sanity control validates the evaluation pipeline behind Table 3: the CNNSpot checkpoint, scored on an untouched GenImage subset (four generators, 1640 images), lands at 0.488 accuracy / 0.451 AUC — near chance, consistent with its widely reported failure to transfer from GAN-era artifacts to diffusion-era generators, and confirming that the low Table 3 entries reflect model limitations rather than an evaluation artifact.

K Multi-Round Local Re-Editing

Section 5.4 frames detection as a race against an erasure channel: each human-in-the-loop edit reruns a bounded diffusion trajectory, and the question is whether repeated re-editing *hides* the edit from a spatial probe. We simulate the loop directly. From GenImage we take 100 untouched real images (25 per generator: ADM, Midjourney, *stable_diffusion_v_1_4*, Wukong), and for $k{=}1, 2, 3$ rounds we re-render the central $256{\times}256$ region with SD-v1.4 inpainting (DDIM, 25 steps, guidance 7.5, neutral prompt "a photograph"), feeding each round's output into the next. Pixel checks confirm the edits stay localized: mean absolute error in the masked region is 55–74 per round versus 2.8–4.4 in the periphery, so the loop changes the bounded area without degrading context.

Table 19 and Figure 4 report the released probe's per-round means. The *windowed* score jumps from 0.18 on the untouched image to 0.94 after a single round (paired $\Delta{=}{+}0.76$, $p{=}8.5{\times}10^{-18}$, 98% of images up), then saturates: rounds $1{\to}2$ move $+0.03$ ($p{=}0.55$, n.s.) and $2{\to}3$ move -0.01 ($p{=}0.11$, n.s.). Of the 94 images detected ($s_{\text{win}}{>}0.5$) at round 1, 93 remain detected at round 3. The *pooled* score rises slowly ($0.03{\to}0.55{\to}0.63{\to}0.71$) and stays well below windowed at every round.

The reading is instructive: re-editing with the *same recognizable generator does not erase the forensic signal* — it re-injects the generator's fingerprint each round, so windowed detection saturates rather than decays. The erasure that closes the detection window is *transmission-level* purification (JPEG, blur; Appendix D) or *switching the editing tool* (Appendix L), not local re-rendering itself. This sharpens Section 5.4: within the window, the spatial probe is the cheapest depth gauge; and the window closes only when the edit leaves the recognizable manifold.

L Cross-Tool Editing Closes the Window

Appendix K shows that re-editing with the same generator saturates the windowed score. The natural escape is to switch tools. We take the 100 round-1 images ($k{=}1$, SD-v1.4 center inpainting)

Table 12: **Fusion over aggregation scales (logistic, calibrated on keling).** AUC.

Score	flux-kontext	BR-Gen cross
pooled	0.928	0.850
windowed	0.937	0.788
max	0.937	0.788
mean	0.937	0.802
fused (logistic, keling-cal.)	**0.937**	0.807

Table 13: **Layer analysis: patch-manifold distance separation (flux-kontext, released backbone).** AUC of per-image mean kNN distance to a real patch bank, edited vs real.

Layer	0	1	2	4	8	12	16	20	24	28	47
AUC	0.536	**0.617**	0.578	0.532	0.459	0.453	0.494	0.496	0.417	0.435	0.470

and re-edit the same 256×256 region with FLUX.1-Fill-dev (fp8 weights, 28 DDIM-like steps, guidance 30, neutral prompt "a photograph"), an architecture distinct from SD-v1.4 and from the probe's release generator. Pixel checks again confirm the edits stay localized. The windowed score collapses from 0.942 to 0.629 (paired $\Delta{=}-0.313$; Wilcoxon $p{=}3.3\times10^{-13}$; 90% of images drop), and pooled falls from 0.545 to 0.316 (Figure 5). The residual signal ($0.63 > 0.5$) is consistent with the FLUX-family results in Section 4.3: the probe detects FLUX-family edits out of the box (0.937 on FLUX-Kontext), so the new tool's own fingerprint is partly visible — but the composite image now mixes two editing fingerprints, and the single-generator hypothesis behind saturation no longer holds. Together with transmission purification, tool switching is the second practical way to close the detection window; both act *outside* the local re-rendering loop, which is why forensics must move early in it.

M Reproducibility

Models. Released MetaCLIP2 probe (frozen *facebook/metaclip-2-worldwide-giant* at 224px + SD1.4-trained linear head); SigLIP2-large for mid-tier comparisons and patch-probe experiments. Preprocessing matches the original: bicubic resize, center crop, CLIP normalization.

Windowed scoring. $s_{\text{win}}(I) = \max_{c\in\mathcal{C}(I)} g(c)$ with $\mathcal{C}(I)$ = center crops at scales $\{0.4, 0.6, 0.8, 1.0\}$; no training.

Patch probing. Per-patch linear head on $[\text{patch token} \,\|\, d_{\text{real}} \,\|\, d_{\text{fake}}]$ with $d.$ = kNN ($k{=}3$) cosine distances to 20k-patch reference banks (300 real flux-dev + 300 fake gpt4o images), trained with pixel-level BCE under per-image class-balanced sampling on composited cut-paste images. Backbones frozen throughout.

Datasets. GenImage (8 generators); AIGI-Now (9 generators $\times$ pixel/semantic); BR-Gen (150k localized forgeries, 5 methods, masks); DDA-COCO (6 VAE variants, exact source real). No target-domain fine-tuning.

Compute. $2\times$RTX 4090, fp16/bfloat16; per-benchmark runtime on the order of tens of minutes.

Code release. All probing/scoring code and per-image scores will be released upon acceptance.

Baseline reproduction. Table 20 reproduces the published benchmark tables of the released MetaCLIP2 probe. All 18 AIGI-Now cells fall within ±0.003 of the paper, and GenImage averages within 0.001, confirming the setup.

Table 14: **Classical high-frequency forensics fail on modern localized edits.** AUC on FLUX-Kontext (n=200/200) and BR-Gen LaMa (n=150/150); hand-crafted residuals with windowed aggregation mirroring our probe. AUC < 0.5 means edited images have *lower* residual energy (over-smoothing) on both benchmarks. Windowed frozen-VFM probe shown for contrast.

Method	FLUX-Kontext	BR-Gen LaMa
Laplacian noise (LNP), windowed	0.142	0.443
Laplacian noise (LNP), pooled	0.141	0.442
Sobel energy, windowed	0.385	0.424
Haar-HH energy, windowed	0.119	0.434
Windowed frozen-VFM probe (ours)	0.937	0.850

Table 15: **Cross-layer patch residuals (P4-style) do not transfer to localized edits.** AUC; fused = released probe + P4-style residual head with β tuned on validation. Training: SD1.4 (2,000+2,000), same split as the released probe.

Variant	FLUX-Kontext pooled	FLUX-Kontext fused	BR-Gen pooled/fused
P4 blocks 8/16/24	0.9281	**0.9193**	0.8495/**0.8398**
P4 early blocks 1/3/5	0.9281	**0.9188**	0.8495/**0.8381**
Windowed (ours, for contrast)	0.9281	0.9366	0.8495/0.7883*

*Windowed lowers BR-Gen AUC: scattered small edits favor global aggregation (Section 3.3); the P4 result differs categorically in that cross-layer residuals do not help on *either* benchmark.

Table 16: **Pooled-feature dilution (flux-kontext, released probe).** Cosine similarity of the pooled feature to the all-patch mean vs the top-15% fake-scored patch mean.

	sim(all-mean)	sim(top-15% fake)	gap
Edited (fake)	0.233	0.100	$+0.133$
Real	0.198	0.056	$+0.142$

Table 17: **Paired significance of aggregation deltas** (windowed minus pooled; 10k bootstrap resamples; two-sided permutation on per-image score swap).

Comparison	Δ	95% CI	perm. p
FLUX-Kontext AUC (n=600)	$+0.008$	$[-0.007, +0.024]$	0.52
FLUX-Kontext TPR@10%FPR	$+0.053$	$[-0.014, +0.104]$	0.11
BR-Gen AUC (n=1000)	-0.060	$[-0.078, -0.044]$	$< 10^{-4}$

Table 18: **BR-Gen crossover analysis: aggregator performance vs true edit-area fraction** (500 forged images binned by mask fraction; shared real pool; released probe).

Mask-fraction bin	n	pooled AUC	windowed AUC	win rate
$[0, 0.05)$	12	0.44	0.37	0.92
$[0.05, 0.15)$	38	0.62	0.55	0.92
$[0.15, 0.30)$	78	0.65	0.55	0.87
$[0.30, 0.50)$	96	0.78	0.66	0.70
$[0.50, 0.75)$	92	0.92	0.86	0.55
$[0.75, 1.00]$	184	0.99	0.98	0.59

Table 19: **Detection across multi-round local re-editing.** Released probe mean scores on 100 real GenImage images re-rendered in the center region for k rounds; paired Wilcoxon versus the previous round.

Round k	Pooled	Windowed	Δ_{win} (p)
0 (untouched)	0.026	0.180	—
1	0.545	0.942	$+0.762$ (8.5×10^{-18})
2	0.627	0.972	$+0.030$ (0.55)
3	0.709	0.961	-0.012 (0.11)

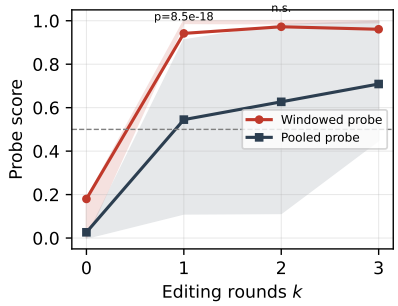

Figure 4: **Multi-round editing saturates windowed detection.** Mean released-probe score vs. editing round k (100 real images, center 256×256 SD-v1.4 inpainting; band = 25–75 percentile). Windowed (max) saturates after one round; pooled stays lower and rises slowly. Dashed line: chance (0.5).

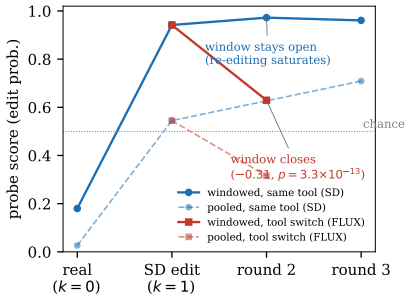

Figure 5: **Tool switching closes the detection window.** Mean released-probe score across editing rounds: same-tool re-editing (SD, blue) keeps the windowed score saturated, whereas switching the second-round editor to FLUX.1-Fill-dev (red) collapses it by 0.31 (paired p=3.3×10^{-13}, 100 real GenImage images). Dashed line: chance (0.5).

Table 20: **Baseline reproduction (released MetaCLIP2 probe, pooled linear head).** Left: GenImage avg accuracy. Right: AIGI-Now avg accuracy (pixel/semantic).

GenImage	Released	Paper	AIGI-Now	Released	Paper
ADM	0.696	0.690	FLUX-dev	0.979/0.940	0.979/0.941
BigGAN	0.804	0.816	FLUX-kera	0.963/0.895	0.963/0.896
GLIDE	0.890	0.887	FLUX-kontext	0.800/0.812	0.799/0.811
Midjourney	0.963	0.959	FLUX-pro	0.977/0.891	0.976/0.892
VQDM	0.815	0.819	gpt4o	0.943/0.887	0.943/0.888
SD-v1.4	0.991	0.993	jimeng	0.966/0.825	0.965/0.825
SD-v1.5	0.989	0.991	keling	0.971/0.902	0.970/0.902
Wukong	0.979	0.980	minimax	0.943/0.849	0.942/0.850
Avg	**0.891**	**0.892**	Nano	0.966/0.817	0.965/0.819
			Avg	**0.907**	**0.907**