BENEATH SEMANTICS: SIGNAL-AXIS RESIDUAL FINGERPRINT FOR AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated image detectors perform well on aggregate benchmarks but exhibit marked declines in performance under distribution shift—whether from new collection pipelines for familiar generators or from entirely unseen generators. We examine image authenticity along two complementary evidence axes: semantic representations, which provide detection cues through pretrained image representations, and signal statistics, which capture local dependencies and spectral organization in residuals. To address both collection shift and unseen-generator shift, we introduce *Signal-axis Residual Fingerprint* (SRF), a compact detector built around a fixed representation of Spatial–Spectral Microstructure. Rather than learning dataset- or generator-specific patterns, SRF uses six fixed resampling operators and an interpretable statistic bank to characterize local dependencies and spectral organization in residuals—traces intrinsic to the generation pipeline rather than specific to a particular collection protocol or generator. The complete five-block configuration, SRF-E, contains only 0.28M fitted parameters, requires no pretrained backbone, and achieves a mean standalone AUROC of 0.9150 across 48 benchmark–generator groups from four benchmark suites. Crucially, its scalar score can be fused with the 12 evaluated frozen-backbone probes to recover performance under both distribution shifts. For the strongest CLIP probe, mean AUROC increases by approximately 2.6 percentage points, worst-group AUROC rises from 0.691 to 0.817, and mean AUROC on the unseen-generator tier improves from 0.855 to 0.899. Controlled transformation experiments show that periodic residual correlations and spectral energy statistics provide processing-related evidence that the current CLIP detection scores do not fully exploit.

1 INTRODUCTION

From generative adversarial networks (Goodfellow et al., 2014; Karras et al., 2017) to diffusion models (Ho et al., 2020; Rombach et al., 2022), generated images have become increasingly realistic, posing challenges to media forensics and platform content safety; rapid advances in generators also make it difficult to keep detectors up to date. Consequently, deployed detectors face two major generalization challenges: *collection shift*, where images from familiar generators are collected through different pipelines, changing their data distributions; and *generator turnover*, where detectors must identify images produced by generators absent from their training data (Yan et al., 2025).

We examine this problem along two evidence axes: semantic representations and signal statistics. The semantic axis provides detection cues through pretrained image embeddings, which encode image content, shape, and contextual features (Radford et al., 2021; Ojha et al., 2023), whereas the signal axis explicitly measures pixel-level dependencies and spectral organization. Even a generated image depicting a plausible scene may exhibit measurable differences in neighboring-pixel correlations or residual spectral distributions: decoding, upsampling, and resampling can alter these statistical structures (Rombach et al., 2022; Odena et al., 2016; Tan et al., 2024), yet such changes may not be fully reflected in pretrained detector scores. Explicit measurements of residual structure therefore have the potential to complement evidence from pretrained detectors.

Low-level cues and their fusion with semantic features have been studied. NPR (Tan et al., 2024) learns upsampling-related neighboring-pixel features, while AIDE (Yan et al., 2025) combines frequency-selected, SRM-filtered patch features with OpenCLIP representations. We investigate

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.

whether fixed, interpretable residual statistics can support standalone detection across generators and collections and improve existing detectors without a pretrained backbone.

To this end, we introduce **SRF** (*Signal-axis Residual Fingerprint*), which measures residuals between images and fixed downsampling–upsampling reconstructions. Six fixed resampling operators span different interpolation kernels and scales. A fixed, interpretable statistic bank summarizes *Spatial–Spectral Microstructure* (SSM)—local statistical dependencies and spectral organization in these residuals. The complete configuration, SRF-E, uses a compact statistical readout with only 0.28M fitted parameters and no pretrained backbone, achieving a mean standalone AUROC of 0.9150 across 48 benchmark–generator groups from four suites. Its scalar output integrates with existing detectors through validation-normalized, equal-weight score fusion, without modifying the semantic partner or training a fusion module (§3).

We partition the 48 evaluation groups relative to our development sources into seen-clean, seen-shifted, and truly-unseen tiers: familiar generators from familiar benchmark releases, familiar generators from other benchmark releases, and generators absent from development. This separates collection shift from generator novelty and identifies where fusion helps.

In addition, to investigate what the residual reader measures, we apply controlled low-level transformations to real photographs. SRF-E responds more strongly than CLIP-LP to periodic perturbations and resampling, supporting the interpretation that measurements of periodic residual correlations and spectral energy distributions supply processing-related evidence that the current CLIP-LP detector scores do not fully exploit. Label-conditioned analysis shows that mean score correlations between SRF-E and the evaluated semantic detectors are lower than those among the semantic detectors themselves, further supporting lower score redundancy between the two types of branches (§5). Our main contributions are:

- **A lightweight detector built on fixed, interpretable residual statistics.** We introduce SRF, which characterizes image microstructure through multi-operator resampling residuals and spatial–spectral statistics. Its complete instantiation, SRF-E, requires only 0.28M fitted parameters and no pretrained backbone (§3).
- **Empirical validation and mechanistic analysis of signal–semantic complementarity.** Plug-and-play score fusion without learned fusion weights validates the incremental value of SRF-E across all 12 evaluated frozen-backbone probes. Controlled transformations and label-conditioned score-correlation analyses provide mechanistic explanations and experimental support for differential responses, lower redundancy, and complementary gains (§4.4, §5).
- **A three-tier evaluation distinguishing collection shift from generator novelty.** Relative to detector-development sources, we divide evaluation groups into familiar sources, images of familiar generators released through other benchmarks, and unseen generators. This enables a more precise assessment of detector generalization and validates the robust generalization of fixed SSM descriptors to unseen generators (§4.3).

2 Related Work

Spatial, frequency, and reconstruction detectors. Spatial and other signal-motivated detectors exploit local generation artifacts but can remain sensitive to the collection pipeline. CNNSpot (Wang et al., 2020) and NPR (Tan et al., 2024) train CNNs on spatial traces of generator upsampling; NPR targets the same physical cue family as SRF but *learns* its extractor, and under our audit it scores 0.993 AUROC on seen-clean groups but 0.516 on the same generators collected through a different pipeline, whereas SRF-E's fixed statistics retain 0.919 (§4.3). Other approaches use frequency spectra (Zhang et al., 2019; Frank et al., 2020; Corvi et al., 2023), steganalysis-style co-occurrence statistics (Fridrich & Kodovsky, 2012; Zhong et al., 2023), or reconstruction error (Ricker et al., 2024); under our protocol, they too degrade sharply on worst-group and truly-unseen metrics (Table 1). Our residual design instead follows the blind-spot intuition (Batson & Royer, 2019)—a resampling reconstruction cannot reuse a pixel's own microstructure—and controlled injections verify that SRF-E responds to multiple forms of SSM associated with generation pipelines rather than a single artifact (§5).

Vision-language and semantic detectors. Semantic detectors capitalize on large pretrained representations: frozen-CLIP probes such as UnivFD (Ojha et al., 2023) exploit image–text pretraining (Radford et al., 2021), adapter and prompt variants such as FatFormer (Liu et al., 2024) and C2P-CLIP (Tan et al., 2025) refine the same semantic axis, and AIDE (Yan et al., 2025) mixes semantic with low-level experts. Such hybrids learn the combination jointly with their features and are evaluated as single models. We instead ask whether a separately built, backbone-free signal-axis score adds to existing semantic detectors without retraining them: with fixed equal-weight fusion, it improves each of 12 frozen backbones, and we compare it against alternative partners under the same rule (§4.4).

Evaluation protocols and leakage audits. Evaluation protocols must distinguish generator novelty from other distribution shifts, yet detection benchmarks can contain duplicates, train–test generator overlap, and format shortcuts (Bammey, 2023; Yan et al., 2025). Our protocol reserves final sets for evaluation, freezes thresholds and score statistics on validation, and includes byte-hash audits (App. A.5); the three-tier partition separates collection-shift robustness from generator novelty under one protocol (§4.3).

3 Reading Spatial–Spectral Microstructure with SRF

Generation pipelines can leave SSM in image residuals: spatial dependencies and spectral organization that provide evidence beyond scene content. SRF reads this structure through six fixed resampling operators and a frozen statistic bank that produces interpretable SSM descriptors, and a linear committee emits one scalar score (Fig. 1). Because the extractor contains no learned parameters, learning is confined to a linear readout over predefined statistics; because the branch exposes a single score, any semantic detector can consume it without being modified.

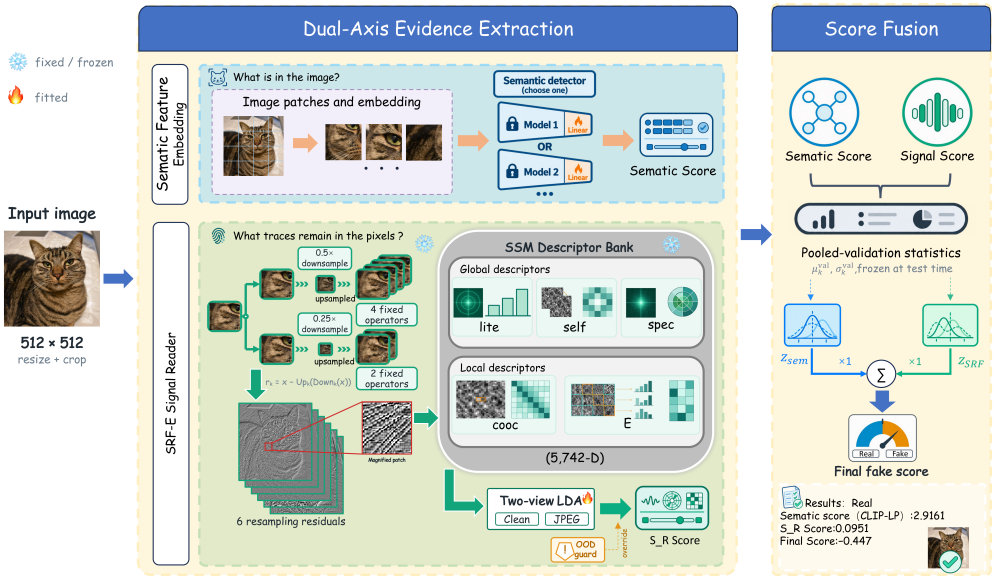

Figure 1: **SRF overview.** *Bottom:* six fixed resampling operators and five statistic blocks produce residual descriptors, scored by a clean/JPEG shrinkage-LDA committee with an OOD guard. *Top:* a frozen-backbone linear probe (CLIP-LP by default) supplies the semantic score. *Right:* validation-standardized scores are summed with equal weights, without learning fusion parameters.

3.1 Preprocessing and Residual Operators

Every input is resized so that its shorter side is 512 pixels, center-cropped to 512^2, and scaled to $[0, 1]$. Rather than estimating artifacts with a learned feature extractor, SRF applies six fixed resampling operators: four downsampling–upsampling pairs at scale 0.5 (nearest–nearest, bilinear–bilinear, bicubic–bicubic, and area–bilinear) and two at scale 0.25 (nearest–nearest and area–

bilinear), and reads the residual of each round trip. Each operator induces

$$r_{\mathrm{op}} \;=\; x - \mathrm{Up}_{\mathrm{op}}\big(\mathrm{Down}_{\mathrm{op}}(x)\big). \tag{1}$$

A channel-averaged residual $\bar{r}_{\mathrm{op}}$ is retained for the spectral blocks. We use fixed resampling reconstructions as references for measuring image microstructure. For a preprocessed image, let $A_k = U_k D_k$ denote a fixed downsampling–upsampling map, yielding the residual $r_k = (I - A_k)x$. This construction cancels components reproduced exactly by the map and bounds the residual of locally smooth components by their variation over the interpolation support. For an additive perturbation δ, the residual change satisfies $r_k(x+\delta) - r_k(x) = (I - A_k)\delta$, enabling direct characterization of responses to local structure under different sampling scales and interpolation rules.

Because downsampling introduces sampling-phase dependence and aliasing, these responses generally cannot be reduced to shift-invariant high-pass filtering. Different kernels therefore provide responses under multiple reconstruction rules and scales (§5). SRF summarizes them with fixed spatial and spectral statistics to exploit distributional differences between real and generated images under a common measurement. The spatial component includes local dependencies and co-occurrences (**self**, **cooc**); the spectral component includes band energies and radial–angular organization (**lite**, **spec**); and **E** measures spatial variation in local spectral statistics. We denote the five blocks in bold throughout.

3.2 Global descriptors

The global blocks describe the residual as a whole: where its energy sits, whether it is periodic, and whether its spectrum carries generator-specific geometry.

lite asks *where residual energy sits*. A two-dimensional FFT of each residual,

$$F = \mathrm{FFT2}(r_{\mathrm{op}}), \qquad P = \log(1 + |F|^2), \tag{2}$$

is partitioned into low, mid, and high radial bands by fixed thresholds on the normalized radius. The block reports per-channel band energies, three band ratios, and spatial moments of $|r|$—a coarse localization of energy that does not commit to a particular spectral geometry.

self asks *whether the residual repeats under small shifts*, the signature of periodic upsampling. It records shifted autocorrelations, sign consistency, and channel coupling over a bank of one- and two-pixel displacements. Camera noise is spatially unstructured: a small shift largely destroys self-similarity. An upconvolution checkerboard, by contrast, produces a regular cadence of agreement and disagreement—the cue to which this block is the dominant responder under controlled injection (§5).

spec asks *whether the spectrum shows rings or directional lines at generator-specific radii*. On the channel-averaged residual it computes a 16-ring + 8-sector FFT descriptor of the spectral component of SSM,

$$\bar{P} = \log\big(1 + \big|\mathrm{FFT2}(\bar{r}_{\mathrm{op}})\big|^2\big), \tag{3}$$

averaging power in sixteen equal radial annuli and, away from the DC neighborhood, in eight angular sectors. Internal resampling of a generator imprints bright rings at characteristic radii and bright lines at characteristic orientations; **spec** is the block that measures this spectral component of SSM.

3.3 Local descriptors

The local blocks look inside the residual: at neighbor co-occurrence, and at whether the microspectrum remains structured after content energy is removed.

cooc asks *whether neighboring residual values co-occur as in camera noise*. Residual values are quantized to seven bins,

$$q = \mathrm{clip}\big(\mathrm{round}(255\, r/2),\, -3,\, 3\big) + 3 \in \{0, 1, \ldots, 6\}, \tag{4}$$

and square-rooted 7×7 co-occurrence histograms are accumulated over four directions on the grayscale and RGB residuals, following a classic steganalysis approach (Fridrich & Kodovsky, 2012). The Hellinger square-root stabilizes the counts. Where **self** detects periodicity, **cooc** detects local texture regularity: pairings of adjacent residual levels that natural sensor noise rarely produces.

E asks *whether the microspectrum is spatially stationary*. The residual is tiled at two grid scales; each tile yields a four-band micro-spectrum. After regressing out local content energy, four statistics—dispersion, Moran's I (Moran, 1950), slope, and R^2—quantify whether the residual field remains structured across the image. Natural photographs are, once content is explained away, spatially unpredictable. Decoder textures leave patches of residual energy that content cannot account for, and Moran's I records whether those patches cluster.

3.4 Linear committee readout

SRF denotes the framework for extracting and reading SSM; **SRF-E** denotes its default complete instantiation, used for every headline result. SRF-E concatenates the global blocks $[\mathrm{lite}, \mathrm{self}, \mathrm{spec}]$ with the local blocks $[\mathrm{cooc}, \mathrm{E}]$. The readout maps the concatenated descriptor to one calibrated score; only linear heads are fitted, so the decision remains an auditable projection rather than an opaque network. Two shrinkage-LDA members (Ledoit & Wolf, 2004) are trained on different domain subsets of the standardized training features, so that no single training domain dictates the projection, and fixed weights combine them. Each member is calibrated only against *real* training projections into a right-tail z-score: the score measures how far an image lies beyond the training-real distribution, a reading whose calibration never references any particular fake generator. Images that look unlike any training real are therefore scored as anomalous regardless of which generator produced them.

Robustness is handled at training time only. A *clean* view and a *JPEG-robust* view both score the same clean test descriptor and are averaged with equal weights. Test images are never recompressed; the robust view is trained on clean training rows plus deterministic JPEG copies of those rows. Finally, a Mahalanobis out-of-distribution switch on the lite block overrides the committee only in a specific contradiction: the LDA predicts "real," yet the residual distribution itself is unlike any training real. This is the failure mode of a novel generator that happens to land on the real side of a linear projection. The operating threshold is selected only on validation. For fusion, the SRF-E score and a frozen semantic detector's score are each standardized with validation statistics and summed with equal weights, without learning fusion parameters (Eq. (5), App. A.1).

4 Experiments

Our experiments address two linked questions: does the signal axis provide forensic evidence on its own, and does it carry evidence that a strong semantic detector still lacks? §4.2 establishes whether the two axes are complementary. §4.3 separates robustness to collection shift from generalization to unseen generators, and locates where the complementarity concentrates. §4.4 tests whether the effect belongs to SRF-E itself or merely to adding a second model. §4.5 reports leave-one-block-out refits and an OOD-switch ablation, assessing component contributions to both standalone and fused detection. Appendix F.3 tests whether the design withstands post-processing and yields a usable operating point. §5 then explains why the two axes complement each other. Unless stated otherwise, all datasets, tiers, baselines, metrics, and evaluation protocols follow App. A.

4.1 A Qualitative View of Spatial–Spectral Microstructure

Figure 2 qualitatively shows the residual structure analyzed by SRF. In this pair, the generated residual is concentrated around contours, while the real residual contains more diffuse background fluctuations. Their spectra also differ: the generated example exhibits more concentrated axis-aligned energy, whereas the real example shows a broader distribution. These observations motivate SRF's spatial and spectral statistics for characterizing SSM beyond visual appearance. The following experiments evaluate how these residual statistics contribute to detection and complement semantic evidence.

4.2 Main comparison

Table 1 first compares the individual detectors. Here CLIP-LP denotes our default semantic branch, a frozen OpenCLIP ViT-L/14 encoder with a logistic-regression head fitted on our development pool following the UnivFD linear-probe recipe (Ojha et al., 2023) (App. A.1); UnivFD denotes the

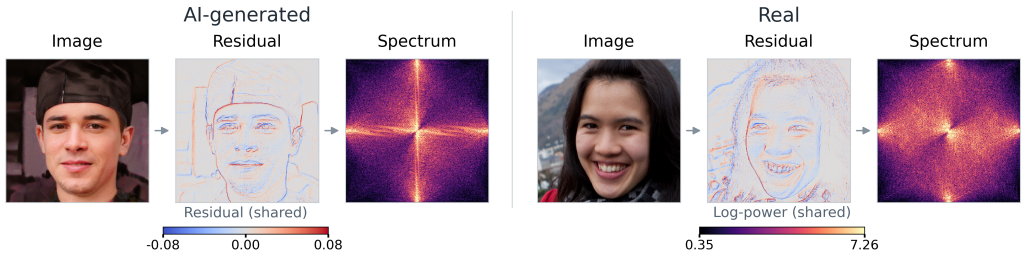

Figure 2: **Illustrative residual structure.** A generated (left) and a real (right) ForenSynths/WhichFaceIsReal face with their residuals and log-power spectra under the same nearest-neighbor $0.5\times$ down/up operator; display scales are shared.

officially released checkpoint. Among the individual detectors, CLIP-LP achieves the highest mean AUROC (0.9363) and worst-group AUROC (0.691), while C2P-CLIP obtains the highest mean AP (0.8016). Despite using only 0.28M fitted parameters and no pretrained backbone, SRF-E ranks third among the individual detectors in both mean AUROC (0.9150) and mean AP (0.7069), and second in worst-group AUROC (0.621).

Table 1: **Overall and tier-wise detection performance.** AUROC and AP are averaged equally over the relevant benchmark–generator groups (counts in parentheses); Worst is the minimum group AUROC. Tiers are defined relative to our development sources, not to each baseline's training data (App. A.1). Blocks separate reference detectors, the two fusion components, and score-fusion systems; CLIP-LP is our frozen CLIP ViT-L/14 linear probe and UnivFD is the officially released checkpoint (App. A.1); in fusion names, C2P is C2P-CLIP. Stored params count checkpoint parameters (App. E). Fusion rows sum the checkpoint sizes of their components. Bold method names identify ours; bold metric values mark column maxima; shading highlights our fusion.

Method	Overall (48 groups) ↑			Tier-wise mean AUROC ↑			Params (M)
	Mean AUROC	Mean AP	Worst AUROC	Seen-clean (19)	Seen-shifted (15)	Truly-unseen (14)	
AEROBLADE	0.6670	0.4099	0.268	0.668	0.722	0.607	151.5
NPR (retrained)	0.7371	0.5419	0.170	**0.993**	0.516	0.627	1.4
CNNSpot	0.7452	0.5629	0.337	0.824	0.807	0.573	23.5
PatchCraft	0.7748	0.5937	0.049	0.906	0.823	0.545	0.12
AIDE (ProGAN)	0.7563	0.5706	0.252	0.917	0.766	0.529	898
AIDE (GenImage)	0.7920	0.5777	0.327	0.957	0.766	0.597	898
FatFormer	0.8304	0.6779	0.372	0.944	0.824	0.684	492.7
UnivFD	0.8353	0.6717	0.361	0.926	0.882	0.661	428
C2P-CLIP	0.9190	0.8016	0.505	0.978	0.921	0.837	304
SRF-E (ours)	0.9150	0.7069	0.621	0.966	0.919	0.842	0.28
CLIP-LP	0.9363	0.7905	0.691	0.982	0.955	0.855	428
CLIP-LP ⊕ C2P	0.9545	**0.8437**	0.681	0.992	0.969	0.888	732
SRF-E ⊕ CLIP-LP (ours)	**0.9622**	0.8336	**0.817**	**0.993**	**0.982**	**0.899**	428.3

The main result of Table 1 is that combining the two evidence axes substantially improves the default semantic detector. Adding SRF-E raises the mean AUROC of CLIP-LP from 0.9363 to 0.9622, a gain of 0.0259. The paired group-bootstrap confidence interval lies entirely above zero (95% CI $[+0.0161, +0.0361]$, $p < 10^{-4}$). Fusion also increases the worst-group AUROC from 0.691 to 0.817, substantially improving coverage of the most difficult group. Section 4.3 further shows that the gain progressively increases from seen-clean to seen-shifted and truly-unseen generators.

We additionally compare against CLIP-LP ⊕ C2P, an ensemble of two semantic detectors. The two systems achieve similar mean AUROC, with SRF-E ⊕ CLIP-LP being numerically higher (0.9622 vs. 0.9545). CLIP-LP ⊕ C2P obtains slightly higher mean AP (0.8437 vs. 0.8336), whereas SRF-E ⊕ CLIP-LP achieves substantially higher worst-group AUROC (0.817 vs. 0.681). Two semantic models therefore yield a higher mean AP, whereas a second *axis* improves worst-group coverage:

SRF-E $\oplus$ CLIP-LP exceeds CLIP-LP $\oplus$ C2P by 13.6 percentage points of worst-group AUROC while adding 0.28M parameters instead of a 304M image encoder.

4.3 Generalization under collection and generator shifts

The tier-wise columns of Table 1 distinguish two generalization challenges that are obscured by aggregate results; the complete tier construction is provided in App. A. Seen-shifted therefore primarily measures sensitivity to collection changes, whereas truly-unseen evaluates generalization to new generator families.

Collection shift has a particularly strong effect on several artifact-oriented and hybrid detectors. NPR drops from 0.993 AUROC on seen-clean groups to 0.516 on seen-shifted groups, while PatchCraft and the two AIDE checkpoints lose between 8.3 and 19.1 percentage points of AUROC. SRF-E retains 0.919 under the same shift (from 0.966), the highest seen-shifted AUROC among signal-oriented and hybrid detectors. Its drop of 4.7 percentage points is comparable to those of the semantic detectors UnivFD (4.4) and C2P-CLIP (5.7); CNNSpot drops less and AEROBLADE even improves, but both start from much lower seen-clean levels (0.824 and 0.668). The contrast with NPR, which targets the same cue family but learns its extractor, is consistent with SRF-E confining learning to a linear readout over fixed statistics.

Generalization to truly-unseen generators remains the main challenge for all individual detectors. Every individual model performs worse on truly-unseen than on seen-clean groups, and even the strongest individual detector, CLIP-LP, reaches only 0.855 AUROC. In contrast, the benefit of fusion increases with the severity of the distribution shift: relative to CLIP-LP, SRF-E $\oplus$ CLIP-LP gains 1.1, 2.7, and 4.5 percentage points of AUROC on seen-clean, seen-shifted, and truly-unseen groups, respectively. The fused model improves 13 of the 14 truly-unseen groups and achieves the best reported performance on the seen-shifted and truly-unseen tiers, reaching 0.982 and 0.899, respectively. These results show that the complementary signal branch becomes increasingly valuable as the test distribution moves further from the training distribution.

4.4 Generality and specificity of signal–semantic complementarity

This section first examines a central question: whether the gains from SRF-E are specific to the default CLIP-L/14 probe (CLIP-LP) or generalize across different types of pretrained representations. We evaluate CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2023), supervised ViT (Dosovitskiy et al., 2020), ConvNeXt (Liu et al., 2022b), Swin-V2 (Liu et al., 2022a), and ResNet (He et al., 2016); the shared linear-probe and fusion protocol is detailed in App. A. Adding SRF-E improves the mean AUROC of every backbone (Fig. 3(a)), with gains ranging from 0.026 to 0.177. Weaker standalone probes tend to gain more, so the fused AUROCs cluster within 0.904–0.962 even though the standalone AUROCs span 0.727–0.936. On the 14 truly-unseen groups, SRF-E also yields positive gains for every backbone, ranging from 0.045 to 0.194; for 10 of the 12 backbones, the gain on truly-unseen groups exceeds the corresponding mean-AUROC gain. These results show that the effectiveness of SRF-E does not depend on a particular CLIP architecture or pretraining objective.

We next examine whether these improvements arise merely from the ensemble effect of adding a second detector. Fig. 3(b) reports the fusion-partner experiment specified in App. A. Across the six partners in the matrix, SRF-E provides the largest gain for every backbone and is the only one that improves all 12 backbones. The distinction is particularly clear for the strongest CLIP-L/14 probe (CLIP-LP): SRF-E increases its mean AUROC by 0.026, whereas each of the five reference partners reduces it by 0.002–0.027. A second detector can still help when it is strong: C2P-CLIP, which is not part of the matrix, raises CLIP-LP by 0.0183 (Table 7). This gain is numerically smaller than that of SRF-E, although the mean-AUROC difference is not significant (+0.0076, 95% CI $[-0.0032, +0.0186]$). Fusion gains therefore depend on which evidence the partner adds, not merely on adding a model. Section 5 examines the underlying mechanism.

4.5 Descriptor and OOD-Guard Ablations

Table 2 reports descriptor-removal refits and a guard-only ablation. Each descriptor ablation removes the named block and refits the readout; fused results use the same validation-normalized, equal-weight fusion with CLIP-LP. Removing lite also disables its dependent OOD guard, whereas w/o

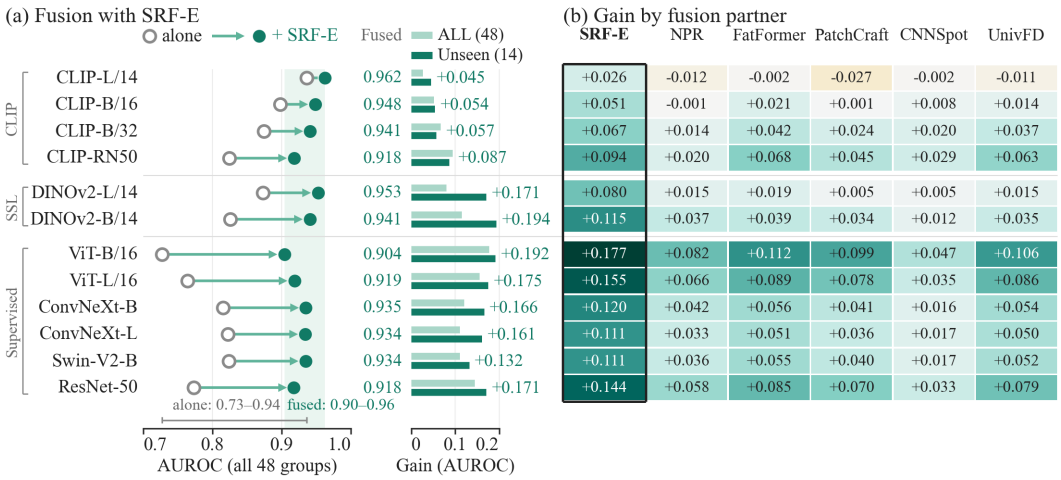

Figure 3: **Cross-backbone fusion with SRF-E. (a)** Twelve frozen-backbone linear probes alone (open circles) and fused with SRF-E (filled); AUROC is the 48-group mean. Bars are raw AUROC gains on all 48 groups and on the 14 truly-unseen groups. Every 48-group gain has group-bootstrap $p < 10^{-4}$ and Holm–Bonferroni-corrected $p \leq 0.0024$; every unseen gain has $p \leq 10^{-4}$. **(b)** Raw AUROC gain of each backbone–partner fusion. SRF-E (boxed) gives the largest gain for every backbone, 1.4–4.3× the strongest positive alternative where one exists (computed from unrounded gains), and is the only partner in the matrix that improves CLIP-L/14. The CLIP-L/14 row is CLIP-LP; the UnivFD column is the official UnivFD checkpoint. C2P-CLIP is evaluated as a CLIP-LP partner in Table 1.

Table 2: **Descriptor and OOD-guard ablations.** Metrics follow Table 1; Fused is equal-weight score fusion with the default CLIP-LP probe. Removing lite also disables the OOD guard. Bold marks column maxima.

Variant	Dim.	OOD	Mean AUROC ↑	Unseen AUROC ↑	Worst-group AUROC ↑	Fused AUROC ↑
Full	5742	on	**0.9150**	0.8419	0.6214	**0.9622**
w/o lite (strict)	5616	off	0.9085	**0.8480**	0.5569	0.9621
w/o self	5166	on	0.8972	0.8092	0.5791	0.9548
w/o spec	5598	on	0.9107	0.8401	0.6202	0.9611
w/o E	5550	on	0.9139	0.8421	**0.6350**	0.9618
w/o cooc	1038	on	0.8785	0.7757	0.4902	0.9488
w/o OOD	5742	off	0.9048	0.8445	0.3914	0.9614

OOD retains all descriptors. Among variants retaining the guard, removing cooc or self causes the largest losses on all four metrics, whereas removing spec or E has small, metric-dependent effects. Removing the lite-dependent pathway (strict w/o lite) reduces mean and worst-group AUROC but slightly improves unseen AUROC, with negligible change in fusion. Enabling the guard raises mean AUROC from 0.9048 to 0.9150 and worst-group AUROC from 0.3914 to 0.6214, while slightly reducing unseen AUROC and improving fused AUROC by only 0.0008.

5 Signal–Semantic Complementarity Analysis

To analyze the fusion gains observed above, we further examine the two branches' sensitivity to controlled low-level transformations, as well as label-conditioned score correlations. Additional qualitative examples of the complementary decisions made by the signal and semantic axes are provided in App. F.1.

Controlled intervention analysis. The complete sampling and transformation protocol is provided in App. A.4. We apply VAE encode–decode, bicubic resampling, and checkerboard perturbation separately to the same 800 real photographs (Fig. 4), as proxies for decoding, resampling, and upsampling traces. All three transformations are common components of contemporary image-

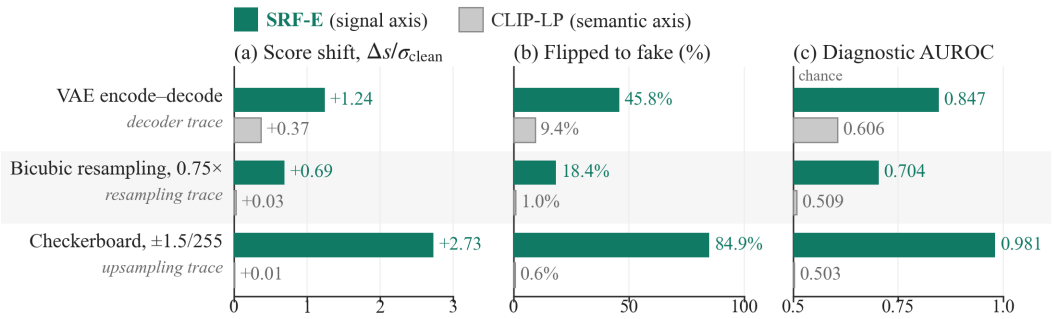

Figure 4: **Sensitivity to controlled processing.** Three transformations are applied separately to the same 800 real photographs. **(a)** Mean paired score change in units of each detector's clean-score SD; **(b)** share of photographs flipped from real to fake at each detector's operating threshold (App. C); **(c)** AUROC separating transformed from original photographs (not benchmark detection AUROC).

generation pipelines, so the traces they induce are widely present in practical generation pipelines; the paired transformations also largely preserve scene content, allowing us to compare the two branches' sensitivity to low-level processing.

Across the three transformations, the normalized mean score change of SRF-E ranges from $+0.69\sigma$ to $+2.73\sigma$, whereas that of CLIP-LP ranges only from $+0.01\sigma$ to $+0.37\sigma$ (Fig. 4(a)). At the operating threshold, SRF-E flips 18.4%–84.9% of the photographs to fake, compared with at most 9.4% for CLIP-LP (Fig. 4(b)). The difference is pronounced when distinguishing transformed photographs from their unmodified originals (Fig. 4(c)): under resampling and checkerboard injection, CLIP-LP attains AUROCs of only 0.509 and 0.503, near chance, whereas SRF-E reaches 0.704 and 0.981. The VAE round trip also affects CLIP-LP, but its response remains weaker than that of SRF-E.

Fig. 6 in App. F further shows that SRF-E responds at every tested resampling strength and already exhibits a large score change at the smallest checkerboard amplitude, whereas CLIP-LP remains near chance at every checkerboard amplitude. Weighted LDA-projection attribution indicates that self-correlation dominates the checkerboard response, that self-correlation and band energy primarily contribute to the VAE response, and that the spectral descriptor of SSM primarily contributes to the resampling response. These differential responses are consistent with the interpretation that SRF-E captures processing-related SSM that CLIP-LP does not fully exploit.

Label-conditioned score correlations. Fig. 7 in App. F reports label-conditioned Spearman score correlations averaged over the canonical 48 groups (protocol in App. A.4). Semantic detectors have a mean correlation of approximately 0.35 with one another, including 0.36 for CLIP-LP–C2P and 0.48 for FatFormer–C2P; in contrast, SRF-E has a mean correlation of approximately 0.11 with the semantic detectors.

Together with the fusion results in Table 1 and the partner comparison in Fig. 3(b), the lower correlation supports lower score redundancy between the two branches and further illustrates the complementarity between SRF-E and semantic detectors.

6 Conclusion

We introduced SRF, a detector that measures spatial–spectral microstructure through fixed resampling residuals. Its complete instantiation, SRF-E, uses 0.28M fitted parameters without a pretrained backbone. Validation-normalized, equal-weight score fusion improves mean AUROC across all 12 frozen-backbone probes. For CLIP-LP, fusion raises mean AUROC from 0.9363 to 0.9622 and worst-group AUROC from 0.691 to 0.817. Controlled transformations and label-conditioned correlations support processing sensitivity and reduced score redundancy between branches.

Our three-tier evaluation distinguishes collection shift from generator novelty, revealing larger fusion gains over CLIP-LP on unseen generators. These findings support fixed residual statistics as a compact complement to pretrained representations. Limited low-FPR separation on unseen generators and sensitivity to strong post-processing remain challenges, motivating improvements in low-FPR detection and reliability-aware fusion.

AI USE STATEMENT

In this work, we used generative AI tools to discuss the interpretation of evaluation metrics. Additionally, we used generative AI tools for language editing and polishing, searching for and identifying relevant literature, drafting and revising portions of the manuscript, and refining the presentation of experimental findings. The authors have reviewed and verified all AI-assisted content, including references. We take full responsibility for the accuracy, originality, and integrity of the final manuscript, its scientific claims, and reported results, including text and claims produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

One canonical grouping library recomputes all reported metrics from frozen per-image score CSVs. The default SRF-E $\oplus$ CLIP-LP protocol separates development fitting and validation-based selection from final full-pool refitting, as specified in App. A.1. The three-tier identity map, leakage and deduplication audit scripts, injection assets, per-group score tables, and evaluation code form a frozen evaluation protocol, which will be released with the code. App. D specifies operators, descriptor dimensions, and committee settings; App. C lists cross-axis normalization statistics and operating thresholds. App. B specifies the statistical procedures, and App. A.1 documents the tier assignment.

ETHICS STATEMENT

This work targets *defensive* media forensics: detecting AI-generated images to support platform integrity and scientific trust. Such detectors carry two risks. False positives can wrongly discredit authentic images. Deployments should therefore treat scores as triage signals, not verdicts. All experiments use publicly released benchmarks under their licenses; we collect no new personal data, and face-containing groups originate from those benchmarks. The injection study modifies only pixel statistics of public photographs and creates no deceptive content. We will release audits rather than generation tools. Generator developers could nevertheless use descriptions of discriminative SSM adversarially, a trade-off shared by published forensic cues.

REFERENCES

Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2023.

Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International conference on machine learning*, pp. 524–533. PMLR, 2019.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pp. 837–845, 1988.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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.

Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on information Forensics and Security*, 7(3):868–882, 2012.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12009–12019, 2022a.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11966–11976. IEEE, 2022b.

Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

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

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8245–8257. IEEE, 2025.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

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.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9130–9140. IEEE, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Chuangchuang Tan, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28130–28139, 2024.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. 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.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8692–8701. IEEE, 2020.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, volume 2025, pp. 70702–70720, 2025.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–6. IEEE, 2019.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in neural information processing systems*, 36:77771–77782, 2023.

A Experimental Settings

This section gives the protocols for the main comparison, fusion studies, descriptor ablations, robustness evaluation, and mechanism analyses, followed by the data audit. Unless specified otherwise, experiments use the canonical 48-group evaluation and the equal-weight standardized fusion in Eq. (5). Experiment-specific sampling and aggregation are stated below.

A.1 Main Comparison and Three-Tier Evaluation

Evaluation data and grouping. We evaluate on 48 benchmark–generator groups from GenImage-8 (Zhu et al., 2023), a 13-group ForenSynths subset (Wang et al., 2020), AIGCDetectBenchmark (Zhong et al., 2023), and Community Forensics (Park & Owens, 2025). Each group pairs a benchmark collection with a generator. Table 3 gives the tier composition, and Table 4 gives the sample counts. The tiers are defined relative to our detector-development sources and are fixed across methods; for checkpoints with different training histories, their names denote benchmark subsets rather than detector-specific exposure.

Table 3: **Composition of the canonical evaluation.** The tiers are defined relative to our detector-development sources; all methods are evaluated on the same groups.

Benchmark	Groups	Seen-clean	Seen-shifted	Truly-unseen
GenImage-8	8	8	0	0
ForenSynths subset	13	11	0	2
AIGCDetectBenchmark	17	0	15	2
Community Forensics	10	0	0	10
Total	48	19	15	14

Table 4: **Sample counts of the canonical evaluation.** Group-level counts describe the real and generated images used to evaluate each benchmark–generator group. File counts count a shared real-image pool once within each benchmark; they do not represent content deduplication.

Benchmark	Groups	Real/ group	Generated/ group	Real files*	Generated files*	Total files*
GenImage-8	8	500	500	500	4,000	4,500
ForenSynths subset	13	Table 8		11,399	11,399	22,798
AIGCDetectBenchmark	17	5,000	1,000†	5,000	16,198	21,198
Community Forensics	10	5,000	500	5,000	5,000	10,000
Total	48	—	—	21,899	36,597	58,496

**GenImage-8 reuses the same pool of 500 real images across its eight groups. AIGCDetectBenchmark and Community Forensics each reuse a pool of 5,000 real images across their respective groups. Consequently, group-level sample counts must not be summed to obtain the number of distinct evaluation files. The total of 58,496 is the sum of the benchmark-level file counts after accounting for this reuse; it is not a count of content-distinct images across or within benchmarks. Exact-overlap and perceptual-hash audits are reported separately in Appendix A.5.*
†AIGCDetectBenchmark generated counts are 1,000 per group except ProGAN (200) and WhichFaceIsReal (998). ForenSynths subset real and generated counts vary by group and are listed in Table 8.

Generator identities and tier assignment. Generator identity distinguishes seen from unseen generators, while the benchmark release separates seen-clean from seen-shifted groups. Seen-clean comprises the eight GenImage groups and 11 ForenSynths groups: their benchmark releases are represented in development, and their images are disjoint from the development pool. Seen-shifted comprises 15 AIGCDetectBenchmark groups whose generators are represented in development but whose images are distributed through a different benchmark release. The 14 truly-unseen groups comprise DALL-E 2 and SDXL from AIGCDetectBenchmark, SAN and SeeingDark from ForenSynths, and the ten Community Forensics groups.

The AIGCDetectBenchmark dataset card maps its anonymized groups to generator identities, which we check against per-image extraction manifests containing 125,026 entries. The identities are ADM, BigGAN, CycleGAN, DALL-E 2, GauGAN, GLIDE, Midjourney, ProGAN, SD1.4, SD1.5, SDXL, StarGAN, StyleGAN, StyleGAN2, VQDM, WhichFaceIsReal (StyleGAN faces), and Wukong. DALL-E 2 and SDXL are assigned to the unseen tier; the remaining 15 groups form the shifted tier.

Detectors and development protocol. We compare SRF-E with CNNSpot (Wang et al., 2020), NPR (Tan et al., 2024), PatchCraft (Zhong et al., 2023), FatFormer (Liu et al., 2024), AEROBLADE (Ricker et al., 2024), the officially released UnivFD checkpoint (Ojha et al., 2023), our CLIP-L/14 linear probe CLIP-LP, C2P-CLIP (Tan et al., 2025), and the ProGAN- and GenImage-trained AIDE checkpoints (Yan et al., 2025). The two fusion systems are CLIP-LP $\oplus$ C2P and SRF-E $\oplus$ CLIP-LP. The three-tier NPR analysis uses our retrained NPR.

For the default SRF-E $\oplus$ CLIP-LP system, the development pool contains 62,814 images: 14,000 from seven GenImage sources (ADM, GLIDE, VQDM, Midjourney, Wukong, SD1.4, and SD1.5), 6,000 from ProGAN, and 42,814 from ForenSynths-all, a sample of the ForenSynths release drawn from ten generator folders (ProGAN, StyleGAN, StyleGAN2, BigGAN, CycleGAN, StarGAN, GauGAN, CRN, IMLE, and Deepfake). Selecting every fifth entry within each source/class bucket gives 12,564 development-validation images and leaves 50,250 fitting images. Development models are fitted on the latter subset. The semantic branch, CLIP-LP, follows the UnivFD linear-probe recipe (Ojha et al., 2023) and uses a frozen OpenCLIP ViT-L/14 encoder, L2-normalized 768-dimensional embeddings, and a logistic-regression head (LBFGS, $C = 1.0$, at most 1,000 iterations). For final evaluation, the CLIP-LP head is refitted on the full development pool and each SRF member on its prescribed full source subsets, including their validation images. The SRF refit re-estimates the LDA heads, feature-standardization statistics, training-real reference calibrators, and lite-based OOD model and its calibration. The component definitions and committee weights are given in App. D.

SRF preprocessing and training views. For SRF, each input is resized to a 512-pixel shorter side, center-cropped to 512^2, and scaled to $[0, 1]$. The clean view is fitted on the prescribed fitting rows; the JPEG-robust view also includes deterministic JPEG copies of those rows, with quality $q = 40 + (\mathrm{crc32}(\mathrm{path}) \bmod 56) \in [40, 95]$. The corresponding full source subsets are used at final refitting. Both views score the same input and their outputs are averaged with weights 0.5/0.5. The default canonical evaluation does not add test-time augmentation or recompression; post-processing conditions are evaluated separately in App. A.3.

Fusion and reporting. The fused score is the equal-weight sum of validation-standardized branch scores,

$$s_{\mathrm{fuse}}(x) = \frac{s_{\mathrm{SRF}}(x) - \mu_{\mathrm{SRF}}}{\sigma_{\mathrm{SRF}}} + \frac{s_{\mathrm{sem}}(x) - \mu_{\mathrm{sem}}}{\sigma_{\mathrm{sem}}}, \tag{5}$$

where s_{sem} is the score of the frozen semantic detector, which is never modified; no fusion parameter is learned, and the same rule applies to every semantic partner. For the default system, branch-score means and population standard deviations are estimated from development-model scores on the pooled validation subset. They are transferred unchanged to final-model scores, which are standardized and summed with both coefficients fixed to one. The cross-axis coefficients are separate from the SRF member weights in App. D and the clean/JPEG view weights above.

Mean AUROC and mean AP are unweighted means of the corresponding group-wise metrics; worst-group AUROC is the minimum group-wise AUROC. Tier and per-benchmark results restrict the same aggregation to their constituent groups, without refitting for individual subsets. For paired fusion comparisons, we resample the evaluation groups with replacement 10,000 times and report the mean gain and percentile 95% confidence interval. The 14-group analysis uses the same procedure on that subset. Complementary pooled DeLong tests and the full statistical reporting procedure are given in App. B.

A.2 Cross-Backbone and Alternative-Fusion Experiments

Backbone transfer. We evaluate 12 frozen representations: CLIP ViT-L/14, ViT-B/16, ViT-B/32, and RN50 (Radford et al., 2021); DINOv2 ViT-L/14 and ViT-B/14 (Oquab et al., 2023); supervised ViT-B/16 and ViT-L/16 (Dosovitskiy et al., 2020); ConvNeXt-B and ConvNeXt-L (Liu et al., 2022b); Swin-V2-B (Liu et al., 2022a); and ResNet-50 (He et al., 2016). Each representation uses the shared linear-probe protocol. Its scalar score is combined with SRF-E using validation-normalized equal-weight fusion. We report standalone and fused mean AUROC on the 48 groups, and the fusion gain on the 14-group unseen subset. The family of 12 backbone comparisons is evaluated with Holm–Bonferroni correction as detailed in App. B.

Fusion-partner comparison. For each backbone, we repeat score-level fusion with the six candidate partners in the fusion-partner matrix. Each backbone–partner pair uses its validation score statistics and equal cross-axis coefficients. Comparing these pairs on the same test groups assesses how gains vary with the added detector under a common fusion rule.

Learned fusion alternatives. The learned score-level and representation-level alternatives use leave-one-benchmark-out evaluation. In each fold, the fusion model is fitted on the other three benchmark suites and evaluated on the held-out suite. The alternatives are logistic fusion of detector scores, a probe combining the CLIP ViT-L/14 embedding with the scalar SRF-E score, and a probe combining the CLIP ViT-L/14 embedding with the raw residual descriptors. Their results are reported separately from the fixed standardized-sum protocol in App. C.

A.3 Post-Processing Robustness and Operating Points

Post-processing. We use a fixed class-balanced subset of 3,200 images: 400 real and 400 generated images from each benchmark suite, with generated images stratified by generator. All methods use the same image IDs. The 17 perturbations in Table 5 are applied to both classes. Fusion retains the original validation normalization statistics and equal weights under every condition.

Table 5: **Post-processing robustness conditions.** The 17 perturbations are applied to both real and generated images.

Perturbation family	Settings
JPEG compression	quality $\in \{95, 85, 75, 60, 40\}$
WebP compression	quality $\in \{75, 50\}$
Resize	scale $\in \{0.75, 0.5\}$
Gaussian blur	$\sigma \in \{0.5, 1.0\}$
Center crop	retained scale $\in \{0.875, 0.75\}$
Gaussian noise	$\sigma \in \{2, 5\}$
Gamma adjustment	$\gamma \in \{0.8, 1.2\}$

For each condition, AUROC is computed by pooling the selected images within each benchmark, then averaged equally over the four benchmarks. The unperturbed condition uses the same subset and aggregation. Thus, these values are four-benchmark means on the robustness subset, while the main comparison reports 48-group means.

Operating points and calibration. Operating-point evaluation uses the canonical test groups. SRF-E and fusion thresholds are selected on development validation by maximizing mean balanced accuracy over nine equally weighted source groups: seven GenImage sources, ProGAN, and pooled ForenSynths-all. The search evaluates 199 score-quantile candidates from the 0.5th to the 99.5th percentile. These thresholds are transferred unchanged to final-model scores. App. C lists the applied thresholds, including the separate CLIP-LP operating threshold.

BalAcc, Acc, and F1 are unweighted means over the 48 test groups at the applied thresholds. Gap compares their mean BalAcc with that obtained at a single global threshold selected retrospectively on the test scores. ECE is computed over pooled test images after applying Platt mappings fitted to development-model validation scores. Separately, TPR at 1% FPR uses each group's test-real score quantile as its threshold, measuring low-FPR separability. App. C reports both evaluations.

A.4 Controlled Transformations and Score Analysis

Interventions. We sample 800 real photographs, 200 from each benchmark suite, and apply the three transformations in Table 6 separately to the same photographs. SRF-E and CLIP-LP score the original and transformed images.

Table 6: **Controlled low-level transformations.** Each transformation is applied to the same source photographs.

Transformation	Setting	Corresponding trace
VAE encode–decode	Stable Diffusion VAE round trip	VAE decoding
Bicubic resampling	down–up at scale 0.75	cross-scale resampling
Checkerboard	amplitude $\pm 1.5/255$	periodic upsampling

The normalized score change is the mean paired score difference divided by the standard deviation of that detector's scores on the original photographs. Flip is the percentage of photographs whose prediction changes from real to fake under each detector's applied operating threshold. Diagnostic AUROC measures discrimination between transformed photographs and unmodified real photographs.

Dose response and attribution. The dose-response study uses the same 800 photographs under the resampling, checkerboard, and VAE settings shown in Fig. 6. Block attribution decomposes the paired change in the weighted committee LDA projection, before empirical calibration and the OOD override. A block's signed share is its mean projection contribution divided by the mean total projection change across the 800 pairs. This attributes the projection response rather than the final nonlinear detector score.

Score correlations. For each detector pair, we compute Spearman correlations separately on real and generated images within each canonical group. We average the two class-conditioned correlations equally, then average across the 48 groups. Conditioning on class separates within-class score agreement from the common real-versus-generated ordering.

Qualitative visualization. The two-axis plot uses 9,000 evaluation images sampled with seed 0, with final SRF-E scores on the horizontal axis and CLIP-LP scores on the vertical axis. Dashed lines mark the applied branch thresholds, and five examples are linked to their positions. For display, we use the RGB-mean nearest-neighbor 0.5-scale residual and stretch its contrast using the 98th percentile of its absolute magnitude. The displayed patch is center-cropped without the descriptor pipeline's resizing; this visualization does not change detector scores.

A.5 Leakage and Deduplication Audit

Audit procedure. For the default SRF-E $\oplus$ CLIP-LP system, the exact-overlap audit compares each canonical benchmark against the complete 62,814-image development pool, including development-validation images. We also inspect exact duplicates within the evaluation collections and perceptual-hash candidate near-duplicates. Duplicate-removal sensitivity is assessed by comparing group-mean AUROC before and after filtering.

Canonical evaluation. The four canonical benchmarks contain no byte-identical matches to the development pool. In the duplicate-removal sensitivity analysis, filtering internal duplicates and perceptual near-duplicates changes mean AUROC by less than 0.004 for all headline systems.

Additional dataset checks. On full ForenSynths, beyond the canonical subset, 57% of raw images are byte-identical to common training data. Removing these matches leaves 38,735 images, on which fusion achieves mean AUROC 0.9733 versus 0.9579 for CLIP-LP, with a worst-group gain of 0.080. We exclude the auxiliary Synthbuster (Bammey, 2023)/COCO (Lin et al., 2014) construction from the main evaluation because file-size and format metadata separate its classes at AUROC 1.0, indicating a confound in this construction.

B Significance Methodology and Full Comparison Table

Procedures. Paired group bootstrap: 10,000 resamples of the 48 groups (14 for the truly-unseen split), reporting mean gain, percentile 95% CI, and $p(\text{gain} \leq 0)$, fixed seed. Paired group bootstrap is the primary analysis for group-mean AUROC gains. Pooled DeLong (DeLong et al., 1988) tests assess pooled-image AUROC, a different aggregation that also includes comparisons across groups. The 12-backbone family is Holm–Bonferroni corrected (Holm, 1979) (all corrected $p \leq 0.0024$).

Table 7: **Validation-frozen fusion comparisons.** AUROC gain is the raw difference of mean AUROC (fusion minus base), not percentage points, under the aggregation in the second column. The 95% CI and bootstrap p come from 10,000 paired group resamples. DeLong z tests pooled-image AUROC, a separate aggregation from the group-mean gain.

Comparison (fusion − base)	Aggregation	AUROC gain	95% CI	boot. p	DeLong z
SRF-E ⊕ CLIP-LP − CLIP-LP	48-group mean	+0.0259	$[+0.0161, +0.0361]$	$< 10^{-4}$	32.7
SRF-E ⊕ CLIP-LP − CLIP-LP	14-group mean	+0.0449	$[+0.0220, +0.0678]$	$< 10^{-4}$	16.7
CLIP-LP ⊕ C2P − CLIP-LP	48-group mean	+0.0183	$[+0.0126, +0.0252]$	$< 10^{-4}$	18.8
SRF-E ⊕ CLIP-LP − CLIP-LP ⊕ C2P	48-group mean	+0.0076	$[-0.0032, +0.0186]$	0.082	19.7

For SRF-E ⊕ CLIP-LP minus CLIP-LP ⊕ C2P, the 95% confidence interval for the 48-group mean AUROC gain includes zero.

Variation across evaluation groups. Fusion improves over CLIP-LP in 42 of the 48 groups and yields lower AUROC in six, including ForenSynths/WhichFaceIsReal ($0.9562 \rightarrow 0.9033$) and Community Forensics/DFGAN ($0.9840 \rightarrow 0.9437$). On the 14 unseen groups, 13 improve and Community Forensics/DFGAN shows a decrease. The complete canonical per-group AUROC results are reported in Table 8.

Table 8: **Per-group AUROC on all 48 canonical benchmark–generator groups.** Fusion denotes the validation-standardized equal-weight combination of SRF-E and CLIP-LP defined in Eq. (5). Delta is the AUROC of fusion minus that of CLIP-LP. Benchmark and generator/group jointly identify an evaluation group.

Benchmark	Generator / group	Tier	N_{real}	$N_{\text{generated}}$	SRF-E	CLIP-LP	Fusion	Delta
GenImage-8	ADM	SC	500	500	0.9970	0.9868	0.9995	+0.0127
GenImage-8	BigGAN	SC	500	500	0.9411	0.9998	0.9989	−0.0009
GenImage-8	GLIDE	SC	500	500	0.9880	0.9933	0.9989	+0.0056
GenImage-8	Midjourney	SC	500	500	0.9887	0.9027	0.9951	+0.0924
GenImage-8	SD1.4	SC	500	500	0.9975	0.9654	0.9997	+0.0343
GenImage-8	SD1.5	SC	500	500	0.9986	0.9674	0.9997	+0.0323
GenImage-8	VQDM	SC	500	500	0.9955	0.9810	0.9996	+0.0186
GenImage-8	Wukong	SC	500	500	0.9979	0.9494	0.9996	+0.0503
ForenSynths	BigGAN	SC	1,000	1,000	0.9478	0.9973	0.9973	+0.0001
ForenSynths	CRN	SC	1,000	1,000	0.9995	0.9978	1.0000	+0.0022
ForenSynths	CycleGAN	SC	1,000	1,000	0.9734	0.9971	0.9993	+0.0022
ForenSynths	Deepfake	SC	1,000	1,000	0.9278	0.9806	0.9777	−0.0029
ForenSynths	GauGAN	SC	1,000	1,000	0.9878	0.9976	0.9998	+0.0022
ForenSynths	IMLE	SC	1,000	1,000	0.9999	0.9989	1.0000	+0.0011
ForenSynths	ProGAN	SC	1,000	1,000	0.9930	0.9998	0.9996	−0.0002
ForenSynths	SAN	TU	219	219	0.8812	0.8045	0.9043	+0.0998
ForenSynths	SeeingDark	TU	180	180	0.7978	0.8457	0.8849	+0.0392
ForenSynths	StarGAN	SC	1,000	1,000	1.0000	0.9999	1.0000	+0.0001
ForenSynths	StyleGAN	SC	1,000	1,000	0.9957	0.9906	0.9998	+0.0092
ForenSynths	StyleGAN2	SC	1,000	1,000	0.9962	0.9963	0.9996	+0.0033
ForenSynths	WhichFaceIsReal	SC	1,000	1,000	0.6214	0.9562	0.9033	−0.0529

Table 8 (continued)

Benchmark	Generator / group	Tier	N_{real}	$N_{\text{generated}}$	SRF-E	CLIP-LP	Fusion	Delta
AIGCDetect	ADM	SS	5,000	1,000	0.9758	0.9567	0.9927	+0.0360
AIGCDetect	BigGAN	SS	5,000	1,000	0.9048	0.9647	0.9825	+0.0178
AIGCDetect	CycleGAN	SS	5,000	1,000	0.9250	0.9876	0.9934	+0.0058
AIGCDetect	DALL-E 2	TU	5,000	1,000	0.7876	0.6912	0.8216	+0.1304
AIGCDetect	GauGAN	SS	5,000	1,000	0.8977	0.9859	0.9895	+0.0036
AIGCDetect	GLIDE	SS	5,000	1,000	0.9104	0.9665	0.9804	+0.0140
AIGCDetect	Midjourney	SS	5,000	1,000	0.9549	0.9193	0.9813	+0.0620
AIGCDetect	ProGAN	SS	5,000	200	0.8436	0.9909	0.9857	−0.0053
AIGCDetect	SD1.4	SS	5,000	1,000	0.9613	0.9617	0.9899	+0.0282
AIGCDetect	SD1.5	SS	5,000	1,000	0.9610	0.9547	0.9890	+0.0343
AIGCDetect	SDXL	TU	5,000	1,000	0.9218	0.9258	0.9702	+0.0444
AIGCDetect	StarGAN	SS	5,000	1,000	0.9325	0.9898	0.9955	+0.0057
AIGCDetect	StyleGAN	SS	5,000	1,000	0.8936	0.9286	0.9659	+0.0373
AIGCDetect	StyleGAN2	SS	5,000	1,000	0.9011	0.8950	0.9574	+0.0625
AIGCDetect	VQDM	SS	5,000	1,000	0.8983	0.8650	0.9481	+0.0831
AIGCDetect	WhichFaceIsReal	SS	5,000	998	0.8884	0.9824	0.9865	+0.0041
AIGCDetect	Wukong	SS	5,000	1,000	0.9391	0.9719	0.9889	+0.0169
Community	DFGAN	TU	5,000	500	0.7538	0.9840	0.9437	−0.0403
Community	DALL-E 2	TU	5,000	500	0.8043	0.8835	0.8850	+0.0015
Community	DALL-E 3	TU	5,000	500	0.8695	0.8566	0.9054	+0.0488
Community	FLUX-schnell	TU	5,000	500	0.7499	0.8130	0.8171	+0.0041
Community	Firefly Image2	TU	5,000	500	0.8967	0.8857	0.9332	+0.0475
Community	IdeogramV2	TU	5,000	500	0.8483	0.8023	0.8772	+0.0748
Community	LCM-LoRA SDXL	TU	5,000	500	0.8370	0.9028	0.9108	+0.0080
Community	Midjourney v5.2	TU	5,000	500	0.9045	0.8180	0.9170	+0.0990
Community	Midjourney v6.1	TU	5,000	500	0.8049	0.8279	0.8591	+0.0312
Community	Stable Cascade	TU	5,000	500	0.9286	0.9212	0.9610	+0.0398

The sample counts describe the canonical evaluation underlying Table 1, before the additional filtering considered in the sensitivity analyses. Shared real-image pools are reused across groups as specified in Table 4. Tier assignments follow Appendix A.1 and are relative to the stated development sources; they do not describe the training exposure of every reference detector. A different benchmark release does not, by itself, establish independent image provenance or content disjointness. SC = seen-clean, SS = seen-shifted, and TU = truly-unseen. ForenSynths denotes the ForenSynths subset, AIGCDetect denotes AIGCDetectBenchmark, and Community denotes Community Forensics. Fusion is SRF-E $\oplus$ CLIP-LP. AUROCs and differences are rounded independently to four decimal places. Differences and improvement counts are computed from unrounded AUROCs; therefore, subtracting the displayed AUROCs may differ from the displayed Delta in the last decimal place.

Aggregation and consistency. Each benchmark–generator group receives equal weight in the overall mean, irrespective of its sample count. Averaging the 48 group AUROCs reproduces the overall values in Table 1: 0.9150 for SRF-E, 0.9363 for CLIP-LP, and 0.9622 for their fusion. The corresponding minimum group AUROCs are 0.6214, 0.6912, and 0.8171. Tier means are computed by restricting the same averaging procedure to the groups assigned to each tier. Fusion improves over CLIP-LP in 42 of the 48 groups and decreases AUROC in six. Among the 14 truly-unseen groups, 13 improve and Community Forensics/DFGAN decreases.

C Score Normalization, Calibration, and Operating Points

Normalization and operating thresholds. The default SRF-E $\oplus$ CLIP-LP fusion uses Eq. (5) with score means and population standard deviations estimated from pooled development-model validation scores and applied unchanged to final-refit scores. The values are $\mu = 1.7517$, $\sigma = 1.5204$ for SRF-E and $\mu = 0.1447$, $\sigma = 4.3104$ for CLIP-LP. The C2P partner uses $\mu = 0.4731$ and $\sigma = 0.4306$.

The operating thresholds used for Table 9 are 1.8591 for SRF-E, 0.0567 for CLIP-LP, and -0.1099 for fusion. The SRF-E and fusion thresholds follow the nine-source-group development-validation procedure in App. A.3. The fusion threshold is expressed on the standardized-sum scale of Eq. (5).

Calibration diagnostics. In a post-hoc sensitivity analysis, using test-pool normalization statistics yields a mean AUROC of 0.9624, compared with 0.9622 under the default development-validation normalization. The default results use the latter protocol. The balanced-accuracy gap to a single global test-oracle threshold is 0.005 for fusion, 0.014 for CLIP-LP, and 0.008 for SRF-E. Platt mappings (Platt et al., 1999) fitted to development-model validation scores and applied to final-model test scores yield pooled ECEs of 0.101, 0.108, and 0.131, respectively.

Learned fusion alternatives. Under the LOBO protocol, logistic fusion of detector scores achieves an AUROC of 0.9635. The default equal-weight standardized sum achieves 0.9622 without fitting a fusion model. These results use different fusion-fitting protocols and are reported as complementary analyses. Within the LOBO experiments, a probe combining CLIP ViT-L/14 embeddings with the scalar SRF-E score achieves 0.9715, compared with 0.9331 when using the raw residual descriptors. This comparison favors the scalar representation in the evaluated setting.

Table 9: **Operating-point and calibration diagnostics.** BalAcc, Acc, and F1 are unweighted 48-group means at the validation-selected thresholds. The SRF-E and fusion thresholds maximize nine-source-group mean balanced accuracy on development validation. Gap is the distance of that 48-group mean BalAcc to a single global test-oracle threshold, not a per-benchmark oracle. ECE is pooled-image expected calibration error after Platt scaling fitted on development validation (Platt et al., 1999). TPR@1% FPR uses a separate threshold from each group's test real-image scores and is reported for all 48 groups and for the 14 truly-unseen groups. Bold marks the best of the three systems (lower is better for gap and ECE).

	Validation threshold (48-group means)			Calibration		TPR@1% FPR (test-real thresholds)	
Method	BalAcc ↑	Acc ↑	F1 ↑	Gap ↓	ECE ↓	All 48 groups ↑	Unseen 14 groups ↑
SRF-E	0.829	0.861	0.684	0.008	0.131	0.427	0.094
CLIP-LP	0.848	0.903	0.736	0.014	0.108	0.529	0.158
Fusion	**0.899**	**0.920**	**0.791**	**0.005**	**0.101**	**0.625**	**0.166**

D Implementation Details

Residual operators. The six round trips use nearest, bilinear, bicubic, and area→bilinear resampling at scale 0.5, and nearest and area→bilinear resampling at scale 0.25. Input preprocessing and JPEG training augmentation are specified in App. A.1.

Descriptor dimensions. lite 126 (6×21: 9 band energies, 3 band ratios, 9 spatial moments), self 576 (6×96), spec 144 (6×24: 16 radial rings + 8 angular sectors), E 192 (6×32: 2 grid scales $\times$ 4 bands $\times$ [dispersion, Moran's I, slope, R^2]), cooc 4704 (6×784, Hellinger square-rooted). SRF-E global = 846 (lite, self, spec); local = 4896 (cooc, E); total 5742. Spectral components total 408 dims (lite bands 72, spec 144, E 192); all else spatial-domain.

Committee configuration. The two shrinkage-LDA members use the source subsets $\{\mathrm{GenImage}, \mathrm{ProGAN}\}$ and $\{\mathrm{GenImage}, \mathrm{ForenSynths}\}$. Their real-reference-calibrated scores are combined as $m = 0.4\, z_{\mathrm{pro}} + 0.6\, z_{\mathrm{fs}}$. Each member uses an empirical right-tail probability against training-real projections, mapped through Φ^{-1} to a z-score. The lite-based OOD model computes Mahalanobis distance on L2-normalized features using a Ledoit–Wolf covariance estimate from training-real images. The switch is activated only when $m < -1.2816$ and the OOD z-score exceeds 1.6449, corresponding to the 10th and 95th standard-normal percentiles, respectively. The development and final-refit stages are specified in App. A.1; cross-axis normalization and operating thresholds are reported in App. C.

E EFFICIENCY

Parameter counting and checkpoint composition. The CLIP-LP and UnivFD rows count the full CLIP ViT-L/14 checkpoint (image+text towers, 428M); only the 304M image tower runs at inference, matching C2P-CLIP's released file. AIDE ships a frozen ConvNeXt-XXL expert in its checkpoint. NPR uses its official truncated ResNet (measured 1.44M). SRF-E uses no pretrained backbone.

SRF-E uses 0.28M fitted parameters and a 2.5 MB checkpoint without a pretrained backbone. In this paper, "lightweight" describes the parameter and storage requirements.

F ADDITIONAL RESULT TABLES AND FIGURES

F.1 QUALITATIVE EXAMPLES OF SIGNAL–SEMANTIC COMPLEMENTARITY

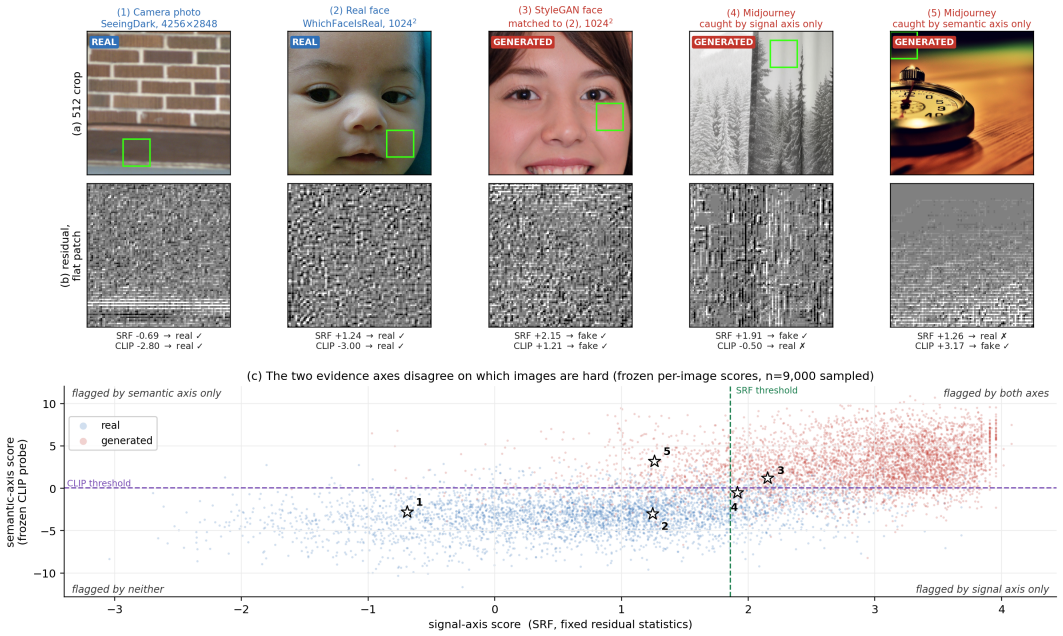

Figure 5: Qualitative examples of signal–semantic complementarity. Top: five real and generated images with enlarged, contrast-enhanced residual patches; green boxes mark the displayed patch locations. Bottom: SRF-E scores versus CLIP-LP scores for 9,000 evaluation images, with dashed lines marking the applied branch thresholds. Numbered markers identify the five examples. Examples (4) and (5) are detected by SRF-E and CLIP-LP alone, respectively. Residual patches are shown for visualization using center crops without the resizing used by the descriptor pipeline; detector scores are unchanged. The visualization protocol is specified in App. A.4.

Figure 5 illustrates the two-axis design. Examples (1)–(2) are real images; (2)–(3) form a resolution-matched real/generated face comparison; (4)–(5) come from the same generator but are detected only by SRF-E or CLIP-LP, respectively. In the scatter plot, moving right indicates higher detection scores under SRF-E and moving upward under CLIP-LP, with dashed lines marking the operating thresholds (App. C). Examples (1)–(2) fall below both thresholds and (3) exceeds both. Examples (4)–(5) illustrate complementary decisions: each generated image exceeds the operating threshold of one branch but not the other.

F.2 Additional Quantitative Results

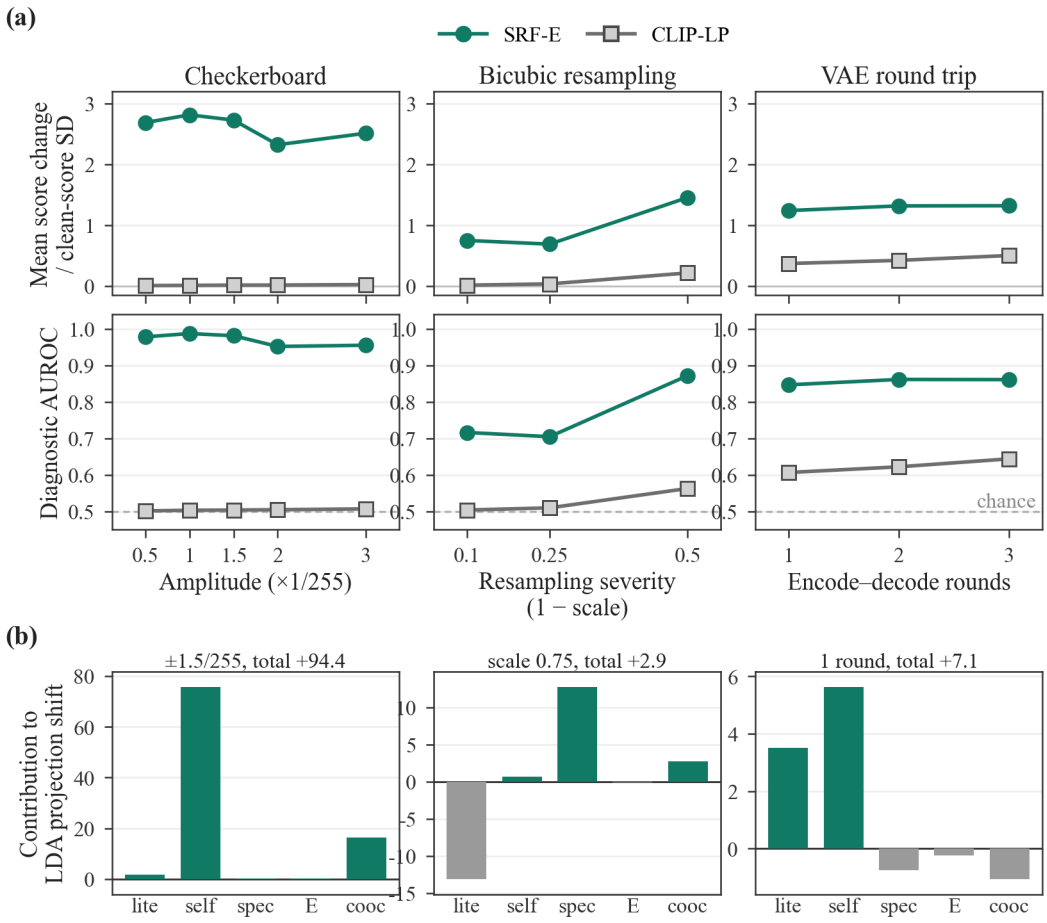

Figure 6: Dose responses and descriptor-block attribution on 800 real photographs. (a) Normalized mean score changes and diagnostic AUROCs under checkerboard perturbations, bicubic resampling, and repeated VAE encode–decode operations. Score changes are normalized by each detector's score standard deviation on the original photographs; diagnostic AUROC distinguishes transformed photographs from their unmodified originals. (b) Signed contributions of the five descriptor blocks to the mean weighted committee LDA-projection change, before empirical calibration and the OOD override. Bars show raw projection contributions, and each panel reports their total. For checkerboard perturbations, the self block accounts for 0.80 of the total projection change. Protocol details are provided in App. A.4. At the smallest checkerboard amplitude, $\pm 0.5/255$, SRF-E changes by $+2.69\sigma$ (diagnostic AUROC 0.978). Resampling at scales 0.9 and 0.5 changes SRF-E by $+0.75\sigma$ and $+1.46\sigma$. Across VAE rounds, CLIP-LP changes by about $+0.37$ to $+0.50\sigma$, while its checkerboard changes remain at most $+0.03\sigma$.

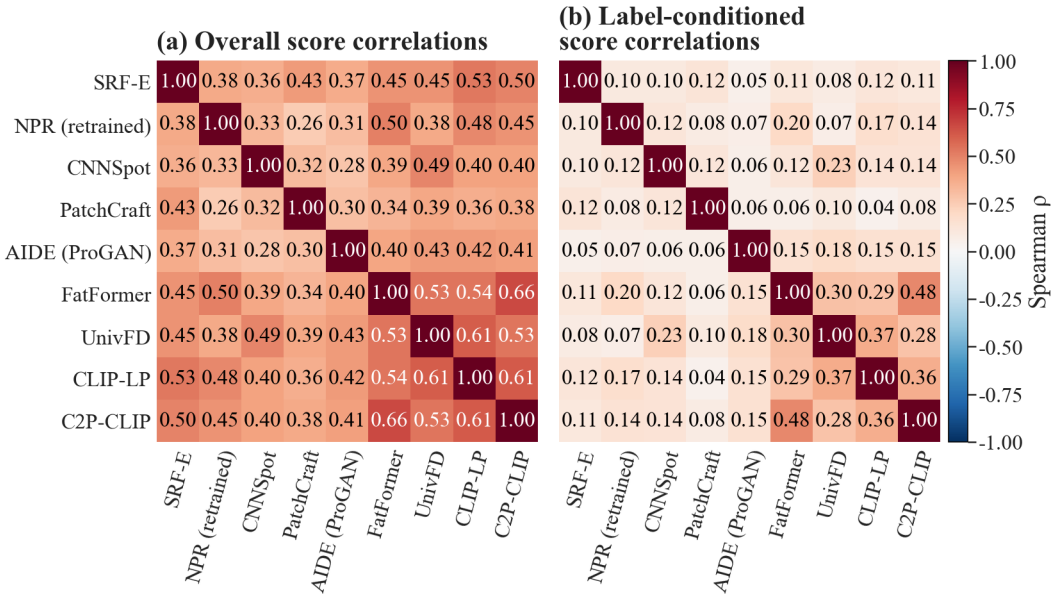

Figure 7: Detector score correlations. (a) Spearman correlations computed across both classes within each evaluation group. (b) Label-conditioned Spearman correlations, computed separately for real and generated images within each group and then averaged with equal class weights. Both panels report unweighted means over the 48 canonical groups and use the same color scale. The mean pairwise correlation among FatFormer, UnivFD, CLIP-LP, and C2P-CLIP is approximately 0.35 in panel (b), excluding self-correlations. SRF-E has a mean correlation of approximately 0.11 with these four detectors (CLIP-LP–C2P-CLIP 0.36; FatFormer–C2P-CLIP 0.48). These correlations are consistent with lower within-class score redundancy. UnivFD denotes the officially released checkpoint, whereas CLIP-LP denotes our frozen CLIP ViT-L/14 linear probe.

Bidirectional coverage (per-group numbers for §5). SRF-E improves CLIP-LP on GenImage-8/Midjourney ($0.903 \rightarrow 0.995$), AIGCDetectBenchmark/DALL-E 2 ($0.691 \rightarrow 0.822$), and GenImage-8/Wukong ($0.949 \rightarrow 1.000$). CLIP-LP improves SRF-E on ForenSynths/WhichFaceIsReal ($0.621 \rightarrow 0.903$) and Community Forensics/DFGAN ($0.754 \rightarrow 0.944$).

Table 10: **Per-benchmark mean AUROC.** Entries are unweighted means of group AUROC within each suite; group counts are in the headers. The last column is the same 48-group unweighted mean as the overall-performance Mean AUROC column of Table 1, not a pooled-image AUROC. Bold marks the highest value in each column.

Method	Mean AUROC ↑				
	GenImage-8 (8)	ForenSynths (13)	AIGCDetect (17)	Community (10)	All groups (48)
AEROBLADE	0.8933	0.5394	0.7236	0.5556	0.6670
NPR (retrained)	0.9971	0.9543	0.5113	0.6308	0.7371
CNNSpot	0.6732	0.9125	0.7685	0.5455	0.7452
PatchCraft	0.9497	0.8917	0.8188	0.4079	0.7748
AIDE (ProGAN)	0.8889	0.8739	0.7493	0.5091	0.7563
AIDE (GenImage)	**0.9999**	0.9145	0.7386	0.5571	0.7920
FatFormer	0.8734	0.9752	0.7964	0.6656	0.8304
UnivFD	0.8752	0.9287	0.8572	0.6448	0.8353
C2P-CLIP	0.9552	0.9832	0.8952	0.8469	0.9190
SRF-E (ours)	0.9881	0.9324	0.9116	0.8398	0.9150
CLIP-LP	0.9682	0.9663	0.9375	0.8695	0.9363
CLIP-LP ⊕ C2P	0.9824	**0.9850**	0.9508	0.8991	0.9545
SRF-E ⊕ CLIP-LP	0.9989	0.9743	**0.9717**	**0.9009**	**0.9622**

F.3 Robustness, calibration, and operating points

Post-processing affects the two branches differently (Table 11). Under JPEG-40 compression, standalone SRF-E decreases from 0.9267 to 0.6797 AUROC, while CLIP-LP decreases from 0.9413 to 0.8733. This pattern is consistent with sensitivity of the residual descriptors to compression.

Fusion exceeds standalone SRF-E under all 17 perturbations and CLIP-LP under 14. It yields lower AUROC than CLIP-LP under JPEG-40 (0.8129 vs. 0.8733), crop 0.875 (0.9296 vs. 0.9328), and crop 0.75 (0.9196 vs. 0.9312). Thus, the benefit of fixed equal-weight fusion depends on the processing condition.

At the validation-frozen threshold, fusion attains a group-mean balanced accuracy of 0.899 (Table 9). With thresholds derived separately from each group's test real-image scores, its mean TPR at 1% FPR is 0.166 on the unseen groups, indicating limited separation in this low-FPR regime.

Table 11: **Post-processing sensitivity under 17 perturbations.** Entries are the unweighted mean of four benchmark-level pooled AUROCs on a fixed, class-balanced 3,200-image subset (400 real and 400 generated images per benchmark; generated images are stratified by generator). This is not the canonical 48-group mean in Table 1. Both classes are perturbed, using the same image IDs for all methods. Fusion uses the validation-frozen equal-weight z-score sum, with validation statistics fixed across conditions. Bold marks the highest value in each row.

Condition	Four-benchmark mean AUROC ↑		
	SRF-E	CLIP-LP	Fusion
Unperturbed			
clean	0.9267	0.9413	**0.9677**
JPEG compression			
quality 95	0.8914	0.9264	**0.9559**
quality 85	0.8620	0.8998	**0.9292**
quality 75	0.8316	0.8910	**0.9087**
quality 60	0.7886	0.8822	**0.8865**
quality 40	0.6797	**0.8733**	0.8129
WebP compression			
quality 75	0.8114	0.8772	**0.8940**
quality 50	0.7724	0.8517	**0.8668**
Resize			
scale 0.75	0.8687	0.9117	**0.9473**
scale 0.5	0.7685	0.8638	**0.8825**
Gaussian blur			
$\sigma = 0.5$	0.9192	0.9347	**0.9667**
$\sigma = 1.0$	0.8253	0.8994	**0.9261**
Center crop			
retained scale 0.875	0.8226	**0.9328**	0.9296
retained scale 0.75	0.7849	**0.9312**	0.9196
Gaussian noise			
$\sigma = 2$	0.8719	0.9122	**0.9489**
$\sigma = 5$	0.7664	0.8807	**0.8880**
Gamma			
$\gamma = 0.8$	0.9250	0.9399	**0.9692**
$\gamma = 1.2$	0.9196	0.9407	**0.9679**

The complete canonical per-group AUROC results are reported in Table 8. Machine-readable versions of these results, together with residual-block ablations and significance results, will accompany the evaluation code.

G LIMITATIONS

Sensitivity to post-processing and variation across groups. The contribution of SRF-E varies across processing conditions and evaluation groups. Under JPEG-40 compression, standalone SRF-E decreases from 0.9267 to 0.6797 AUROC, and fusion reaches 0.8129 compared with 0.8733 for CLIP-LP alone (Table 11). Fusion improves over CLIP-LP in 42 of the 48 canonical groups and yields lower AUROC in six, including ForenSynths/WhichFaceIsReal and Community Forensics/DFGAN (Table 8). Fixed equal-weight fusion does not adapt to variation in branch reliability; reliability-aware fusion is a direction for future evaluation.

Detection at low false-positive rates. Using thresholds derived from each group's test real-image scores, fusion achieves a mean TPR of 0.166 at 1% FPR across the 14 unseen groups, with a TPR of 0.03 on Community Forensics/FLUX-schnell. The mean unseen-group AUROC of 0.899 summarizes performance across the ROC curve and does not ensure high recall in this low-FPR regime. Improving score separation at low FPR remains an important direction.

Evaluation scope. Our conclusions are limited to the evaluated generators, collections, and processing conditions. The unseen tier denotes generators absent from our detector-development sources; transfer to future generators and different image-processing pipelines requires further evaluation. Adversarial robustness has not been evaluated and remains an open question under explicit threat models.