GENERATOR FINGERPRINT ISOLATION: ADDRESSING BLIND SPOTS IN AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated image detection is an open-set generalization problem: detectors trained on a limited set of generators may fail to generalize to unseen generators, resulting in generator-specific blind spots. A common response is to adapt the detector using samples from a missed generator, but this modifies its decision function and may affect generators it already detects reliably. We instead ask whether such blind spots can be repaired by adding generator-specific evidence while preserving the original detector. We propose **Generator Fingerprint Isolation (GFI)**, a training-free framework that augments a frozen detector with such evidence.

GFI exploits an empirical property of pretrained representations such as CLIP and DINO: the principal components that account for most variation among real-image embeddings can be removed without eliminating generator-discriminative signal. Across the blind-spot generators considered in our evaluation, removing the principal components accounting for 95% of real-image variance yields a residual space that preserves 78.8% of the measured generator-discriminative signal. GFI uses this residual space to extract a generator-specific fingerprint for each blind-spot generator. At inference, alignment with the fingerprint provides generator-specific evidence that is fused with frozen detector's prediction. Fingerprints can be registered independently as new blind-spot generators are encountered.

We evaluate GFI on AIGCDetectBench, with 17 GAN, diffusion, and commercial generators, across three off-the-shelf modern detectors—DDA, Effort, and SPAI. For each detector, we identify its own blind-spot generators and repair them using generator-specific fingerprints. Across these targeted generators, GFI improves mean AUC from 0.827 to 0.972 for DDA, 0.747 to 0.914 for Effort, and 0.762 to 0.900 for SPAI, while also substantially reducing FPR@95% Recall. AUC improvements are statistically significant for every detector on its identified blind-spot generators under paired DeLong tests. For generators not targeted for repair, GFI shows no systematic degradation. These results demonstrate that generator-specific evidence can repair targeted blind spots while preserving the detector.

1 INTRODUCTION

AI-generated image detection is inherently an *open-set generalization* problem. A detector may perform well on generators represented during training yet fail on unseen generators. A natural response is to adapt the detector using examples from the missed generator, but this changes its decision function and can affect generators that were already well detected. We ask: *can a detector blind spot be addressed by adding generator-specific evidence while keeping the detector unchanged?*

Our key insight is that pretrained image representations provide a suitable space for extracting such evidence. Modern detectors commonly use pretrained vision encoders such as CLIP (Radford et al., 2021) and DINO (Siméoni et al., 2025), which capture variation from many aspects of image content, including scene structure and appearance. Such variation can dominate the feature space even when it is not useful for distinguishing real from generated images, while generator-discriminative signal can remain in lower-variance directions. Across the blind-spot generators we evaluate, removing the principal components accounting for 95% of real-image feature variance retains 78.8% of the measured generator-discriminative signal in the resulting residual space.

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



Figure 1: GFI overview. First, real images R_{fit} are used to identify and remove dominant directions of variation from pretrained image features, yielding a residual feature space. Second, generated images S_k from the blind-spot generator G_k are projected into this space to obtain a generator fingerprint e_k. Third, held-out real images R_{cal} calibrate alignment with e_k relative to real images. Finally, this generator-specific evidence is combined with frozen detector score to produce a GFI score.

Based on this observation, we propose **Generator Fingerprint Isolation (GFI)**, a training-free framework for adding generator-specific evidence to a frozen detector. Figure 1 provides an overview of the approach. GFI first extracts pretrained embeddings (e.g., CLIP or DINO) from real images, identifies the principal components capturing dominant variation, and constructs a residual space by removing these components. For a generator corresponding to a detector blind spot, generated examples are projected into this residual space to extract a characteristic fingerprint. Held-out real images are then used to calibrate alignment with each fingerprint. At inference, alignment with a fingerprint provides generator-specific evidence, which is combined with the frozen detector score to repair the blind spot. Fingerprints and their calibration are stored in a reusable bank, allowing new generators to be registered independently.

We evaluate GFI on **AIGCDetectBench** (Zhong et al., 2023), a cross-generator benchmark spanning 17 image generators, including GANs, diffusion, and commercial models. Detection performance varies substantially across generators, exposing generator-specific blind spots across detectors. For three modern detectors (DDA, Effort, and SPAI), GFI targets each detector's own blind-spot generators and improves mean AUC – 0.827 to 0.972 for DDA, 0.747 to 0.914 for Effort, and 0.762 to 0.900 for SPAI. FPR@95% Recall decreases by 0.368, 0.337, and 0.313 for DDA, Effort, and SPAI, respectively. On generators not targeted for repair, GFI shows no systematic degradation.

2 Related Work

Generalized detection. Early approaches trained classifiers on real images and outputs from a limited set of generators and studied how well their learned cues transferred across architectures (Wang et al., 2020; Chai et al., 2020). Later work showed that conventional classifiers can fail on newer generator families and that frozen pretrained vision–language features can improve cross-generator transfer (Ojha et al., 2023; Cozzolino et al., 2024). DDA (Chen et al., 2025) aligns real and synthetic data in pixel and frequency domains, B-Free (Guillaro et al., 2025) constructs semantically

aligned synthetic data, and AIDE (Yan et al., 2025b) combines CLIP with low-level patch features. These methods seek a shared decision function that transfers across generators. GFI considers the complementary setting of a trained detector with generator-specific blind spots, where additional evidence is needed without modifying the detector.

Spectral and reconstruction evidence. A complementary body of work extracts cues associated with image formation. Frequency-based methods exploit spectral artifacts (Frank et al., 2020), while spatial and gradient-based approaches model neighboring-pixel, texture, or derivative statistics (Tan et al., 2024; Zhong et al., 2023; Tan et al., 2023). SPAI (Karageorgiou et al., 2025) models real-image spectral structure through spectral reconstruction similarity. Reconstruction-based methods measure discrepancies from diffusion (Wang et al., 2023), autoencoder (Ricker et al., 2024), or latent-space (Luo et al., 2024) reconstruction. These approaches provide forensic signals complementary to semantic classification, but typically rely on image statistics or reconstruction procedures; reconstruction methods may also require additional generative-model inference at test time.

Residual representations and generator fingerprints. Several prior works isolate forensic information from semantic content. RINE (Koutlis & Papadopoulos, 2024) combines projected CLS representations from intermediate CLIP blocks, NS-Net (Yan et al., 2025a) projects CLIP features onto the null space of a text-derived semantic subspace, while LoRC (Yan et al., 2026) removes an image-specific semantic component from DINO features to detect residual structure. Effort (Yan et al., 2025c) instead operates in parameter space, adapting residual components of pretrained weights with labeled data.

Another line of work exploits generator fingerprints for detection and attribution (Marra et al., 2019; Yu et al., 2019). ManiFPT (Song et al., 2024) models artifacts relative to a real-image manifold, Deep Image Fingerprint (Sinitsa & Fried, 2024) learns a CNN-based image-space fingerprint from generated and real-image residuals, enabling low-data detection of images from a known generator and model-lineage analysis, and GAPL (Qin et al., 2026) learns a compact set of canonical forgery prototypes to manage variation across generators.

These approaches highlight two complementary directions. Residual-representation methods reduce the influence of semantic content to make forensic cues more discriminative, while fingerprint-based methods characterize artifacts associated with particular generators for detection or attribution. However, these works mainly develop standalone detection or attribution systems. GFI instead studies whether generator-associated evidence provides additional value when fused post hoc with an already trained detector.

3 Problem Formulation

Let $x \in \mathcal{X}$ denote an image and $\mathbf{f}(x) = \mathcal{E}(x) \in \mathbb{R}^d$ its representation from a fixed pretrained image encoder $\mathcal{E}$. This representation captures both semantic content (objects, scene layout, appearance) and distributional differences from the generating process. Let P_{real} and P_{G_k} denote the representation distributions of real images and images generated by G_k, respectively.

A pretrained detector D assigns each image a scalar score $b(x) \in \mathbb{R}$, with larger values indicating stronger evidence that x is generated.[1] We measure how well this score separates images from generator G_k from real images via its ranking performance,

$$A(b; G_k) = \Pr\big[\, b(X_g) > b(X_r) \,\big], \qquad X_g \sim P_{G_k},\ X_r \sim P_{\text{real}}, \tag{1}$$

which corresponds to the AUC of b against G_k. A value near 1 indicates reliable separation; a value near 0.5 indicates chance-level discrimination between G_k and real images.

A detector may achieve high $A(b; G_j)$ across many generators G_j yet, for a particular G_k, yield $A(b; G_k)$ far below the rest. We call such a G_k a *blind-spot generator (for D)*: the differences between G_k's generated images and real images are not adequately captured by D's decision function, even though D generalizes well elsewhere.

[1] We treat D as a black box: our formulation does not assume $b(x)$ is computed from $\mathbf{f}(x)$ specifically, only that both are available for a given input image.

This motivates a post-hoc setting in which the existing detector D is treated as a source of complementary evidence rather than replaced. The goal is to identify generator-specific structure already present in the representation $\mathbf{f}(x)$, and to incorporate it *without* changing D itself.

Objective. Given a frozen detector D with score function $b(x)$, and let G_k be a blind-spot generator identified for D. Given a set $\mathcal{S}_k$ of images produced by G_k, we construct generator-specific evidence for G_k and combine it with D's existing score $b(x)$ to obtain a decision score $s_k(x)$ satisfying

$$A(s_k; G_k) > A(b; G_k), \tag{2}$$

that is, s_k separates G_k's generated images from real images more reliably than D's own score b does. Critically, constructing s_k must leave D's parameters unchanged, and should avoid systematic degradation on other generators G_j ($j \neq k$) and limit false-positive increases on real images. The generator-specific evidence is therefore meant to complement D's existing score, not replace it. When multiple blind-spot generators are identified, their respective generator-specific evidence can be incorporated jointly. Section 4 develops this construction.

4 Method: Generator Fingerprint Isolation

GFI consists of three stages, as illustrated in Figure 1: (1) *isolate* generator fingerprints by removing dominant real-image variation from pretrained features, (2) *calibrate* fingerprint alignment relative to held-out real images, and (3) *fuse* the resulting generator-specific evidence with the frozen detector D's score $b(x)$. No parameters of base detector D are optimized.

4.1 Isolating Generator Fingerprints

Constructing the residual projection. We use the real-image set $\mathcal{R}_{\text{fit}}$ to identify the dominant directions of variation in the pretrained feature space. Its mean feature is

$$\boldsymbol{\mu}_{\text{real}} = \frac{1}{|\mathcal{R}_{\text{fit}}|} \sum_{x \in \mathcal{R}_{\text{fit}}} \mathcal{E}(x). \tag{3}$$

After centering the features of $\mathcal{R}_{\text{fit}}$, PCA yields orthonormal components $\{\mathbf{v}_i\}_{i=1}^{d}$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$. For a variance-removal threshold γ, we choose the smallest $K(\gamma)$ satisfying

$$\frac{\sum_{i=1}^{K(\gamma)} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geq \gamma. \tag{4}$$

The residual projection is

$$Q = \sum_{i=K(\gamma)+1}^{d} \mathbf{v}_i \mathbf{v}_i^{\top}. \tag{5}$$

Thus, Q removes the dominant directions of variation in real-image features and retains the remaining lower-variance directions. Recalling $\mathbf{f}(x) = \mathcal{E}(x)$ from Section 3, the centered feature is

$$\mathbf{u}(x) = \mathbf{f}(x) - \boldsymbol{\mu}_{\text{real}},$$

and the residual representation is

$$\tilde{\mathbf{f}}(x) = Q\,\mathbf{u}(x). \tag{6}$$

Measuring retained generator information. We first test whether generator-discriminative signal survives removal of the dominant real-image variation. Let $\boldsymbol{\mu}_{G_k}$ denote the mean feature of images generated by G_k, and define its displacement from the real-image mean as

$$\mathbf{d}_k = \boldsymbol{\mu}_{G_k} - \boldsymbol{\mu}_{\text{real}}, \qquad d_{k,i} = \mathbf{v}_i^{\top} \mathbf{d}_k. \tag{7}$$

We quantify the variance-normalized generator–real separation removed by discarded components as

$$L_k(\gamma) = \frac{\sum_{i=1}^{K(\gamma)} d_{k,i}^2/\lambda_i}{\sum_{i=1}^{d} d_{k,i}^2/\lambda_i}. \tag{8}$$

$1 - L_k(\gamma)$ measures the fraction retained outside the removed subspace. This tests whether removing dominant real-image variation also removes the signal that distinguishes generated images.

Extracting a generator fingerprint. For each registered generator G_k, i.e. a generator for which a fingerprint has been constructed and added to the bank (in our experiments, the identified blind-spot generators of Section 3) – we use its self-generated image set $\mathcal{S}_k$ (as defined in the Objective, Section 3) to estimate the characteristic direction in the residual space. Their mean residual feature is

$$\hat{\boldsymbol{\mu}}_k^Q = \frac{1}{|\mathcal{S}_k|} \sum_{x \in \mathcal{S}_k} \tilde{\mathbf{f}}(x). \tag{9}$$

We normalize this mean residual to obtain the unit-norm fingerprint

$$\mathbf{e}_k = \frac{\hat{\boldsymbol{\mu}}_k^Q}{\left\|\hat{\boldsymbol{\mu}}_k^Q\right\|_2}, \qquad \|\mathbf{e}_k\|_2 = 1. \tag{10}$$

4.2 Fingerprint Calibration

Calibrating fingerprint evidence. The fingerprint direction alone does not indicate whether an image is unusual. A real image may also align with $\mathbf{e}_k$. GFI therefore calibrates fingerprint alignment against the held-out real-image set $\mathcal{R}_{\text{cal}}$. For a test image x, its alignment with fingerprint $\mathbf{e}_k$ is

$$a_k(x) = \mathbf{e}_k^\top \tilde{\mathbf{f}}(x). \tag{11}$$

The alignment scores on the held-out real images,

$$\{a_k(r) : r \in \mathcal{R}_{\text{cal}}\},$$

define the empirical cumulative distribution function F_k. We convert the percentile of the test image's alignment into a normal-score evidence value:

$$z_k(x) = \Phi^{-1}\left(F_k(a_k(x))\right), \tag{12}$$

where Φ^{-1} denotes the inverse CDF of the standard normal distribution. Probabilities are clipped away from 0 and 1 before applying the transformation. Larger $z_k(x)$ indicates stronger evidence that the image has unusually high alignment with fingerprint $\mathbf{e}_k$ relative to held-out real images.

The fingerprint bank is the collection of fingerprints and their calibration distributions:

$$\mathcal{B} = \{(\mathbf{e}_k, F_k) : k \in \mathcal{K}\}, \tag{13}$$

where $\mathcal{K}$ denotes the set of registered generators.

At inference, GFI evaluates an image against all fingerprints in the bank. Because fingerprints capture generator-dependent evidence, a test image may exhibit strong alignment with some fingerprints, while unrelated fingerprints may contribute little evidence. Hence, we aggregate strongest calibrated fingerprint scores. Let $z_{(1)}(x) \geq z_{(2)}(x) \geq \cdots \geq z_{(|\mathcal{K}|)}(x)$ denote calibrated scores $\{z_k(x)\}_{k \in \mathcal{K}}$ sorted in decreasing order, and let $m' = \min(m, |\mathcal{K}|)$. We define the aggregated generator-specific evidence for x – the quantity referred to abstractly in the Objective of Section 3 – as

$$g(x) = \frac{1}{m'} \sum_{j=1}^{m'} z_{(j)}(x). \tag{14}$$

We use $m = 2$, aggregating the two strongest fingerprint matches. Thus, $g(x)$ captures the strongest evidence rather than averaging unrelated fingerprints.

4.3 FUSION

Fusion with the frozen detector. Recall from Section 3 that D denotes the frozen detector and $b(x)$ its score. We calibrate $b(x)$ against the held-out real images $\mathcal{R}_{\text{cal}}$ in the same way as the fingerprints above. Let F_{base} denote the empirical CDF of

$$\{b(r) : r \in \mathcal{R}_{\text{cal}}\}.$$

The calibrated detector evidence is

$$z_{\text{base}}(b(x)) = \Phi^{-1}\left(F_{\text{base}}(b(x))\right). \tag{15}$$

The final GFI score combines calibrated detector evidence with aggregated fingerprint evidence:

$$s(x) = z_{\text{base}}(b(x)) + g(x). \tag{16}$$

This single score $s(x)$ instantiates the objective score $s_k(x)$ of Section 3 for every blind-spot generator G_k simultaneously: since s does not depend on which generator is being evaluated, $s_k \equiv s$ for all k, and equation 2 is tested as $A(s; G_k) > A(b; G_k)$ for each target G_k in turn.

The final prediction is

$$\hat{y}(x) = \mathbf{1}[s(x) > \tau], \tag{17}$$

where τ is selected according to the evaluation protocol.

A new blind-spot generator can therefore be enrolled independently by estimating its fingerprint from self-generated images and calibrating its alignment using the held-out real-image set, without recomputing the residual subspace or previously estimated fingerprints.

5 EXPERIMENTS

We organize the experiments around three questions: *(1) Do detectors exhibit generator-specific blind spots? (2) Can fingerprints repair these blind spots? (3) Does GFI preserve generalization for non-target generators and false positives under real-image distribution shift.*

5.1 EXPERIMENTAL SETUP

We use AIGCDetectBench as the **evaluation** benchmark: it lets us identify detector-specific blind spots, and evaluate GFI on both targeted and non-target generators.

GFI uses two separate image sets: real images to build a generic "real-image" reference and independently generated fake images from each target generator to build its fingerprint. ***Neither set overlaps with AIGCDetectBench, ensuring disjoint fingerprint construction and evaluation images.***

The pool of real images is assembled from FFHQ, COCO, and CNNDetection/LSUN, covering faces, general scenes, and objects totalling 6000 images. We split them 2:1 into $\mathcal{R}_{\text{fit}}$ and $\mathcal{R}_{\text{cal}}$: the former estimates the real-image mean and PCA subspace, while the latter is held out for calibrating frozen detector and fingerprint scores. The generator-specific fake pool is built independently for each target: we self-generate fresh set of images for each target generator for fingerprint construction.

We evaluate seven detectors: SPAI (Karageorgiou et al., 2025), Corvi2023 (Corvi et al., 2023), UFD (Ojha et al., 2023), PatchCraft (Zhong et al., 2023), DDA (Chen et al., 2025), B-Free (Guillaro et al., 2025), and Effort (Yan et al., 2025c). We report AUC and FPR@R95, the false-positive rate at 95% recall on generated images. We use CLIP ViT-L/14 as the frozen feature encoder.

5.2 DO EXISTING DETECTORS EXHIBIT GENERATOR-SPECIFIC BLIND SPOTS?

Table 1 shows that detector performance varies substantially across generators. Critically, this variation is not consistent across detectors: a generator that one detector detects poorly is often detected well by another. Hence, we define a *blind spot* at the detector–generator *pair* level: a specific detector failing on a specific generator. To test whether the pattern holds regardless of how a detector is built, we select three detectors for evaluation that each address generalization differently:

Table 1: Frozen detector performance across 17 AIGCDetectBench generators. Cells report **AUC (↑) / FPR@R95 (↓)**. Red denotes AUC < 0.90; **bold red** marks GFI targets. Unsupported generators are grouped at the bottom. DDA, Effort, and SPAI are the main evaluation set.

	Main			Additional			
Generator	**DDA**	**Effort**	**SPAI**	**Corvi2023**	**UFD**	**PatchCraft**	**B-Free**
ProGAN	.99 / .06	.94 / .20	.90 / .47	.92 / .30	1.00 / .00	1.00 / .00	.99 / .05
StyleGAN	.94 / .32	.98 / .07	**.79** / .69	.90 / .36	.95 / .23	.90 / .70	.86 / .52
StyleGAN2	.98 / .11	**.69** / .82	.90 / .41	.93 / .24	.98 / .10	1.00 / .00	.96 / .18
CycleGAN	**.83** / .54	.97 / .06	**.87** / .45	.56 / .75	1.00 / .00	.85 / .48	.91 / .31
StarGAN	**.72** / .71	1.00 / .01	.90 / .45	.42 / .98	1.00 / .01	1.00 / .00	.92 / .27
WhichFaceIsReal	**.87** / .37	**.75** / .50	**.58** / .79	.78 / .64	.98 / .10	.98 / .07	1.00 / .00
ADM	.96 / .20	**.80** / .49	**.81** / .64	.82 / .52	.91 / .34	.93 / .34	.86 / .56
GLIDE	.97 / .14	.90 / .30	.96 / .17	.96 / .14	.87 / .49	.91 / .50	.93 / .38
SD 1.4	1.00 / .00	1.00 / .02	1.00 / .00	.95 / .19	.72 / .68	.99 / .06	1.00 / .00
SD 1.5	1.00 / .00	.99 / .02	1.00 / .00	.95 / .18	.74 / .64	.99 / .06	1.00 / .00
SD-XL	1.00 / .00	.95 / .18	.99 / .04	.69 / .74	.67 / .86	1.00 / .01	1.00 / .00
VQDM	**.89** / .36	.90 / .40	.96 / .18	.98 / .09	.95 / .23	.95 / .32	.94 / .34
Wukong	1.00 / .00	.98 / .07	1.00 / .01	.95 / .21	.78 / .62	.97 / .16	1.00 / .00
Excluded: unavailable/unsupported pipelines to create new images							
BigGAN	.94 / .32	.92 / .17	.81 / .75	.80 / .57	.99 / .03	.99 / .02	.98 / .09
GauGAN	.96 / .22	.88 / .32	.89 / .50	.90 / .36	1.00 / .00	.84 / .47	.99 / .03
Midjourney	.99 / .05	.83 / .52	.97 / .14	.74 / .74	.53 / .93	.97 / .17	.99 / .04
DALL·E 2	.99 / .03	.81 / .59	.89 / .47	.95 / .21	.70 / .76	1.00 / .01	.98 / .11

- **DDA** intervenes at the *data* level: it aligns generated and real images in pixel and frequency domains, removing incidental statistical differences between them, and only then trains its classifier. Hence, it reduces the opportunity to rely on pixel- or frequency-domain biases.
- **Effort** intervenes at the *representation* level: it freezes the dominant singular components of a pretrained model's weights and adapts only a complementary residual subspace to learn forgery patterns.
- **SPAI** intervenes at the *spectral* level: it learns the spectral distribution of real images via self-supervised masked spectral learning, then flags generated images as out-of-distribution based on how poorly their spectrum reconstructs.

Because these mechanisms are largely independent, they test whether generator-dependent evidence remains useful across detectors with fundamentally different generalization behavior.

We identify blind spots separately for each detector, flagging detector–generator pairs with AUC below 0.90. We exclude candidates whose generator pipeline is not publicly available. This yields 11 targets: DDA (CycleGAN, StarGAN, VQDM, WhichFaceIsReal), Effort (StyleGAN2, WhichFaceIsReal, ADM), and SPAI (CycleGAN, StyleGAN, WhichFaceIsReal, ADM).

5.3 Can Generator-Dependent Fingerprints Repair These Blind Spots?

Having identified these blind spots, we evaluate GFI in its intended post-hoc repair setting. To construct a fingerprint, GFI assumes that generator-specific evidence need not lie in the highest-variance directions of real-image features, which often reflect factors unrelated to forgery (e.g., scene content). We progressively remove these directions up to γ, the fraction of real-image feature variance removed, and measure generator-dependent signal remaining in the *residual space*.

Table 2 reports retained generator-dependent signal and standard deviation of fingerprint alignment scores on held-out real images. For $\gamma = 0.95$ (146 principal components), 78.8% of generator-dependent signal remains, and real-image score standard deviation falls to 0.040. Increasing γ to 0.99 reduces score dispersion but lowers retained signal to 58.1%. We therefore use $\gamma = 0.95$ throughout.

We next evaluate GFI on the 11 target detector–generator pairs identified in Section 5.2. Table 3 shows the resulting recovery: mean AUC increases from 0.827 to 0.972 for DDA, from 0.747 to 0.914 for Effort, and from 0.762 to 0.90 for SPAI. Mean FPR@R95 falls correspondingly, from 0.495

Table 2: Percentage of signal retained after removing top-K principal components of real-image features. Final column reports AUC of the residual fingerprint alignment score at $\gamma = 0.95$. σ_{real} is the mean standard deviation of residual fingerprint scores on held-out real images.

	Signal retained (%)						AUC
Generator	K=0	K=5	K=17	K=75	K=146	K=403	@K=146
CycleGAN	100.0	99.7	96.4	90.2	82.8	64.7	1.00
StarGAN	100.0	99.1	97.2	88.1	78.9	58.7	1.00
StyleGAN	100.0	99.6	98.0	89.4	83.2	62.9	1.00
StyleGAN2	100.0	99.4	98.3	87.1	78.9	55.1	1.00
WhichFaceReal	100.0	98.7	96.2	86.2	77.2	57.0	1.00
VQDM	100.0	99.7	97.8	85.1	75.3	52.5	1.00
ADM	100.0	99.0	96.9	85.7	75.1	56.1	.887
Mean retained	100.0	99.3	97.3	87.4	**78.8**	58.1	**.984**
σ_{real}	.630	.276	.179	.067	**.040**	.017	—

Table 3: Repair of targeted detector blind spots. Base is the frozen detector, FT is matched-data fine-tuning, and GFI augments Base with generator fingerprints. Cells report AUC (↑) / FPR@R95 (↓); Δ is GFI's AUC gain over Base.

Detector	**Blind Spot Generator**	**Base**	**FT**	**GFI**	**Δ AUC**
DDA	CycleGAN	.83/.54	.91/.38	**.98/.10**	+.15
DDA	StarGAN	.72/.71	.91/.36	**.99/.01**	+.27
DDA	VQDM	.89/.36	.97/.10	**.98/.11**	+.09
DDA	WhichFaceIsReal	.87/.37	**.96/.14**	.93/.29	+.06
Effort	StyleGAN2	.69/.82	.71/.81	**.98/.09**	+.29
Effort	WhichFaceIsReal	.75/.50	.75/.49	**.84/.39**	+.09
Effort	ADM	.80/.49	.81/.48	**.92/.32**	+.12
SPAI	CycleGAN	.87/.45	.88/.44	**.97/.14**	+.10
SPAI	StyleGAN	.79/.69	.77/.69	**.97/.24**	+.18
SPAI	WhichFaceIsReal	.58/.79	.58/.79	**.72/.69**	+.14
SPAI	ADM	.81/.64	.81/.64	**.94/.25**	+.13
Mean	*11 targets*	.78/.58	.82/.47	**.93/.24**	+.15

to 0.127 for DDA, from 0.603 to 0.266 for Effort, and from 0.643 to 0.33 for SPAI. All 11 target pairs show statistically significant AUC improvements under paired DeLong tests (Appendix G.2).

Compared with fine-tuning (FT) on the same target-generator images, FT recovers many DDA blind spots but yields little improvement for Effort and SPAI. GFI improves all three detectors without updating detector parameters. The exception is DDA-WhichFaceIsReal, where FT slightly outperforms GFI (0.96 vs. 0.93 AUC). We also note SPAI-WhichFaceIsReal and Effort-WhichFaceIsReal, though substantially improved, remain below the 0.90 threshold.

5.4 Does GFI Preserve Generalization Beyond the Targets?

Improving blind spots should not come at the cost of performance elsewhere. We therefore check GFI along two further dimensions: its effect on *non-target generators*, i.e., generators outside the 11 blind-spot targets from Section 5.2, and false positives under real-image distribution shift.

Non-target generators. For each detector, we take the *fingerprint bank*, built in Section 5.3 for that detector's target generators, and apply it to the detector-generator pairs *excluded* from fingerprint construction. This tests whether equipping a detector with fingerprints for its known blind spots introduces false alarms on generators it did not need help with.

GFI preserves non-target performance overall, with mean AUC improving for all three detectors. Mean AUC changes from 0.98 to 0.99 for DDA, from 0.93 to 0.94 for Effort, and from 0.94 to 0.96 for SPAI. Mean FPR@R95 changes from 0.11 to 0.05 for DDA, stays at 0.21 for Effort, and improves from 0.28 to 0.16 for SPAI. The full per-generator breakdown is in Appendix G.1, Table 9.

Table 4: False-positive rate on OOD real domains (↓), under a 5%-calibrated fixed threshold on $\mathcal{R}_{\text{cal}}$. **Base** is the absolute FPR of the frozen detector; **ΔFT** and **ΔGFI** are the change relative to Base.

	DDA			Effort			SPAI		
	Base	ΔFT	ΔGFI	Base	ΔFT	ΔGFI	Base	ΔFT	ΔGFI
CelebA	.10	−.03	+.04	.00	±.00	±.00	.22	−.01	+.05
ImageNet-val	.03	−.01	+.03	.01	±.00	+.04	.05	±.00	+.10
LAION-face	.06	+.01	+.05	.11	±.00	+.04	.17	−.02	+.05
LAION-web	.20	+.22	+.11	.00	±.00	+.01	.06	±.00	+.12
RAISE	.10	−.05	−.02	.39	±.00	+.06	.04	±.00	+.02
WildRF	.08	−.03	±.00	.16	±.00	+.03	.08	±.00	+.04
SynthWildX	.09	−.03	+.01	.29	±.00	+.01	.09	±.00	+.05
B-Free viral	.01	−.01	±.00	.04	±.00	−.02	.12	±.00	+.05
Mean	.08	+.01	+.03	.13	±.00	+.02	.10	±.00	+.06

Robustness to real-image distribution shift. A separate concern is whether GFI's added fingerprint evidence causes more false positives when the *real* images themselves come from an unfamiliar distribution. We test this on eight out-of-distribution (OOD) real-image domains: CelebA, ImageNet-Val, LAION-Face, LAION-Web, RAISE, WildRF, SynthWildX, and B-Free viral images.

Unlike the FPR@R95 metric used elsewhere in this paper, here we use a single fixed decision threshold, calibrated once on the held-out real calibration set $\mathcal{R}_{\text{cal}}$, and apply it unchanged across all eight OOD domains. (FPR@R95 is not meaningful here, since these domains contain no generated images against which to measure recall.)

Table 4 shows a modest increase in mean FPR under the fixed threshold: from 0.08 to 0.11 for DDA, from 0.13 to 0.15 for Effort, and from 0.10 to 0.16 for SPAI. Degradation is concentrated in a few domains, chiefly LAION-Web (+0.11 DDA, +0.12 SPAI) and ImageNet-Val (+0.10 SPAI). Overall, GFI is not fully invariant to real-image distribution shift, but degradation is concentrated in specific domains, highlighting calibration as an important deployment consideration.

6 Limitations

1. **Adversarial robustness:** GFI relies on generator-dependent signal that remains in the residual feature space. As with other feature-based detection methods, an adversary that specifically targets this signal could reduce fingerprint-based evidence while preserving image semantics. Robustness to adversarial manipulation remains an open direction.
2. **First-order fingerprint representation:** GFI represents each generator by the normalized mean displacement between generated and real images in the residual feature space. Our experiments show that this representation captures sufficient generator-dependent signal across the evaluated generators. However, it does not explicitly model higher-order distributional structure, which may provide additional signal for generators whose characteristics are not well represented by a mean displacement.
3. **Real-image distribution shift:** GFI calibrates scores on a fixed real-image reference set, so absolute fingerprint scores and operating thresholds depend on the calibration population. Under distribution shift, calibration can change and affect absolute FPR even when the underlying fingerprint signal is unchanged. Our diverse evaluation shows generally low FPR, with larger variation on some domains; broader calibration data can reduce this dependence by better representing the deployment population.

7 Conclusion

We present GFI, a training-free mechanism for addressing generator-specific blind spots. GFI removes dominant real-image feature directions, represents each target generator through a fingerprint, and adds the resulting evidence to a frozen detector. Across AIGCDetectBench, GFI improves blind-spot detection without systematic degradation on non-target generators, showing that generator-dependent evidence can be extracted from pretrained features and incorporated without modifying the detector.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability and presentation of the manuscript, including editing and restructuring text. We did not use generative AI assistants to design, implement, or analyze the experiments. Generative image models were used to produce the enrollment images described in Section 5. We reviewed all AI-assisted text and independently verified the technical content, experimental claims, and final manuscript. We take full responsibility for the final content of this work, including all claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies the detection of AI-generated images and is intended to improve the reliability of existing detection systems. We use publicly available datasets and do not collect or process personally identifiable information. The proposed method is designed as a post-hoc augmentation of existing detectors and does not introduce new capabilities for generating or manipulating images. We acknowledge that AI-generated image detectors can exhibit biases across image distributions and generators; our evaluation therefore includes non-target generators and diverse real-image distributions to assess potential false-positive degradation.

REPRODUCIBILITY STATEMENT

The paper provides the complete description of GFI, including fingerprint construction, residual-subspace estimation, score calibration, and inference in Section 4. The experimental setup, datasets, detector configurations, evaluation protocol, and target selection are described in Section 5, with additional implementation and ablation details provided in the appendix. We will release the implementation and configuration files used to obtain the reported results as supplementary material.

REFERENCES

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *ECCV*, 2020.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes AI-generated image detector easier generalizable. In *NeurIPS*, 2025.

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.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of AI-generated image detection with CLIP. In *CVPR Workshops*, 2024.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *ICML*, 2020.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general AI-generated image detection. In *CVPR*, 2025.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution AI-generated image detection by spectral learning. In *CVPR*, 2025.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *ECCV*, pp. 394–411, 2024.

Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. LaRE^2: Latent reconstruction error based method for diffusion-generated image detection. In *CVPR*, 2024.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs leave artificial fingerprints? In *MIPR*, 2019.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023.

Ziheng Qin, Yuheng Ji, Renshuai Tao, Yuxuan Tian, Yuyang Liu, Yipu Wang, and Xiaolong Zheng. Scaling up ai-generated image detection with generator-aware prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *CVPR*, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.

Sergey Sinitsa and Ohad Fried. Deep image fingerprint: Towards low budget synthetic image detection and model lineage analysis. In *WACV*, 2024.

Hae Jin Song, Mahyar Khayatkhoei, and Wael AbdAlmageed. ManiFPT: Defining and analyzing fingerprints of generative models. In *CVPR*, 2024.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: generalized artifacts representation for GAN-generated images detection. In *CVPR*, 2023.

Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In *CVPR*, 2024.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *CVPR*, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *ICCV*, 2023.

Haozhen Yan, Ruoxin Chen, Jiahui Zhan, Bo Wang, Youchang Xiao, Shouhong Ding, Liqing Zhang, Taiping Yao, and Jianfu Zhang. LoRC: Detecting AI-generated images via low-rank collapse in semantic residuals. In *Computer Vision–ECCV 2026*, volume 17035 of *Lecture Notes In Computer Science*. Springer, 2026. doi: 10.1007/978-3-032-37464-6_12.

Jiazhen Yan, Fan Wang, Weiwei Jiang, Ziqiang Li, and Zhangjie Fu. NS-Net: Decoupling CLIP semantic information through NULL-space for generalizable AI-generated image detection. *arXiv preprint arXiv:2508.01248*, 2025a.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In *ICLR*, 2025b.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *ICML*, 2025c.

Ning Yu, Larry S. Davis, and Mario Fritz. Attributing fake images to GANs: learning and analyzing GAN fingerprints. In *ICCV*, 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.

A Experimental Setup and Reproducibility

This appendix provides implementation details and controlled analyses that complement the main-paper experiments. In particular, we examine the construction of the residual representation, the distribution of measured generator-discriminative signal across the feature spectrum, the dependence and transferability of generator fingerprints, and the robustness of the resulting evidence to changes in the representation and input images. We also provide additional statistical and fusion analyses.

The main paper already reports the benchmark composition, detector comparisons, identified blind spots, the choice of $\gamma = 0.95$, the performance of GFI on the targeted detector–generator pairs, and its effect on non-target generators. We therefore focus here on experiments that provide additional evidence for the mechanism, statistical reliability, and practical behavior of GFI.

The analyses in this appendix focus on the seven generators identified as blind spots for at least one detector in the main paper—CycleGAN, StarGAN, StyleGAN, StyleGAN2, WhichFaceIsReal, VQDM, and ADM. Unless otherwise noted, mechanism experiments use CLIP ViT-L/14 features at $\gamma = 0.95$ on the fixed canonical real-image fit/calibration split.

GFI performs no gradient-based optimization. The base detector and feature encoder remain frozen throughout; the residual subspace, generator fingerprints, and calibration distributions are estimated directly from data without updating detector or encoder parameters.

A.1 What data and representations are used?

The roles of the real-image and generated-image sets used throughout the experiments are summarized in Section 5.1 of the main paper. In particular, $\mathcal{R}_{\text{fit}}$ is used to estimate the real-image mean and residual PCA subspace, while the disjoint calibration set $\mathcal{R}_{\text{cal}}$ is used to calibrate fingerprint alignment and the frozen detector score. The enrollment images $\mathcal{S}_k$ are self-generated for each target generator and are separate from the images used by AIGCDetectBench for final evaluation.

For the mechanism analyses, we construct a real-image reference distribution from up to $\sim$6,000 images drawn from three domains. Each source is capped at 2,000 images to prevent any single domain from dominating the estimated feature distribution. The resulting images are merged, sorted by file path, and globally shuffled using seed 42.

We use two frozen pretrained visual encoders, CLIP ViT-L/14 and DINOv3 ViT-L. For both encoders, we extract the `[CLS]` token from the 8th transformer block (0-indexed), giving a feature dimension of $D = 1024$. All backbone parameters remain frozen.

The reported set sizes follow the capping and splitting rules described above. The pooled real images are partitioned into a canonical fit set $\mathcal{R}_{\text{fit}}$ and a disjoint calibration set $\mathcal{R}_{\text{cal}}$; this split is fixed across experiments to ensure a consistent evaluation protocol. Generator enrollment is performed directly from the self-generated set $\mathcal{S}_k$ for each generator, without gradient-based optimization or updates to the detector or feature encoder. The AIGCDetectBench evaluation images are kept separate from these reference and enrollment sets.

A.2 How are the experiments evaluated?

All experiments use a fixed canonical real-image fit/calibration split. Images are sorted by file path before splitting to ensure deterministic assignment. The fit set $\mathcal{R}_{\text{fit}}$ is used to estimate the real-image mean and residual PCA subspace, while the disjoint calibration set $\mathcal{R}_{\text{cal}}$ is used for fingerprint-score calibration and calibration of the frozen detector score. Generator enrollment is performed directly from the self-generated enrollment set $\mathcal{S}_k$, without gradient-based optimization.

AUC and FPR@R95 are the primary evaluation metrics. AUC is the area under the ROC curve computed from the detection scores, while FPR@R95 denotes the false-positive rate on real images at the operating point corresponding to 95% recall on generated images. The operating rule is fixed across methods, while the resulting threshold may differ across detector variants and generators.

For the analysis of the variance-removal threshold, we additionally use the sensitivity index

$$d' = \frac{\bar{s}_g - \bar{s}_r}{\sigma_{\text{pooled}}}, \qquad \sigma_{\text{pooled}} = \sqrt{\frac{\sigma_r^2 + \sigma_g^2}{2}},$$

where $\bar{s}_r$ and $\bar{s}_g$ denote the mean real and generated scores.

We also report the standard deviation of scores assigned to real images, denoted by σ_{real}. This quantity measures the dispersion of the real-image score distribution and is used as a diagnostic of score stability.

For statistical comparisons, paired AUC tests use the DeLong test with the Sun–Xu midrank implementation.

The random seeds are fixed throughout: COCO sampling uses seed 99, CNNDetection sampling uses seed 77, the global shuffle uses seed 42, and feature-space intervention experiments use seed 42. Bootstrap experiments use 1,000 resamples with seeds 43–1042.

B GFI SPECIFICATION

The main paper introduces GFI as a three-stage procedure: *isolate*, *calibrate*, and *fuse*. This section gives the precise computational specification used in the experiments. The procedure is summarized in Algorithm 1.

The residual subspace is estimated once for a given real-image reference distribution. Once this subspace has been constructed, generator fingerprints can be estimated independently without recomputing the subspace or modifying previously estimated fingerprints.

B.1 WHAT EXACTLY DOES GFI COMPUTE?

The fingerprint bank is

$$\mathcal{B} = \{(\mathbf{e}_k, F_k) : k \in \mathcal{K}\},$$

where $\mathcal{K}$ denotes the set of enrolled generators. The bank therefore stores both the generator-dependent direction and the real-image calibration distribution associated with that direction.

At inference, the query representation is first centered and projected into the residual subspace. It is then aligned with every enrolled fingerprint. These raw alignments are converted into calibrated evidence scores and aggregated by averaging the $m' = \min(m, |\mathcal{K}|)$ strongest calibrated matches,

$$g(x) = \frac{1}{m'} \sum_{j=1}^{m'} z_{(j)}(x), \qquad z_{(1)} \geq z_{(2)} \geq \cdots,$$

with $m = 2$ used throughout. Top-m pooling keeps the evidence responsive to the strongest generator matches rather than diluting a strong match across the full bank. The final GFI score is

$$s(x) = z_{\text{base}}(x) + g(x).$$

Thus, GFI transforms generator-dependent evidence through three distinct stages:

$$a_k(x) \;\longrightarrow\; z_k(x) \;\longrightarrow\; g(x) \;\longrightarrow\; s(x).$$

Here, $a_k(x)$ is the raw alignment of x with the fingerprint of generator G_k; $z_k(x)$ converts this alignment into calibrated evidence relative to held-out real images; $g(x)$ aggregates the calibrated evidence across the enrolled generators; and $s(x)$ combines this fingerprint evidence with the calibrated score of the frozen base detector. This separates *fingerprint matching*, *evidence calibration*, *evidence aggregation*, and *detector fusion* into explicit, sequential operations.

C WHY DOES FINGERPRINT ISOLATION WORK?

The main paper establishes that removing the highest-variance real-image directions at $\gamma = 0.95$ improves the residual-space representation. The analyses below examine why this occurs: whether the removed directions contain most of the measured generator-discriminative signal, whether the retained signal is concentrated in the residual spectrum, and whether the result can be explained by numerical effects or arbitrary dimensionality reduction.

Algorithm 1 Generator Fingerprint Isolation (GFI)

Stage 1: Isolate *[once, from $\mathcal{R}_{\text{fit}}$]*

Require: Real features $\mathbf{F}_r \in \mathbb{R}^{n_r \times D}$, variance threshold γ

$\boldsymbol{\mu}_{\text{real}} \leftarrow \frac{1}{n_r} \sum_j \mathbf{f}_{r,j}$

$\mathbf{Z} \leftarrow \mathbf{F}_r - \mathbf{1}\boldsymbol{\mu}_{\text{real}}^\top$

Compute PCA basis $\{\mathbf{v}_i, \lambda_i\}_{i=1}^D$ from $\mathbf{Z}$

Find the smallest $K(\gamma)$ satisfying

$$\frac{\sum_{i=1}^{K(\gamma)} \lambda_i}{\sum_{i=1}^{D} \lambda_i} \geq \gamma$$

$\mathbf{Q} \leftarrow \sum_{i=K(\gamma)+1}^{D} \mathbf{v}_i \mathbf{v}_i^\top$ ▷ residual projection matrix

return $(\boldsymbol{\mu}_{\text{real}}, \mathbf{Q})$

Stage 2: Calibrate *[per generator G_k; base once]*

Require: Enrollment features $\mathbf{F}_{G_k}$; real calibration set $\mathcal{R}_{\text{cal}}$; base scorer $b(\cdot)$

$\boldsymbol{\mu}_{G_k} \leftarrow \frac{1}{n_g} \sum_j \mathbf{f}_{g,j}$

$\boldsymbol{\delta}_k \leftarrow \mathbf{Q}(\boldsymbol{\mu}_{G_k} - \boldsymbol{\mu}_{\text{real}})$

$\mathbf{e}_k \leftarrow \boldsymbol{\delta}_k / \|\boldsymbol{\delta}_k\|_2$

$F_k \leftarrow \text{ECDF}\left\{\mathbf{e}_k^\top \mathbf{Q}(\mathbf{f}(r) - \boldsymbol{\mu}_{\text{real}}) : r \in \mathcal{R}_{\text{cal}}\right\}$

$F_{\text{base}} \leftarrow \text{ECDF}\{b(r) : r \in \mathcal{R}_{\text{cal}}\}$

return $(\mathbf{e}_k, F_k), F_{\text{base}}$

Stage 3: Fuse *[per query image x]*

$\mathbf{u}(x) \leftarrow \mathbf{f}(x) - \boldsymbol{\mu}_{\text{real}}$

$\tilde{\mathbf{f}}(x) \leftarrow \mathbf{Q}\mathbf{u}(x)$

$a_k(x) \leftarrow \mathbf{e}_k^\top \tilde{\mathbf{f}}(x)$ ▷ raw fingerprint alignment

$z_k(x) \leftarrow \Phi^{-1}(F_k(a_k(x)))$

Sort $\{z_k(x)\}_{k \in \mathcal{K}}$ as $z_{(1)} \geq z_{(2)} \geq \cdots$;

$m' \leftarrow \min(m, |\mathcal{K}|)$ ▷ $m = 2$ throughout

$g(x) \leftarrow \frac{1}{m'} \sum_{j=1}^{m'} z_{(j)}(x)$ ▷ top-m pooling

$z_{\text{base}}(x) \leftarrow \Phi^{-1}(F_{\text{base}}(b(x)))$

return

$$s(x) \leftarrow z_{\text{base}}(x) + g(x)$$

C.1 Do dominant real-image directions contain most of the generator-discriminative signal?

Let

$$d_i = v_i^\top (\mu_{G_k} - \mu_{\text{real}})$$

denote the generator–real mean displacement along principal component i. The cumulative generator–real displacement energy contained in the first k components is

$$E_k = \frac{\sum_{i=1}^{k} d_i^2}{\sum_{i=1}^{D} d_i^2}.$$

Figure 2 compares this cumulative generator displacement with the cumulative variance of the real-image feature distribution. The dominant principal components explain most of the real-image variance and also contain a large fraction of the raw generator mean displacement.

As shown in Figure 2, at $\gamma = 0.95$ the first $K = 146$ principal components account for 95% of the variance in CLIP real-image features and also contain a large fraction of the raw generator mean displacement. Raw displacement, however, does not directly determine how strongly a direction separates generated from real images because the principal directions have different levels of variability among real images. In particular, λ_i measures the variance of real-image features along principal direction i, whereas d_i measures the generator–real mean shift along that direction. We therefore

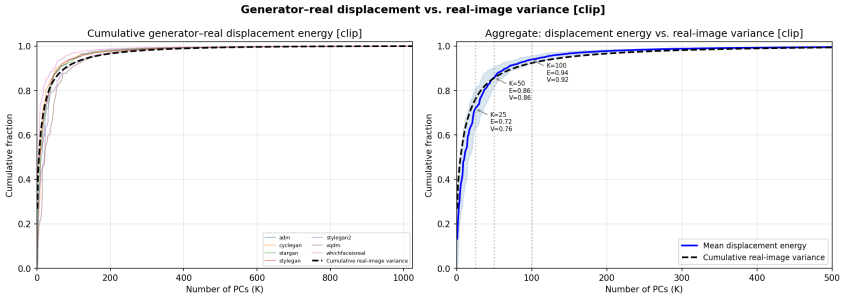

Figure 2: **Generator–real displacement vs. real-image variance (CLIP ViT-L/14).** Cumulative generator–real mean-displacement energy and cumulative real-image feature variance across the PCA spectrum. The dashed line marks $\gamma = 0.95$.

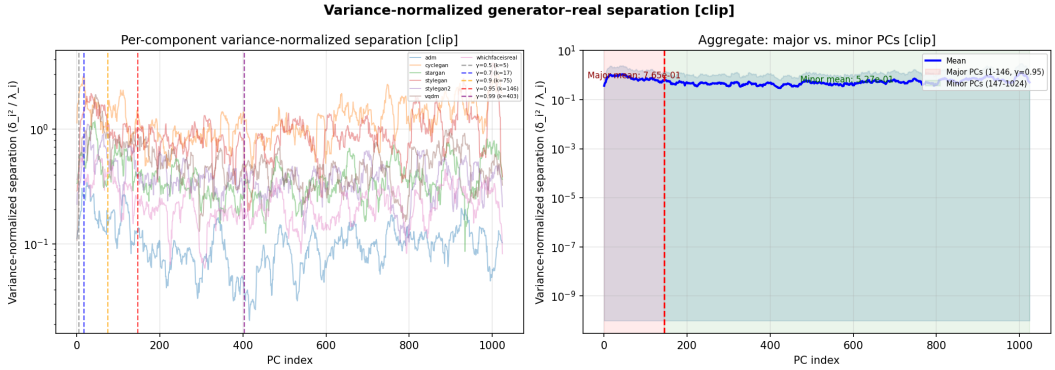

Figure 3: **Per-component discriminability profile (CLIP ViT-L/14).** $J_i = d_i^2/\lambda_i$ across the variance-ordered PCA spectrum.

normalize the squared shift by the corresponding real-image variance,

$$J_i = \frac{d_i^2}{\lambda_i},$$

which measures generator–real separation relative to the variability naturally present in real images.

Figure 3 shows the resulting per-component discriminability profile across the variance-ordered PCA spectrum. Only about 21% of the total variance-normalized generator–real separation lies in the first 146 components, while roughly 79% remains in the residual 878 components.

Thus, the directions that account for most real-image variance contain most of the raw generator displacement but only a minority of the variance-normalized generator–real separation.

This energy–discriminability mismatch provides the central empirical motivation for the isolate stage: removing high-variance real-image directions suppresses a large amount of real-image variation while retaining substantial measured generator-discriminative signal.

C.2 Is the result explained by numerical instability in tiny eigenvalues?

Because J_i normalizes displacement by λ_i, and the residual subspace corresponds to the lower-variance tail of the spectrum, one concern is that the preserved discriminability could be driven by numerically tiny eigenvalues.

We therefore clamp all eigenvalues to prescribed variance floors before computing the discriminability decomposition. Table 5 reports the resulting discriminability loss and preserved fraction.

As shown in Table 5, the preserved discriminability remains unchanged at 78.7% (loss 21.3%) from a floor of 10^{-12} through 10^{-6}. It decreases only at 10^{-4}, where the floor begins to remove directions

Table 5: Preserved discriminability under variance flooring (CLIP).

Floor ϵ	Mean disc. loss	Preserved fraction
10^{-12}	0.2128	0.7872
10^{-10}	0.2128	0.7872
10^{-8}	0.2128	0.7872
10^{-6}	0.2128	0.7872
10^{-4}	0.3287	0.6713

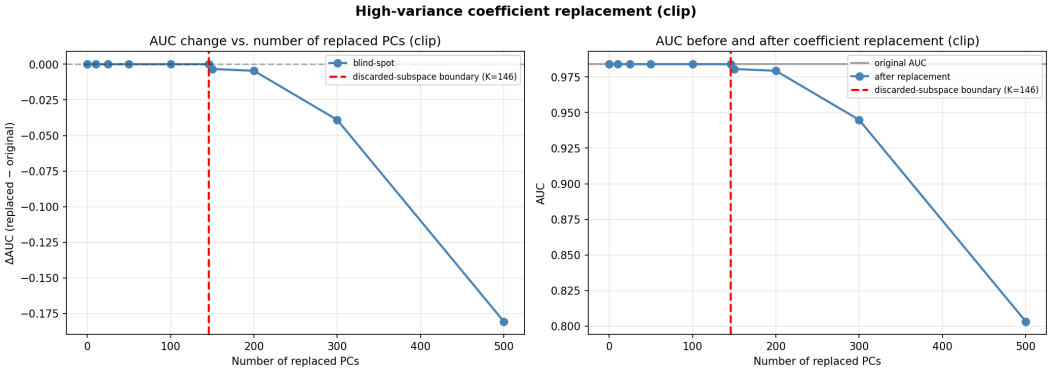

Figure 4: **High-variance coefficient replacement.** AUC as the number of replaced principal components increases at $\gamma = 0.95$. Performance remains unchanged while modifications are confined to the discarded subspace and begins to decrease once the intervention reaches the retained residual components.

carrying measurable discriminative signal, increasing the loss to 32.9%. Thus, the observed residual signal is not explained by numerical effects associated with near-zero eigenvalues.

D What Does the Residual Representation Capture?

D.1 Is residual evidence insensitive to the discarded high-variance directions?

We perform a controlled feature-space intervention to determine whether the residual evidence depends on the discarded principal components. For each generated image, we replace the coefficients of the top-K principal components with those of a randomly selected ImageNet image while leaving all residual components unchanged.

The resulting detection performance is shown in Figure 4.

As shown in Figure 4, replacing the coefficients of all principal components up to the discarded boundary, $K = 146$, has negligible effect on fingerprint evidence, with $\Delta\text{AUC} < 0.001$. Performance begins to decrease only when the intervention extends into the retained residual subspace.

This behavior follows directly from

$$\mathbf{Q}\mathbf{v}_i = 0, \qquad i \leq K(\gamma).$$

Consequently, modifications confined to the discarded subspace do not contribute to the residual representation and therefore do not affect fingerprint alignment. The intervention provides a controlled test that the residual evidence is insensitive to changes in the removed high-variance directions.

D.2 Does performance depend on which directions are removed?

To determine whether the effect of PCA filtering is simply a consequence of reducing dimensionality, we compare three masking strategies at matched removed dimensionality:

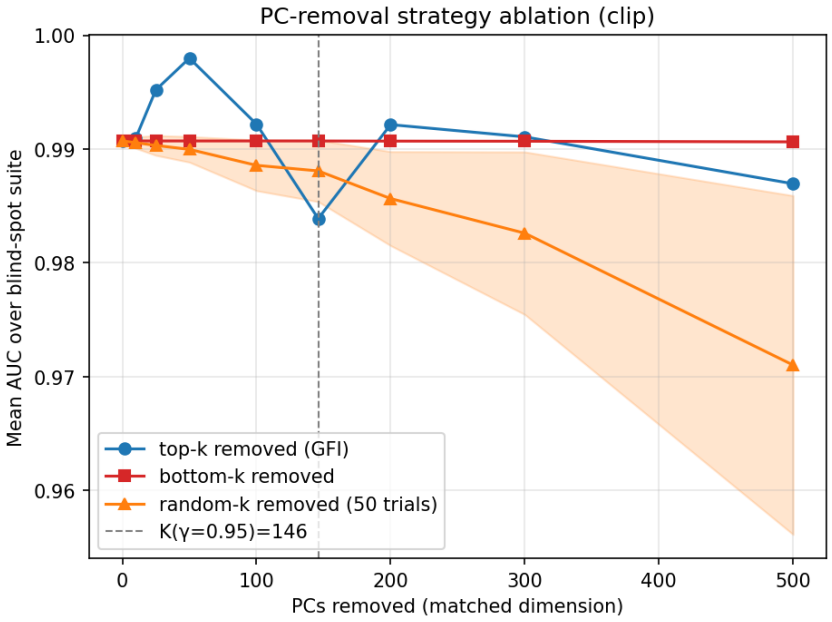

Figure 5: **PC-removal strategy control.** Mean AUC as a function of the number of removed PCs under top-k, random-k, and bottom-k masking. The dashed line marks $K = 146$, corresponding to $\gamma = 0.95$ for this analysis. The fingerprint direction is re-estimated in each retained subspace.

1. **Top-k:** remove the k highest-variance PCs;
2. **Random-k:** remove k uniformly sampled PCs, averaged over repeated trials;
3. **Bottom-k:** remove the k lowest-variance PCs.

For each strategy and value of k, the fingerprint direction is re-estimated in the retained subspace using the same training data and evaluated on the same held-out images. Thus, this experiment tests whether a discriminative direction can be recovered after different subsets of PCs are removed; it does not evaluate a fixed fingerprint under post-hoc masking.

The resulting AUC curves are shown in Figure 5.

Starting from a full-spectrum mean AUC of 0.991, the three strategies produce different behavior. Top-k removal reaches its highest AUC of 0.998 near $k = 50$ and exceeds the full-spectrum baseline at several evaluated dimensionalities. Its behavior is not monotonic, however: at $K = 146$, corresponding to $\gamma = 0.95$, mean AUC decreases to 0.984. Bottom-k removal remains close to the full-spectrum baseline throughout the sweep. Random-k removal degrades more consistently as additional components are removed, reaching an AUC of 0.971 at $k = 500$.

The strong performance under bottom-k removal indicates that generator-discriminative information is not confined exclusively to the lowest-variance PCs. Because the fingerprint is re-estimated after masking, the detector can exploit alternative signal in the remaining high- and mid-variance components. Conversely, the degradation under random removal shows that indiscriminately discarding dimensions is more likely to remove useful discriminative structure.

These results demonstrate that performance depends on which components are removed, but they do not establish that top-k removal is uniformly superior to bottom-k removal. In particular, top-k removal does not maximize AUC at the selected $K = 146$ operating point. Our choice of top-k removal at $\gamma = 0.95$ is instead motivated by the objective of suppressing dominant real-image variation while preserving strong generator discrimination.

Table 6: Full cross-generator transfer matrix at $\gamma = 0.95$ (AUC, CLIP). Mean off-diag 0.788.

train \ eval	adm	cyclegan	stargan	stylegan	stylegan2	whichfaceisreal	vqdm
adm	0.887	0.964	0.957	0.979	0.674	0.507	0.991
cyclegan	0.692	1.000	0.967	0.978	0.664	0.692	0.871
stargan	0.740	0.998	1.000	0.954	0.660	0.989	0.972
stylegan	0.742	0.976	0.857	1.000	0.609	0.675	0.938
stylegan2	0.572	0.765	0.631	0.595	1.000	0.510	0.842
whichfaceisreal	0.436	0.907	1.000	0.875	0.528	1.000	0.479
vqdm	0.751	0.903	0.881	0.982	0.925	0.478	1.000

Table 7: Residual fingerprint analysis across CLIP ViT-L/14 layers at $\gamma = 0.95$.

Block	K removed	mean AUC	mean OOD FPR	mean d'	mean σ_{real}
0	14	0.866	0.051	1.976	0.012
4	69	0.983	0.051	3.826	0.019
8	146	0.984	0.074	6.301	0.040
12	343	0.997	0.131	6.136	0.048
16	402	0.998	0.126	5.021	0.051
20	640	0.982	0.122	3.972	0.115
23	646	0.983	0.112	4.228	0.224

E Is the Residual Evidence Really Generator-Dependent?

E.1 Does a fingerprint represent a specific generator or generic synthetic images?

A key question is whether the enrolled direction captures a generic "generated-image" signature or instead retains generator-dependent structure. We therefore enroll a fingerprint using one generator and evaluate that fingerprint against images from other generators.

The cross-generator transfer results are reported in Table 6.

As shown in Table 6, the mean off-diagonal AUC is 0.788. Thus, an enrolled fingerprint is not simply a universal synthetic-image direction, although it can carry substantial cross-generator signal. Transfer is strongest for several related GAN generators—for example, StarGAN's fingerprint scores WhichFaceIsReal at 0.989 and, in the reverse direction, WhichFaceIsReal's fingerprint scores StarGAN at 1.000. CycleGAN also transfers strongly to StarGAN and StyleGAN, with AUCs of 0.967 and 0.978, respectively. StyleGAN2 is comparatively distinct from the other GAN fingerprints: most other fingerprints score it in the 0.51–0.67 range, while the VQDM fingerprint is an exception at 0.925.

This pattern indicates that the enrolled direction captures structured, generator-dependent properties rather than a single generic synthetic-image direction. It also motivates the fingerprint-bank formulation used by GFI: each generator contributes its own evidence channel, while related generators can partially reinforce one another.

F How Robust Is Fingerprint Isolation?

F.1 Does the phenomenon depend on feature-extraction layer?

The default representation used throughout the paper is extracted from block 8. To determine whether the observed residual fingerprint structure is specific to this choice, we evaluate the residual fingerprint score across the transformer blocks of CLIP ViT-L/14 at $\gamma = 0.95$.

Table 7 reports the number of removed components, mean AUC, mean OOD FPR, mean sensitivity index d', and mean real-score dispersion at the evaluated layers.

Table 8: Backbone comparison at $\gamma = 0.95$ (per-generator AUC).

Generator	CLIP	DINOv3
adm	0.887	0.950
cyclegan	1.000	1.000
stargan	1.000	1.000
stylegan	1.000	1.000
stylegan2	1.000	1.000
whichfaceisreal	1.000	0.994
vqdm	1.000	1.000
Mean	0.984	0.992

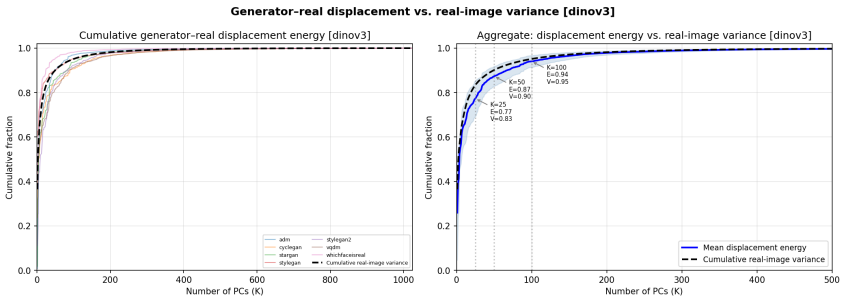

Figure 6: **Generator–real displacement vs. real-image variance (DINOv3).** Cumulative generator–real mean-displacement energy and cumulative real-image feature variance across the PCA spectrum. The dashed line marks $\gamma = 0.95$.

As shown in Table 7, the earliest block provides insufficiently discriminative residual features (AUC 0.866 at block 0). From block 4 onward, AUC is uniformly high (≥ 0.979), peaking at 0.998 at block 16. However, deeper representations also exhibit greater real-image score dispersion and higher OOD FPR: OOD FPR rises from 0.074 at block 8 to 0.12–0.13 at blocks 12–16, while σ_{real} grows from 0.040 at block 8 to 0.224 at block 23.

Block 8 therefore provides a favorable trade-off between discriminability and stability—near-peak AUC (0.984) with low OOD FPR among the strong layers and a tight real-image score distribution—and is retained as the default representation throughout the paper.

F.2 Does fingerprint isolation persist across pretrained backbones?

We next compare CLIP ViT-L/14 and DINOv3 ViT-L at $\gamma = 0.95$. This test asks whether the variance–discriminability relationship is a property of a particular pretrained representation or whether it also appears in a different pretrained backbone.

The comparison is summarized in Table 8.

As shown in Table 8, both backbones yield strong generator discriminability, with mean AUC of 0.984 for CLIP and 0.992 for DINOv3. The difference is concentrated on ADM (0.887 for CLIP versus 0.950 for DINOv3). The two backbones differ mainly in the amount of variance they must remove and in the resulting real-image score distribution. At $\gamma = 0.95$, CLIP removes $K = 146$ components, which carry about 21% of the total variance-normalized discriminability (79% retained), whereas DINOv3 removes only $K \approx 97$ components, carrying roughly 11% of the discriminability ($\approx 89\%$ retained).

The corresponding DINOv3 cumulative analysis is shown in Figure 6.

Figure 6 shows that the qualitative variance–discriminability asymmetry also occurs with DINOv3. Although DINOv3 attains a slightly higher mean AUC, its real-score dispersion is roughly $4.6\times$ larger than CLIP's ($\sigma_{\text{real}} \approx 0.18$ versus 0.04). This broader real-image score distribution provides

Table 9: Performance on non-target generators. Each cell reports **AUC (↑) / FPR@R95 (↓)** for the frozen detector (Base) and GFI. Non-target generators are those not selected as detector-specific blind spots.

DDA			Effort			SPAI		
Generator	**Base**	**GFI**	**Generator**	**Base**	**GFI**	**Generator**	**Base**	**GFI**
ProGAN	.99/.06	1.00/.00	ProGAN	.94/.20	.99/.06	ProGAN	.90/.47	.99/.05
StyleGAN	.94/.32	.97/.17	StyleGAN	.98/.07	.99/.02	StyleGAN2	.90/.41	.98/.11
StyleGAN2	.98/.11	.99/.06	BigGAN	.92/.17	.91/.27	BigGAN	.81/.75	.90/.45
BigGAN	.94/.32	.99/.05	GauGAN	.88/.32	.91/.29	GauGAN	.89/.50	.97/.14
GauGAN	.96/.22	.99/.02	GLIDE	.90/.30	.93/.25	GLIDE	.96/.17	.96/.15
ADM	.96/.20	.99/.06	SD 1.4	1.00/.02	.99/.03	Midjourney	.97/.14	.91/.36
GLIDE	.97/.14	.99/.05	SD 1.5	.99/.02	.99/.04	SD 1.4	1.00/.00	.99/.02
Midjourney	.99/.05	.98/.12	SD-XL	.95/.18	.91/.33	SD 1.5	1.00/.00	.99/.02
SD 1.4	1.00/.00	1.00/.01	VQDM	.90/.40	.94/.28	SD-XL	.99/.04	.96/.19
SD 1.5	1.00/.00	1.00/.00	Wukong	.98/.07	.98/.09	VQDM	.96/.18	.98/.09
SD-XL	1.00/.00	1.00/.01	DALL-E 2	.81/.59	.86/.52	Wukong	1.00/.01	.99/.04
Wukong	1.00/.00	1.00/.00	CycleGAN	.97/.06	.97/.12	DALL-E 2	.89/.47	.89/.46
DALL-E 2	.99/.03	.99/.04	StarGAN	1.00/.01	1.00/.01	StarGAN	.90/.45	1.00/.03
			Midjourney	.83/.52	.77/.65			
Mean	.98/.11	**.99/.05**	*Mean*	.93/.21	**.94/.21**	*Mean*	.94/.28	**.96/.16**

less favorable operating-point stability under real-image distribution shift and motivates the use of CLIP as the primary backbone.

Thus, the qualitative phenomenon is not unique to CLIP, although the choice of backbone affects the stability of the resulting real-image score distribution.

G Does GFI Improve and Extend Existing Detectors?

The main paper already reports the principal detector-level results in Table 3, including the comparison between the frozen detector, parameter efficient fine-tuning, and GFI on each detector's targeted blind spots. We do not repeat those results here. Instead, this section provides the statistical analysis underlying the reported improvements and isolates the contribution of the fingerprint channel.

G.1 Does GFI Degrade Non-Target Generators?

A natural concern is whether adding fingerprints for targeted blind spots could adversely affect generators that are not enrolled in the fingerprint bank. We therefore evaluate GFI on the non-target generators for each detector separately. A generator is considered non-target for a given detector if it is not among that detector's targeted blind spots in Section 5.2. These generators are not used to estimate the corresponding detector-specific fingerprints.

Table 9 reports the complete per-generator results. Overall, GFI does not systematically degrade performance on non-target generators. For DDA, mean AUC increases from 0.98 to 0.99 while mean FPR decreases from 0.11 to 0.05. For Effort, mean AUC increases from 0.93 to 0.94 while mean FPR remains at 0.21. For SPAI, mean AUC increases from 0.94 to 0.96 while mean FPR decreases from 0.28 to 0.16.

The per-generator results show that the effect is not uniformly positive. GFI improves several non-target generators while producing small decreases on others, such as Midjourney for DDA and Effort. Thus, we do not interpret these results as uniform improvement across all non-target generators. Rather, the aggregate results indicate that enrolling fingerprints for identified blind spots does not introduce a systematic performance penalty on the remaining generators. Improvements on some non-target generators are consistent with generator-dependent evidence transferring across related generation processes.

Table 10: Significance of the training-free GFI layer on each detector's registered targets (CLIP).

Detector	Target	base	GFI	ΔAUC	DeLong p	ΔFPR
DDA	cyclegan	0.832	0.980	+0.148	3.3e-105	-0.435
DDA	stargan	0.723	0.997	+0.274	1.6e-269	-0.695
DDA	VQDM	0.890	0.976	+0.085	3.5e-89	-0.253
DDA	whichfaceisreal	0.875	0.934	+0.059	1.6e-16	-0.078
Effort	stylegan2	0.689	0.980	+0.290	$(p < 10^{-300})$	-0.738
Effort	whichfaceisreal	0.748	0.845	+0.097	1.4e-40	-0.106
Effort	ADM	0.804	0.917	+0.113	4.3e-110	-0.168
SPAI	cyclegan	0.875	0.971	+0.097	1.2e-74	-0.309
SPAI	stylegan	0.795	0.966	+0.171	5.0e-221	-0.456
SPAI	whichfaceisreal	0.581	0.722	+0.141	1.3e-62	-0.108
SPAI	ADM	0.811	0.943	+0.132	7.4e-133	-0.390

Table 11: Fingerprint-only contribution on registered generators (CLIP).

Detector bank	Target	fp-only AUC	fp-only FPR@R95
DDA	cyclegan	0.923	0.336
DDA	stargan	1.000	0.000
DDA	VQDM	0.952	0.268
DDA	whichfaceisreal	0.708	0.742
Effort	stylegan2	0.993	0.032
Effort	whichfaceisreal	0.767	0.675
Effort	ADM	0.876	0.606
SPAI	cyclegan	0.921	0.275
SPAI	stylegan	0.968	0.172
SPAI	whichfaceisreal	0.722	0.757
SPAI	ADM	0.911	0.511

G.2 Are the observed detector improvements statistically significant?

We evaluate whether the improvements produced by GFI are statistically distinguishable from the corresponding frozen detector scores. The comparison uses identical evaluation images for the base and GFI scores, so the AUC comparison is paired.

The complete detector–target significance results are reported in Table 10.

As shown in Table 10, all eleven detector–target comparisons show statistically significant improvements under the paired DeLong test. The AUC gains range from $+0.059$ (DDA/WhichFaceIsReal) to $+0.290$ (Effort/StyleGAN2), and the associated p-values are below conventional significance thresholds for all eleven comparisons. Every pair also shows a reduction in FPR@R95, by 0.078 to 0.738.

These results complement the aggregate improvements reported in the main paper by showing that the gains are not explained by sampling variation in the paired evaluation.

G.3 Does the fingerprint score contain useful detection signal on its own?

The fused GFI score is $s(x) = z_{\text{base}}(x) + g(x)$. To understand what the fingerprint channel contributes, we evaluate the aggregated fingerprint score g *on its own*—with the base detector switched off—on each detector's targeted generators. For these isolated fingerprint-channel diagnostics (this subsection and §G.4), g is aggregated as the unweighted mean over the detector's bank, which reports the average per-specialist signal; the deployed fused scores in the main paper use the top-2 pooling specified in the GFI algorithm.

Table 11 reports the fingerprint-only AUC and FPR@R95.

Table 12: Per-image Pearson correlation between base z_b and GFI g (CLIP).

Detector	Target	real	fake	all
DDA	cyclegan	-0.30	+0.06	+0.29
DDA	stargan	-0.05	-0.03	+0.35
DDA	VQDM	+0.05	-0.03	+0.51
DDA	whichfaceisreal	+0.03	-0.02	+0.22
Effort	stylegan2	+0.08	-0.05	+0.30
Effort	whichfaceisreal	-0.07	+0.06	+0.17
Effort	ADM	+0.05	+0.16	+0.39
SPAI	cyclegan	-0.35	+0.21	+0.26
SPAI	stylegan	+0.08	-0.04	+0.40
SPAI	whichfaceisreal	-0.10	+0.06	+0.01
SPAI	ADM	+0.10	-0.07	+0.35

The decomposition in Table 11 shows that the fingerprint channel carries substantial generator-dependent detection signal by itself. On the GAN-family blind spots it is strong to near-perfect—for example, DDA's bank scores StarGAN at AUC 1.000 and CycleGAN at 0.923, SPAI's bank scores StyleGAN at 0.968, and Effort's bank scores StyleGAN2 at 0.993. The weakest fingerprint-only targets are the WhichFaceIsReal images (0.708–0.767), which are also difficult targets for the base detectors.

The fingerprint-only score differs across detectors because each detector's bank aggregates its own set of enrolled specialists. WhichFaceIsReal, for instance, is scored at 0.708 by the DDA bank, 0.767 by the Effort bank, and 0.722 by the SPAI bank. The fingerprint channel is therefore not a detector-independent synthetic-image score: it reflects the generator evidence represented in the corresponding fingerprint bank.

G.4 Does fingerprint evidence provide complementary information?

The motivation for fusion is that fingerprint evidence may capture generator-dependent information that is not fully reflected in the base detector's decision function. We therefore measure the dependence between the calibrated base score and the aggregated fingerprint score on the evaluation images.

Specifically, we compute the per-image correlation between $z_{\mathrm{base}}(x)$ and $g(x)$ *within each class*—first on real images and then on generated images—for every targeted generator. This within-class analysis is important: pooling real and generated images would induce a positive correlation simply because both scores separate the two classes. Conditioning on the class instead asks whether the two channels make similar rankings among images of the same type.

Table 12 reports the Pearson correlations for all three base detectors; the Spearman correlations show the same qualitative pattern.

As shown in Table 12, the within-class correlations are small: they lie in $[-0.35, +0.21]$, and most have magnitude below 0.1. The fingerprint channel therefore exhibits substantially different image-level behavior from the base detector rather than simply reproducing its ranking.

The most negative within-class values occur on real CycleGAN images for DDA (-0.30) and SPAI (-0.35), where weaker fingerprint alignment does not track the base detector's ordering. In contrast, the pooled "all" column is consistently positive ($+0.01$ to $+0.51$), as expected when both channels separate real from generated images; this class-mixing effect is what the within-class conditioning removes.

The low within-class correlations indicate that the fingerprint and base scores exhibit different image-level rankings, consistent with complementary evidence.

H SUMMARY OF SUPPLEMENTARY FINDINGS

The supplementary experiments provide evidence for several properties of the GFI construction. First, the variance–discriminability analysis shows that the dominant principal components explain most of the real-image variance but only a minority of the measured variance-normalized generator–real discriminability. At $\gamma = 0.95$, the first 146 components remove 95% of real-image variance while retaining about 79% of the measured generator-discriminative signal.

Second, the masking experiments show that performance depends on which components are removed rather than on dimensionality reduction alone. The coefficient-replacement experiment confirms that changes confined to the discarded subspace do not alter residual fingerprint scores.

Third, cross-generator transfer demonstrates that fingerprints are generator-dependent rather than a generic synthetic-image direction. The mean off-diagonal AUC is 0.788, with stronger transfer among several related GAN generators. This supports the use of a fingerprint bank containing independently estimated generator evidence.

Finally, the backbone experiment shows that the qualitative variance–discriminability relationship persists across pretrained encoders, while the detector-level analyses show that the fingerprint channel is individually informative, statistically complementary to the base detector, and associated with statistically significant improvements on all eleven targeted detector–generator pairs.

Together, these results provide supplementary evidence for the intended post-hoc use of GFI as an additional generator-dependent evidence channel for an existing frozen detector.