THE TELL-TALE DENOISER: DIFFUSION MODEL ATTRIBUTION VIA SPECTRAL GEOMETRY SIGNATURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Attributing a generated image to its source diffusion model is a fundamental challenge in provenance verification and intellectual property protection. This problem is particularly difficult because diffusion models trained on different datasets can converge to similar score functions and thus similar output distributions, making the generated images themselves unreliable as attribution evidence. Existing non-invasive methods either fail on architecturally similar variants or rely on signals that vanish when models share the same autoencoder. We propose *Spectral Geometry Signatures* (SGS), a non-invasive attribution method that fingerprints each candidate model's denoising behavior rather than its outputs. Our key insight is that a model's score function exhibits a distinctive spectral geometry (how it redistributes energy across spatial frequency bands during denoising) that is intrinsic to the denoiser and invariant to the choice of autoencoder. SGS extracts this signature via frequency-controlled perturbations and standard forward passes, with no inversion, optimization, or generation-time enrollment. Across closely related diffusion checkpoints and models differing in architecture and training procedure, SGS provides highly discriminative attribution signals, remains effective under cross-domain prompt shift, and is stable across the autoencoder choices we evaluate. Our code is available at `https://anonymous.4open.science/r/SGS-1B08`.

1 INTRODUCTION

Diffusion models have become the dominant paradigm for large-scale generative modeling, with systems such as Stable Diffusion (Rombach et al., 2022), SDXL (Podell et al., 2023), and PixArt-α (Chen et al., 2023) enabling high-fidelity image synthesis at unprecedented scale. As these models are widely distributed and fine-tuned in open ecosystems, the need to verify model provenance has become increasingly critical for ownership verification, provenance tracking, and misuse detection. In particular, *given a generated image, can we reliably determine which model produced it?*

This problem is fundamentally challenging. Diffusion models generate images through a stochastic, multi-step denoising process in which noise is gradually transformed into a final image via a learned score function. Prior work has shown that even models trained on different datasets can converge to remarkably similar score functions, producing visually indistinguishable outputs (Kadkhodaie et al., 2023). Attribution methods that rely solely on generated images therefore face an inherent ambiguity, with limited discriminative signal in the output space.

Existing approaches to model attribution fall into two broad families, each with fundamental limitations. *Invasive methods* such as watermarking require retraining or fine-tuning and may degrade generation quality (Zhao et al., 2023; Fernandez et al., 2023; Yuan et al., 2024). *Non-invasive methods* avoid model modification but face two distinct limitations. Classifier-based approaches (Xu et al., 2025; Corvi et al., 2023) operate on generated images and struggle to separate architecturally similar variants. Inversion-based methods such as RONAN (Wang et al., 2023a) and LATENTTRACER (Wang et al., 2024) query candidate models directly but require iterative optimization per candidate. They also lose discriminative power when models share the same autoencoder, because reconstruction losses become uninformative across candidates (Wang et al., 2024). The fundamental gap is the same across all non-invasive approaches: no existing method probes the denoising

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.

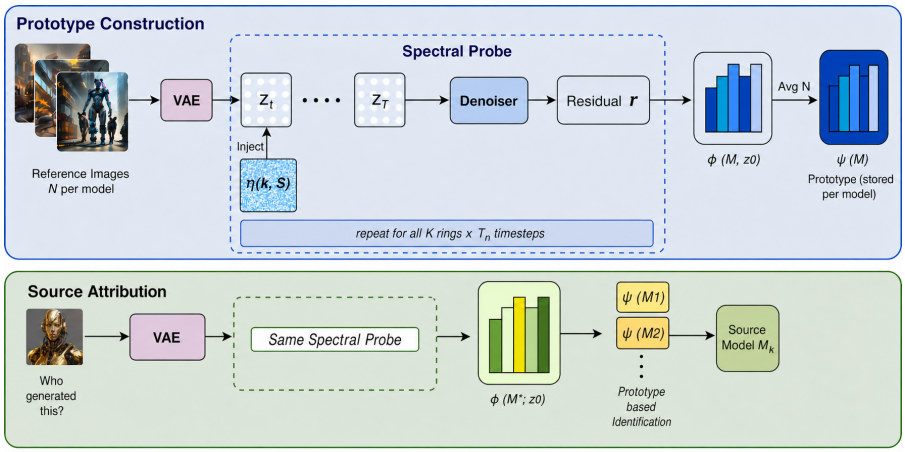

Figure 1: **Pipeline for Spectral Geometry Signatures (SGS)**. Prototype Construction: for each candidate model M, reference images are encoded via a shared VAE to latent z_0. The spectral probe injects band-limited noise $\eta(k, S)$ into frequency ring k, runs a single denoiser forward pass, and measures cross-band energy in the residual $\mathbf{r}$, forming coupling matrix. Repeating over all K rings and T_n timesteps yields signature $\phi(M, z_0)$; averaging over N such signatures produces prototype $\psi(M)$. Source Attribution: a query image of unknown origin is encoded and probed identically. The resulting signature $\phi(M^*, z_0)$ is compared against stored prototypes using classification techniques; the source model M_k is identified as the nearest prototype.

function itself without costly inversion or reliance on output-space signals that vanish for similar models.

A complementary line of work targets model ownership verification rather than source attribution. Teng et al. (2025) show that model-specific information can be recovered from denoiser behavior for ownership verification, while Floros et al. (2026) exploit temporal denoiser dynamics for single-model origin attribution. These results establish the denoiser as a promising source of attribution signal, but they do not yet provide a stable, efficient representation for fine-grained n-way discrimination among closely related candidate models.

In this work, we take a different perspective. Instead of attributing images by analyzing *what a model generates*, we attribute them by characterizing *how a model denoises*. To this end, we propose **Spectral Geometry Signatures** (SGS), a non-invasive model attribution method that characterizes diffusion models through the spectral structure of their score function (see Fig. 1). Our key insight is that each model exhibits a distinctive *spectral geometry*, reflected in how it redistributes energy across spatial frequency bands during the denoising process. By injecting frequency-controlled perturbations and measuring cross-band energy transfer over timesteps, SGS extracts a stable signature that is intrinsic to the model and, as we show empirically, invariant to the choice of autoencoder used during probing. The method requires only standard forward passes, with no inversion, optimization, or model modification.

We evaluate SGS on a challenging closed-set[1] task spanning fifteen diffusion models that differ in training data, architecture, and training procedure . Accuracy holds at 99.98% in a 19-way task that pools four additional flow-matching models against the full diffusion registry, showing the spectral coupling principle generalizes across training objectives. This enables us to answer key questions left open by prior work: can models with identical architectures but different datasets be distinguished? Can fingerprints persist across different architectures trained on similar data? Does distillation leave a detectable imprint on an otherwise identical model? Our results show that SGS reliably captures all three sources of variation, providing a unified view of how diffusion models encode their identity.

In summary, we make four contributions: (1) we introduce SGS, a non-invasive attribution framework that characterizes candidate generative models through their responses to controlled spectral perturbations, requiring only forward evaluations and no inversion, optimization, or model modification; (2) we show that these response signatures enable fine-grained attribution across both architecturally distinct models and closely related checkpoints that are difficult to distinguish from

[1]The true source model is assumed to be present in the candidate registry.

generated outputs alone; (3) we establish that the resulting attribution signal remains highly discriminative under changes in prompt distribution, image post-processing, probing autoencoder, and increased same-family competition; and (4) we demonstrate that the same spectral-response principle extends beyond standard diffusion denoisers to the flow-matching models.

2 Background & Related Work

Diffusion denoising. Latent diffusion models (Rombach et al., 2022) define a forward process that gradually corrupts a latent z_0 with Gaussian noise across T timesteps (Ho et al., 2020), producing $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \eta$ where $\eta \sim \mathcal{N}(0, I)$. A denoising network $\hat{\epsilon}_\theta(z_t, t)$ is trained to reverse this process by estimating the noise component at each noise level, which is equivalent to approximating the score function $\nabla_{z_t} \log p_t(z_t)$ of the data distribution (Song et al., 2020b). At inference, a sampler such as DDIM (Song et al., 2020a) iteratively applies the denoiser from $t{=}T$ down to $t{=}0$ to produce a clean image from pure noise. Because this network parameterizes the score function of the training distribution, any model-specific inductive bias must manifest through it, making it the right object to probe for attribution. A parallel paradigm, *flow matching* (Lipman et al., 2022; Liu et al., 2022), replaces this stochastic trajectory with a deterministic velocity field transporting noise to data via an ODE, and underlies recent large-scale image transformers (Esser et al., 2024). Despite its different forward process and training objective, it is likewise parameterized by a single network fit to the training distribution – the same structural property SGS exploits, which we show lets the probe extend beyond the denoising formulation to this distinct paradigm as well.

Fingerprinting diffusion models. Diffusion models trained on non-overlapping subsets of a dataset converge to nearly identical score functions (Kadkhodaie et al., 2023), suggesting that output-based attribution is fundamentally limited when training distributions are similar. Teng et al. (2025) propose FingerInv for model ownership verification, which identifies models via inversion trajectories of generated images. Other approaches embed fingerprints directly into models through weight modulation or decoder fine-tuning for user attribution (Tripathi et al., 2025; Fei et al., 2025), but these require modifying model parameters and training pipelines. In contrast, we show that despite apparent convergence, diffusion models exhibit distinct spectral dynamics during denoising, enabling intrinsic and non-invasive fingerprinting.

Non-invasive model attribution. A growing line of work addresses source attribution without modifying the model or its outputs. Classifier-based methods (Xu et al., 2025; Bonechi et al., 2025) train on output-space artifacts to identify source generators, achieving strong performance across diverse models but struggling to separate architecturally similar variants that produce visually similar outputs. Inversion-based detectors such as RONAN (Wang et al., 2023a) and LatentTracer (Wang et al., 2024) attribute an image to the model that reconstructs it with the smallest calibrated loss; they are alteration-free yet slow (one optimisation per candidate) and lose discriminative power when models share the same auto-encoder. In contrast, SGS probes the denoising function directly, succeeding where decoder-level and output-level signals collapse.

In summary, existing non-invasive methods exploit output-space or decoder-level signals that collapse for shared-autoencoder models, and no existing method probes the denoising function directly without costly inversion. This gap motivates SGS, introduced after establishing the empirical basis for denoiser-level probing in the next section. Additional background on watermarking and neural network fingerprinting is in Appendix B.

3 Where Does Model Identity Reside?

We identify a deliberately hard attribution setting to probe where discriminative signal resides: SD v1.4[2] vs. SD v1.5 [3], two models sharing nearly all of their weights, the same UNet-860M backbone, encoder, and KL-f8 VAE (Rombach et al., 2022). We ask: *which representation most clearly exposes model-specific signal for this near-identical pair?*

The denoiser is the right object of study. The answer lies in the denoising function itself (Teng et al., 2025). The denoiser parameterizes the score function $\nabla_z \log p(z)$, which encodes the full statistical geometry of the training distribution, including its cross-frequency covariance structure. Any model-specific inductive bias arising from training data, architecture, or fine-tuning must manifest through this function. Unlike output images, which are mediated by specific prompts and noise

[2] https://huggingface.co/CompVis/stable-diffusion-v1-4

[3] https://huggingface.co/runwayml/stable-diffusion-v1-5

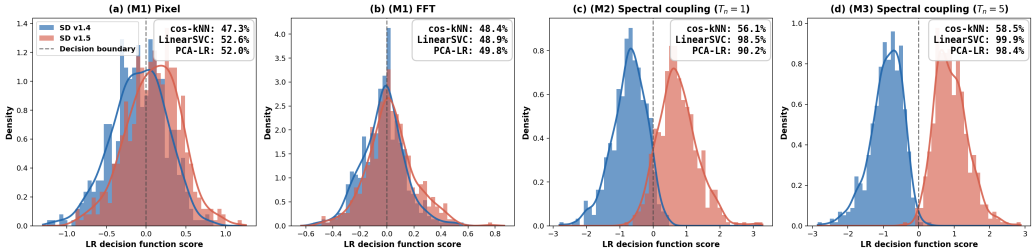

Figure 2: **Spectral coupling signatures are linearly separable; image features are not.** LR decision score distributions for SD v1.4 (blue) vs. SD v1.5 (red) after PCA-64 + L_2 normalisation. Pixel and FFT scores overlap completely, confirming the absence of any linear signal. Spectral Coupling T_n=1 and T_n=5 show clearly separated distributions, directly visualising the global mean-shift that enables $\geq$ 98% classification accuracy from a 192-dimensional signature.

seeds, the denoiser is an *intrinsic, input-independent object*: it is shared across all generations and directly reflects what the model has learned. This motivates us to probe the denoiser directly, rather than analyzing its outputs.

Setup. We generate N=1,000 images using 500 prompts per model from the Stable Diffusion Prompts dataset[4] at 512×512 resolution with 50 DDIM steps and classifier-free guidance (CFG) scale 7.5. We compare three feature families: pixel-space statistics, FFT power spectrum, and spectral coupling signatures derived from denoiser responses. For fair comparison, all features are preprocessed identically: standardization, projected to 64 principal components as a reasonable low-dimensional representation , and ℓ_2 normalization. We evaluate three classifiers: LinearSVC on raw features, PCA-LR on compressed features, and cosine-kNN (k=5). Results in Figure 2 shows Logistic-regression decision score distributions for all four feature types.

P_1: Output features carry no signal. We evaluate pixel-space features (786,000 dimensions) and FFT power spectrum features (262,000 dimensions) as standard output-based representations. Despite their dimensional advantage, all three classifiers achieve near-chance performance. The failure is structural: SD v1.4 and SD v1.5 share the same VAE, so their output distributions are statistically indistinguishable at the level of spatial and spectral statistics that these features capture.

Takeaway 1. For models sharing the same autoencoder, output-space features provide little discriminative signal despite their high dimensionality. This difficulty is not specific to linear probes: an end-to-end ResNet-50 trained on generated images shows similar difficulty for closely related model variants (Appendix L). .

P_2: Model identity emerges in denoiser responses. We probe the denoiser's response to controlled frequency perturbations. For each spatial frequency band k, we inject band-limited noise at a single probe timestep (T_n=1, τ=25, where T_n denotes the number of probe timesteps, distinct from the total diffusion steps T=50), chosen to sit in the mid-noise regime where the denoiser Jacobian $J_\theta = \partial\hat{\epsilon}_\theta/\partial z_t$ (formalized in Appendix C) captures both global and local structure, and measure how the injected energy is redistributed across output bands. This yields a spectral coupling matrix $C \in \mathbb{R}^{K\times K}$, where K is the number of frequency rings (a design parameter, set to K=8 here) and $C[k, j]$ quantifies how energy injected in ring k is routed to ring j by the denoiser. We probe at S=3 perturbation amplitudes $\{1.0, 1.25, 1.5\}$, yielding a 192-dimensional signature ($S{\cdot}K^2 = 3{\cdot}64 = 192$). Detailed method is provided in §4.2.

LinearSVC reaches **98.5**% accuracy and PCA-LR reaches 90.2%, confirming that the discriminative signal is preserved under strong compression. In contrast, cosine-kNN performs substantially worse (56.1%), indicating that separation does not arise from tight local clustering. Instead, the signal manifests as a *global mean shift*: class centroids are well separated, while individual samples overlap locally. As a result, linear classifiers recover the signal, whereas neighborhood-based methods fail to capture it. More broadly, these results show that the discriminative information is not in raw outputs themselves, but in how the denoiser *transforms structured perturbations*.

[4] https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts

Takeaway 2. A single-timestep spectral response linearly separates two nearly identical models from a 192-dimensional signature. The signal arises from the denoiser's Jacobian structure and is therefore intrinsic to the model, not an artifact of the probing procedure.

P_3: Identity accumulates across timesteps. We extend the probe to T_n=5 evenly spaced timesteps spanning $[0.2T, T]$, concatenating coupling matrices at each step to yield a 960-dimensional signature ($S \cdot T_n \cdot K^2 = 3 \cdot 5 \cdot 64 = 960$). LinearSVC reaches $\mathbf{99.9}\%$ and PCA-LR reaches 98.4%, both improving consistently over the single-timestep result. The gain reflects genuine complementarity rather than added dimensionality.

Takeaway 3. Aggregating responses across T_n=5 timesteps improves attribution accuracy because different timesteps contribute complementary model-specific information along the denoising trajectory. This also suggests a substantially cheaper single-timestep operating point with little loss.

Together, these results show that model identity is encoded in the spectral geometry of the denoiser's Jacobian, not in generated outputs, motivating SGS: forward-pass-only spectral coupling probed across timesteps, with no inversion or model modification.

4 Spectral Geometry Signatures

We formalize the extraction of spectral coupling signatures from diffusion model denoisers into a complete attribution pipeline, translating the empirical observations of § 3 into a principled method.

4.1 Problem Formulation

Task. We consider *closed-set model attribution*: given a generated image x_0 and a finite registry of candidate models $\mathcal{M} = \{M_1, \ldots, M_n\}$, identify which model in $\mathcal{M}$ produced x_0. . The registry may include models differing in training data, architecture, or training procedure, including closely related fine-tuned variants.

Verifier access. SGS assumes white-box access to candidate models: the verifier queries the denoiser forward pass $\hat{\epsilon}_\theta(z_t, t, \varnothing)$ with arbitrary inputs, where $\varnothing$ denotes the empty text embedding used for unconditional denoising, requiring only forward passes with no gradients, optimization, or knowledge of generation parameters.

4.2 Our Approach

Preliminaries. Before probing, we standardize conditions across all candidate models to ensure signatures reflect genuine model-specific behavior rather than artifacts of mismatched latent spaces or sampling parameters. Each input image x_0 is encoded to a latent $z_0 = \mathcal{E}(x_0)$ using a single shared VAE, regardless of which candidate model is being probed, yielding consistent latent dimensions across all models. A shared scheduler is used for all models, ensuring timestep indices refer to the same noise levels. The denoiser is called with an empty text embedding $\varnothing$ throughout, removing prompt-dependent signal.

1) Partition. We partition the $H \times W$ frequency plane of the latent into K concentric radial rings. The normalised radial coordinate of frequency (u, v) is $\rho(u, v) = \frac{\sqrt{u^2+v^2}}{\sqrt{H^2+W^2}}$; for $k \in \{0, \ldots, K-1\}$, ring k contains all frequencies whose normalised radius ρ falls in $\left[\frac{k}{K}, \frac{k+1}{K}\right)$, giving binary mask $\Omega_k \in \{0, 1\}^{H \times W}$. Ring 0 covers the DC component and lowest spatial frequencies (global structure); ring $K-1$ covers the highest frequencies (fine texture).

2) Probe. As established in §3, the spectral coupling matrix $C[k, j]$ is a frequency-domain projection of the denoiser Jacobian $\partial\hat{\varepsilon}_\theta/\partial z_t$, directly measuring how the score function's local geometry routes energy across frequency bands. We now formalize this measurement. A formal derivation connecting $C[k, j]$ to the denoiser Jacobian is provided in Appendix C.

Band-limited probe noise. Given i.i.d. Gaussian noise $\eta \sim \mathcal{N}(0, I)$, we confine its energy to ring k by masking in the Fourier domain and renormalising to match the original noise energy:

$$\eta^{(k)} = \mathcal{F}^{-1}\big(\mathcal{F}(\eta) \odot \Omega_k\big) \cdot \frac{\sigma(\eta)}{\sigma(\mathcal{F}^{-1}(\mathcal{F}(\eta) \odot \Omega_k)) + \delta} \tag{1}$$

The renormalisation ($\delta = 10^{-8}$) ensures all rings are probed at comparable signal levels. A scaled version with amplitude s is then $\eta^{(k,s)} = s \cdot \eta^{(k)}$.

Denoiser query. We construct the noisy latent and query the denoiser:

$$z_\tau = \sqrt{\bar{\alpha}_\tau}\, z_0 + \sqrt{1-\bar{\alpha}_\tau}\, \eta^{(k,s)}, \tag{2}$$

$$\hat{\eta} = \hat{\epsilon}_\theta(z_\tau,\, \tau,\, \varnothing) \tag{3}$$

The residual $\mathbf{r} = \hat{\eta} - \eta^{(k,s)}$ measures how the model's predicted noise deviates from the injected perturbation. If the model were a perfect noise estimator for band k, the residual would be zero; any nonzero residual reveals which frequencies the denoiser maps energy into or out of.

Coupling matrix entry. For a fixed latent z_0, the denoiser response to a band-limited perturbation is stochastic; we therefore average over R independently drawn noise samples to obtain a stable estimate of ring-k to ring-j energy routing:

$$C_{s,\tau}[k,j] = \frac{1}{R}\sum_{i=1}^{R}\left\langle |\mathcal{F}(\mathbf{r}_i^{(k,s,\tau)})|^2,\, \Omega_j \right\rangle \tag{4}$$

Row k of $C_{s,\tau}$ encodes the denoiser's output energy distribution when ring k is perturbed, capturing which output frequency bands this model routes that energy to.

Row normalisation. Each row is ℓ_1-normalised so that it sums to one:

$$C_{s,\tau}[k,\cdot] \leftarrow \frac{C_{s,\tau}[k,\cdot]}{\|C_{s,\tau}[k,\cdot]\|_1 + \delta} \tag{5}$$

This converts absolute energies into a distribution over output bands, removing content-dependent scale variation so that the coupling pattern reflects the model's learned behavior rather than image brightness.

3) Aggregate. We collect coupling measurements across multiple timesteps and perturbation amplitudes to build a complete model signature that captures spectral routing behavior along the full denoising trajectory.

Timestep selection. We sample T_n evenly spaced probe timesteps τ in $[0.2T, T]$, spanning overall denoising behavior as motivated by Takeaway 3 in §3. Note that τ indexes the probe timesteps and is distinct from the general diffusion timestep $t \in \{0, \ldots, T\}$.

Multi-scale probing. We evaluate the denoiser's response across S perturbation amplitudes, chosen to span a range of noise magnitudes while remaining within the model's operational noise regime. Aggregating across scales captures how spectral coupling behaves under different perturbation strengths, providing complementary discriminative signal with minimal computational overhead.

Signature tensor and final vector. The full measurement is collected into:

$$\mathbf{A} \in \mathbb{R}^{S\times T_n \times K \times K}, \qquad \mathbf{A}[s,\, \tau,\, k,\, j] = C_{s,\tau}[k,j] \tag{6}$$

The SGS signature is the flattened tensor:

$$\Phi(M,\, z_0) = \mathrm{vec}(\mathbf{A}) \;\in\; \mathbb{R}^{S\cdot T_n \cdot K^2} \tag{7}$$

Each dimension encodes one specific query: *at probe timestep τ and amplitude s, if energy is injected in ring k, what fraction of the denoiser residual appears in ring j?*

4) Identify. Given a query image x with $z_0 = \mathcal{E}(x)$, we probe z_0 through *every* candidate model's own denoiser in the registry, producing one signature per candidate: $\{\Phi(M_i, z_0)\}_{i=1}^n$, each computed using that candidate's own weights. The source model is identified by selecting the candidate whose probe response is most self-consistent with its own stored prototype.

Prototype construction. Given N reference images $\{x_0^{(i)}\}$ generated by model M, the model prototype is the mean signature:

$$\Psi(M) = \frac{1}{N}\sum_{i=1}^{N}\Phi\Big(M,\, \mathcal{E}(x_0^{(i)})\Big) \tag{8}$$

Averaging suppresses content-dependent noise, recovering the model's intrinsic centroid in signature space. The row normalisation in Eq. 5 removes content-scale variation by design, which is why a small N suffices.

Zero-training identification (argmin). We assign the query to the candidate whose own probe response is most self-consistent with its own stored prototype:

$$\hat{M} = \underset{M_i\in\mathcal{M}}{\arg\min}\; d_{\cos}(\Phi(M_i,\, \mathcal{E}(x)),\; \Psi(M_i))\,, \tag{9}$$

where $d_{\cos}(u,v) = 1 - \frac{u^\top v}{\|u\|\|v\|}$. Each $\Phi(M_i, \mathcal{E}(x))$ is computed using M_i's own denoiser weights; the source model is unknown at query time and is never assumed.

Classifier-based identification. When labelled signatures are available, a LinearSVC trained on raw signatures substantially improves over argmin, particularly when model weight overlap is high. For classifier-based identification, we apply the trained multi-class LinearSVC to each $\Phi(M_i, \mathcal{E}(x))$ in turn and select the candidate whose signature receives the highest self-class decision-function margin. Both variants therefore probe all n candidate denoisers at query time, since the source model is by definition unknown. A linear classifier suffices because SGS signatures exhibit the global mean-shift structure observed in §3; The full algorithm is in Appendix A.

5 Experimental Setup

Our primary closed-set evaluation uses a core registry of eight diffusion models spanning differences in training data, architecture, and training procedure. We additionally expand the registry to fifteen models with seven fine-tuned variants to test robustness to increased same-family competition. Finally, we extend the evaluation to four flow-matching models (Table 1) to test whether the spectral coupling signal generalizes beyond standard diffusion denoisers.

Model	Backbone	Variation	Registry
SD v1.4, v1.5, v2.1	UNet-860M	Training data	Core (8)
Dreamshaper-8, Realistic Vision v5	UNet-860M	Training data	Core (8)
SDXL	UNet-2.6B	Architecture	Core (8)
PixArt-α	DiT + T5-XXL	Architecture	Core (8)
SDXL-Turbo	UNet-2.6B	Training procedure	Core (8)
RV v6, epiCRealism, AbsoluteReality v1.8.1	UNet-860M	Fine-tune depth	Extended (15)
Dreamshaper-7, Deliberate v2	UNet-860M	Fine-tune depth	Extended (15)
RealVisXL V4.0, Dreamshaper-XL v2 Turbo	UNet-2.6B	Fine-tune depth	Extended (15)
InstaFlow, AuraFlow	Flow-matching	Training objective	FM extension (19)
CogView4-6B, SANA-1600M	Flow-matching	Training objective	FM extension (19)

Table 1: **Model registry.** Together these models form a challenging attribution setting that spans the full range of model-level variation present in the open-weight diffusion ecosystem. Repository URLs are provided in Appendix M.

Data. For each model we generate signatures from N=50 reference images and evaluate on N_{test}=500 held-out images from the SD Prompts dataset[5], using disjoint prompt sets. Sensitivity to the number of reference images used to construct each prototype is studied in Appendix J. We additionally test cross-domain generalization on MS-COCO prompts (Lin et al., 2014) and cross-semantic attribution on DiffusionDB prompts (Wang et al., 2023b).

Metrics. We report Top-1 attribution accuracy for all classifiers across all experiments.

SGS implementation. We use the SD v1.5 KL-f8 VAE as the shared encoder (512×512 input, 64×64 latents). The probe scheduler is DDIM with T=50 steps. We use K=8 frequency rings, S=3 amplitude scales $\{1.0, 1.25, 1.5\}$, R=10 noise repeats, and T_n=5 evenly spaced probe timesteps τ in $[0.2T, T]$, giving signature dimension $S{\cdot}T_n{\cdot}K^2$=960. . LinearSVC is trained on raw signatures using 5-fold cross-validation. The choice of K and S is studied in Appendix E.1; R=10 and T_n=5 were selected based on convergence of the coupling estimate and accuracy saturation respectively. Replacing DDIM with DPM-Solver++ yields nearly identical headline accuracy, indicating that SGS is robust to scheduler choice (Appendix E.3). A native-VAE control testing whether the shared probing encoder drives separability is reported in Appendix F. For completeness, we include a preliminary open-set rejection analysis in Appendix K.

Baselines. We compare SGS against two non-invasive attribution methods that, like SGS, attribute generated images to their source model without modifying model parameters. LatentTracer (Wang et al., 2024) attributes images by inverting the VAE decoder and measuring reconstruction fidelity under each candidate model. RONAN (Wang et al., 2023a) attributes images via gradient-based input inversion with random initialization, reverse-engineering the latent input of each candidate model.

6 Results

Our evaluation addresses the following core questions:

1. **Attribution accuracy and registry scaling** (§6.1): Can SGS reliably attribute images in our core eight-model benchmark, and does performance persist when the registry is expanded to more models?

[5] https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts

2. **Generalization** (§6.2, §6.3): Does SGS remain informative under prompt-distribution shifts and image post-processing?
3. **Baseline comparison** (§6.4): How does SGS compare to non-invasive attribution baselines on the eight-model closed-set task?

6.1 Attribution Accuracy and Registry Scaling

Our primary evaluation uses the core eight-model registry. SDXL, PixArt-α, SDXL-Turbo, SD v2.1, Dreamshaper-8, and Realistic Vision v5 are attributed perfectly by both classifiers (Figure 3). The remaining PCA-LR errors are concentrated between SD v1.4 and SD v1.5, where approximately 1.5% residual confusion remains; LinearSVC resolves this pair and achieves near-perfect overall attribution.

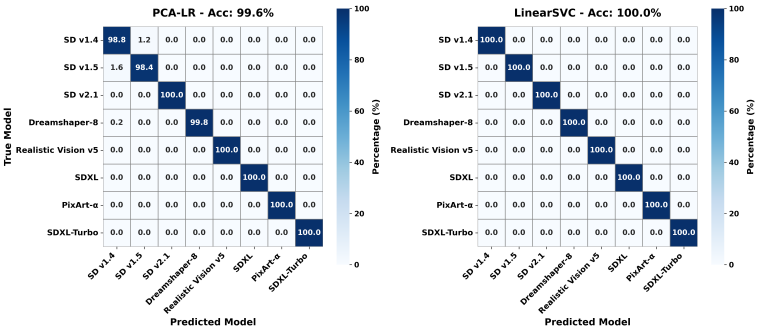

Figure 3: **SGS attribution on the core eight-model registry** (N_{test}=500 per model, T_n=5). PCA-LR and LinearSVC both separate architecturally distinct models and closely related fine-tunes; the remaining PCA-LR errors are concentrated between SD v1.4 and SD v1.5.

Registry scaling (M=8 → 15). We next add seven fine-tuned checkpoints, substantially increasing same-family competition. LinearSVC remains at $\mathbf{99.99}\%$ Top-1 accuracy on the 15-model registry, indicating that the added near-family competitors do not materially degrade attribution. Full per-model results and margin analyses are reported in Appendix G.

Beyond diffusion. We separately evaluate four flow-matching models (InstaFlow, AuraFlow, CogView4-6B, and SANA-1600M) using probes adapted to their respective forward-process conventions. All four are perfectly distinguished from one another, and LinearSVC reaches $\mathbf{99.98}\%$ in a pooled 19-model evaluation with the 15 diffusion models. This provides evidence that the spectral-coupling signal extends beyond standard diffusion denoisers to the flow-matching models evaluated here. Probe adaptations, the corresponding Jacobian analysis, and full classifier results are provided in Appendices C and H.

6.2 Cross-Domain Generalization

We test whether SGS signatures transfer across prompt distributions by building prototypes on SD-Prompts and evaluating them unchanged on MS-COCO (Table 2). The Argmin rule (Eq. 9) is essentially unaffected (+0.07 pp), while LinearSVC drops only 3.8 pp (99.98%→96.20%); other classifiers degrade more substantially. This suggests that the underlying signatures remain largely model-specific across prompt distributions, while some learned decision boundaries are partially distribution-dependent.

Classifier	SD-Prompts	MS-COCO	Drop (pp)
Argmin†	78.33%	78.40%	+0.07
LinearSVC	**99.98%**	**96.20%**	−3.80
PCA-LR	99.62%	92.50%	−7.12
Random Forest	96.23%	83.60%	−12.63
PCA-SVM (RBF)	91.83%	80.30%	−11.53
PCA-kNN (k=5)	82.37%	74.20%	−8.17

Table 2: **Top-1 accuracy across prompt distributions.** Prototypes from SD-Prompts (N=50, T_n=5, M=8, N_{test}=500), evaluated unchanged on MS-COCO. Drop: absolute Top-1 decrease. † training-free.

Cross-semantic attribution. To test the adversarial arrangement where each model's test images come from the opposite prompt domain to the registry prototypes, we evaluate SD v1.4 vs. SD v1.5 with prototypes anchored on one domain (MS-COCO or DiffusionDB) and query images from the other. Argmin degrades to 31-38% under this adversarial mismatch (vs. 53-56% in-domain), a real gap. However, LinearSVC and PCA-LR hold at 95-99% test accuracy under the same conditions. Full results are reported in Appendix I.

6.3 Robustness to Image Distortions

We evaluate SGS against six strong post-processing distortions applied to query images before signature extraction (Table 3). Prototypes are always built from clean images.

	SD-Prompts			MS-COCO		
Distortion	**argmin**	**PCA-LR**	**LinearSVC**	**argmin**	**PCA-LR**	**LinearSVC**
Rotation (75°)	77.5%	90.6%	91.9%	62.5%	90.0%	90.6%
JPEG (Q=25)	76.2%	91.2%	93.1%	55.6%	86.9%	85.0%
Crop (75%) + resize	78.1%	91.2%	90.6%	56.3%	86.9%	87.5%
Brightness ($\times 1.5$)	75.6%	88.8%	88.8%	48.8%	86.3%	85.6%
Noise (σ=0.1)	61.9%	92.5%	93.8%	58.1%	93.8%	92.5%
Blur (σ=8)	51.9%	95.6%	95.6%	21.3%	99.4%	**99.9%**

Table 3: **Robustness of SGS to strong image-level distortions** (M=8, N=50 per model, 20 test images per model, N_{test}=160 total per distortion). Each distortion is applied to the query image before signature extraction; prototypes are always built from clean images.

The six distortions split into two behaviours. Rotation, JPEG, cropping, and brightness cause no meaningful degradation ($\leq$1.3pp Argmin drop; LinearSVC $\geq$88%) because they preserve latent structure. Noise and blur hurt Argmin substantially (61.9% and 51.9%), but classifier accuracy is *higher* than baseline: LinearSVC reaches 93.8% under noise and **99.9**% under blur on MS-COCO. This is not paradoxical: strong distortions add isotropic noise that blurs individual-sample distances but leaves class centroids intact, and LinearSVC operates on centroids. SGS measures denoiser spectral routing, not pixel statistics, so pixel-space distortions cannot erase the fingerprint regardless of strength. Sensitivity to prototype set size is studied in Appendix J; probe design sensitivity in Appendix E.1.

6.4 Baseline Comparison

Model	**RONAN**	**LatentTracer**	**SGS (ours)**
SD v1.4	21.4%	22.1%	**100.0%**
SD v1.5	20.2%	19.7%	**100.0%**
SD v2.1	22.2%	26.6%	**100.0%**
Dreamshaper-8	21.2%	22.1%	**99.8%**
Realistic V. v5	20.4%	20.9%	**100.0%**
SDXL	48.4%	59.2%	**100.0%**
PixArt-α	55.4%	79.0%	**100.0%**
SDXL-Turbo	35.0%	67.0%	**100.0%**
Overall Top-1	30.5%	44.8%	**99.9%**

Table 4: **Baseline comparison on the 8-model closed-set task** (N_{test}=500 per model). Both baselines exploit decoder reconstruction fidelity and collapse for all five models sharing the SD KL-f8 VAE

Table 4 reveals a structural gap. Both baselines rely on decoder reconstruction as their discriminative signal, effective when autoencoders differ (LatentTracer: 59-79%; RONAN: 35-55% on SDXL, PixArt-α, SDXL-Turbo), but collapsing to 20-27% for all five shared VAE models, indistinguishable from the 20% random baseline. This is structural: shared decoders produce identical reconstruction losses regardless of which candidate is queried. SGS probes the denoiser instead, achieving perfect attribution on all eight models, a 55pp gain over LatentTracer and 70pp over RONAN, with the entire gap on exactly the model pairs that matter most.

We evaluate these baselines on the core eight-model registry, which already captures the shared-VAE setting central to their failure mode. Both methods also require costly per-candidate iterative optimization, making larger-registry evaluation less practical.

7 Conclusion and Discussion

We introduced SGS, a non-invasive attribution framework that identifies diffusion models through their denoiser responses to controlled spectral perturbations. SGS achieves strong closed-set attribution, including for closely related checkpoints that are difficult to separate from generated outputs alone. The signal remains stable under prompt shift, image post-processing, and native-VAE substitution. It also remains highly discriminative as the registry grows and extends to the flow-matching models evaluated in this work. These results suggest that model identity is reflected in the internal response geometry of the generative model, not only in its outputs.

SGS is designed for settings where candidate-model denoisers are accessible, making it well suited to open-weight model ecosystems. It focuses on source attribution and can naturally complement post-generation tamper detection. Future work can extend SGS to broader attribution settings, improve robustness across deployment conditions, and explore how response-based signatures can support other forms of model provenance and analysis. More broadly, non-invasive attribution can support provenance and ownership verification without requiring model modification or prior enrollment.

AI Use Disclosure

Generative AI tools were used to assist with manuscript editing, organization, literature summarization, and feedback on experimental design and interpretation. AI-assisted suggestions were reviewed

and verified by the authors, who take full responsibility for the final content, claims, analyses, and presentation of this work.

REFERENCES

Simone Bonechi, Paolo Andreini, and Barbara Toniella Corradini. Who made this? fake detection and source attribution with diffusion features. *arXiv preprint arXiv:2510.27602*, 2025.

Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Ipguard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary. In *Proceedings of the 2021 ACM asia conference on computer and communications security*, pp. 14–25, 2021.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 973–982, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Jianwei Fei, Yunshu Dai, Zhihua Xia, Fangjun Huang, and Jiantao Zhou. Omnimark: Efficient and scalable latent diffusion model fingerprinting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16550–16558, 2025.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22466–22477, 2023.

Andreas Floros, Seyed-Mohsen Moosavi-Dezfooli, and Pier Luigi Dragotti. Tracing the roots: Leveraging temporal dynamics in diffusion trajectories for origin attribution. *Advances in Neural Information Processing Systems*, 38:99888–99907, 2026.

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

Yuepeng Hu, Zhengyuan Jiang, Moyang Guo, and Neil Gong. Stable signature is unstable: Removing image watermark from diffusion models. *arXiv preprint arXiv:2405.07145*, 2024.

Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. *arXiv preprint arXiv:2310.02557*, 2023.

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.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Nils Lukas, Yuxuan Zhang, and Florian Kerschbaum. Deep neural network fingerprinting by conferrable adversarial examples. *arXiv preprint arXiv:1912.00888*, 2019.

Sen Peng, Yufei Chen, Cong Wang, and Xiaohua Jia. Intellectual property protection of diffusion models via the watermark diffusion process. In *International Conference on Web Information Systems Engineering*, pp. 290–305. Springer, 2025.

Zirui Peng, Shaofeng Li, Guoxing Chen, Cheng Zhang, Haojin Zhu, and Minhui Xue. Fingerprinting deep neural networks globally via universal adversarial perturbations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13430–13439, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Yuhui Quan, Huan Teng, Ruotao Xu, Jun Huang, and Hui Ji. Fingerprinting deep image restoration models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13285–13295, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Huan Teng, Yuhui Quan, Chengyu Wang, Jun Huang, and Hui Ji. Fingerprinting denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28811–28820, 2025.

Subarna Tripathi et al. Paladin: Robust neural fingerprinting for text-to-image diffusion models. *arXiv preprint arXiv:2506.03170*, 2025.

Zhenting Wang, Chen Chen, Yi Zeng, Lingjuan Lyu, and Shiqing Ma. Where did i come from? origin attribution of ai-generated images. *Advances in neural information processing systems*, 36: 74478–74500, 2023a.

Zhenting Wang, Vikash Sehwag, Chen Chen, Lingjuan Lyu, Dimitris N Metaxas, and Shiqing Ma. How to trace latent generative model generated images without artificial watermark? *arXiv preprint arXiv:2405.13360*, 2024.

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 893–911, 2023b.

Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Detecting origin attribution for text-to-image diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8775–8785. IEEE, 2025.

Zihan Yuan, Li Li, Zichi Wang, and Xinpeng Zhang. Watermarking for stable diffusion models. *IEEE Internet of Things Journal*, 11(21):35238–35249, 2024.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023. NEW

APPENDIX

A ALGORITHM

Algorithm 1 summarizes the complete SGS extraction procedure described in §4.2, including spectral perturbation, denoiser evaluation, cross-band energy aggregation, and normalization.

Algorithm 1: Spectral Geometry Signature (SGS) extraction for diffusion models. The procedure injects band-limited perturbations at selected noise levels and measures cross-frequency energy redistribution in the denoiser response. Flow-matching probe adaptations are described in Appendix H.

Input: Denoiser $\hat{\epsilon}_\theta$, latent z_0, rings K, probe timesteps $\mathcal{T}$, scales $\mathcal{S}$, repeats R
Output: Signature $\Phi \in \mathbb{R}^{S \cdot T_n \cdot K^2}$

Compute ring masks $\{\Omega_k\}_{k=0}^{K-1}$;
$\mathbf{A} \leftarrow \mathbf{0}^{S \times T_n \times K \times K}$;
for *each repeat* $r = 1, \ldots, R$ **do**
  for *each* $(s,\ \tau) \in \mathcal{S} \times \mathcal{T}$ **do**
    for *each ring* $k = 0, \ldots, K-1$ **do**
      $\eta^{(k,s)} \leftarrow s \cdot \mathcal{F}^{-1}(\mathcal{F}(\eta) \odot \Omega_k)$; // band-limited perturbation
      $z_\tau \leftarrow \sqrt{\bar{\alpha}_\tau}\, z_0 + \sqrt{1-\bar{\alpha}_\tau}\, \eta^{(k,s)}$; // inject into diffusion
      $\hat{\eta} \leftarrow \hat{\epsilon}_\theta(z_\tau,\ \tau,\ \varnothing)$; // denoiser forward pass
      $\mathbf{r} \leftarrow \hat{\eta} - \eta^{(k,s)}$; // residual = model response
      $\mathbf{A}[s, \tau, k, \cdot]\ +\!= \left\langle |\mathcal{F}(\mathbf{r})|^2,\ \Omega_j \right\rangle_{j=0}^{K-1}$; // energy redistribution
$\mathbf{A} \leftarrow \mathbf{A}/R$;
Row-normalise: $\mathbf{A}[s, \tau, k, \cdot] \leftarrow \mathbf{A}[s, \tau, k, \cdot] \,/\, \|\mathbf{A}[s, \tau, k, \cdot]\|_1$;
return $\mathrm{vec}(\mathbf{A})$;

B Extended Related Work

Watermarking for diffusion models. A dominant line of work on protecting diffusion model intellectual property is *watermarking*, which embeds identifiable signals into model parameters or generated outputs. Existing approaches introduce watermarks through training-time modifications, such as embedding triggers in the data (Zhao et al., 2023; Peng et al., 2025), fine-tuning model components (Yuan et al., 2024), or altering the decoder to encode persistent signatures in generated images (Fernandez et al., 2023). While effective, these methods are inherently *invasive*, requiring retraining or fine-tuning and potentially affecting generation quality or model usability. Moreover, watermarks can be removed or degraded through post-processing or model modifications (Hu et al., 2024), motivating the need for non-invasive alternatives.

Model fingerprinting in neural networks. Non-invasive fingerprinting aims to identify intrinsic characteristics of a model without modifying its parameters. In classification models, prior work exploits decision boundary geometry or adversarial examples to construct model-specific signatures (Cao et al., 2021; Lukas et al., 2019; Peng et al., 2022). In image restoration, fingerprinting methods identify *critical points* near the performance boundary that produce distinctive responses across models (Quan et al., 2023). These approaches rely on the observation that models differ in how they behave on carefully chosen inputs. However, extending these ideas to diffusion models is non-trivial, as diffusion models are probabilistic generative systems that map noise distributions to images rather than deterministic input-output mappings.

C Theoretical Grounding

Score function and spectral geometry. The denoising network $\hat{\epsilon}_\theta(z_t, t)$ is related to the score function via $\nabla_{z_t} \log p_t(z_t) \propto -\hat{\epsilon}_\theta(z_t, t)$ (Song et al., 2020b). The score field $\nabla_{z_t} \log p_t(z_t)$ encodes the statistical geometry of the (noisy) data distribution at noise level t, including how energy is correlated across spatial frequency bands. Concretely, the Hessian $\nabla^2_{z_t} \log p_t(z_t)$ characterizes the local covariance structure: its off-diagonal blocks in the frequency domain reflect how variations in one frequency band couple to variations in another. Different training datasets imprint different covariance structures, and different architectures impose different inductive biases on how this structure is represented (Kadkhodaie et al., 2023). Both effects manifest in the Jacobian of the denoiser $J_\theta(z_t, t) = \partial \hat{\epsilon}_\theta(z_t, t)/\partial z_t$, which inherits the spectral geometry of the learned score function.

Generalization to flow-matching objectives. The derivation above requires only that the network $f_\theta(z_t, t)$ has a well-defined input-output Jacobian $J_\theta = \partial f_\theta / \partial z_t$ and that we can inject a band-limited perturbation and observe the residual response. Flow-matching models learn a velocity field $v_\theta(z_t, t)$ that plays an analogous role to the score estimate, and the first-order expansion in Eq. 10 applies to any such vector-field parameterization, not only $\hat{\epsilon}_\theta$. Appendix H confirms this empirically across four distinct flow-matching architectures.

Spectral coupling as a Jacobian projection. The coupling matrix $C[k, j]$ is a Monte Carlo estimate of this Jacobian structure in the frequency domain. To see this, consider a first-order expansion of the denoiser response to a band-limited perturbation $\eta^{(k,s)}$ confined to ring k:

$$\hat{\epsilon}_\theta(z_t, t, \varnothing) \approx \hat{\epsilon}_\theta(z_0, t, \varnothing) + J_\theta(z_0, t)\sqrt{1-\bar{\alpha}_t}\,\eta^{(k,s)}, \tag{10}$$

where $z_t = z_0 + \sqrt{1-\bar{\alpha}_t}\,\eta^{(k,s)}$. The output residual $\delta\hat{\eta} = \hat{\epsilon}_\theta(z_t, t, \varnothing) - \hat{\epsilon}_\theta(z_0, t, \varnothing)$ therefore reflects how J_θ maps ring-k inputs to outputs across all rings. The coupling entry $C[k, j]$ measures the energy of this residual in ring j, averaged over noise samples. If $\eta^{(k,s)}$ is isotropic within ring k, then

$$C[k, j] \;\propto\; \left\| P_j\, J_\theta(z_0, t)\, P_k \right\|_F^2, \tag{11}$$

where P_k denotes projection onto ring k. Since J_θ is determined by the learned score function, the expected couplings $C[k, j]$ are intrinsic to the model and, after averaging over images, largely independent of image content.

For a flow-matching model, the generative dynamics are defined by

$$\frac{dz_t}{dt} = v_\theta(z_t, t),$$

where v_θ is the learned time-conditioned velocity field. Under a small band-limited perturbation δz,

$$v_\theta(z_t + \delta z, t) \approx v_\theta(z_t, t) + J_\theta(z_t, t)\delta z,$$

so the same cross-frequency coupling statistic $\|P_j J_\theta P_k\|_F^2$ applies directly.

D Signature Separability Across Models

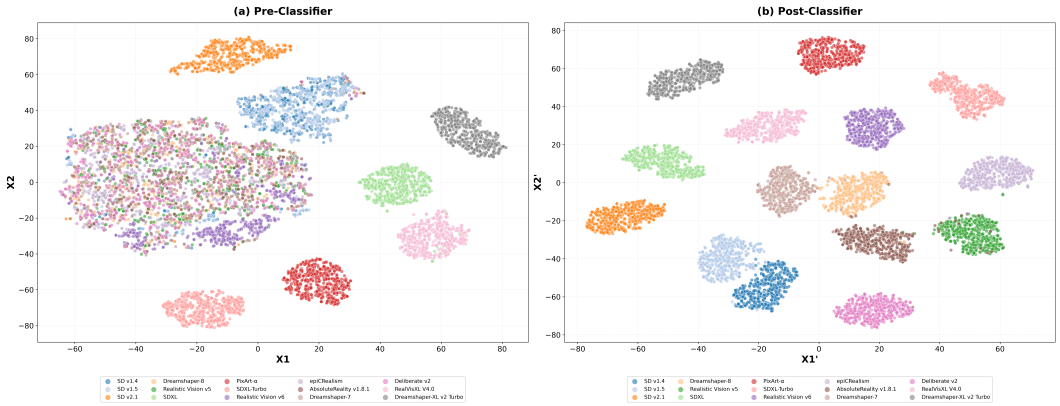

Figure 4: **t-SNE visualization of SGS signatures before and after classifier projection.** (a) *Pre-classifier*: t-SNE of dimensionally reduced SGS signatures, with no classifier applied. Seven of the fifteen models form visually distinct clusters, confirming that the signature space carries strong intrinsic structure without any learned transformation. (b) *Post-classifier*: t-SNE of signatures after LR projection. All fifteen clusters are cleanly separated, consistent with the high attribution accuracy reported in confusion matrix 3.

The t-SNE visualization in Figure 4 provides direct geometric evidence for the quality of SGS signatures. Even before any classifier is applied, the raw signature space already separates most of the models into distinct, compact clusters. This intrinsic separability confirms that the spectral coupling representation captures genuine model-specific structure rather than fitting to classifier artifacts. A linear classifier resolves this residual ambiguity entirely, as shown in panel (b), consistent with the global mean-shift structure established in §3.

E Probe Design and Sensitivity

E.1 Frequency Rings and Perturbation Amplitudes

Table 5 reports attribution accuracy across all combinations of K and S, evaluated at $T_n{=}1$ to isolate the effect of spatial resolution and multi-scale probing from temporal aggregation. Both

K	S	Sig. dim	Argmin	LinearSVC
2	1	4	52.3%	52.4%
	2	8	53.5%	65.3%
	3	12	60.9%	77.4%
4	1	16	65.1%	90.2%
	2	32	66.8%	94.6%
	3	48	70.3%	97.2%
6	1	36	69.8%	94.3%
	2	72	72.4%	97.0%
	3	108	75.0%	99.6%
8	1	64	71.9%	94.2%
	2	128	74.2%	97.5%
	3*	192	76.4%	**99.9%**

Table 5: **Ablation over frequency rings K and perturbation amplitudes S** at a single probe timestep (T_n=1, τ=25), evaluated on all 8 models (N_{test}=500 per model). S=1: amplitude $\{1.0\}$; S=2: $\{1.0, 1.25\}$; S=3: $\{1.0, 1.25, 1.5\}$ (full baseline*). Sig. dim $= S \times K^2$. * baseline used in main experiments.

parameters independently and substantially affect performance. At K=2, even the full S=3 configuration reaches only 77.4% with LinearSVC, substantially below the 99.9% achieved with K=8. Increasing K from 2 to 4 with S=3 jumps performance to 97.2%, confirming that spectral resolution is critical: coarser frequency partitions cannot distinguish the subtle cross-band routing differences that encode model identity. Performance saturates beyond K=6, with K=6, S=3 already achieving 99.6%. The baseline configuration (K=8, S=3) operates in this saturated regime, offering a favorable trade-off between performance and signature dimensionality. The effect of S is consistent across all values of K: adding perturbation amplitudes improves performance by 10–25 pp at fixed K, with diminishing returns from S=2 to S=3, suggesting three amplitudes is a practical optimum.

E.2 Per-Timestep Informativeness

Probe step	DDIM idx / 49	Noise remaining	Slice dim	Argmin	LinearSVC
step 1/5	10	79.6%	192	62.23%	$99.97\% \pm 0.03\%$
step 2/5	19	61.2%	192	64.19%	$99.92\% \pm 0.10\%$
step 3/5	29	40.8%	192	63.59%	$98.79\% \pm 0.41\%$
step 4/5	39	20.4%	192	53.21%	$92.52\% \pm 0.13\%$
step 5/5	49	0.0%	192	35.33%	$64.43\% \pm 1.19\%$
Full 5-step	all 5	—	960	65.56%	$\mathbf{99.99}\% \pm 0.03\%$

Table 6: **Per-timestep informativeness** (M=15, N_{test}=500 per model, 5-fold CV). Each row uses only the named timestep's 192-dimensional sub-block of the full 960-dim signature, sliced post hoc with no new denoiser passes. Noise remaining is the DDIM schedule's own noise fraction (1.0 = pure noise, 0.0 = fully denoised). The final, nearly-clean step alone loses 35.6pp of LinearSVC accuracy relative to the earliest step.

The 960-dimensional signature already contains one 192-dimensional sub-block per probe timestep ($S{\cdot}K^2$ each). Table 6 reports LinearSVC and Argmin accuracy using only each timestep's sub-block in isolation, sliced post hoc with no new denoiser passes. Signal concentrates strongly at the high-noise end: step 1/5 (DDIM step 10/49, 79.6% noise remaining) alone achieves 99.97% LinearSVC accuracy. Accuracy degrades monotonically as noise decreases, reaching 64.43% at step 5/5 (0.0% noise remaining), a 35.6pp drop. The full 5-step aggregate (99.99%) essentially matches the best single timestep, confirming that later timesteps add robustness and redundancy rather than new discriminative content. This quantifies what Takeaway 3 in §3 previously argued only qualitatively: different timesteps probe orthogonal aspects of the denoiser's spectral geometry, with the high-noise regime contributing the dominant signal.

K	S	Sig. dim	Argmin		LinearSVC	
			DDIM	DPM++	DDIM	DPM++
2	1	4	52.3%	54.6%	52.4%	54.7%
	2	8	53.5%	55.2%	65.3%	67.4%
	3	12	60.9%	62.0%	77.4%	80.1%
4	1	16	65.1%	66.8%	90.2%	91.4%
	2	32	66.8%	70.1%	94.6%	95.2%
	3	48	70.3%	73.0%	97.2%	97.7%
6	1	36	69.8%	71.4%	94.3%	95.0%
	2	72	72.4%	74.1%	97.0%	97.1%
	3	108	75.0%	75.6%	99.6%	99.8%
8	1	64	71.9%	72.5%	94.2%	94.9%
	2	128	74.2%	74.9%	97.5%	97.9%
	3*	192	76.4%	76.8%	**99.9%**	**100.0%**

Table 7: **Scheduler ablation: DDIM vs. DPM-Solver++ across the full $K\times S$ grid** (M=8, T_n=1, N_{test}=500 per model). DPM-Solver++ is never worse than DDIM by more than 0.2pp on any classifier at the headline config, and is consistently 0.4–2.7pp better on Argmin across the full grid. * headline config.

Model	Shared VAE	Native VAE	Δ	VAE changed?
SD v1.4	53.8%	53.8%	0.0pp	no
SD v1.5	39.6%	39.6%	0.0pp	no
SD v2.1	79.4%	79.2%	−0.2pp	no
Dreamshaper-8	30.8%	30.8%	0.0pp	no
Realistic Vision v5	41.4%	41.4%	0.0pp	no
SDXL	97.2%	97.4%	+0.2pp	**yes**
PixArt-α	100.0%	100.0%	0.0pp	**yes**
SDXL-Turbo	100.0%	100.0%	0.0pp	**yes**
RealVisXL V4.0	99.2%	97.2%	−2.0pp	**yes**
Dreamshaper-XL v2	100.0%	100.0%	0.0pp	**yes**
Aggregate classifiers (M=15, all classifiers $<$ 1pp shift)				
LinearSVC	99.99%	99.97%	−0.02pp	
PCA-LR	99.13%	99.12%	−0.01pp	

Table 8: **Native-VAE ablation** Per-model rows report Argmin attribution accuracy when the five models with distinct VAEs are probed using either the shared SD v1.5 VAE or their native VAE. Aggregate LinearSVC and PCA-LR performance is shown over the 15-model registry

E.3 Scheduler Robustness

Table 7 repeats the $K\times S$ probe-design sweep from Appendix E.1 with the probe scheduler swapped from DDIM to DPM-Solver++, holding all other hyperparameters fixed. DPM-Solver++ tracks DDIM within 1–3pp at every configuration on Argmin, and within 0.2pp at the headline config (K=8, S=3) on all trained classifiers. LinearSVC achieves 100.0% under DPM-Solver++ at the headline config vs. 99.9% under DDIM. The coupling signature is therefore not an artifact of DDIM's particular stepping math; the signal is robust to the choice of probe scheduler.

F Native-VAE Ablation

A natural concern is whether SGS separability is driven by the shared SD v1.5 encoder rather than the candidate denoiser. We therefore recompute signatures for the five models with distinct native VAEs using their own autoencoders.

Table 8 reports two views of the effect. Argmin directly measures whether changing the encoder alters each model's prototype compatibility, while LinearSVC and PCA-LR measure the resulting change in overall attribution. Four of the five models shift by at most 0.2pp under Argmin; only RealVisXL V4.0 changes by 2.0pp. Aggregate classifier accuracy changes by less than 0.02pp. Thus, the strong attribution performance is not primarily explained by the choice of probing autoencoder.

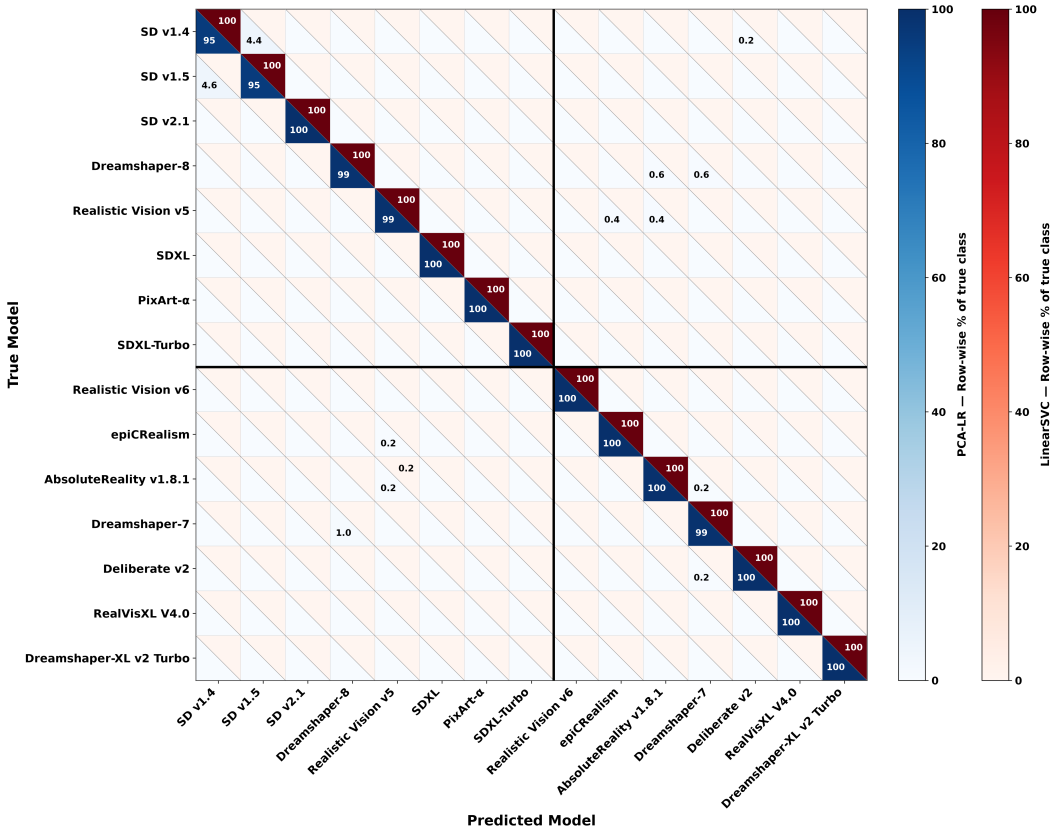

Figure 5: **Registry scaling from 8 to 15 models.** Each cell reports PCA-LR (blue, lower triangle) and LinearSVC (red, upper triangle) row-wise accuracy (N_{test}=500 per model, T_n=5). Black lines separate the original eight models from seven additional fine-tuned checkpoints. PCA-LR achieves 99.1% Top-1 accuracy, while LinearSVC reaches 99.99%.

Metric	Macro average	Worst class
AUROC	1.0000	0.9999
TPR@1% FPR	0.9999	0.9980
TPR@5% FPR	1.0000	1.0000

Table 9: **Closed-set decision-margin separation on the 15-model registry.** Statistics are computed from out-of-fold one-vs-rest LinearSVC decision margins.

G Registry Scaling and Closed-Set Margin Analysis

We evaluate whether the attribution performance observed on the core eight-model benchmark persists as the candidate registry becomes more densely populated with related models. We therefore expand the registry with seven additional fine-tuned checkpoints, increasing the number of candidates from M=8 to M=15. This extension introduces several same-family competitors rather than only unrelated architectures, providing a targeted stress test of registry density.

Figure 5 shows that increasing the registry from eight to fifteen models introduces little degradation for the trained classifiers. PCA-LR remains highly accurate, with residual errors concentrated among closely related variants, while LinearSVC makes only one error across all 7,500 test predictions. In contrast, the training-free Argmin rule drops from 78.3% to 65.5%, indicating that direct prototype matching is more sensitive to increased same-family competition.

Decision-margin separation. We additionally examine whether the near-perfect Top-1 accuracy masks weak class separation by computing one-vs-rest ROC statistics from out-of-fold LinearSVC decision margins (Table 9). Fourteen of the fifteen models achieve empirical AUROC = 1.0; the

remaining model, AbsoluteReality v1.8.1, reaches 0.9999. At a fixed 1% false-positive rate, the minimum class-wise TPR is 0.998. Thus, the strong Top-1 result is accompanied by consistently large closed-set separation across the expanded registry.

H Flow-Matching Generalization

We evaluate four flow-matching models: InstaFlow, AuraFlow, CogView4-6B, and SANA-1600M, using probes adapted to each model's forward-process convention (Table 10). We consider two settings: (i) attribution among the four flow-matching models alone, and (ii) a pooled 19-model task combining the four flow-matching models with the 15-model diffusion registry.

Probe adaptation. Following the formulation in Appendix C, we probe each flow-matching model through its learned velocity field $v_\theta(z_t, t)$ using the model's native forward-process interpolation. We use the same spectral partition, perturbation amplitudes, noise averaging, and normalized probe times as in the diffusion setting. The resulting responses are aggregated into the same SGS representation used for the pooled attribution experiment. All flow-matching experiments use the same $K{=}8$, $S{=}3$, $R{=}10$, and $T_n{=}5$ configuration.

Classifier	**Pooled** ($M{=}19$)	**Flow-only** ($M{=}4$)
PCA-SVM (RBF)	89.66%	100.00%
PCA-LR	99.29%	100.00%
PCA-kNN ($k{=}5$)	76.20%	100.00%
LinearSVC (raw)	**99.98%**	**100.00%**
Random Forest	93.31%	100.00%

Table 10: **Attribution with flow-matching models.** Flow-only evaluates InstaFlow, AuraFlow, CogView4-6B, and SANA-1600M as a four-model attribution task. Pooled evaluation combines these models with the 15-model diffusion registry. All classifiers perfectly distinguish the four flow-matching models in the flow-only setting; LinearSVC achieves 99.98% Top-1 accuracy in the pooled 19-model setting.

All evaluated classifiers achieve 100% accuracy when attribution is restricted to the four flow-matching models, indicating that their SGS responses are readily distinguishable from one another. The pooled 19-model setting is more demanding because flow-matching models must also be distinguished from the diffusion registry. LinearSVC remains nearly perfect at 99.98%, while PCA-LR also retains high accuracy at 99.29%. Other decision rules degrade more substantially, with PCA-SVM, Random Forest, and PCA-kNN reaching 89.66%, 93.31%, and 76.20%, respectively.

These results provide evidence that the spectral-response structure used by SGS is not specific to conventional diffusion denoisers. Rather, it remains discriminative for the four flow-matching models evaluated here, including when they are pooled with a heterogeneous diffusion-model registry.

I Cross-Semantic Attribution: Full Results

	Scenario A (registry = MS-COCO)			**Scenario B** (registry = DiffusionDB)		
Method	**v1.4 match.**	**v1.5 mism.**	**Overall**	**v1.5 match.**	**v1.4 mism.**	**Overall**
Argmin	56.0%	38.4%	47.2%	56.0%	31.2%	43.6%
PCA-SVM (RBF)	58.2%	35.8%	47.0%	56.4%	51.8%	54.1%
PCA-LR	96.2%	95.6%	95.9%	98.6%	95.8%	97.2%
PCA-kNN	52.6%	45.2%	48.9%	39.4%	58.2%	48.8%
LinearSVC	**98.4%**	**98.0%**	**98.2%**	**99.8%**	**97.8%**	**98.8%**
Random Forest	53.4%	44.0%	48.7%	64.8%	52.2%	58.5%

Table 11: **Cross-semantic attribution: adversarial domain mismatch on the SD v1.4 vs. SD v1.5 pair.** Registry prototypes are anchored on one prompt domain; one model's test images match that domain, the other's come from the opposite domain. "Matched" columns use test images from the registry's anchor domain; "mismatched" columns are the adversarial condition. Overall pools both classes at their true 50/50 split. LinearSVC and PCA-LR hold at 95–99% even under full domain mismatch; Argmin and neighborhood methods degrade substantially.

Table 11 gives the full classifier breakdown for the cross-semantic experiment described in §6.2. The key finding is a classifier-level split: linear classifiers (LinearSVC, PCA-LR) hold at 95-99% under both adversarial scenarios, while training-free argmin and neighborhood methods (PCA-kNN, Random Forest) degrade to near-chance. This reflects the global mean-shift structure: linear decision boundaries exploit the model-centroid separation that is stable across domains, while centroid-matching and local-clustering methods are sensitive to the per-image distribution shift introduced by the adversarial domain assignment.

J Sensitivity to Reference-Set Size

Figure 6 shows how attribution accuracy scales with N, the number of reference signatures available per model for prototype construction or classifier training. At N=8 per model (64 total training signatures), LinearSVC achieves 84.0% on SD-Prompts and 80.7% on MS-COCO, already well above the 12.5% random baseline. Accuracy improves steadily with N: at N=25, LinearSVC reaches 94.8% on SD-Prompts and 92.4% on MS-COCO, and at N=50 it reaches 99.9% and 96.2% respectively. The gap between SD-Prompts and MS-COCO remains stable across all values of N (≈3–4pp), confirming that the cross-domain drop reflects a fixed distribution shift rather than amplified estimation error. PCA-LR follows the same trend, consistently 2–3pp below LinearSVC. These results confirm that N=50 prototypes per model is a practical and sufficient operating point for reliable attribution.

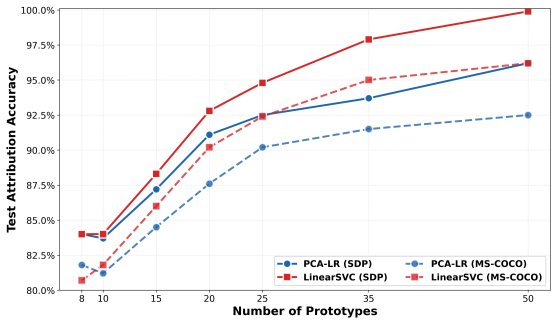

Figure 6: **Attribution accuracy vs. number of prototype images per model (N).** LinearSVC and PCA-LR evaluated on SD-Prompts (in-distribution) and MS-COCO (cross-domain).

K Open-Set Rejection: Full LOMO Results

Although our primary evaluation is closed-set, we additionally study whether SGS can reject images whose source model is absent from the enrolled registry. We use a leave-one-model-out (LOMO) protocol over the 15-model registry: for each run, one model is held out as unknown, while the remaining fourteen define the enrolled candidate set. We evaluate four rejection scores derived from the SGS representation: cosine distance, LinearSVC margin scoring (MLS), Mahalanobis distance, and k-NN distance.

Table 12 reveals a clear dependence on how closely the held-out model resembles models already present in the registry. Architecturally distinct or otherwise well-separated models, including PixArt-α, SDXL-Turbo, and the XL-family variants, are rejected reliably by several scoring rules. In contrast, closely related fine-tuned checkpoints such as SD v1.4, SD v1.5, Dreamshaper-8, and Dreamshaper-7 remain substantially more difficult to distinguish from enrolled relatives.

The choice of rejection score also matters. k-NN provides the most consistent performance, with a mean AUROC of 0.843 and no below-chance held-out cases. Mahalanobis distance performs strongly for several architecturally distinct models but degrades for some same-family variants. MLS is particularly unstable: although it achieves near-perfect separation for several held-out models, it falls below chance for five others, indicating that large closed-set classification margins do not necessarily translate into reliable open-set rejection.

Overall, these results show that open-set attribution is substantially more difficult than closed-set identification when the unknown source is closely related to models already enrolled in the registry. Distinguishing unseen same-family fine-tunes therefore remains an important open challenge.

Held-out model	Cosine	MLS	Mahalanobis	k-NN (k=5)
SD v1.4	0.542	0.703	0.430	0.656
SD v1.5	0.506	0.867	0.439	0.663
SD v2.1	0.801	0.071	0.984	0.978
Dreamshaper-8	0.471	0.805	0.445	0.649
Realistic Vision v5	0.536	0.956	0.578	0.793
SDXL	0.929	0.722	0.950	0.997
PixArt-α	1.000	0.000	1.000	1.000
SDXL-Turbo	0.983	0.072	0.999	1.000
Realistic Vision v6	0.711	0.959	0.865	0.944
epiCRealism	0.692	0.090	0.785	0.912
AbsoluteReality v1.8.1	0.469	1.000	0.486	0.644
Dreamshaper-7	0.442	0.999	0.409	0.629
Deliberate v2	0.560	0.183	0.614	0.781
RealVisXL V4.0	0.954	0.998	0.932	0.998
Dreamshaper-XL v2	0.929	0.999	0.995	0.996
Mean	0.701	0.628	0.727	**0.843**

Table 12: **Full leave-one-model-out AUROC (M=15), all 15 held-out models across all 4 scoring methods.** MLS (maximum likelihood scoring via LinearSVC margin) hides a bimodal split: near-perfect ($\geq$0.96) on 6 models but inverting below chance on 4 others (SD v2.1, SDXL-Turbo, epiCRealism, Deliberate v2). k-NN is the only method with no below-chance cells.

L ResNet-50 Per-Model Results

To test whether the difficulty of output-space attribution is specific to linear classifiers, we train a ResNet-50 directly on generated images from the 15-model registry. For each model, we use 1,000 generated images for training and evaluate on 50 held-out images, yielding a balanced 15-way classification task. Table 13 reports the resulting per-model accuracy and dominant confusion. To provide a strong output-space control, we train ResNet-50 with 1,000 labeled images per model, substantially more than the reference set used by SGS. Despite this larger training set, output-space classification remains difficult for several closely related model variants.

Model	Correct/50	Accuracy	Most confused with
SD v1.4	24	48.0%	SD v1.5 · 14/50 (28%)
SD v1.5	13	26.0%	SD v1.4 · 24/50 (48%)
SD v2.1	31	62.0%	SD v1.5 · 8/50 (16%)
Dreamshaper-8	18	36.0%	Dreamshaper-7 · 17/50 (34%)
Realistic Vision v5	34	68.0%	Realistic Vision v6 · 7/50 (14%)
SDXL	36	72.0%	RealVisXL V4.0 · 5/50 (10%)
PixArt-α	41	82.0%	SD v1.4 · 2/50 (4%)
SDXL-Turbo	48	96.0%	SD v1.5 · 1/50 (2%)
Realistic Vision v6	40	80.0%	Realistic Vision v5 · 5/50 (10%)
epiCRealism	14	28.0%	AbsoluteReality · 18/50 (36%)
AbsoluteReality v1.8.1	25	50.0%	Dreamshaper-8 · 6/50 (12%)
Dreamshaper-7	25	50.0%	Dreamshaper-8 · 7/50 (14%)
Deliberate v2	29	58.0%	AbsoluteReality · 6/50 (12%)
RealVisXL V4.0	31	62.0%	SDXL · 6/50 (12%)
Dreamshaper-XL v2	50	100.0%	-
Overall		**61.2%**	

Table 13: **ResNet-50 per-model accuracy on the 15-model task.** Every confusion above 20% falls on exactly the same-lineage fine-tune pairs that SGS identifies as hardest (v1.4↔v1.5, Dreamshaper-7↔8, epiCRealism→AbsoluteReality), confirming the task's genuine difficulty rather than an artifact of the spectral probe method.

M Model Registry: Full URLs

All models are publicly available on the Hugging Face Hub. Table 14 lists the full repository URLs for all nineteen models in the registry.

Model	URL
Core registry (8 models)	
SD v1.4	https://huggingface.co/CompVis/stable-diffusion-v1-4
SD v1.5	https://huggingface.co/runwayml/stable-diffusion-v1-5
SD v2.1	https://huggingface.co/stabilityai/stable-diffusion-2-1
Dreamshaper-8	https://huggingface.co/Lykon/dreamshaper-8
Realistic Vision v5	https://huggingface.co/SG161222/Realistic_Vision_V5.1_noVAE
SDXL	https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0
PixArt-α	https://huggingface.co/PixArt-alpha/PixArt-XL-2-1024-MS
SDXL-Turbo	https://huggingface.co/stabilityai/sdxl-turbo
Extended registry (7 additional fine-tunes)	
Realistic Vision v6	https://huggingface.co/SG161222/Realistic_Vision_V6B1_noVAE
epiCRealism	https://huggingface.co/emilianJR/epiCRealism
AbsoluteReality v1.8.1	https://huggingface.co/Lykon/AbsoluteReality
Dreamshaper-7	https://huggingface.co/Lykon/dreamshaper-7
Deliberate v2	https://huggingface.co/XpucT/Deliberate
RealVisXL V4.0	https://huggingface.co/SG161222/RealVisXL_V4.0
Dreamshaper-XL v2 Turbo	https://huggingface.co/Lykon/dreamshaper-xl-v2-turbo
Flow-matching extension	
InstaFlow	https://huggingface.co/XCLiu/instaflow-0.9B-from-sd-1-5
AuraFlow	https://huggingface.co/fal/AuraFlow
CogView4-6B	https://huggingface.co/THUDM/CogView4-6B
SANA-1600M	https://huggingface.co/Efficient-Large-Model/Sana_1600M_1024px

Table 14: Full Hugging Face repository URLs for all models used in this work.