PRISM: POST-HOC RECOGNITION OF IMAGE SOURCE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Image generative models (GMs) now produce photorealistic content that is virtually indistinguishable from natural images. This creates a critical need for reliable provenance to distinguish between synthetic vs authentic images and attribute generated content to its source. While diffusion models (DMs) have been the dominant paradigm for visual content generation, recent image autoregressive models (IARs) match DMs' generation quality and substantially improve the sampling speed by adapting the next-token prediction paradigm from large language models. We propose a unified post-hoc attribution method that identifies the source GM of a given image by exploiting model-specific artifacts (fingerprints) that arise in both DMs and IARs. Our key insight is to combine the fingerprints from the two main constituent parts of GMs: (1) the autoencoder that maps between pixel and latent spaces and (2) the latent generator that defines the image distribution in DMs and produces visual tokens in IARs. In the first stage, we identify the autoencoder family using reconstruction and quantization errors. In the second stage, we attribute the image to the latent generator within the identified model family using conditional log-likelihoods for DMs and token log-probabilities for IARs. Across diverse model families and versions, our two-stage framework enables accurate model attribution, providing a unified approach to the provenance of synthetic images.

1 INTRODUCTION

The rapid advancement in image generation blurs the boundary between real and generated content. The photorealistic images produced by generative models (GMs), which are indistinguishable from human-created images, undermine misinformation detection (Kramer et al., 2025). They are also published online, then scraped, and reused as training data for new GMs, causing performance degradation known as model collapse (Alemohammad et al., 2023; Shumailov et al., 2024). This process amplifies existing models' errors, biases, and unfairness by embedding them into future model versions (Wyllie et al., 2024). These consequences call for image generative model attribution that not only enables the distinction between generated and natural content, but also traces the harmful synthetic images to the exact source model that produced the content to prevent the training of a new model on data generated from its previous versions.

One of the solutions to trace the exact source of harmful content is to embed invisible watermarks into the GMs and detect the watermarks to verify if the image comes from a given model (Kirchenbauer et al., 2023; Fernandez et al., 2023; Wen et al., 2023; Gunn et al., 2025; Li et al., 2025; Kerner et al., 2025; Tong et al., 2025; Wang et al., 2025). However, such watermarks introduce a pattern during the training or generation pipeline of GMs, and therefore cannot be applied when the image is already generated and posted on the internet. Different from watermarking-based methods, recently proposed post-hoc image attribution methods (Wang et al., 2026; Damm et al., 2025; Wang et al., 2024) do not require any modifications to the generative pipeline and perform image attribution based only on the signals introduced during the generation process. However, they consider solely either DMs (Wang et al., 2024; 2026) or IARs (Damm et al., 2025), separately, lacking a unified model attribution across the image GMs. While the main building blocks of these GMs are jointly an autoencoder and a latent image generator, most existing post-hoc image attribution methods are built on the artifacts introduced only either by the autoencoder (Wang et al., 2024; 2026) or the autoregressive image generator (Damm et al., 2025), instead of considering the model holistically.

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.

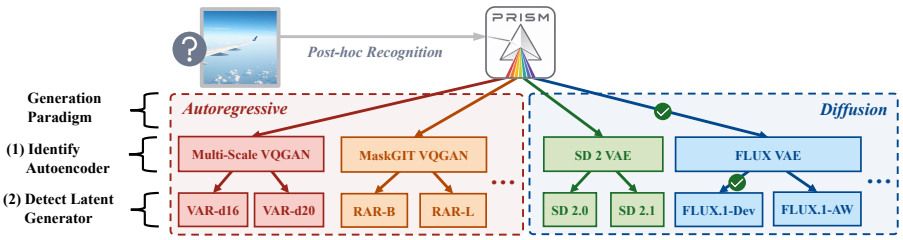


Figure 1: **Overview of our Post-hoc Recognition of Image Source Models (PRISM) framework.** ***Supported Generation Paradigm.*** Our method is designed to cater to both autoregressive and diffusion generation paradigms, achieving **high accuracy on 24 models**. It consists of the following two main stages: **(1)** ***Identify Autoencoder.*** We use the characteristics of the autoencoder based on the image reconstruction loss to identify the specific autoencoder version, which defines a model family. This is followed by stage: **(2)** ***Detect Latent Generator.*** We identify the specific latent image generative models, such as autoregressive models generating image tokens for IARs and the neural backbones of the latent diffusion models.

To the best of our knowledge, we propose the first **P**ost-hoc **R**ecognition of **I**mage **S**ource **M**odels (PRISM) framework for tracing an image to its source generative model, being universally applicable to both DMs and IARs, operating entirely post-hoc, and identifying specific autoencoder and model versions. We provide a graphical overview of our method in Figure 1. The central insight of our work is that attributing an image to a particular GM cannot rely solely on either the autoencoder or the latent generator. To address this issue, we propose a two-stage unified approach that combines signals from both components. In Stage 1, we focus on the autoencoder, measuring the image reconstruction loss twice for accurate identification. Additionally, we leverage quantization loss for IARs due to their discrete image generation process. After detecting the model family, Stage 2 involves latent generator attribution. Here, we compute conditional scores using log-likelihoods for DMs and log-probabilities for IARs, and attribute the image to the latent generator with the maximum score. Overall, our method advances provenance for generated images by unifying attribution methods to identify the source diffusion or autoregressive models using artifacts inherently left by the generation pipeline.

Our contributions can be summarized as follows:

1. We propose the first unified attribution method for image GMs, which is effective for both DMs and IARs, and traces an image back to its exact source model.
2. We identify fingerprints for both autoencoders and latent generators in DMs and IARs, and integrate them into a unified two-stage provenance framework.
3. Our approach significantly outperforms six other image provenance methods that we test on both IARs and DMs, and across a diverse set of 24 image GMs.

2 Background and Related Work

2.1 Generative Models

The most prominent families of image generative models (GMs) are diffusion models (DMs) and image autoregressive models (IARs). These models typically utilize an autoencoder, which compresses an image from the pixel space into a latent space. The generative process then operates in this latent space, synthesizing new image representations that are subsequently decoded back into the pixel space.

Diffusion Models (DMs). DMs generate images by gradually refining random Gaussian noise into structured outputs through a learned denoising process (Song & Ermon, 2020; Ho et al., 2020). During generation, a sample is drawn as an initial noise image $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, followed by T iterative steps that progressively remove predicted noise $\epsilon_{\mathcal{G}}(\mathbf{x}_t, t, \mathbf{c})$ at each time step $t = T, \ldots, 1$.

Image Autoregressive Models (IARs). IARs generate images leveraging the next-token prediction similarly to LLMs. IARs consist of a vector-quantized variational autoencoder (VQVAE) that

encodes images into discrete latent tokens, and an autoregressive transformer that models the distribution over these tokens to later generate the next tokens. For instance, LlamaGen (Sun et al., 2024) employs an autoencoder to downsample images into a quantized latent, which is then modeled using a Llama-based autoregressive transformer. VAR (Tian et al., 2024) designs a scale-wise VQVAE and a coarse-to-fine latent generation. Randomized autoregressive model (RAR) (Yu et al., 2025b) further generalizes next-token prediction by introducing randomized token permutations into the training procedure.

2.2 Attribution for Image Generative Models

There are mainly two types of methods for image attribution: *watermark-based* and *post-hoc* methods. Watermark-based attribution infers the source of a generated image via signals embedded by the model owner during training or inference (Wen et al., 2023; Li et al., 2025; Kerner et al., 2025; Meintz et al., 2025). Such methods require modifying the generative pipeline and are therefore inapplicable to models or images that have already been released. In contrast, post-hoc provenance methods rely on inherent fingerprints of the model and can be applied to existing models and images without any modification. Existing post-hoc provenance methods detect the model family based on a given autoencoder in DMs. AEROBLADE (Ricker et al., 2024) uses reconstruction error as a simple signal, while RONAN (Wang et al., 2023) and LatentTracer (Wang et al., 2024) derive stronger gradient-based scores from reconstruction losses. AEDR (Wang et al., 2026) further improves robustness of the attribution via double reconstruction ratio. In contrast, PRADA (Damm et al., 2025) targets the attribution of only model version in IARs by leveraging IAR-specific membership inference attacks (Kowalczuk et al., 2025; Yu et al., 2025a), followed by a weighting mechanism that combines multiple signals. D^3QE (Zhang et al., 2025) and Zhao et al. (2026) introduce the quantization error of a VQVAE codebook as a detection signal for provenance within the IAR paradigm. Overall, prior post-hoc attribution methods tend to specialize in a single generative paradigm (DM or IAR) and support only a subset of attribution tasks. We compare the existing post-hoc methods in Table A12.

3 Method

In this section, we formulate the problem of image attribution in Section 3.1, and define the generation process in Section 3.2. We then introduce the two types of fingerprints for the autoencoder and latent generators in Section 3.3 and Section 3.4, respectively. In these two sections, we also design the instantiations for both fingerprints for IARs and DMs. Motivated by our observations in Figure 2, we propose the unified provenance framework of two-stage model attribution in Section 3.6.

3.1 Problem Formulation

We introduce the following two settings for our task.

(i) Owner auditing. A model owner holds every version they have released, and aims to identify the source models of generated images for two reasons. First, in *data curation*, the owner must exclude images produced by their own earlier versions from a new training corpus, breaking the self-consumption loop that drives model collapse. Second, in *incident response*, a harmful image is traced to the specific released version that produced it, which is the version to be patched, retrained, or withdrawn. Model version attribution serves both tasks: it separates the owner's own outputs from a third party's fine-tune of the same base, and it traces the unsafe checkpoint to fix.

(ii) External auditing. An external auditor aim to identify the source models of a given image. The auditor holds no weights and cannot inspect any model, but can require the enrolled model owners to participate. Each owner evaluates the query image locally, running Stage 1 against their private reference populations and Stage 2 over their own generators, and returns only a categorical decision without releasing model weights. We give the detailed protocol in Appendix J. The auditor thereby learns which enrolled version produced the image, and can suspend that version from the registry until it is remediated, while learning nothing else about any participant's models.

For both cases, the core problem can be formalized with the following notations. Given an image $\mathbf{x}$, we consider a set of suspect generative models $\{\mathcal{M}_i^j \mid i = 1, \ldots, m;\ j = 1, \ldots, n_i\}$. Each

model $\mathcal{M}_i^j$ consists of an autoencoder $\mathcal{A}_i$ (with encoder $\mathcal{E}_i$ and decoder $\mathcal{D}_i$) and a latent generator $\mathcal{G}_i^j$. Multiple latent generators $\{\mathcal{G}_i^j\}_{j=1}^{n_i}$ share the same autoencoder $\mathcal{A}_i$, forming a *model family*. Our goal is to determine whether $\mathbf{x}$ was generated by one of these models and, if so, to identify both the model family (shared autoencoder) and the specific model version (latent generator).

3.2 Preliminaries for Image Generative Models

The generation process for a generative model $\mathcal{M}$ can be formalized into the following two steps.

1) Latent Sampling: Based on the conditioning $\mathbf{c}$, the latent generator predicts a latent prior $P_{\mathcal{G}}(\mathbf{z}|\mathbf{c})$, from which a latent $\mathbf{z}$ is sampled via $\mathbf{z} \sim P_{\mathcal{G}}(\mathbf{z}|\mathbf{c})$. Particularly, $\mathcal{G}$ refers to the denoising model in DMs, and the autoregressive transformer in IARs.

2) Latent Decoding: The latent $\mathbf{z}$ is further decoded from the latent space to the pixel space via $\mathbf{x} = \mathcal{D}(\mathbf{z})$. Note that, $\mathcal{E}$ and $\mathcal{D}$ are trained in a pairwise manner, such that $\mathbf{x}$ can also be mapped to the latent space with $\mathcal{E}$. Concretely, the autoencoder is instantiated as a continuous VAE in diffusion models, and a VQVAE (Vector Quantized VAE) in IARs.

3.3 Autoencoder Fingerprint

Since a generated image $\mathbf{x}$ is decoded from the latent representation $\mathbf{z}$, we observe that $\mathbf{z}$ resides in a special latent feature space defined by the encoder $\mathcal{E}$ and decoder $\mathcal{D}$. Since the compression and reconstruction mappings between $\mathbf{x}$ and $\mathbf{z}$ are specific to the pair $(\mathcal{E}, \mathcal{D})$, images not generated by this model exhibit high reconstruction error when processed by $\mathcal{E}$ and $\mathcal{D}$, which serves as an effective attribution signal. Given an image $\mathbf{x}$ and an autoencoder $\mathcal{A}$ comprising encoder $\mathcal{E}$ and decoder $\mathcal{D}$, we define the reconstruction loss as:

$$\mathcal{L}_{\text{Rec}}(\mathbf{x}, \mathcal{A}(\mathbf{x})) = \text{MSE}(\mathbf{x}, \mathcal{D} \circ \mathcal{E}(\mathbf{x})), \tag{1}$$

where $\mathcal{D} \circ \mathcal{E}(\mathbf{x}) = \mathcal{D}(\mathcal{E}(\mathbf{x}))$ denotes the reconstructed image obtained by encoding $\mathbf{x}$ into the latent space and subsequently decoding it, and $\text{MSE}(\cdot, \cdot)$ denotes the mean squared error. Inspired by AEDR (Wang et al., 2026), we reconstruct the image twice and calculate the double reconstruction ratio.

$$\rho_{\text{Rec}}(\mathbf{x}) = \frac{\mathcal{L}_{\text{Rec}}(\mathbf{x}, \mathcal{A}(\mathbf{x}))}{\mathcal{L}_{\text{Rec}}(\mathcal{A}(\mathbf{x}), \mathcal{A} \circ \mathcal{A}(\mathbf{x}))}. \tag{2}$$

VAE Attribution for DMs. For diffusion models, we directly use the above double reconstruction ratio ρ_{Rec} as the AE attribution signal, where we define $\mathcal{L}_{\mathcal{A}} = \rho_{\text{Rec}}$.

VQVAE Attribution for IARs. Since VQVAE is the main autoencoder architecture for most IARs, we can formulate the autoencoder reconstruction for IARs as follows:

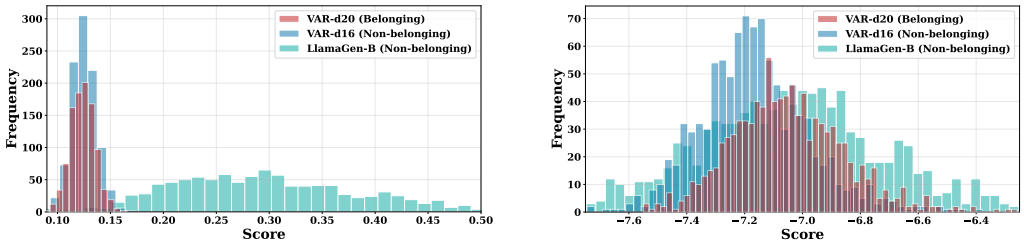
$$\mathcal{A}(\mathbf{x}) = \mathcal{D} \circ \mathcal{Q}^{-1} \circ \mathcal{Q} \circ \mathcal{E}(\mathbf{x}), \tag{3}$$

where $\mathcal{Q}$ denotes the quantization operation, which maps a continuous feature map to discrete tokens by looking up in a codebook $\mathcal{Z}$. Correspondingly, the dequantization operation $\mathcal{Q}^{-1}$ embeds discrete tokens into the vectors in $\mathcal{Z}$. When mapped back to the latent space, an image generated by a given IAR model has a smaller quantization error than non-belonging images. Inspired by Zhao et al. (2026) and D^{3}QE (Zhang et al., 2025), we leverage the quantization error as another attribution signal for IARs, which can be formulated as $\mathcal{L}_{\text{Q}} = \text{MSE}(\mathcal{E}(\mathbf{x}), \mathcal{Q}^{-1} \circ \mathcal{Q} \circ \mathcal{E}(\mathbf{x}))$.

The overall fingerprint is a composition of reconstruction and quantization errors, formalized as $\mathcal{L}_{\mathcal{A}} = \rho_{\text{Rec}} \mathcal{L}_{\text{Q}}$. For IARs, Stage 1 uses an encoder finetuned to act as an inverse decoder (Zhao et al., 2026): with the decoder $\mathcal{D}$ frozen, $\hat{\mathcal{E}}$ is optimized to minimize $\text{MSE}(\hat{\mathcal{E}} \circ \mathcal{D}(\mathbf{z}), \mathbf{z})$ on synthetic latents. The encoder finetuning is a one-time, per-family offline step, and the inference remains training-free.

3.4 Latent Generator Fingerprint

For an image $\mathbf{x}$ generated by latent generator $\mathcal{G}$ based on the condition $\mathbf{c}$, its latent representation $\mathbf{z}$ is sampled by the latent generator from a prior of $P_{\mathcal{G}}(\mathbf{z}|\mathbf{c})$. Therefore, the latent $\mathbf{z}$ has a high predicted probability on the latent generator $\mathcal{G}$. Based on this observation, we obtain the latent by encoding $\mathbf{x}$



(a) Autoencoder fingerprint.

(b) Latent generator fingerprint.

Figure 2: **Comparisons of the autoencoder and latent generator fingerprints.** We show the score distributions for ***(a)*** *autoencoder fingerprints* $\mathcal{L}_{\mathcal{A}}$, and ***(b)*** *the latent generator fingerprints* $\mathcal{L}_{\mathcal{G}}$. The evaluated model is VAR-d20. We use VAR-d20 generated data as the belonging data, and VAR-d16/LlamaGen-B generated data as non-belonging data.

via $\tilde{\mathbf{z}} = \mathcal{E}(\mathbf{x})$ and calculate the conditional log probability with $\mathcal{G}$ as $\log p_{\mathcal{M}}(\mathbf{x}|\mathbf{c}) = \log p_{\mathcal{G}}(\mathcal{E}(\mathbf{x})|\mathbf{c})$. Particularly, we obtain the label for an image with an ImageNet classifier EVA02 (Fang et al., 2024) for class-to-image models, and estimate the prompts with an image-to-text model BLIP2 (Li et al., 2023b) for text-to-image GMs. We provide comparisons of performance between using ground-truth conditions and estimated conditions in Appendix F. We find that our proposed conditional-likelihood signal is simple but effective, outperforming carefully designed calibration mechanisms proposed by Kowalczuk et al. (2025), ICAS (Yu et al., 2025a) and PRADA (Damm et al., 2025).

Denoising Model Attribution for DMs. For diffusion models, the probabilities of an image can be estimated via the evidence lower bound (ELBO) (Ho et al., 2020; Li et al., 2023a) of the conditional log-likelihood:

$$\log p_{\mathcal{M}}(\mathbf{x}_0 \mid \mathbf{c}) \geq -\mathbb{E}_{t,\epsilon}\left[\|\epsilon - \epsilon_{\mathcal{M}}(\mathbf{x}_t, t, \mathbf{c})\|^2\right] + C, \tag{4}$$

where $\epsilon \sim \mathcal{N}(0, \mathcal{I})$, $\epsilon_{\mathcal{M}}(\mathbf{x}_t, t, \mathbf{c})$ is a noise predictor, $t \sim \text{Uniform}(1, ..., T)$, and C is a constant. Therefore, we estimate $\log p_{\mathcal{M}}(\mathbf{x}_0 \mid \mathbf{c})$ with the noise prediction error of the denoising probabilistic model $-\mathbb{E}_{t,\epsilon}\left[\|\epsilon - \epsilon_{\mathcal{M}}(\mathbf{x}_t, t, \mathbf{c})\|^2\right]$. More precisely, the noise is predicted by $\mathcal{G}$ in the latent space $\mathbf{z} = \mathcal{E}(\mathbf{x})$, so we use $-\mathbb{E}_{t,\epsilon}\left[\|\epsilon - \epsilon_{\mathcal{G}}(\mathbf{z}_t, t, \mathbf{c})\|^2\right]$.

Autoregressive Transformer Attribution for IARs. For IARs, the image is encoded and quantized into a token sequence $\mathbf{r} = \mathcal{Q} \circ \mathcal{E}(\mathbf{x})$ of T discrete codebook indices. The autoregressive transformer $\mathcal{G}$ generates the tokens by sequentially predicting the probability of the next token based on the contextual (previous) tokens and the conditioning $\mathbf{c}$, and models the conditional log-probability as $\log p_{\mathcal{G}}(\mathbf{r}|\mathbf{c}) = \sum_{t=1}^{T} \log p_{\mathcal{G}}(\mathbf{r}_t|\mathbf{r}_{<t}, \mathbf{c})$.

3.5 Comparison of Two Fingerprints

The autoencoder fingerprint can be formulated as follows:

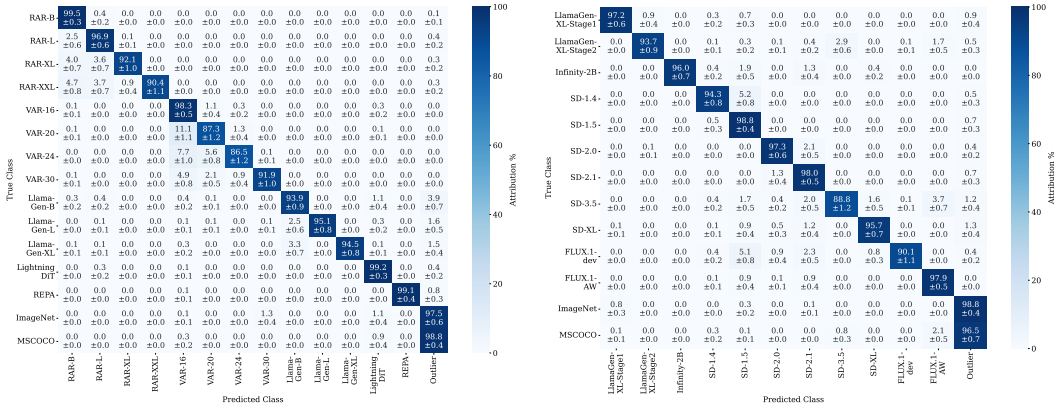
$$\mathcal{L}_{\mathcal{A}} = \begin{cases} \rho_{\text{Rec}}, & \text{if } \mathcal{M} \text{ is DM}, \\ \rho_{\text{Rec}} \cdot \mathcal{L}_{\text{Q}}, & \text{if } \mathcal{M} \text{ is IAR}, \end{cases} \tag{5}$$

and the latent generator fingerprint can be formulated as

$$\mathcal{L}_{\mathcal{G}} = \begin{cases} -\mathbb{E}_{t,\epsilon}\left[\|\epsilon - \epsilon_{\mathcal{G}}(\mathbf{z}_t, t, \mathbf{c})\|^2\right], & \text{if } \mathcal{M} \text{ is DM}, \\ \sum_{t=1}^{T} \log p_{\mathcal{G}}(\mathbf{r}_t|\mathbf{r}_{<t}, \mathbf{c}), & \text{if } \mathcal{M} \text{ is IAR}. \end{cases} \tag{6}$$

We plot the distributions of autoencoder fingerprints $\mathcal{L}_{\mathcal{A}}$ and latent generator fingerprints $\mathcal{L}_{\mathcal{G}}$ in Figure 2 (a) and (b), respectively. Low $\mathcal{L}_{\mathcal{A}}$ indicates low double reconstruction ratio ρ_{Rec} and/or low quantization error $\mathcal{L}_{\text{Q}}$, and therefore the image is more likely to be generated by this autoencoder $\mathcal{A}$ of the given model $\mathcal{M}$. High $\mathcal{L}_{\mathcal{G}}$ means that the latent generator $\mathcal{G}$ predicts a high probability for the image, indicating a higher likelihood that the image belongs to this generator $\mathcal{G}$ of the model $\mathcal{M}$.

Strengths and Limitations of Autoencoder Fingerprints. As shown in Figure 2 (a), the autoencoder exhibits significantly lower $\mathcal{L}_{\mathcal{A}}$ for images decoded from its own decoder compared to images



(a) Class-to-image models.

(b) Text-to-image models.

Figure 3: **Attribution results.** We present the attribution accuracy (%) for class-to-image (a) and text-to-image models (b).

from other model families. This clear separation enables reliable *model family* identification, as images generated by different autoencoders produce distinct reconstruction patterns. However, the autoencoder fingerprint is fundamentally rooted in the decoder $\mathcal{D}$: any latent representation $\mathbf{z}$ sampled by *any* latent generator $\mathcal{G}$, as long as it uses $\mathcal{D}$ for decoding, will exhibit the same low $\mathcal{L}_{\mathcal{A}}$. Consequently, autoencoder fingerprints alone cannot distinguish between different model versions within the same family (*e.g.,*, VAR-d16 vs. VAR-d20), as they share identical autoencoders.

Strengths and Limitations of Latent Generator Fingerprints. As shown in Figure 2 (b), the latent generator fingerprint $\mathcal{L}_{\mathcal{G}}$ for images generated by VAR-d20 are clearly higher than that generated by VAR-d16, despite both models sharing the same autoencoder. This demonstrates that $\mathcal{L}_{\mathcal{G}}$ captures model-specific generation patterns at the latent generator level, enabling fine-grained *model version* attribution. However, latent generator fingerprints alone are insufficient for reliable attribution. First, images from different model families may produce similar $\mathcal{L}_{\mathcal{G}}$ values, since the fingerprint does not account for autoencoder-specific artifacts. Second, real images may also receive high predicted probabilities under certain generators, as the learned distributions can overlap with the natural image manifold. As illustrated in Figure 2 (b), the score distributions for VAR-d16 and LlamaGen-B show significant overlap, making family-level disambiguation challenging with $\mathcal{L}_{\mathcal{G}}$ alone.

Complementary Nature of Both Fingerprints. The analysis above reveals that autoencoder fingerprints excel at model family identification but fail at version-level attribution, while latent generator fingerprints enable version-level discrimination but struggle with family-level separation and real image rejection. This complementary nature motivates our two-stage attribution framework, which leverages both fingerprints in a hierarchical manner to achieve unified model attribution.

3.6 Unified Provenance Framework

Motivated by the complementary strengths and limitations of both fingerprints discussed above, we propose a unified two-stage provenance framework that systematically combines autoencoder and latent generator fingerprints to trace an image to its source generative model. Our framework addresses the problem formulation in Section 3.1 and can successfully identify both the model family (autoencoder $\mathcal{A}_i$) and the specific model version (latent generator $\mathcal{G}_i^j$). The framework is also summarized in Appendix A.

Stage 1: Autoencoder Identification. In the first stage, we identify which autoencoder family the image belongs to by leveraging the autoencoder fingerprint $\mathcal{L}_{\mathcal{A}}$. For each candidate autoencoder $\mathcal{A}_i$ in our suspect set, we compute $\mathcal{L}_{\mathcal{A}}(\mathbf{x}; \mathcal{A}_i)$ according to Equation (5).

We employ Grubbs' test Grubbs (1949) to statistically test the following hypothesis:

$$\mathcal{H}_0^{(i)} : \text{The image } \mathbf{x} \text{ was not generated by autoencoder } \mathcal{A}_i. \tag{7}$$

Table 1: **Comparison between post-hoc image attribution methods.** We present the F1 scores.

Type	Method	RAR				VAR				LlamaGen			LightningDiT	REPA	Natural		Macro-F1
		B	L	XL	XXL	d16	d20	d24	d30	B	L	XL	XL	XL	ImageNet	MSCOCO	
Autoencoder Attribution	AEROBLADE	2.7	2.7	0.0	0.0	0.6	0.0	0.0	0.2	0.6	0.0	0.0	14.2	87.4	0.8	0.0	7.3
	AEDR	6.8	8.4	0.0	0.0	13.7	0.0	0.0	0.0	2.3	0.0	0.0	0.0	0.0	0.0	0.2	2.1
	Autoencoder only	40.0	0.2	0.0	0.0	39.8	0.0	0.0	0.0	48.8	0.0	0.0	98.6	99.0	94.3	95.0	34.4
Generator Attribution	Generator only	4.1	2.6	2.1	2.1	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	12.6	0.0	0.0	1.6
	Delta	43.2	62.8	80.9	76.3	38.2	40.3	50.5	60.9	70.1	61.9	56.6	11.0	15.8	0.0	0.0	44.6
	PRADA	85.0	92.3	90.6	92.3	40.7	32.2	44.2	62.7	97.3	97.7	97.0	0.0	0.0	1.1	42.7	58.4
PRISM		**94.1**	**94.4**	**95.4**	**95.0**	**88.0**	**89.0**	**91.6**	**95.1**	**94.0**	**97.5**	**97.2**	**97.7**	**99.5**	**94.3**	**95.0**	**94.5**

The Grubbs' test is well-suited to our Stage 1 attribution, as it directly tests whether an individual observation deviates significantly from a reference distribution, enabling rejection of non-belonging images. Sensitivity to the significance level α and the population size N, and the Gaussianity assumption, are examined in Appendix I. Given a reference population of N images known to be generated by $\mathcal{A}_i$, with mean μ_i and standard deviation σ_i of their $\mathcal{L}_\mathcal{A}$, we reject $\mathcal{H}_0^{(i)}$ if:

$$\frac{\mathcal{L}_\mathcal{A}(\mathbf{x};\mathcal{A}_i)-\mu_i}{\sigma_i} < \frac{N-1}{\sqrt{N}}\sqrt{\frac{(t_{\alpha/N,N-2})^2}{N-2+(t_{\alpha/N,N-2})^2}}, \tag{8}$$

where $t_{\alpha/N,N-2}$ denotes the critical value of the t-distribution with $N-2$ degrees of freedom at significance level α/N. We test all candidate autoencoders $\{\mathcal{A}_1,\ldots,\mathcal{A}_m\}$ and identify the set of autoencoders for which $\mathcal{H}_0$ is rejected:

$$\mathcal{S}_{\text{AE}} = \{\mathcal{A}_i \mid \mathcal{H}_0^{(i)} \text{ is rejected}\}. \tag{9}$$

If $\mathcal{S}_{\text{AE}} = \emptyset$, the image is classified as *not generated* by any of the suspect models (potentially a real image or from an unknown image GM). Otherwise, we proceed to Stage 2 with the identified autoencoder.

Stage 2: Latent Generator Attribution. Having identified the autoencoder family $\mathcal{A}_{i^*}$ in Stage 1, we now determine which specific latent generator $\mathcal{G}_{i^*}^j$ within this family produced the image. All latent generators $\{\mathcal{G}_{i^*}^1,\ldots,\mathcal{G}_{i^*}^{n_{i^*}}\}$ sharing the autoencoder $\mathcal{A}_{i^*}$ are candidates.

For each candidate latent generator $\mathcal{G}_{i^*}^j$, we compute the latent generator fingerprint $\mathcal{L}_\mathcal{G}(\mathbf{x};\mathcal{G}_{i^*}^j)$ according to Equation (6). Since images generated by their source generator exhibit the *highest* predicted probability, we attribute the image to the latent generator with the maximum $\mathcal{L}_\mathcal{G}$ score:

$$j^* = \underset{j\in\{1,\ldots,n_{i^*}\}}{\arg\max}\ \mathcal{L}_\mathcal{G}(\mathbf{x};\mathcal{G}_{i^*}^j). \tag{10}$$

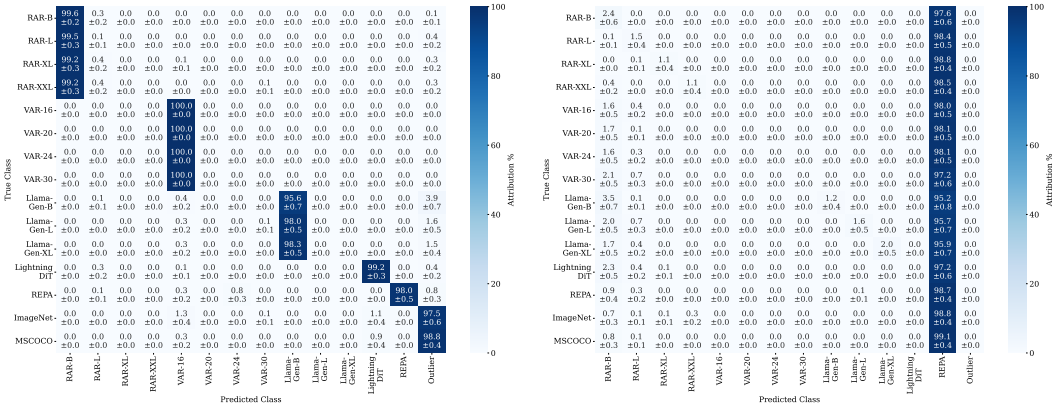
The final attribution result can be formalized as the model tuple $\mathcal{M}_{i^*}^{j^*} = (\mathcal{A}_{i^*}, \mathcal{G}_{i^*}^{j^*})$.

4 Experiments

4.1 Experimental Setup

Models. For image autoregressive models, we evaluate 4 state-of-the-art model families: LlamaGen (Sun et al., 2024), RAR (Yu et al., 2025b), VAR (Tian et al., 2024), and Infinity (Han et al., 2025). We include both class-to-image and text-to-image variants for LlamaGen. We also experiment with 7 diffusion model families: Stable Diffusion 1 (Rombach et al., 2022), Stable Diffusion 2, Stable Diffusion 3.5 (Esser et al., 2024), SD XL (Podell et al., 2024), LightningDiT (Yao et al., 2025), REPA (Yu et al., 2025c), and FLUX (Labs et al., 2025). We assess 24 different model versions for the above model families, and a full list of evaluated models is described in Appendix B.

Datasets. We build an evaluation dataset consisting of 26,000 images, including both generated and real images. Specifically, we generate 1,000 images for each of the 24 aforementioned generative models. For class-to-image models, we randomly sample labels from 1000 ImageNet classes and use them as inputs for generation. For text-to-image models, we randomly sample 1000 text captions



(a) Only autoencoder fingerprints.

(b) Only latent generator fingerprints.

Figure 4: **Attribution results with fingerprints only for: (a) autoencoder and (b) latent generator.** We present the confusion matrices normalized by each row.

from MS-COCO as prompts to the models. We also include real datasets, ImageNet (Deng et al., 2009) and MS-COCO (Lin et al., 2014), in our evaluation, where we randomly sample 1000 images from each dataset. For the 1000 images in each subset, we randomly select 250 images as our population set for the Grubbs' test in autoencoder attribution, and use the other 750 images for evaluation. The hardware and software configuration used for every experiment is listed in Appendix O.

4.2 Model Attribution Results

Class-to-image Attribution. Figure 3a presents the confusion matrix for class-to-image model attribution. Our method successfully distinguishes images across model families with different autoencoders, model versions, and real images. Cross-family confusion is negligible, with near-zero misattribution between RAR, VAR, LlamaGen, LightningDiT, and REPA families, validating the effectiveness of our autoencoder fingerprint in the first attribution stage. Within each family, our latent generator fingerprint enables fine-grained version attribution. Importantly, real images from ImageNet and MS-COCO are correctly classified as outliers with high accuracy, demonstrating that our framework reliably avoids false attribution of natural images to any suspect models.

Text-to-image Attribution. Figure 3b shows the attribution results for text-to-image models spanning both IARs and DMs. Our method achieves consistently high attribution accuracy across all evaluated models, with LlamaGen-XL-Stage1 and Stage2 reaching 97.2% and 93.7% accuracy, respectively. For diffusion models, we observe strong performance across all Stable Diffusion variants and FLUX variants. Notably, cross-paradigm confusion between IARs and DMs remains negligible, demonstrating that our unified framework effectively handles both generation paradigms. Real images from ImageNet and MS-COCO are correctly identified as outliers, confirming robust rejection of natural images even in the text-to-image setting.

4.3 Comparisons between Provenance Methods

We compare different provenance approaches in Table 1.

Different Autoencoder Fingerprints. We test three different attribution methods based on autoencoder fingerprints. *AEROBLADE* achieves only 7.3% Macro-F1, as simple reconstruction error fails to distinguish between model families effectively. *AEDR* reaches 2.1% Macro-F1, improving on certain datasets (RAR-B, RAR-L, VAR-d16) but struggling with fine-grained attribution overall. The *Autoencoder only* approach based on Equation (5) achieves 34.4% Macro-F1 with strong performance in distinguishing between model families, but failing at detecting model versions within the same family due to its inability to differentiate model versions sharing a single autoencoder.

Different Latent Generator Fingerprints. We also evaluate three latent generator attribution methods. We start with applying the conditional probability for the *Generator only* method based on

Equation (6), which achieves only 1.6% Macro-F1, failing across nearly all model families because the latent generator fingerprint alone cannot perform reliable outlier rejection. We further extend baseline methods originally designed for IARs in (Damm et al., 2025) to DMs. The *Delta* score based on the differences in membership inference scores between conditional vs unconditional generations for both IARs (Kowalczuk et al., 2025) and DMs (Zhai et al., 2024) improves the overall performance to 44.6% Macro-F1, but fails on most VAR variants, the diffusion-based class-to-image models, and natural images. We provide more detailed results for both class-to-image and text-to-image models in Appendix C. *PRADA* achieves 58.4% Macro-F1, but has very limited performance on VAR variants, LightningDiT, REPA, and natural images.

In contrast to the above approaches, our combined PRISM method featuring fingerprints for both autoencoders and latent generators achieves robust identification across all tested model architectures. Appendix L additionally compares against classification baselines, and Appendix K contrasts PRISM with a generic synthetic-image detector.

4.4 Ablation Studies

In this section, we mainly compare our unified PRISM framework against only using one type of fingerprint. Further ablation studies and detailed confusion matrices are included in Appendix D.

Autoencoder Attribution. Figure 4a presents attribution results using only autoencoder fingerprints based on Equation (5). The autoencoder fingerprint achieves near-perfect family-level separation. Crucially, real images from ImageNet and MS-COCO are reliably rejected as outliers. However, the autoencoder fingerprint cannot distinguish between model versions within the same family since all model versions sharing the same autoencoder receive identical scores. This limitation motivates our two-stage approach where autoencoder fingerprints handle model family identification while latent generator fingerprints enable attribution to a specific model version.

Latent Generator Attribution. Figure 4b shows attribution results using fingerprints only from the latent generator based on Equation (6). Without the autoencoder stage, the latent generator fingerprint $\mathcal{L}_G$ does not provide a sufficiently discriminative signal across model families: across virtually every row, more than 95% of images collapse to the REPA prediction, while the outlier route is essentially never triggered. This collapse occurs because $\mathcal{L}_G$ values from REPA are systematically the lowest across nearly all inputs, so a per-class minimum-loss decision on $\mathcal{L}_G$ alone is dominated by REPA. These results confirm that latent generator fingerprints alone cannot reliably perform model-family identification or outlier rejection, validating our design choice to use initial model family identification before applying latent generator attribution.

In addition to the above experiments, Appendix E reports robustness across augmentation types and strengths, Appendix G evaluates the performance under the open-world setting, inference cost is reported in Appendix H, and Appendix M examines performance under unseen model versions.

5 Conclusions

We introduce PRISM: the first unified provenance framework for attributing a generated image to its source GM across the two dominant paradigms of DMs and IARs. Our central observation is that accurate provenance tracing must account for the two constituent components of modern image GMs, namely the autoencoder and the latent generator. Prior post-hoc methods typically use fingerprints only from one of these components and are largely confined to a single paradigm (either DM or IAR). We address this gap with a two-stage procedure that (1) identifies the underlying autoencoder version and (2) attributes the image to the specific latent generator. We strengthen the autoencoder identification by leveraging double reconstruction loss for DMs and extend the signal to IARs by combining it with the quantization loss. In the second stage, we show that scores based on the log-likelihoods for DMs and log-probabilities for IARs provide sufficient discrimination signal to pinpoint the latent generator version and, consequently, the final version of the image GM. Our comprehensive empirical evaluation on 24 widely used state-of-the-art class-to-image and text-to-image DMs and IARs demonstrates consistently high attribution accuracy, outperforming six other methods by a large margin. Overall, our novel PRISM provenance framework provides a practical foundation for general post-hoc attribution of synthetic images to their source models, advancing toward more transparency in generated content.

AI USE STATEMENT

Large language models were used solely to assist with editing, including improving grammar, clarity, and phrasing of text originally written by the authors. They were not used for research ideation, methodology design, analysis, or generation of technical content. We have reviewed all AI-assisted text and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The two-stage attribution procedure is given in full as pseudo-code in Appendix A. All 24 evaluated generative models, their autoencoders, and the checkpoints used are listed in Appendix B, and the hardware and software environment is described in Appendix O. The conditioning-estimation models used for label and prompt recovery are documented in Appendix F, and the sensitivity of the Grubbs' test to its hyperparameters is analyzed in Appendix I. Complete confusion matrices and per-version scores backing every number in the main text are reported in Appendix C, Appendix D, and Appendix E.

REFERENCES

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go mad. In *The Twelfth International Conference on Learning Representations*, 2023.

Michael George Bulmer. *Principles of statistics*. Courier Corporation, 2012.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

Simon Damm, Jonas Ricker, Henning Petzka, and Asja Fischer. Prada: Probability-ratio-based attribution and detection of autoregressive-generated images. *arXiv preprint arXiv:2511.20068*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

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 (ICCV)*, pp. 22466–22477, October 2023.

Frank Ephraim Grubbs. *Sample criteria for testing outlying observations*. University of Michigan, 1949.

Sam Gunn, Xuandong Zhao, and Dawn Song. An undetectable watermark for generative image models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15733–15744, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6840–6851, 2020.

Louis Kerner, Michel Meintz, Bihe Zhao, Franziska Boenisch, and Adam Dziedzic. Bitmark: Watermarking bitwise autoregressive image generative models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pp. 17061–17084. PMLR, 2023.

Antoni Kowalczuk, Jan Dubiński, Franziska Boenisch, and Adam Dziedzic. Privacy attacks on image autoregressive models. In *Forty-second International Conference on Machine Learning*, 2025.

Robin SS Kramer, Alex L Jones, Daniel Fitousi, and Jeremy J Tree. Ai-generated images of familiar faces are indistinguishable from real photographs. *Cognitive Research: Principles and Implications*, 10(1):70, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2206–2217, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Kecen Li, Zhicong Huang, Xinwen Hou, and Cheng Hong. Gaussmarker: Robust dual-domain watermark for diffusion models. In *Forty-second International Conference on Machine Learning*, 2025.

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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs leave artificial fingerprints? In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 506–511, 2019.

Michel Meintz, Jan Dubiński, Franziska Boenisch, and Adam Dziedzic. Radioactive watermarks in diffusion and autoregressive image generative models. *arXiv preprint arXiv:2506.23731*, 2025.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Organika. SDXL detector. Hugging Face model repository, `https://huggingface.co/Organika/sdxl-detector`, 2023.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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. In *The Twelfth International Conference on Learning Representations*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021.

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

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 (CVPR)*, pp. 10684–10695, June 2022.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Yang Song and Stefano Ermon. Improved Techniques for Training Score-Based Generative Models. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 12438–12448, 2020.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Yu Tong, Zihao Pan, Shuai Yang, and Kaiyang Zhou. Training-free watermarking for autoregressive image generation. *arXiv preprint arXiv:2505.14673*, 2025.

Chao Wang, Zijin Yang, Yaofei Wang, Weiming Zhang, and Kejiang Chen. Aedr: Training-free ai-generated image attribution via autoencoder double-reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9675–9683, 2026.

Zhenting Wang, Chen Chen, Yi Zeng, Lingjuan Lyu, and Shiqing Ma. Where did i come from? origin attribution of AI-generated images. In *Advances in Neural Information Processing Systems*, 2023.

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? In *International Conference on Machine Learning*, pp. 51396–51414. PMLR, 2024.

Ziyi Wang, Songbai Tan, Gang Xu, Xuerui Qiu, Hongbin Xu, Xin Meng, Ming Li, and Fei Richard Yu. Safe-var: Safe visual autoregressive model for text-to-image generative watermarking. *arXiv preprint arXiv:2503.11324*, 2025.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. *Advances in Neural Information Processing Systems*, 36: 58047–58063, 2023.

Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. Fairness feedback loops: training on synthetic data amplifies bias. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2113–2147, 2024.

Andrew Chi-Chih Yao. Protocols for secure computations. In *23rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 160–164. IEEE, 1982.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.

Hongyao Yu, Yixiang Qiu, Yiheng Yang, Hao Fang, Tianqu Zhuang, Jiaxin Hong, Bin Chen, Hao Wu, and Shu-Tao Xia. Icas: Detecting training data from autoregressive image generative models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11209–11217, 2025a.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18431–18441, 2025b.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*, 2025c.

Shengfang Zhai, Huanran Chen, Yinpeng Dong, Jiajun Li, Qingni Shen, Yansong Gao, Hang Su, and Yang Liu. Membership inference on text-to-image diffusion models via conditional likelihood discrepancy. *Advances in Neural Information Processing Systems*, 37:74122–74146, 2024.

Yanran Zhang, Bingyao Yu, Yu Zheng, Wenzhao Zheng, Yueqi Duan, Lei Chen, Jie Zhou, and Jiwen Lu. D^3QE: Learning discrete distribution discrepancy-aware quantization error for autoregressive-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Bihe Zhao, Louis Kerner, Michel Meintz, Tameem Bakr, Franziska Boenisch, and Adam Dziedzic. Data provenance for image auto-regressive generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

A Algorithm for Unified Provenance

We present the detailed algorithm for our PRISM unified provenance for image generative models in Algorithm 1.

Algorithm 1 Unified Model Attribution Framework

Input: Image $\mathbf{x}$, suspect models $\{\mathcal{M}_i^j = (\mathcal{A}_i, \mathcal{G}_i^j)\}$, each with autoencoder $\mathcal{A}_i$ and the latent image generator $\mathcal{G}_i^j$, significance level α

Output: Attributed model $\mathcal{M}_{i^*}^{j^*}$ or NON-BELONGING

// Stage 1: Autoencoder Identification
$\mathcal{S}_{\text{AE}} \leftarrow \emptyset$
for each autoencoder $\mathcal{A}_i \in \{\mathcal{A}_1, \ldots, \mathcal{A}_m\}$ **do**
  Compute $\mathcal{L}_{\mathcal{A}}(\mathbf{x}; \mathcal{A}_i)$ via Equation (5)
  Compute Grubbs' statistic via Equation (8)
  if $\mathcal{H}_0^{(i)}$ is rejected **then**
    $\mathcal{S}_{\text{AE}} \leftarrow \mathcal{S}_{\text{AE}} \cup \{\mathcal{A}_i\}$
  end if
end for
if $\mathcal{S}_{\text{AE}} = \emptyset$ **then**
  return NON-BELONGING
end if
// Stage 2: Latent Image Generator Attribution
// candidates are all generators whose autoencoder passed Stage 1
for each $\mathcal{G}_i^j$ with $\mathcal{A}_i \in \mathcal{S}_{\text{AE}}$ **do**
  Compute $\mathcal{L}_{\mathcal{G}}(\mathbf{x}; \mathcal{G}_i^j)$ via Equation (6)
end for
$(i^*, j^*) \leftarrow \underset{i,\, j\,:\, \mathcal{A}_i \in \mathcal{S}_{\text{AE}}}{\arg\max}\ \mathcal{L}_{\mathcal{G}}(\mathbf{x}; \mathcal{G}_i^j)$
return $\mathcal{M}_{i^*}^{j^*}$

B Details of Evaluated Models

We present the details of evaluated models in Table A1, including the generative paradigms, conditioning type, autoencoder architecture, and model checkpoints. For example, RAR model family shares an autoencoder with the architecture of MaskGIT-VQGAN (Chang et al., 2022), where the potential latent generators are RAR-B, RAR-L, RAR-XL, and RAR-XXL.

Table A1: **Evaluated model families and checkpoints.**

Model Type	Model Family	Conditioning Type	Autoencoder	Model Checkpoints
IAR	RAR	Class-to-image	MaskGIT-VQGAN	RAR-B, RAR-L, RAR-XL, RAR-XXL
	VAR	Class-to-image	Multi-scale VQVAE	VAR-d16, VAR-d20, VAR-d24, VAR-d30
	LlamaGen (C2I)	Class-to-image	VQGAN-16384	LlamaGen-B, LlamaGen-L, LlamaGen-XL
	LlamaGen (T2I)	Text-to-image	VQGAN-16384	LlamaGen-XL-Stage1, LlamaGen-XL-Stage2
	Infinity	Text-to-image	Infinity-VAE	Infinity-2B
DM	Stable Diffusion 1	Text-to-image	SD1-VAE	Stable Diffusion 1.4, Stable Diffusion 1.5
	Stable Diffusion 2	Text-to-image	SD2-VAE	Stable Diffusion 2.0, Stable Diffusion 2.1
	Stable Diffusion 3.5	Text-to-image	SD3-VAE	Stable Diffusion 3.5
	Stable Diffusion XL	Text-to-image	SDXL-VAE	Stable Diffusion XL Base
	FLUX	Text-to-image	FLUX-VAE	FLUX.1-dev, FLUX.1-AW
	LightningDiT	Class-to-image	VA-VAE	LightningDiT-XL
	REPA	Class-to-image	SD-VAE	REPA SiT-XL

C Detailed Comparisons between Provenance Methods

We compare against six baselines spanning autoencoder and latent generator attribution. For autoencoder-based methods, we include AEROBLADE (Ricker et al., 2024), which uses recon-

struction error, and AEDR (Wang et al., 2026), which extends reconstruction-based attribution with double reconstruction and calibration. For generator-based methods, we include Delta (Kowalczuk et al., 2025; Zhai et al., 2024), which compares conditional vs. unconditional membership scores, and PRADA (Damm et al., 2025), which combines token-wise weighted membership signals. We also evaluate single-stage ablations of our own framework: autoencoder only (Equation (5)) and generator only (Equation (6)). We do not include LatentTracer (Wang et al., 2024) as AEDR supersedes it across all models, and LatentTracer additionally requires per-image gradient optimization, which is computationally prohibitive.

C.1 Detailed Comparison on Class-to-image Models

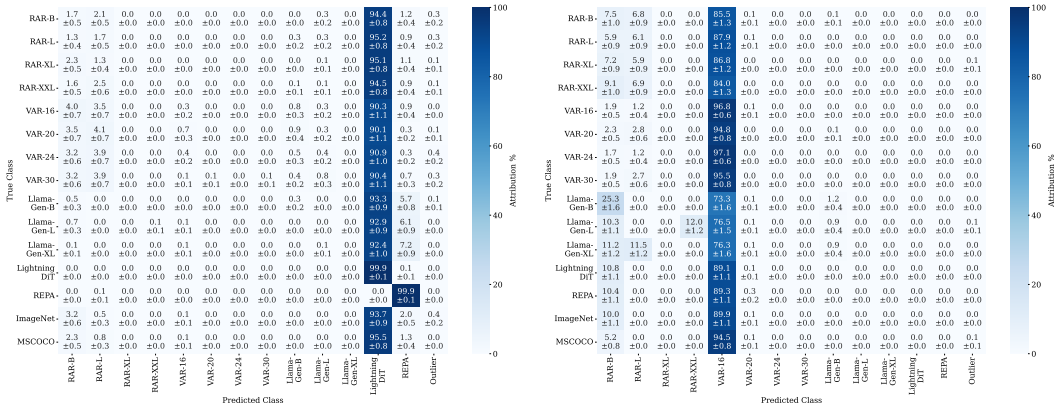
Figure A1 presents confusion matrices comparing six provenance approaches on class-to-image models. AEROBLADE (Figure A1a) collapses most predictions toward LightningDiT regardless of the true source, while AEDR (Figure A1b) skews predictions toward VAR-d16, and both achieve only marginal diagonal accuracy. The autoencoder-only method (Figure A1c) achieves near-perfect family-level discrimination across all five class-to-image families (RAR, VAR, LlamaGen, LightningDiT, REPA), yet fundamentally cannot distinguish model versions within a family that shares the same autoencoder. Conversely, the generator-only approach (Figure A1d) cannot perform reliable outlier rejection, collapsing almost every prediction onto a single candidate while the outlier route is essentially never triggered. Delta (Figure A1e) and PRADA (Figure A1f) show intermediate intra-family discrimination but maintain substantial confusion patterns, particularly disperse predictions for VAR variants across RAR and the diffusion-based LightningDiT/REPA columns, and exhibit non-negligible false attribution rates for natural images. These results empirically validate our two-stage framework design: autoencoder fingerprints provide robust family-level separation and outlier rejection, while latent generator fingerprints enable fine-grained version attribution within identified families.

C.2 Detailed Comparison on Text-to-image Models

Table A2 and Figure A2 present comprehensive comparisons across text-to-image models spanning both IARs (LlamaGen, Infinity) and DMs (Stable Diffusion variants and FLUX). Among autoencoder attribution methods, the autoencoder-only approach achieves the strongest performance on models with distinctive VAEs (66.3% Macro-F1) but cannot distinguish model versions within the same family. AEROBLADE (19.5% Macro-F1) and AEDR (50.1% Macro-F1) exhibit complementary failure modes: AEROBLADE collapses many predictions toward FLUX.1-dev, while AEDR scores well on a few targets (e.g., LlamaGen Stage1/Stage2, SD-3.5) but fails on others. The generator-only approach (26.7% Macro-F1) achieves moderate success on select DM variants but completely fails on LlamaGen Stage1 (3.7%) and on FLUX models (0.0%). Delta (4.0% Macro-F1) and PRADA (7.9% Macro-F1), originally designed for IARs, fail to generalize when extended to DMs and produce highly dispersed predictions across the FLUX and Stable Diffusion families. In contrast, our unified PRISM framework reaches 95.5% Macro-F1 across all paradigms, demonstrating that the complementary integration of autoencoder and latent generator fingerprints is essential for robust cross-paradigm provenance.

Table A2: **Comparison of different autoencoder and latent generator attribution methods for text-to-image models.**

Type	Method	LlamaGen		Infinity	Stable Diffusion						FLUX.1		Natural		Macro-F1
		Stage1	Stage2	2B	1.4	1.5	2.0	2.1	3.5	XL	Dev	AW	ImageNet	MSCOCO	
AE Attribution	AEROBLADE	0.0	0.0	0.9	42.3	0.2	50.9	0.0	50.2	78.1	26.4	4.2	0.0	0.2	19.5
	AEDR	95.7	96.7	47.6	65.9	0.2	64.9	0.6	97.1	59.5	61.9	0.0	42.7	18.5	50.1
	Autoencoder only	93.8	99.1	100.0	66.0	0.2	66.2	1.0	95.7	88.0	60.7	0.0	96.2	95.1	66.3
Gen Attribution	Generator only	3.7	0.0	51.3	59.7	44.4	42.4	20.8	95.9	28.6	0.0	0.0	0.0	0.0	26.7
	Delta	0.7	21.4	12.8	1.0	0.0	0.6	0.2	1.3	0.6	7.0	6.2	0.0	0.0	4.0
	PRADA	9.7	71.4	0.0	0.0	5.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.1	7.9
PRISM		98.1	96.3	98.0	95.8	91.5	97.1	94.1	92.3	96.5	94.7	95.4	96.2	95.1	95.5

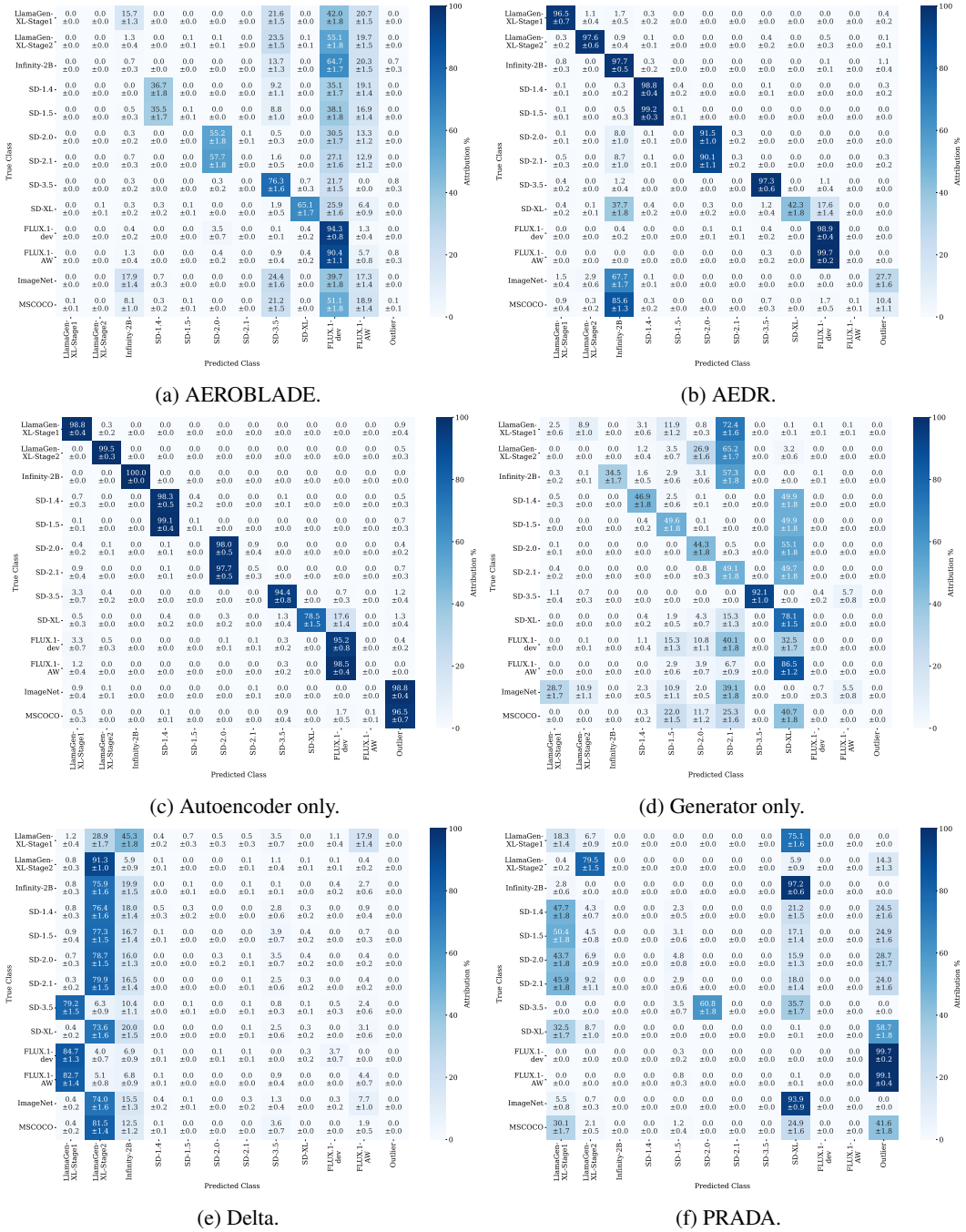


(a) AEROBLADE. (b) AEDR.

(c) Autoencoder only. (d) Generator only.

(e) Delta. (f) PRADA.

Figure A1: **Comparisons between different image provenance approaches on class-to-image models.**



(a) AEROBLADE. (b) AEDR.

(c) Autoencoder only. (d) Generator only.

(e) Delta. (f) PRADA.

Figure A2: **Comparison between image provenance approaches on text-to-image models.**

D Further Ablation Studies

D.1 Token-wise Score Weighting Strategies

PRADA (Damm et al., 2025) proposes token-wise score weighting based on the observation that attribution signals vary in magnitude across token positions. To investigate whether this strategy benefits our framework, we analyze the cumulative density functions of delta scores in Figure A3. When comparing VAR-d30 against VAR-d16 (Figure A3a), we observe substantial separation at larger delta values, indicating scenarios where PRADA's score mapping could potentially improve discrimination. However, when comparing VAR-d20 against VAR-d16 (Figure A3b), the distributions remain nearly identical across all delta values, suggesting that token-wise weighting provides no additional discriminative power. This analysis explains our empirical findings in Figure A5: uniform weighting across token positions yields more robust attribution than sophisticated weighting strategies, particularly when distinguishing between closely related model variants.

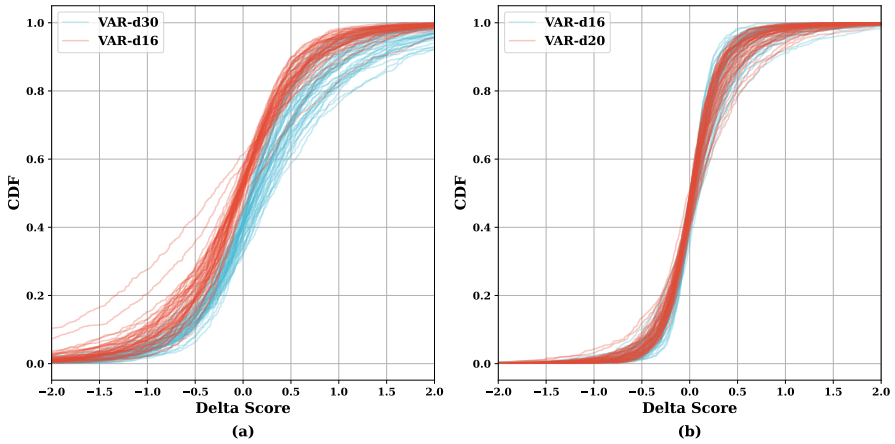


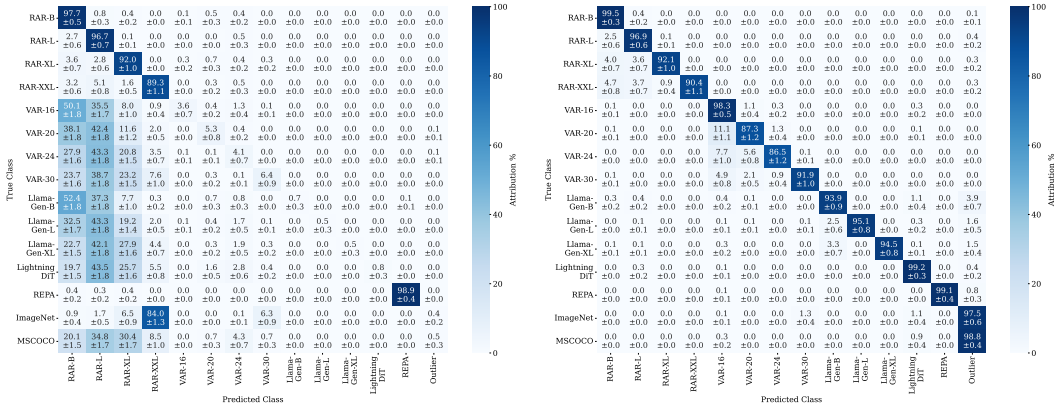
Figure A3: **Cumulative density function (CDF) for different delta values of different datasets.** We show the score distributions for ***(a)*** *VAR-d30 vs VAR-d16* $\mathcal{L}_{\mathcal{A}}$, and ***(b)*** *VAR-d16 vs VAR-d20* $\mathcal{L}_{\mathcal{G}}$. The evaluated model is VAR-d20. We plot the distributions for 50 data points in each dataset.

D.2 Comparison between Original and Finetuned Encoders

We observe that adapting the encoder to approximate the inverse of the decoder consistently improves attribution performance, as demonstrated in Figure A4. Specifically, we construct a set of synthetic feature maps $\mathbf{z}$ by sampling from the latent space (disjoint from the evaluation set) and optimize a modified encoder $\hat{\mathcal{E}}$ using the inverse loss $\mathcal{L}_{\text{Inv}} = \text{MSE}(\hat{\mathcal{E}} \circ \mathcal{D}(\mathbf{z}), \mathbf{z})$, where the decoder weights $\mathcal{D}$ remain frozen. This is a lightweight offline preprocessing step that operates entirely on synthetic latents produced by the frozen decoder and requires no real training data or any modification to the generative pipeline. At inference time, our method remains training-free: attribution is performed in a single forward pass through the adapted encoder without any optimization. This adaptation encourages the encoder to better recover the latent representation from the decoded image, leading to improved attribution accuracy.

D.3 Ablation Studies for Different Latent Generator Fingerprints

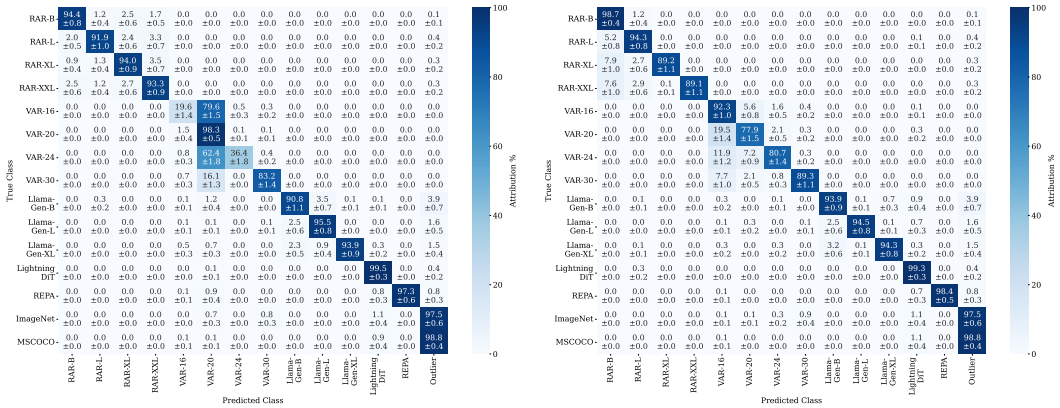
We compare three strategies for extracting attribution signals from autoregressive transformers: PRADA's token-wise weighted scores (Damm et al., 2025), conditional probability delta (Kowalczuk et al., 2025), and simple conditional probability (Ours), as shown in Figure A5. Interestingly, the simple approach of directly using conditional probabilities without weighting or delta computations achieves the best attribution performance. We hypothesize that this occurs because both PRADA and delta-based methods introduce additional variance through their weighting or differencing operations, which can amplify noise when the underlying attribution signal is already strong. In contrast, our approach leverages the raw conditional probabilities that naturally encode how well



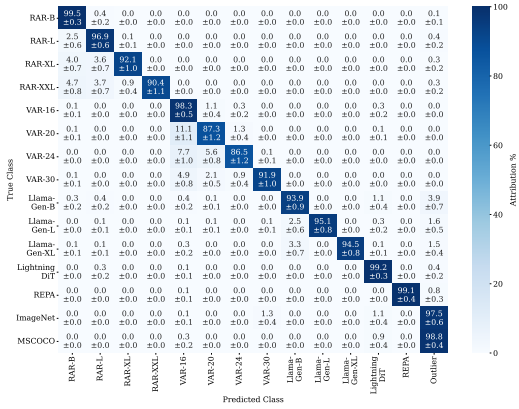
(a) Original encoder. (b) Finetuned encoder.

Figure A4: **Comparisons between the original encoder and finetuned encoders.**

each token fits the model's learned distribution, providing a more direct and stable fingerprint of the generating model.



(a) PRADA. (b) Delta.



(c) Generator only (Ours).

Figure A5: **Comparisons between different latent generator fingerprints when inserted into our framework.**

E ROBUSTNESS EVALUATIONS

We perform extensive robustness evaluations across five augmentation types at several strengths each (Table A3). The evaluated types of transforms include JPEG compression, resizing, Gaussian blurring, applying Gaussian noise, and saturation change. The evaluated models are the RAR family with the augmentation-finetuned encoders. Notably, our approach is robust to a wide range of transformations and strengths.

Table A3: **Robustness to post-processing, by augmentation type and strength.** We report F1 per model version with the augmentation-finetuned encoder.

Augmentation	Strength	RAR-B	RAR-L	RAR-XL	RAR-XXL	Macro-F1
None	—	94.5	95.0	95.5	94.7	96.2
JPEG	q=90	93.5	92.1	94.3	92.6	93.4
	q=80	91.6	90.9	92.7	90.2	90.4
	q=70	89.0	85.2	90.3	88.1	86.8
	q=60	86.5	82.2	86.6	82.7	81.9
Resize	$\times$0.9	94.5	94.7	96.4	94.6	96.2
	$\times$0.8	93.5	93.2	95.2	93.3	95.2
	$\times$0.7	93.6	94.0	95.6	92.9	94.9
	$\times$0.6	91.9	91.0	94.5	90.5	92.7
	$\times$0.5	86.5	83.7	88.3	84.6	84.4
Gaussian blur	r=0.5	94.8	94.9	96.0	94.7	96.4
	r=1.0	91.6	92.2	95.1	91.8	93.0
Gaussian noise	std=1	94.1	93.9	95.3	93.6	95.3
	std=2	93.9	93.1	94.6	92.9	94.2
	std=3	92.0	92.0	93.3	91.1	92.2
	std=4	92.0	90.7	92.8	89.7	90.5
	std=5	90.1	87.9	90.1	88.4	87.4
Saturation	$\times$1.1	94.5	94.1	95.3	94.4	95.5
	$\times$1.2	93.1	93.7	94.4	93.4	93.6
	$\times$1.3	91.5	91.2	93.1	92.2	91.1
	$\times$1.4	89.2	88.4	90.5	89.5	87.3
	$\times$1.5	85.7	85.1	87.0	86.0	82.9

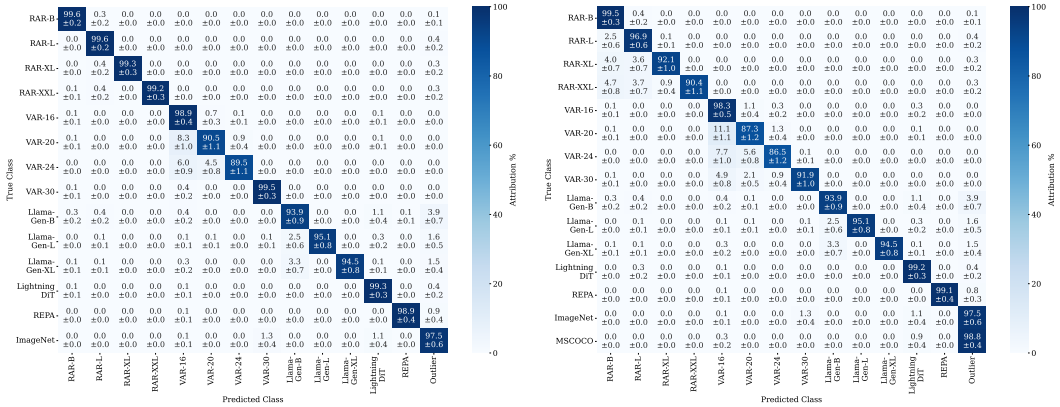
F EVALUATION FOR CONDITIONING ESTIMATION

F.1 LABEL ESTIMATION FOR CLASS-TO-IMAGE MODELS

We compare ground truth labels and EVA02-predicted labels for class-to-image models in Figure A6. As shown in Figures A6a and A6b, using EVA02-predicted labels instead of ground truth labels results in minimal performance degradation across most model families. These results demonstrate that our framework effectively handles the practical scenario where ground truth class labels are unavailable and must be estimated from the image content.

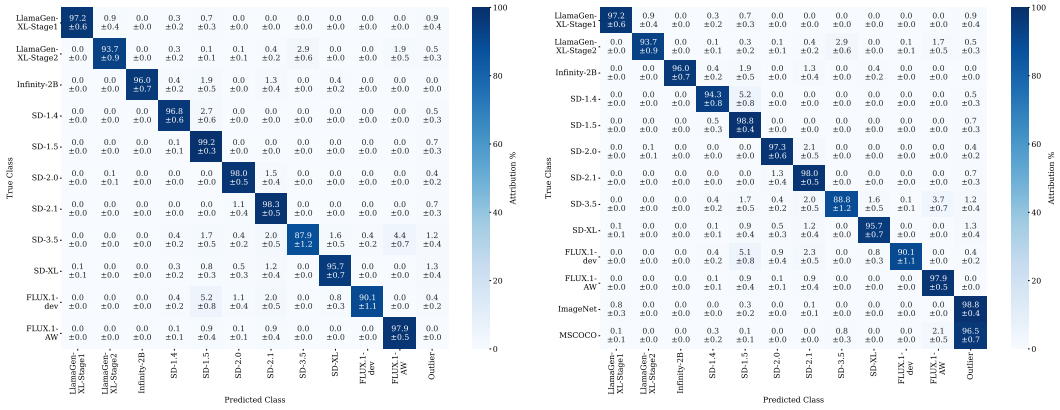
F.2 PROMPT ESTIMATION FOR TEXT-TO-IMAGE MODELS

We compare ground truth prompts and predicted prompts for text-to-image models in Figure A7, evaluating captions generated by BLIP-2 (Li et al., 2023b) and LLaVA (Liu et al., 2023). As illustrated in Figures A7a to A7c, both caption estimation methods achieve attribution accuracy comparable to using ground truth prompts. These results confirm that our unified provenance framework is resilient to caption estimation quality, enabling reliable attribution in real-world scenarios where original generation prompts are unknown.

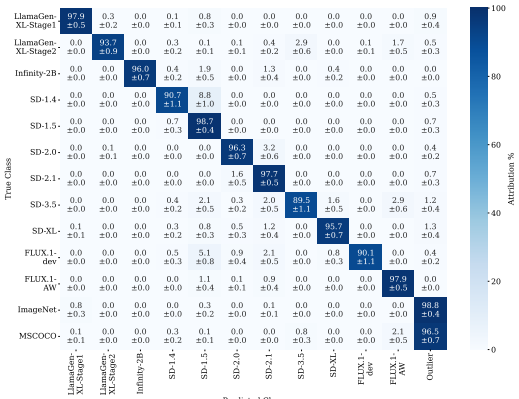


(a) Groundtruth labels. (b) EVA02-generated labels.

Figure A6: **Comparisons between the ground truth labels and predicted labels for class-to-image models.**



(a) Ground-truth captions. (b) BLIP2 captions.



(c) LLaVA captions.

Figure A7: **Comparisons between the ground truth captions and predicted captions for text-to-image models.**

F.3 Examples for Prompt Recovery

We note that the ground-truth, BLIP-2-generated, and LLaVA-generated captions have very different styles and lengths, but our attribution accuracy remains nearly identical. For example, consider the following captions for the same image:

Example Captions for the Same Image

Original: "two female sitting at a table across from each other eating pizza"

BLIP-2: "a group of young women eating pizza in a restaurant"

LLaVA: "The image depicts two individuals engaged in a shared activity: eating pizza. They are seated opposite each other at what appears to be an indoor dining table covered with a white tablecloth. Both of them have black hair, which stands out against their attire—one is wearing a blue top while the other has on a gray one. Their faces show expressions of enjoyment as they savor their slices of pizza. The pizza itself is prominently displayed on a wooden tray that rests on the table between them."

Therefore, our signal depends on overall semantic content rather than exact phrasing, enabling reliable provenance when original prompts are unavailable.

G Open-World Attribution

PRISM performs well in the open-world setting, where query images may originate from models outside the candidate pool. We note that open-world attribution is fundamentally a model family identification problem, which our Stage 1 directly addresses. Specifically, if an image does not belong to any of the enrolled model families, our autoencoder identification stage marks it as NON-BELONGING, thereby handling open-world cases without any architectural modification. In contrast, flat-attribution approaches such as PRADA cannot reliably reject images that do not belong to any known candidate model.

To validate this, we construct a candidate pool containing only the IAR class-to-image model families (RAR, VAR, LlamaGen) and test on images generated by nine out-of-pool sources: natural images from ImageNet and MSCOCO, and synthetic images from five Stable Diffusion variants, SD-3 Medium, and FLUX.1-dev. An ideal method should classify all out-of-pool images as NON-BELONGING. As shown in Table A4, PRISM achieves above 92% rejection accuracy across all out-of-pool sources, while PRADA misclassifies the vast majority of them into one of the enrolled models.

Table A4: **Open-world rejection accuracy (%) on out-of-pool sources.** The candidate pool contains only IAR class-to-image models (RAR, VAR, LlamaGen). Higher accuracy is better, as the method should classify every out-of-pool image as NON-BELONGING.

Method	ImageNet	MSCOCO	SD-1.4	SD-1.5	SD-2.0	SD-2.1	SD-3-Med.	SD-XL-Base	FLUX.1-dev	Macro-F1
PRADA	1.7	54.2	50.2	49.0	50.8	50.8	46.1	52.9	50.5	46.94
PRISM	**95.4**	**96.0**	**95.6**	**95.5**	**95.3**	**94.8**	**94.2**	**95.1**	**93.7**	**94.75**

H Computational Overhead Analysis

Our hierarchical design reduces overhead: Stage 1 narrows candidates to a specific model family, then Stage 2 evaluates only generators within that family, reducing the number of model instantiations from multiplicative in the pool size to additive. Table A5 shows that our method achieves 2–3.5$\times$ speedup over PRADA, with the advantage growing as model count increases.

Theoretical complexity. Consider an attribution pool with m model families, where each family shares a single autoencoder and contains n generator versions, so the pool covers $m \cdot n$ candidate models in total. Let C_{AE} denote the per-image cost of one autoencoder reconstruction (used to compute $\mathcal{L}_{\mathcal{A}}$) and C_{LG} the per-image cost of one latent generator forward pass (used to compute $\mathcal{L}_{\mathcal{G}}$). A flat, latent-generator-only attribution baseline must, for each query image, instantiate every

Table A5: **Average inference time (seconds per image) comparison with PRADA.**

# Models	PRADA	PRISM
4 (RARs only)	0.350	0.159
8 (RARs + VARs)	1.019	0.291
11 (RARs + VARs + LlamaGen)	1.367	0.387

candidate model: it pairs each of the m autoencoders with each of the n generators in that family, yielding

$$T_{\text{flat}} \;=\; \mathcal{O}(m \cdot n) \cdot \big(C_{\text{AE}} + C_{\text{LG}}\big). \tag{11}$$

In contrast, our hierarchical PRISM decouples the two stages. Stage 1 evaluates the m autoencoder fingerprints once to identify (or reject) the model family, and Stage 2 only evaluates the n latent generators inside the selected family:

$$T_{\text{PRISM}} \;=\; \mathcal{O}(m) \cdot C_{\text{AE}} \;+\; \mathcal{O}(n) \cdot C_{\text{LG}}. \tag{12}$$

This converts a multiplicative $m \cdot n$ cost into an additive $m + n$ cost, and avoids running any latent generator outside the identified family. The speedup grows with both the number of families and the number of versions per family, which is consistent with the trend observed in Table A5.

I Analysis of Grubbs' Test

I.1 Sensitivity to Significance Level and Population Size

Our method is not sensitive to the hyperparameters of Grubbs' test. As shown in Table A6a, varying the significance level α across more than an order of magnitude changes F1 by only 0.6 points (we use $\alpha = 0.01$). Similarly, Table A6b shows that Macro-F1 remains stable across a wide range of population sizes N, varying by only 1.2 points (we use $N = 250$).

Table A6: **Sensitivity of attribution performance to Grubbs' test hyperparameters:** (a) significance level α and (b) population size N.

(a) Significance level α.

α	0.03	0.02	0.01	0.005	0.001
Macro-F1	93.8	94.0	94.1	94.3	94.4

(b) Population size N.

N	100	200	250	300	400
Macro-F1	93.2	94.4	94.1	94.0	93.7

I.2 Gaussianity of Reconstruction Scores

Grubbs' test assumes the underlying population follows a Gaussian distribution. To verify this assumption, we compute the skewness of the autoencoder reconstruction score distributions for representative model pairs. An absolute skewness below 1 indicates no significant deviation from Gaussianity (Bulmer, 2012). As shown in Table A7, all but one configuration exhibit $|\text{skewness}| < 1$, confirming that the Gaussian assumption underlying Grubbs' test is well justified in practice.

Table A7: **Skewness of autoencoder reconstruction score distributions.** $|\text{Skewness}| < 1$ indicates no significant deviation from Gaussianity.

Tested AE	RAR-B	VAR-d30	LlamaGen-L
RAR-B	0.9725	0.7195	1.0132
VAR-d30	0.5449	0.2348	0.6232
LlamaGen-L	−0.0232	−0.1417	−0.1733

J PRIVACY-PRESERVING COLLABORATIVE AUDITING

In practice, model owners may wish to participate in attribution without exposing their model weights or internal fingerprint scores to other parties. The two-stage design of PRISM naturally supports a *federated attribution protocol* in which each model owner performs the full attribution decision locally and reveals only the final categorical outcome.

Protocol. Consider a central coordinator (e.g., a platform or auditor) who receives a query image $\mathbf{x}$ and wishes to determine its source model. The coordinator distributes $\mathbf{x}$ to all participating model owners $\{O_1, \ldots, O_m\}$. Each owner O_i may hold *multiple* autoencoders $\{\mathcal{A}_{i,k}\}_{k=1}^{K_i}$, for example a provider who releases several model families each built on a different autoencoder architecture. Each autoencoder $\mathcal{A}_{i,k}$ is paired with its own set of latent generators (model versions) $\{\mathcal{G}_{i,k}^j\}_{j=1}^{n_{ik}}$ and a private reference population for Grubbs' test. The protocol proceeds as follows:

1. The coordinator sends $\mathbf{x}$ to every model owner.
2. Each owner O_i locally iterates over their autoencoders. For each $\mathcal{A}_{i,k}$, they compute the autoencoder fingerprint $\mathcal{L}_{\mathcal{A}}(\mathbf{x}; \mathcal{A}_{i,k})$ and apply Grubbs' test (Equation (8)) against the corresponding private reference distribution.
3. If none of the owner's autoencoders reject their respective $\mathcal{H}_0$ (i.e., the image is unlikely to originate from any autoencoder held by O_i), the owner returns $(\texttt{null}, \texttt{null})$ to the coordinator.
4. For each autoencoder $\mathcal{A}_{i,k}$ whose $\mathcal{H}_0^{(i,k)}$ is rejected, the owner proceeds to Stage 2: they compute $\mathcal{L}_{\mathcal{G}}(\mathbf{x}; \mathcal{G}_{i,k}^j)$ for every associated generator and identify the best match $j_k^* = \arg\max_j \mathcal{L}_{\mathcal{G}}(\mathbf{x}; \mathcal{G}_{i,k}^j)$. Among all passing autoencoders, the owner selects the $(k^*, j_{k^*}^*)$ pair with the highest Stage 2 score and returns $(\mathcal{A}_{i,k^*}, \mathcal{G}_{i,k^*}^{j_{k^*}^*})$ to the coordinator.

At the end of the protocol, the coordinator collects all responses. If exactly one owner returns a non-null result, the attribution is complete. If all owners return null, the image is classified as NON-BELONGING.

Privacy Properties. Under this protocol, model weights never leave the owner's premises. Crucially, no numerical fingerprint scores ($\mathcal{L}_{\mathcal{A}}$ or $\mathcal{L}_{\mathcal{G}}$) are communicated, and each owner reveals only a categorical decision. This prevents other parties from inferring properties of the model's latent space or generation distribution from the fingerprint magnitudes. The private reference distributions used for Grubbs' test also remain local, so calibration statistics (μ_{ik}, σ_{ik}) are never shared. The number of autoencoders each owner holds (K_i) is also not disclosed to the coordinator.

Conflict Resolution. When multiple owners claim the image, the conflict may arise from closely related autoencoders (e.g., two providers whose models share a similar VAE architecture). In this case, a lightweight tiebreaker can be employed: only the conflicting owners engage in a secure pairwise comparison of their Stage 2 scores, for example via a standard secure comparison protocol (Yao, 1982), so that only the identity of the higher-scoring model is revealed, without exposing the raw $\mathcal{L}_{\mathcal{G}}$ values to either party. Note that within a single owner, conflicts across their own autoencoders are resolved locally in Step 4 and never reach the coordinator.

K COMPARISON TO GENERATED IMAGE DETECTION

Generated image detection is a task related to our image attribution task, but with different and orthogonal goals. We provide qualitative and quantitative comparisons between the two tasks in this section.

K.1 QUALITATIVE COMPARISONS

Generated image detection and our targeted image attribution are *different but complementary* tasks. PRISM addresses *enrolled-set* attribution: given an image and a registry of models whose weights the verifier (model owner or auditor) can run, it traces the image to its source model or returns NON-BELONGING. Our task is not a universal synthetic-image detector, and it is complementary to one: the two tasks answer different questions over different label spaces, and the reject branch of our method is exactly where a detector composes. Synthetic-image detector maps an image to

$\{\text{real}, \text{synthetic}\}$ and our method maps it to $\{m_1, \ldots, m_k, \text{NON-BELONGING}\}$. The two tasks can be composed to achieve a stronger goal rather than compete with each other. PRISM can trace an image to its source model accurately, and a detector applied to the rejected non-belonging model can further separate real images from images generated by unknown models. The composition order matters. Moreover, a detector's errors concentrate on generators outside its training distribution, and therefore needs to be sufficiently retrained with existing datasets whenever a new model appears.

K.2 Quantitative Comparisons

Evaluated detectors. We compare against three state-of-the-art detectors listed in Table A8. Organika (Organika, 2023) is an open-source Swin-B model finetuned on Wikimedia images paired with SDXL generations. UniversalFakeDetect (Ojha et al., 2023) is a frozen CLIP ViT-L/14 with a linear real-versus-fake head trained on ProGAN outputs. Community Forensics (Park & Owens, 2025) is a ViT-S/16 trained on 2.7M images from $\sim$4800 generators, the widest-coverage open detector we are aware of and, by a large margin, the strongest of the three.

Evaluation setup. We use 22 generators across 11 families together with the two real datasets, at 1000 images per source. We evaluate each detector in the two different settings as follows.

Setting 1: Binary real-versus-fake detection. Here we use the detector for the task it was designed for, namely whether an image is real or generated. We choose the decision threshold by allowing a 5% false positive rate, which fixes the threshold at the 95th percentile of the scores on the 2000 pooled real images from ImageNet and MS-COCO. We then report, for each generator, the fraction of its 1000 images that the detector catches. Nothing is trained on our data in this setting. Specifically, almost every published detector is trained on SD-1.x and SD-2.x but likely not the other nine families. We therefore report the two groups separately.

Setting 2: Attributing to the source model. Here we evaluate a general detector for our question: does a general detector's internal representation already contain the identity of the generator? We freeze the network, take the features from its second-last layer, and train an initialized logistic regression on top of them. The layer is trained using our labels on the first 250 images per source and evaluated on the remaining 750, which is the split PRISM calibrates on. Real images are excluded, so this classifier only has to tell generators apart. At version granularity, each of the 22 generators is its own class, and at family granularity the versions of a family are all labeled into one of 11 family classes. Note that the detector was never trained for attribution, so we design this adaptation for the missing training ourselves, which gives an extra advantage to the original detectors. The attribution setting is therefore close to an upper bound on what these detectors could contribute to attribution.

Detector performance on the two settings. PRISM attributes an image to its exact model version at 95.0 Macro-F1 and selects the correct family for 99.3% of images. None of the evaluated detectors can achieve a close level of attribution performance, even when granted a supervised probe on its own features. The strongest of the three detectors reaches only 48.5 Macro-F1 at version granularity, and the weaker two fall to 34.9 and 33.9. The detectors perform better on the binary detection they were built for, but mostly when they are trained on an enormous amount of images and generators. Community Forensics flags 99.8% of images from unseen families only because it was trained on $\sim$4800 of them and 2.7M images, whereas UniversalFakeDetect flags 5.7% of FLUX.1-Dev and Organika 2.7% of VAR-d16. Every detector therefore inherits a coverage requirement that grows with each model release, while PRISM calibrates Stage 1 on each enrolled model's own outputs and needs no corpus of other generators at all. Therefore, the best strategy will be to combine the two tasks to achieve the best overall attribution results.

L Comparison to Classification Baselines

We further compare PRISM against four classification baselines and provide a the results as follows.

L.1 Experimental Setup

Every baseline is trained exactly on the data PRISM calibrates on, namely the same 250 images per source, and is evaluated on the same held-out 750, giving 15 class-to-image and 13 text-to-image classes. We fine-tune an ImageNet-pretrained ResNet-50 (He et al., 2016) end-to-end, and we train

Table A8: **Three released detectors under the same protocol.** We evaluate two settings: *1) Real-versus-fake detection:* measured at a threshold calibrated to 5% false positives on real images, reported for the families plausibly seen in training (SD-1 and SD-2) and for the nine remaining families. *2) Source attribution:* We perform multinomial logistic regressions on frozen detector features, trained on the same 250 images per source that PRISM calibrates on and evaluated on the held-out 750, over the identical 22 sources at both granularities.

Detector	Backbone	Training data	Real-vs-fake detection accurary (%)		Source attribution macro-F1	
			SD-1/2	Other 9 families	Family	Version
Organika (Organika, 2023)	Swin-B	Wikimedia vs. SDXL pairs	31.0	21.4	60.3	33.9
UniversalFakeDetect (Ojha et al., 2023)	CLIP ViT-L/14	ProGAN, LSUN 20-class	35.8	55.6	63.3	34.9
Community Forensics (Park & Owens, 2025)	ViT-S/16	**2.7M** imgs, ∼**4800** gens	100.0	99.8	85.6	48.5
PRISM	-	-		*Not its task*	**99.3**	**95.0**

Table A9: **Classification baselines under one protocol**, all trained on the same 250 images per source that PRISM calibrates on and evaluated on the same held-out 750, over the identical sources at both granularities.

Method	Version Macro-F1		Family Macro-F1		Reject
	c2i	t2i	c2i	t2i	option
ResNet-50 (He et al., 2016)	20.9	59.5	47.8	85.2	✗
Marra et al. (2019)	7.1	18.1	18.7	41.9	✗
CLIP ViT-L/14 (Radford et al., 2021)	24.3	64.7	48.8	88.6	✗
DINOv2 ViT-B/14 (Oquab et al., 2024)	16.9	45.8	36.9	69.6	✗
PRISM	**94.5**	**95.5**	**99.2**	**99.4**	✓

linear probes on two frozen foundation backbones, DINOv2 ViT-B/14 (Oquab et al., 2024) and CLIP ViT-L/14 (Radford et al., 2021), each with its own published preprocessing. We also evaluate the hand-crafted PRNU-style fingerprint of Marra et al. (2019), using the per-class mean noise residual and maximum correlation.

L.2 Comparison Results

Table A9 reports the results. The fine-tuned ResNet-50 reaches 20.9 Macro-F1 on c2i and 59.5 on t2i, the CLIP probe 24.3 and 64.7, the DINOv2 probe 16.9 and 45.8, and the fingerprint introduced by Marra et al. (2019) 7.1 and 18.1. In contrast, our method achieves 94.5 and 95.5. Notably, every baseline is relatively better at attributing to a family than the version. Coarse provenance is therefore partially possible with the classification backbones, but the classifiers cannot identify versions that share a decoder, which is exactly the granularity our method targets. In addition, the non-belonging rejection mechanism is not provided by these backbones, whereas PRISM can effectively reject images not from a given suspect model pool.

M Unseen Model Versions Rejection Within a Model Family

By default, Stage 2 assumes that all possible model versions are known, as is typical for model owners auditing proprietary models. It therefore always returns a version. In certain applications for open-source models, users may finetune private model versions from an open-source VAE without sharing their checkpoints with the auditor. For this case, we extend Stage 2 with an optional rejection test that can reject an image from an unseen version. The test uses the winner's margin defined as follows:

$$\Delta(\mathbf{x}) = \mathcal{L}_{\mathcal{G}}(\mathbf{x}; \mathcal{G}^{j^*}) - \max_{j \neq j^*} \mathcal{L}_{\mathcal{G}}(\mathbf{x}; \mathcal{G}^{j}). \tag{13}$$

The margin is the gap between the score of the best-matching version and that of the runner-up. It flags an image as *family known, version unknown* when $\Delta(\mathbf{x})$ is below the q-quantile of the margins on $\mathcal{G}^{j^*}$'s own calibration images.

Table A10: **Unseen versions within an enrolled family.** We first report LlamaGen as the IAR setting and then the FLUX derivatives as the diffusion setting. We remove one version from the candidate set and query with its images. Under *No rejection mechanism*, the table reports the enrolled sibling selected by Stage 2 and the percentage of images assigned to it. Under *With proposed rejection mechanism*, it reports the percentage for which the margin test returns *family known, version unknown* at q=1%. The cost on correctly attributed enrolled images is 1.9% for LlamaGen and 1.3% for FLUX.

Held-out version	No rejection mechanism		With proposed rejection mechanism
	Mostly labelled model	Labelled rate	True rejection rate
LlamaGen IARs, three enrolled versions			
LlamaGen-B	LlamaGen-L	97.9	98.3
LlamaGen-L	LlamaGen-XL	98.8	99.3
LlamaGen-XL	LlamaGen-L	100.0	71.1
FLUX diffusion derivatives, seven enrolled, identical VAE			
FLUX.1-Dev	MidJourney-LoRA	66.0	100.0
AWPortrait-LoRA	AWPortrait-Full	99.2	100.0
AWPortrait-Full	AWPortrait-LoRA	98.9	99.9
KreaDev	AWPortrait-Full	83.2	100.0
HyperSD-LoRA	MidJourney-LoRA	37.9	100.0
MidJourney-LoRA	AWPortrait-Full	68.1	100.0
schnell	AWPortrait-Full	66.0	100.0

We evaluate this test in two settings chosen to cover the two dominant generation paradigms. LlamaGen represents image autoregressive models with three model versions. The FLUX.1-Dev together with six community derivatives represent diffusion models. They share the *identical* FLUX VAE, so Stage 1 cannot separate them even in principle, and Stage 2 makes the whole decision. The derivatives span distillation, full fine-tuning, and LoRA adaptation. In each setting, we remove one version from the candidate set and query the system with its 750 held-out images. This setup represents an auditor encountering a community fine-tune of an enrolled model.

Table A10 shows that the rejection test already works well for LlamaGen. Without rejection, every held-out version is assigned to an enrolled sibling. LlamaGen-XL is assigned to LlamaGen-L in 100% of cases. With rejection at q=1%, the test rejects LlamaGen-B at 98.3%, LlamaGen-L at 99.3% and LlamaGen-XL at 71.1%. The cost on enrolled images is only 1.9%. For the FLUX derivatives, rejection is near perfect even though their identical VAE makes Stage 1 uninformative. Without rejection, every held-out version is assigned to an enrolled sibling. AWPortrait-LoRA and AWPortrait-Full claim each other about 99% of the time because they are the same adaptation at two merge strengths. At q=1%, six derivatives are rejected 100.0% of the time, and AWPortrait-Full is rejected 99.9% of the time. The cost on enrolled images is only 1.3%. Every derivative stays at or above 99.9% at every operating point from q=1% through q=30%, so the threshold need not be tuned to the query distribution.

N Clarification of Main Contributions

PRISM introduces the first framework for tracing a generated image to its exact source *model version* across both dominant generation paradigms, with the option to reject an image not from the suspect model pool. Existing post-hoc methods identify either a model family but not a version, or a version within only one paradigm. We formulate this task and design a two-stage framework that separates family attribution from version attribution. This task and framework are the central contribution we claim.

In addition to the core contribution, we design new components wherever the literature provides none. Diffusion models had no version-level score, so we derive the probability estimation using ELBO, enabling paradigm-agnostic version attribution. We also introduce version-level rejection, allowing the system to answer "family known, version unknown" instead of silently assigning an image to an enrolled sibling. Both components are new and necessary for the framework.

Where existing components already work well, we instantiate them in our framework. Replacing solved components would not strengthen the framework contribution. Stage 1 therefore uses reconstruction-based signals for diffusion autoencoders and quantization-based signals for autoregressive tokenizers.

Notably, we do not merely adopt the IAR version score, but *improve* it for our framework. Prior version-level scores require carefully engineered calibration mechanisms (Kowalczuk et al., 2025; Yu et al., 2025a; Damm et al., 2025), but we show that this machinery is not only unnecessary, but even harms the performance, which is a new finding by itself. Specifically, we find that estimating the conditioning with an off-the-shelf captioner and using the plain conditional probability already outperforms those calibrated scores by a wide margin, as Table 1 shows. This simpler signal is stronger and allows one Stage 2 to serve both paradigms.

Table A11 identifies the origin of every component and distinguishes what we propose, improve and instantiate.

Table A11: **Origin of each component of PRISM.** We *propose* the framework and components that did not exist. We *improve* the component to achieve better results in our framework. We *instantiate* components that already work well.

Component	Prior counterpart	Our role
Two-stage task and framework	None	**Proposed**
Version score for DMs (conditional ELBO)	None	**Proposed**
Version-level rejection (margin test)	None	**Proposed**
Version score for IARs (likelihood)	ICAS (Yu et al., 2025a), PRADA (Damm et al., 2025)	**Improved**
Double-reconstruction ratio ρ_{Rec}	AEDR (Wang et al., 2026)	Instantiated
Quantization error $\mathcal{L}_{\text{Q}}$ for IARs	D^3QE (Zhang et al., 2025)	Instantiated
$\rho_{\text{Rec}} \cdot \mathcal{L}_{\text{Q}}$ and encoder finetuning	Zhao et al. (2026)	Instantiated
Grubbs test	RONAN (Wang et al., 2023), LatentTracer (Wang et al., 2024)	Instantiated

O Environment

We conducted all experiments on an Ubuntu 22.04 system equipped with an Intel Xeon Gold 6330 CPU and an NVIDIA A40 GPU (48 GB memory), using CUDA 12.5, Python 3.12, and PyTorch 2.9.0.

P Evaluation on Additional Model Versions

We further extend the text-to-image evaluation substantially by adding five diverse FLUX derivatives: FLUX-Schnell, FLUX-KreaDev, FLUX-HyperSD-LoRA, FLUX-MidJourney-LoRA, and FLUX-AWPortrait-Full. The result is shown in Figure A8. These additions span different forms of model adaptation, including distillation, full-model fine-tuning, and LoRA-based adaptation, rather than merely neighboring official checkpoints. Together with FLUX.1-dev and FLUX.1-AW, they form a challenging set of seven FLUX versions that share the same FLUX VAE. Consequently, Stage 1 can identify only their common model family, and fine-grained separation must be provided by the latent-generator fingerprints in Stage 2.

Our method attributes each additional FLUX version with 97.3–99.7% accuracy, averaging 98.7%, while retaining 98.7% and 99.6% accuracy on FLUX.1-dev and FLUX.1-AW, respectively. Confusion among the closely related FLUX derivatives and with other text-to-image model families remains negligible. Meanwhile, ImageNet and MS-COCO images are correctly rejected as outliers with 98.8% and 94.7% accuracy. These results demonstrate that the framework scales to a substantially denser within-family candidate set and remains effective across heterogeneous model-adaptation mechanisms, supporting fine-grained attribution of community model versions even when they share an identical autoencoder.

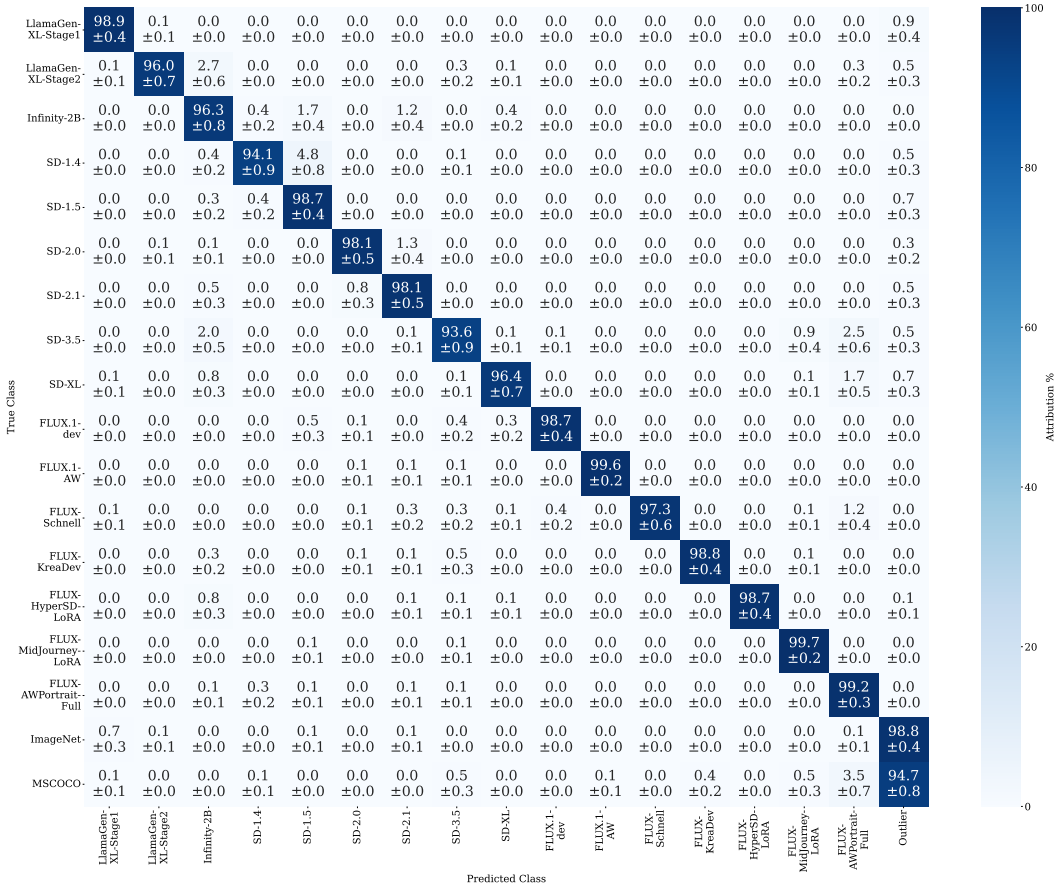


Figure A8: **Attribution on the expanded text-to-image model attribution with additional FLUX versions.** We extend the original 11-version text-to-image pool with five FLUX derivatives: FLUX-Schnell, FLUX-KreaDev, FLUX-HyperSD-LoRA, FLUX-MidJourney-LoRA, and FLUX-AWPortrait-Full. Together with FLUX.1-dev and FLUX.1-AW, the seven evaluated FLUX versions cover diverse model modifications, including distillation, full-model fine-tuning, and LoRA adaptation, while sharing the same FLUX VAE. The row-normalized confusion matrix shows 97.3–99.7% attribution accuracy on the five additional versions and reliable rejection of natural images as outliers.

Table A12: **Capabilities of post-hoc image attribution methods.** Only PRISM supports attribution to both autoencoder families and model versions across DM and IAR paradigms.

Key Capability		AEROBLADE	LatentTracer	AEDR	PRADA	**Ours (PRISM)**
Attribution Level	Model Family	✓	✓	✓	✗	✓
	Model Version	✗	✗	✗	✓	✓
Supported Paradigm	Diffusion (DM)	✓	✓	✓	✗	✓
	Autoregressive (IAR)	✗	✗	✗	✓	✓