LEARNING GENERALIZABLE MODEL FINGERPRINTS FOR ATTRIBUTION OF UNSEEN GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we propose a synthetic image attribution method based on generative model fingerprints that requires no synthetic images from real-world generative models during training. Instead, we train the fingerprint extractor entirely on a large and controllable bank of reconstruction models spanning diverse architectures and independently perturbed parameterizations, using the fingerprints embedded in their reconstructed images as supervisory signals. The proposed framework learns content-invariant fingerprints that distinguish not only different architectures but also different parameterizations of the same architecture, enabling direct attribution to previously unseen generative models. We further adopt an agent-guided collaborative training strategy to adaptively coordinate sampling, local evidence selection, loss configuration, frequency scheduling, and model-bank expansion during training. We evaluate the learned fingerprints on 26 unseen source models collected from four public benchmarks. Our method achieves a macro-average verification AUC of 97.01%, outperforming the strongest baseline by 5.46 percentage points, and shows clear separation between intra-source and inter-source fingerprints. These results demonstrate that fingerprints learned solely from controlled reconstruction models can generalize effectively to diverse unseen generative models.

1 INTRODUCTION

Generative models such as DALL-E 2 Ramesh et al. (2022) and Midjourney Midjourney, Inc. (2023) can now produce highly realistic images at unprecedented scale, raising growing concerns over the provenance and misuse of synthetic visual content. Beyond detecting whether an image is generated, synthetic image attribution aims to identify the model responsible for its creation, with important applications in intellectual property protection Fei et al. (2025a;c) and digital forensics Dai et al. (2024; 2025).

Existing attribution methods can be broadly divided into proactive and passive approaches. Proactive methods modify the generation process, for example by embedding watermarks into generated images Fei et al. (2022); Fernandez et al. (2023); Fei et al. (2025b), whereas passive methods infer the source model directly from the image Yang et al. (2022a); Abady et al. (2024). A major challenge for passive attribution is generalization to unseen generative models. Most existing methods rely on synthetic images produced by a predefined collection of generators for training, and their attribution capability is therefore closely tied to the models and image distributions available during training. As new generators, architectures, and parameterizations continue to emerge, repeatedly collecting synthetic images and retraining the attribution model becomes increasingly difficult to scale. Enlarging the simulated open space can improve coverage Yang et al. (2023b), but still depends on predefined training sources and schedules.

We address this limitation by learning generative model fingerprints without using synthetic images from real-world generative models during training. Instead, we construct a large and controllable bank of reconstruction models spanning diverse architectures and independently perturbed parameterizations. Different reconstruction models process the same real images, allowing image content to be controlled while exposing architecture- and parameter-dependent traces. These reconstructions provide the supervisory signal for learning fingerprints that are invariant to image content yet discriminative across model identities, including different parameterizations of the same architecture.

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.

To strengthen fingerprint learning, we combine supervised contrastive, prototype cross-entropy, structure-aware ranking, and local counterfactual objectives. We further introduce an agent-guided collaborative training strategy in which a shared frozen language model periodically analyzes training diagnostics and adjusts sampling, local evidence weighting, loss configuration, frequency emphasis, and model-bank expansion. This allows the training process to respond to the evolving difficulty of different model identities rather than relying on a fixed schedule. Our contributions are summarized as follows:

- We propose a synthetic image attribution framework that learns generalizable model fingerprints without requiring synthetic training images from real-world generative models.We construct a controllable reconstruction-model bank with diverse architectures and parameterizations, and jointly optimize supervised contrastive, prototype cross-entropy, structure-aware ranking, and local counterfactual objectives to learn content-invariant and source-discriminative fingerprints.
- We introduce an agent-guided collaborative training strategy that adaptively coordinates sampling, local evidence, loss and frequency configurations, and model-bank expansion according to training diagnostics.
- Extensive experiments on 26 unseen source models across four public benchmarks show that our method achieves a macro-average verification AUC of 97.01%, outperforming the strongest baseline by 5.46 percentage points, demonstrating strong generalization without training on any real-world generator's outputs.

2 Related work

2.1 Generative Model Fingerprints and Attribution

Synthetic image attribution aims to identify the source generative model rather than merely distinguish real from synthetic content, providing finer-grained evidence for visual forensics Yu et al. (2019). Existing approaches can be broadly divided into proactive and passive attribution. Proactive methods embed identifiable signals into the generation process, such as model watermarks Fernandez et al. (2023), whereas passive methods exploit intrinsic artifacts introduced during image synthesis. Marra et al. Marra et al. (2019) showed that GAN-generated images contain model-specific traces in noise residuals, while Yu et al. Yu et al. (2019) learned fingerprints that separate source-specific artifacts from image content. Frequency-domain studies further revealed systematic spectral artifacts associated with generative models Frank et al. (2020). Subsequent works extended fingerprint-based attribution to architecture-level identification and improved robustness across content domains and image transformations Yang et al. (2022b); Bui et al. (2022). These studies collectively establish generative fingerprints as effective forensic cues for source attribution and related synthetic-image forensic tasks.

2.2 Open-Set and Generalizable Source Attribution

A central challenge in synthetic image attribution is generalization to sources that are not observed during training. Early open-set approaches mainly extend closed-set classifiers with rejection mechanisms. For example, Wang et al. Wang et al. (2023) employ a ViT-based classifier with threshold-based rejection, while Fang et al. Fang et al. (2023) identify unknown sources by measuring distances to known-class representations. POSE Yang et al. (2023a) further expands the open space by using augmentation models to simulate diverse unseen fingerprints.

Beyond rejection-based recognition, recent studies formulate attribution as source verification. Abady et al. Abady et al. (2024) use a Siamese framework to determine whether two synthetic images originate from the same source architecture, enabling verification beyond a fixed classifier. Other works improve attribution generalization through transferable visual representations. Cioni et al. Cioni et al. (2025) investigate pretrained CLIP features for synthetic-image origin attribution across GAN and diffusion models, while Nguyen et al. Nguyen et al. (2025) learn forensic representations from real images for zero-shot detection and open-set source attribution. More recently, One for All Fei et al. (2026) introduces a stochastic reconstruction-based fingerprint simulator and trains a fingerprint extractor using supervised contrastive learning, enabling image-to-model attribution

for previously unseen generative models without retraining the attribution system. Together, these studies shift synthetic image attribution from fixed closed-set classification toward more scalable and generalizable source verification.

3 METHOD

3.1 DEFINITION OF GENERATIVE MODEL FINGERPRINT

A generative model fingerprint is a source-specific pattern in synthetic images that is independent of image content. Such patterns appear in the amplitude spectrum Frank et al. (2020), shaped primarily by model architecture and modulated by model weights Corvi et al. (2023). We use reconstruction models to generate training images with known architectures and parameterizations and learn their fingerprints from the amplitude spectrum. The overview of our method is shown in Fig. 1.

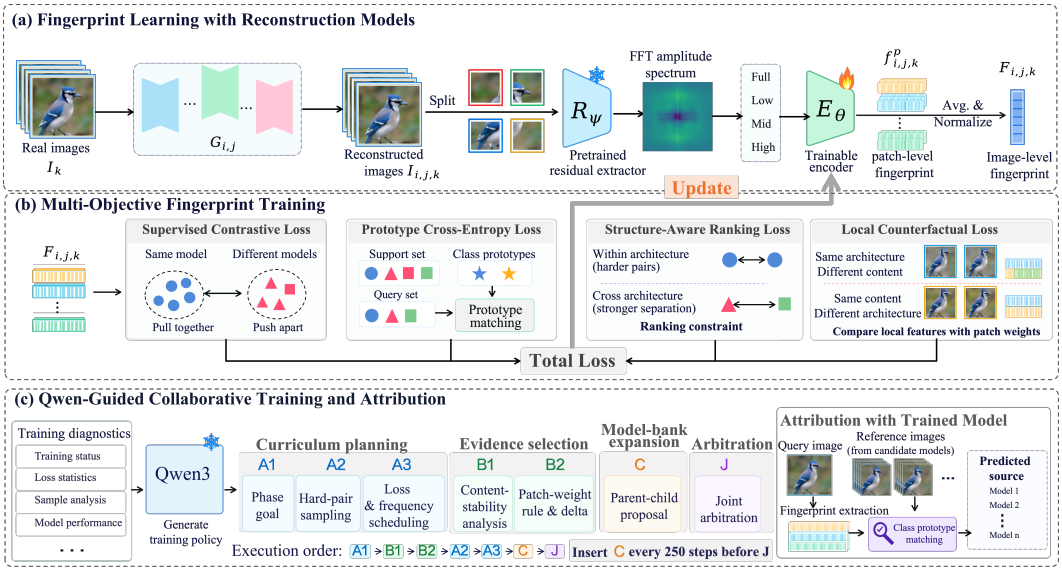


Figure 1: Pipeline of our method.

3.2 FINGERPRINT LEARNING WITH RECONSTRUCTION MODELS

We build an initial bank of 288 architectures with 20 independent parameterizations each, yielding 5,760 reconstruction models. For each architecture, 18 parameterizations are used for fingerprint learning and the remaining two for diagnostics on disjoint image contents. Each update samples four architectures and two parameterizations. All eight models reconstruct the same $J = 16$ real images. Let $G_{i,j}$ denote the model with architecture i and parameterization j, whose reconstruction of $I \in \mathbb{R}^{256\times256}$ is denoted by $I_{i,j}$. The reconstruction is split into four 128×128 corner patches $I^p_{i,j}$ that cover the full image, with $p \in \{1, 2, 3, 4\}$. We pretrain a residual network R_ψ and keep it frozen during fingerprint learning. Each corner patch $I^p_{i,j}$ is first processed by R_ψ, followed by a two-dimensional Fourier transform of the extracted residual. Since fingerprint evidence from different generative models is unevenly reliable across low-, mid-, and high-frequency bands, we apply four fixed, non-trainable soft masks M_v with $v \in \{\text{full}, \text{low}, \text{mid}, \text{high}\}$ to the FFT amplitude spectrum. A shared trainable ResNet-50 E_θ then encodes the masked Fourier amplitude spectra:

$$f^p_{i,j} = E_\theta\big(\text{Resize}_{128}\big(\big|\text{FFT2}(R_\psi(I^p_{i,j}))\big| \odot M_{v_t}\big)\big)\,. \tag{1}$$

Here, $f^p_{i,j} \in \mathbb{R}^{2048}$, M_{v_t} is the fixed soft mask for one band v_t sampled for the entire batch from $v \in \{\text{full}, \text{low}, \text{mid}, \text{high}\}$, and $\odot$ denotes element-wise multiplication. We average the local features and normalize the result to obtain the image-level fingerprints $F_{i,j}$:

$$F_{i,j} = \frac{\bar{f}_{i,j}}{\|\bar{f}_{i,j}\|_2}, \qquad \bar{f}_{i,j} = \frac{1}{4}\sum_{p=1}^{4} f^p_{i,j}. \tag{2}$$

We train E_θ with four losses and use a shared frozen Qwen3-4B-Instruct-2507 Team (2025) to adjust training from numerical diagnostics. Its roles cover curriculum planning (A), evidence selection (B), model-bank expansion (C), and joint arbitration (J).

Supervised contrastive loss. Since the procedure above is applied to each of the J real images independently, we write $F_{i,j,k}$ for the fingerprint obtained from the k-th image ($k \in \{1, \dots, J\}$) when this dependence needs to be made explicit. To enforce content-invariant model fingerprints, reconstructions generated by the same $G_{i,j}$ from different input images are treated as positives, while reconstructions from all other sampled models are treated as negatives, including different parameterizations of the same architecture. We optimize

$$\mathcal{L}_{\text{sup}} = \frac{1}{8J} \sum_{i,j,k} \left[\log \sum_{\substack{i',j',k' \\ (i',j',k') \neq (i,j,k)}} e^{F_{i,j,k}^\top F_{i',j',k'}/\tau} - \frac{\sum_{k' \neq k} F_{i,j,k}^\top F_{i,j,k'}}{(J-1)\tau} \right]. \tag{3}$$

where $\tau = 0.07$ is the temperature. This objective increases the similarity between fingerprints of the same reconstruction model across different contents while separating fingerprints produced by different models.

Prototype cross-entropy loss. To classify different fingerprints, we use the first $J/2$ contents as support and the remainder as queries. The support prototype is $\mu_{i,j} = \frac{2}{J} \sum_{k=1}^{J/2} F_{i,j,k}$, giving

$$\mathcal{L}_{\text{proto}} = -\frac{1}{4J} \sum_{i,j} \sum_{k > J/2} \log \frac{e^{F_{i,j,k}^\top \mu_{i,j}/\tau}}{\sum_{i',j'} e^{F_{i,j,k}^\top \mu_{i',j'}/\tau}}. \tag{4}$$

Each prototype score averages the cosine similarities between a query fingerprint and the support fingerprints of a model. This objective encourages each query to match the prototype of its source model while separating it from the prototypes of other sampled models.

Structure-aware ranking loss. To further separate models with similar fingerprints, we explicitly rank query scores on the same prototype. For the prototype $\mu_{i,j}$ of model $G_{i,j}$, $F_{i,j,k}^\top \mu_{i,j}$ is the score of its own query k, while $F_{i',j',k'}^\top \mu_{i,j}$ is the score of a query k' from another model $G_{i',j'}$ on the same prototype. We define $\ell_{i,j;i',j'}$ as the mean of $\log\big(1 + \exp\big((F_{i',j',k'} - F_{i,j,k})^\top \mu_{i,j}/\tau\big)\big)$ over all query pairs $k, k' > J/2$, which penalizes cases where queries from $G_{i',j'}$ obtain scores comparable to or higher than queries from $G_{i,j}$. Since models sharing the same architecture differ only in parameterization and are more difficult to distinguish, we separate same-architecture and cross-architecture model pairs:

$$\mathcal{L}_{\text{rank}} = \frac{\lambda_s}{8} \sum_{\substack{i,j,j' \\ j' \neq j}} \ell_{i,j;i,j'} + \frac{1-\lambda_s}{48} \sum_{\substack{i,j,i',j' \\ i' \neq i}} \ell_{i,j;i',j'}. \tag{5}$$

Here, the first term contains eight ordered pairs formed by the two parameterizations within each of the four sampled architectures, while the second contains 48 ordered pairs across different architectures. The coefficient λ_s controls their relative contribution. This objective encourages each model's queries to score higher on its own prototype than queries from other models. λ_s controls the emphasis on same-architecture discrimination: increasing λ_s assigns more weight to parameterization-level comparisons, while decreasing it emphasizes cross-architecture separation.

Local counterfactual loss. In addition to supervising the aggregated fingerprint $F_{i,j,k}$ via the image-level losses above, we further constrain the four local features $f_{i,j,k}$ extracted from the four equal image blocks. Each local feature is normalized as $\hat{f}^p_{i,j,k} = f^p_{i,j,k}/\|f^p_{i,j,k}\|_2$. For the same block position p, we compare two controlled changes. Replacing the image content while keeping the reconstruction model fixed gives $d^+_{i,j,k,p} = \|\hat{f}^p_{i,j,k} - \hat{f}^p_{i,j,k^-}\|_2^2$, where $k^- \neq k$. Replacing the parameterization while keeping the architecture and image content fixed gives $d^-_{i,j,k,p} = \|\hat{f}^p_{i,j,k} - \hat{f}^p_{i,j',k}\|_2^2$, where j' is the other sampled parameterization of architecture i. For each frequency band

v and image block p, we assign an evidence weight $w_{v,p}$ to reflect the reliability of that local feature for fingerprint learning. A larger $w_{v,p}$ gives the corresponding block a larger contribution to the local counterfactual loss. For the sampled band v_t, the weights are normalized as $\widetilde{w}_{v,p} = (w_{v,p} + 0.025)/(\sum_{p'=1}^{4} w_{v,p'} + 0.1)$. We then optimize

$$\mathcal{L}_{\mathrm{cf}} = \frac{1}{8J} \sum_{i,j,k,p} \widetilde{w}_{v_t,p} \left(d^{+}_{i,j,k,p} + [m + d^{+}_{i,j,k,p} - d^{-}_{i,j,k,p}]_{+} \right), \tag{6}$$

where $[x]_+ = \max(0, x)$ and $m = 0.1$. The first term reduces the variation caused by image content, while the second requires a change in model parameterization to produce a larger local-feature difference than a change in content.

The four objectives are combined as $\mathcal{L} = \alpha\mathcal{L}_{\mathrm{sup}} + \beta\mathcal{L}_{\mathrm{proto}} + \gamma\mathcal{L}_{\mathrm{rank}} + \delta\mathcal{L}_{\mathrm{cf}}$. The total loss is back-propagated to update the encoder E_θ, while the residual network R_ψ and the sampled reconstruction models remain frozen.

3.3 Agent-Guided Collaborative Training

Every 50 encoder updates, training diagnostics are summarized and passed to a shared frozen Qwen3-4B-Instruct-2507. Each role selects from predefined options using the diagnostics and preceding decisions. The roles are executed in the order A1→B1→B2→A2→A3→J; every 250 updates, C is inserted before J.

Phase planning (A1). A1 selects the phase goal from {*balance*, *identity*, *content*} using recent losses, ranking margins, gradient statistics, held-out performance, model-bank coverage, expansion feedback, and recent decisions. The selected goal is provided to the subsequent roles in the same decision round.

Local evidence selection (B1–B2). We evaluate the four image blocks under the four frequency masks on shared held-out contents. For each frequency band v and block p, $V_{v,p}$ measures the mean squared deviation from each model's mean local feature across contents, while $D_{v,p}$ measures the mean squared distance between paired parameterizations under matched content. We further compute $R_{v,p} = D_{v,p}/(V_{v,p} + 10^{-6})$ and $A_{v,p}$, the AUC averaged over ordered same-architecture model pairs. B1 analyzes these measurements together with the phase goal and recommends one evidence rule. B2 selects the final rule from {*balanced*, *identity*, *content*, *ratio*} and chooses $\delta \in \{0.05, 0.1, 0.2\}$. The balanced rule sets $w^{(0)}_{\mathrm{full},p} = 1/4$ and zero weights for the other bands. The identity, content, and ratio rules score block–band combinations by $A_{v,p}$, $-V_{v,p}$, and $R_{v,p}$, respectively. For these rules, four combinations are selected greedily while penalizing redundancy with previously selected combinations. The resulting weights are $w_{v,p} = \frac{1}{2}w^{(0)}_{v,p} + \frac{1}{8}$ for selected combinations and $w_{v,p} = \frac{1}{2}w^{(0)}_{v,p}$ otherwise. These weights and δ determine the local counterfactual supervision.

Hard-pair sampling (A2). A2 selects $r \in \{0, 0.25, 0.5\}$, which determines how many of the four sampled architecture pairs are drawn from difficult same-architecture pairs. Pair difficulty is maintained by an exponential moving average of the same-architecture ranking loss. Candidate values of r are evaluated with full-band probes using the losses, ranking margins, and the cosine similarity between supervised contrastive and same-architecture ranking gradients. The selected ratio assigns $4r$ of the four architecture-pair slots to difficult pairs, while the remaining slots are drawn from the coverage queue.

Loss and frequency scheduling (A3). A3 selects one predefined loss profile and one predefined frequency profile. The loss profile determines $(\alpha, \beta, \gamma, \lambda_s)$ in the training objective, while the frequency profile determines the sampling probabilities over the full, low, mid, and high masks. One band v_t is sampled for the entire training batch. These settings remain unchanged until a new policy is committed.

Model-bank expansion (C). Every 250 updates, C proposes a parent model from the difficult training identities, a predefined parameter-perturbation strength, and a parent–child insertion interval of two or four encoder updates. These choices are included in the proposed settings submitted to J.

Joint arbitration (J). J compares the newly proposed settings with the previous settings and a conservative variant. The conservative variant retains the proposed loss profile, limits the hard-pair ratio and counterfactual weight to $r \leq 0.25$ and $\delta \leq 0.1$, uses the balanced frequency profile and a four-step parent–child insertion interval, and averages the proposed evidence weights with $w^{(0)}$.

On 250-step rounds, each candidate is additionally evaluated by a one-step temporary update of the final convolution in the last residual block. All candidates start from the same encoder weights. After the temporary update, the change in held-out AUC is measured and the encoder weights are restored before evaluating the next candidate. J uses these measurements together with the training diagnostics and recent decisions to select the settings for subsequent training.

The selected settings take effect from the next encoder update. On expansion rounds, the parent specified by the committed settings is copied and its multidimensional weight tensors are perturbed with Gaussian noise scaled by their standard deviations. The child is then trained for 100 steps using reconstruction mean-squared error. It is admitted to the model bank if its held-out reconstruction error satisfies the predefined parent-relative criterion and its output differs from the parent under matched content. An accepted child receives a new parameterization identity within the same architecture, remains frozen during fingerprint learning, and replaces one sampled model pair at the selected insertion interval.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. For training the fingerprint simulator and extractor, we used real images from the train split of COCO 2017 dataset Lin et al. (2014). We evaluate source-model attribution on AIGIBench Li et al. (2025), DRAGON Bertazzini et al. (2025), FakeBench, the fully synthetic subset of DailyBench Jiang et al. (2026), and ImageAttributionBench Mou et al. (2026). Since several source models overlap across these datasets, we consolidate the duplicated sources and select 26 distinct models for source-model attribution.

Metrics. We employ the following metrics to evaluate verification and attribution performance. **Area Under the ROC Curve (AUC)** is used for verification tasks. It measures the ability of the verification score to distinguish positive samples from negative samples across all possible decision thresholds. **Accuracy (ACC)** is used for attribution tasks and is defined as the fraction of accepted known-model query images that are correctly assigned to their source models. **False Rejection Rate (FRR)** is the fraction of known-model query images incorrectly rejected as unknown. **False Acceptance Rate (FAR)** is the fraction of unknown-model query images incorrectly accepted as belonging to the candidate set.

Training details. The residual extractor R_ψ is pretrained for 6,000 steps with AdamW using a batch size of 32, an initial learning rate of 10^{-3}, weight decay 10^{-4}, and cosine learning-rate decay, and is frozen during fingerprint learning. We train the encoder for 5,000 steps using AdamW with a base learning rate of 3×10^{-5} for E_θ, weight decay 10^{-4}, gradient clipping at 1, and a 100-step warm-up followed by cosine decay. The four loss coefficients are initialized as $(\alpha, \beta, \gamma, \delta) = (1, 0.5, 0.5, 0.1)$, with $\lambda_s = 0.5$ for the ranking loss. During training, these coefficients are adaptively adjusted according to Qwen's decisions and remain unchanged until the next decision is committed. The Agent updates the training policy every 50 encoder updates, with model-bank expansion considered every 250 updates. Experiments tested on one NVIDIA A100 GPU.

4.2 TASK FORMULATION

Task 1: Image-to-Model Verification.

Given a query image $\mathbf{I}^q$ and a target model M^t, the objective is to verify whether $\mathbf{I}^q$ was synthesized by M^t. Since the identity of a generative model can be characterized using a collection of images it generates, we construct a *reference set* $\mathcal{R} = \{\mathbf{I}_1^r, \mathbf{I}_2^r, \ldots, \mathbf{I}_k^r\}$ consisting of k images generated by M^t. The image-to-model similarity is then computed by averaging the pairwise scores between $\mathbf{I}^q$ and all images in the reference set $\mathcal{R}$:

$$\texttt{score}(\mathbf{I}^q, M^t) = \frac{1}{k} \sum_{\mathbf{I}^r \in \mathcal{R}} \texttt{score}(\mathbf{I}^q, \mathbf{I}^r), \tag{7}$$

where a higher $\texttt{score}(\mathbf{I}^q, M^t)$ indicates that $\mathbf{I}^q$ is more likely to have been generated by M^t.

To turn this similarity into a decision, we compare it against a reference threshold that reflects how similar the images generated by M^t are among themselves. Concretely, we estimate the intra-model similarity (IMS) of M^t as the mean pairwise score over all distinct image pairs in its reference set:

$$\texttt{IMS}(M^t) = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i}}^{k} \texttt{score}(\mathbf{I}_i^r, \mathbf{I}_j^r). \tag{8}$$

Following the same calibration principle as in the image-to-image case, the acceptance threshold is this baseline scaled by α, i.e., $\delta = \texttt{IMS}(M^t)\,(1+\alpha)$. The query is accepted as generated by M^t if $\texttt{score}(\mathbf{I}^q, M^t) \geq \delta$, and rejected as not generated by M^t otherwise. For evaluation, positive samples are query images generated by M^t but excluded from the reference set $\mathcal{R}$, while images generated by all other models are treated as negative samples.

Task 2: Open-Set Attribution.

Given a query image $\mathbf{I}^q$ and a candidate set of known models $\mathcal{C} = \{M^1, M^2, \ldots, M^{|\mathcal{C}|}\}$, the goal is to either attribute $\mathbf{I}^q$ to its source model within $\mathcal{C}$ or reject it as originating from an unknown model outside the candidate set. Similar to the verification setting, each candidate model M^j is represented by a reference set $\mathcal{R}^j$ consisting of images it generates. The similarity between the query image and each candidate model, $\texttt{score}(\mathbf{I}^q, M^j)$, is computed using the image-to-model score defined in Eq. 7. We first select the best-matching candidate:

$$j^* = \arg \max_{j \in \{1,\ldots,|\mathcal{C}|\}} \text{score}(\mathbf{I}^q, M^j), \tag{9}$$

and denote its score by $\texttt{score}(\mathbf{I}^q, M^{j^*})$. To decide whether this best match is trustworthy, we adopt a *per-model* acceptance threshold derived from each model's intra-model similarity. For every candidate M^j, we estimate the expected intra-model similarity $\texttt{IMS}(M^j)$ from its reference set using Eq. 8, which characterizes the typical similarity among images generated by M^j and serves as a model-specific baseline. The acceptance threshold for M^j is this baseline scaled by α, i.e., $\delta_j = \texttt{IMS}(M^j)\,(1+\alpha)$. The query image is attributed to M^{j^*} only if its score reaches that model's own threshold, i.e., $\texttt{score}(\mathbf{I}^q, M^{j^*}) \geq \delta_{j^*}$; otherwise it is rejected as originating from an unknown model outside $\mathcal{C}$.

4.3 Our performance on Task 1

We evaluate fingerprint discriminability through image-to-model verification on 26 sources, comparing our method with Siamese Abady et al. (2024), Uni. Att. Cioni et al. (2024), and FP Yang et al. (2025). All methods use the same fixed support/query split with 256 support images per source. Table 1 reports the mean and population standard deviation of each source's 25 directed pairwise AUC against the remaining sources. Our method achieves a macro-averaged AUC of 97.01%, outperforming FP and Uni. Att. by 5.46 and 15.46 percentage points, respectively. It exceeds 90% AUC on all 26 sources and 95% on 21. Our method also has the lowest standard deviation across per-source AUC at 2.36 , indicating more consistent discrimination across the evaluated sources.

To examine fingerprint separability, Table 2 compares intra- and inter-source cosine similarities. The margin is the difference between the intra- and inter-source means. Standard deviations describe score variability within each source comparison and variation across sources in the final row. Intra-source similarity exceeds inter-source similarity for all 26 sources, with averages of 0.7243 and 0.5664, respectively, yielding a mean margin of 0.1579. The margin exceeds 0.1 for 23 sources

Table 1: Image-to-model verification AUC (%) on 26 sources.

Source	Siamese	Uni. Att.	FP	Ours	Source	Siamese	Uni. Att.	FP	Ours
CogView3-Plus-3B	$62.10_{\pm 14.14}$	$70.24_{\pm 14.24}$	$77.67_{\pm 17.49}$	$97.01_{\pm 4.73}$	FaceSwap	$68.52_{\pm 8.28}$	$96.52_{\pm 11.29}$	$97.28_{\pm 8.95}$	$95.54_{\pm 8.25}$
FLUX.1-dev	$63.48_{\pm 12.66}$	$78.53_{\pm 13.54}$	$92.24_{\pm 8.13}$	$99.45_{\pm 1.16}$	Imagen 3	$59.13_{\pm 14.22}$	$88.30_{\pm 8.98}$	$89.78_{\pm 9.55}$	$98.46_{\pm 4.30}$
Infinity-2B	$67.07_{\pm 10.55}$	$76.85_{\pm 12.03}$	$98.83_{\pm 0.98}$	$96.17_{\pm 6.19}$	StyleSwin	$65.16_{\pm 12.12}$	$94.71_{\pm 11.78}$	$95.62_{\pm 9.50}$	$97.19_{\pm 5.27}$
R3GAN	$73.53_{\pm 4.09}$	$96.29_{\pm 11.39}$	$99.88_{\pm 0.38}$	$97.83_{\pm 4.98}$	GPT-4o	$63.86_{\pm 16.32}$	$84.08_{\pm 11.80}$	$90.65_{\pm 9.68}$	$93.97_{\pm 10.56}$
SD 1.5	$59.98_{\pm 8.87}$	$64.22_{\pm 11.99}$	$91.04_{\pm 7.82}$	$99.56_{\pm 0.90}$	Gemini 2.0 Flash	$62.71_{\pm 14.51}$	$73.72_{\pm 12.36}$	$81.66_{\pm 14.82}$	$98.74_{\pm 2.70}$
SD 3 Medium	$63.95_{\pm 15.02}$	$66.83_{\pm 13.99}$	$86.44_{\pm 12.58}$	$96.58_{\pm 6.01}$	Grok 3	$59.56_{\pm 12.50}$	$81.07_{\pm 8.40}$	$99.44_{\pm 0.58}$	$99.96_{\pm 0.07}$
Stable Cascade	$61.82_{\pm 13.33}$	$79.79_{\pm 12.58}$	$92.82_{\pm 6.66}$	$100.00_{\pm 0.00}$	Ideogram v1 Turbo	$74.54_{\pm 12.81}$	$74.50_{\pm 12.60}$	$94.04_{\pm 5.76}$	$96.90_{\pm 5.30}$
StyleGAN3-T	$54.34_{\pm 11.19}$	$90.75_{\pm 5.68}$	$85.98_{\pm 13.55}$	$99.28_{\pm 1.68}$	Janus-Pro	$45.49_{\pm 6.40}$	$78.19_{\pm 10.80}$	$94.39_{\pm 6.54}$	$100.00_{\pm 0.01}$
StyleGAN-XL	$76.12_{\pm 10.57}$	$98.46_{\pm 4.03}$	$98.80_{\pm 2.31}$	$100.00_{\pm 0.01}$	Kling Image v1	$67.21_{\pm 14.68}$	$76.38_{\pm 14.00}$	$89.32_{\pm 10.69}$	$97.27_{\pm 5.52}$
SD 2.1	$63.97_{\pm 14.55}$	$61.97_{\pm 14.02}$	$75.39_{\pm 16.93}$	$97.96_{\pm 4.21}$	E4S	$69.35_{\pm 9.01}$	$97.46_{\pm 7.42}$	$97.64_{\pm 7.68}$	$96.14_{\pm 10.66}$
SDXL-Base-1.0	$62.87_{\pm 13.99}$	$68.74_{\pm 15.98}$	$89.65_{\pm 8.53}$	$93.89_{\pm 12.27}$	DeepFloyd IF	$49.87_{\pm 14.61}$	$73.34_{\pm 13.29}$	$81.44_{\pm 11.87}$	$96.28_{\pm 4.35}$
BlendFace†	$74.20_{\pm 7.60}$	$97.04_{\pm 9.48}$	$98.31_{\pm 5.76}$	$95.50_{\pm 11.56}$	Mobius	$73.21_{\pm 10.17}$	$78.22_{\pm 13.12}$	$92.91_{\pm 6.48}$	$94.49_{\pm 10.28}$
DALL·E 3	$65.94_{\pm 16.27}$	$84.57_{\pm 11.85}$	$89.19_{\pm 9.80}$	$93.31_{\pm 11.17}$	Step1X-Edit v1.2	$41.11_{\pm 4.69}$	$89.40_{\pm 3.97}$	$99.99_{\pm 0.02}$	$90.77_{\pm 10.67}$
Average AUC over 26 sources:	Siamese $63.43_{\pm 8.40}$	Uni. Att. $81.55_{\pm 10.84}$	FP $91.55_{\pm 6.76}$	Ours **97.01**$_{\pm 2.36}$					

Table 2: Intra- and inter-source fingerprint similarities on 26 sources. Margin = Intra − Inter.

Source	Intra	Inter	Margin	Source	Intra	Inter	Margin
CogView3-Plus-3B	$0.7212_{\pm 0.0329}$	$0.5737_{\pm 0.0820}$	0.1475	FaceSwap	$0.7465_{\pm 0.0532}$	$0.5447_{\pm 0.0996}$	0.2018
FLUX.1-dev	$0.7111_{\pm 0.0163}$	$0.5953_{\pm 0.0557}$	0.1158	Imagen 3	$0.6917_{\pm 0.0292}$	$0.5427_{\pm 0.0669}$	0.1489
Infinity-2B	$0.6977_{\pm 0.0251}$	$0.5713_{\pm 0.0741}$	0.1264	StyleSwin	$0.7167_{\pm 0.0179}$	$0.6118_{\pm 0.0730}$	0.1049
R3GAN	$0.7463_{\pm 0.0332}$	$0.5568_{\pm 0.0904}$	0.1895	GPT-4o	$0.6929_{\pm 0.0320}$	$0.5633_{\pm 0.0763}$	0.1296
SD 1.5	$0.7631_{\pm 0.0238}$	$0.5861_{\pm 0.0792}$	0.1770	Gemini 2.0 Flash	$0.7060_{\pm 0.0171}$	$0.5816_{\pm 0.0676}$	0.1245
SD 3 Medium	$0.7135_{\pm 0.0254}$	$0.5930_{\pm 0.0741}$	0.1205	Grok 3	$0.7946_{\pm 0.0147}$	$0.5705_{\pm 0.0821}$	0.2241
Stable Cascade	$0.7041_{\pm 0.0104}$	$0.4130_{\pm 0.0598}$	0.2912	Ideogram v1 Turbo	$0.6913_{\pm 0.0211}$	$0.5802_{\pm 0.0643}$	0.1111
StyleGAN3-T	$0.7699_{\pm 0.0252}$	$0.6014_{\pm 0.0785}$	0.1685	Janus-Pro	$0.7863_{\pm 0.0228}$	$0.5181_{\pm 0.0899}$	0.2682
StyleGAN-XL	$0.8389_{\pm 0.0231}$	$0.5496_{\pm 0.0903}$	0.2893	Kling Image v1	$0.7173_{\pm 0.0192}$	$0.5851_{\pm 0.0767}$	0.1322
SD 2.1	$0.7178_{\pm 0.0278}$	$0.5733_{\pm 0.0762}$	0.1445	E4S	$0.7324_{\pm 0.0429}$	$0.5385_{\pm 0.0939}$	0.1939
SDXL-Base-1.0	$0.6866_{\pm 0.0249}$	$0.5774_{\pm 0.0704}$	0.1092	DeepFloyd IF	$0.7200_{\pm 0.0231}$	$0.6217_{\pm 0.0649}$	0.0983
BlendFace	$0.7260_{\pm 0.0426}$	$0.5461_{\pm 0.0906}$	0.1799	Mobius	$0.6686_{\pm 0.0182}$	$0.5770_{\pm 0.0681}$	0.0916
DALL·E 3	$0.6942_{\pm 0.0378}$	$0.5578_{\pm 0.0799}$	0.1364	Step1X-Edit v1.2	$0.6769_{\pm 0.0303}$	$0.5963_{\pm 0.0598}$	0.0806
Average over 26 sources:	Intra $0.7243_{\pm 0.0389}$	Inter $0.5664_{\pm 0.0387}$	Margin **0.1579**$_{\pm 0.0574}$				

and is largest for Stable Cascade (0.2912), StyleGAN-XL (0.2893), and Janus-Pro (0.2682). Intra-source score distributions also have smaller standard deviations than inter-source distributions for every source, indicating consistent matching scores among images generated by the same model. These results support the separability of the fingerprints extracted by our method: they preserve consistent same-source similarities while producing lower cross-source similarities on average, providing discriminative evidence for distinguishing images generated by different models. Besides, we present the score distributions in Fig. 2, where queries from the target model consistently obtain higher cosine similarity scores than those from the other model, with limited overlap between the two distributions. It demonstrates that our method yields highly separable fingerprints for image-to-model verification.

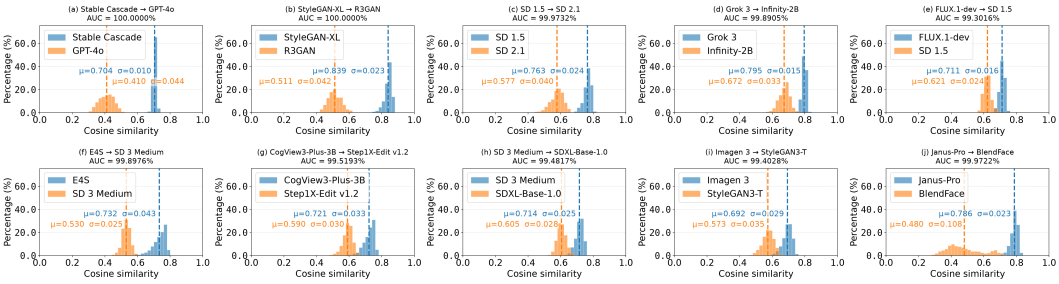


Figure 2: Similarity distributions for image-to-model verification.

4.4 Our performance on Task 2

To evaluate open-set attribution under different levels of difficulty, we construct multiple splits with varying numbers of known and unknown source models, ranging from $K/U = 3/3$ to $13/13$. Table 3 reports the attribution accuracy on known models together with the detection performance

Table 3: Open-set attribution under varying numbers of known and unknown models. K/U denotes the numbers of known and unknown models. $\text{Acc}_{\text{known}}$ is the attribution accuracy on known models, and AUC_{un} measures the detection of unknown models.

K/U	Method	$\text{Acc}_{\text{known}}\uparrow$	$\text{AUC}_{\text{un}}\uparrow$	FRR↓	FAR↓	K/U	Method	$\text{Acc}_{\text{known}}\uparrow$	$\text{AUC}_{\text{un}}\uparrow$	FRR↓	FAR↓
3/3	Uni. Att.	$96.46_{\pm 4.06}$	$75.23_{\pm 12.52}$	$29.46_{\pm 10.20}$	$32.79_{\pm 22.05}$	9/9	Uni. Att.	$85.49_{\pm 6.41}$	$67.03_{\pm 9.49}$	$27.96_{\pm 4.62}$	$43.82_{\pm 14.55}$
	FP	$\mathbf{99.22}_{\pm 1.58}$	$93.79_{\pm 5.25}$	$\mathbf{15.11}_{\pm 4.78}$	$12.83_{\pm 15.67}$		FP	$87.41_{\pm 4.37}$	$84.64_{\pm 6.06}$	$25.83_{\pm 1.38}$	$34.88_{\pm 8.25}$
	Ours	$99.06_{\pm 1.54}$	$\mathbf{94.14}_{\pm 5.48}$	$16.68_{\pm 3.82}$	$\mathbf{9.76}_{\pm 10.42}$		Ours	$\mathbf{90.97}_{\pm 4.77}$	$\mathbf{84.84}_{\pm 3.19}$	$\mathbf{15.96}_{\pm 0.99}$	$\mathbf{26.72}_{\pm 4.99}$
5/5	Uni. Att.	$90.10_{\pm 6.26}$	$74.77_{\pm 15.49}$	$31.46_{\pm 6.68}$	$31.16_{\pm 22.17}$	11/11	Uni. Att.	$85.86_{\pm 4.77}$	$64.02_{\pm 9.15}$	$34.07_{\pm 7.60}$	$45.56_{\pm 14.30}$
	FP	$92.90_{\pm 5.21}$	$89.40_{\pm 7.33}$	$\mathbf{15.12}_{\pm 3.40}$	$17.74_{\pm 15.53}$		FP	$83.62_{\pm 4.86}$	$80.55_{\pm 6.67}$	$26.24_{\pm 1.93}$	$\mathbf{26.95}_{\pm 11.12}$
	Ours	$\mathbf{95.56}_{\pm 5.67}$	$\mathbf{90.10}_{\pm 6.94}$	$17.54_{\pm 3.42}$	$\mathbf{17.50}_{\pm 10.71}$		Ours	$\mathbf{90.65}_{\pm 5.55}$	$\mathbf{89.51}_{\pm 4.36}$	$\mathbf{17.29}_{\pm 1.82}$	$27.08_{\pm 7.87}$
7/7	Uni. Att.	$89.00_{\pm 5.24}$	$69.99_{\pm 2.75}$	$30.58_{\pm 8.44}$	$40.23_{\pm 6.48}$	13/13	Uni. Att.	$85.29_{\pm 3.24}$	$59.47_{\pm 3.10}$	$32.16_{\pm 4.48}$	$54.12_{\pm 4.18}$
	FP	$\mathbf{97.58}_{\pm 3.88}$	$83.58_{\pm 6.08}$	$\mathbf{15.86}_{\pm 2.35}$	$27.83_{\pm 9.65}$		FP	$84.65_{\pm 3.66}$	$78.56_{\pm 4.79}$	$19.90_{\pm 1.51}$	$36.06_{\pm 9.68}$
	Ours	$95.01_{\pm 4.47}$	$\mathbf{87.81}_{\pm 3.94}$	$16.62_{\pm 2.26}$	$\mathbf{20.14}_{\pm 6.38}$		Ours	$\mathbf{90.26}_{\pm 5.10}$	$\mathbf{87.36}_{\pm 4.42}$	$\mathbf{13.76}_{\pm 1.40}$	$\mathbf{29.08}_{\pm 7.84}$

Table 4: Ablation results on 26-source image-to-model verification.

Setting	$\mathcal{L}_{\text{sup}}$	$\mathcal{L}_{\text{proto}}$	$\mathcal{L}_{\text{rank}}$	$\mathcal{L}_{\text{cf}}$	AUC (%)	Δ
All objectives	✓	✓	✓	✓	**96.86**	–
$\mathcal{L}_{\text{sup}}$	×	✓	✓	✓	94.32	-2.54
$\mathcal{L}_{\text{proto}}$	✓	×	✓	✓	92.66	-4.20
$\mathcal{L}_{\text{rank}}$	✓	✓	×	✓	93.75	-3.10
$\mathcal{L}_{\text{cf}}$	✓	✓	✓	×	94.04	-2.82

for unknown models. Our method achieves the highest AUC_{un} in all six settings while maintaining competitive or best $\text{Acc}_{\text{known}}$. It also provides favorable FRR/FAR trade-offs, especially in the larger splits, demonstrating robust attribution when both known and unknown source sets become more diverse.

Table 4 shows that removing any single objective lowers the AUC, and the full model achieves the best result of 96.86%. The largest performance drop comes from removing $\mathcal{L}_{\text{proto}}$ (4.20), followed by $\mathcal{L}_{\text{rank}}$ (3.10), $\mathcal{L}_{\text{cf}}$ (2.82), and $\mathcal{L}_{\text{sup}}$ (2.54). This indicates that prototype supervision is the most critical component, while the other three objectives provide complementary gains.

5 Conclusion

In this work, we propose a synthetic image attribution method based on generative model fingerprinting that requires no synthetic images from any real-world generative model during training. Instead, we train a fingerprint extractor entirely on a large, controllable bank of reconstruction models spanning diverse architectures and independently perturbed parameterizations, using the model-specific traces embedded in their reconstructions as a proxy supervisory signal. We adopt an agent-guided collaborative training procedure. Extensive experiments on 26 unseen source models across four public benchmarks show that our method achieves a macro-average verification AUC of 97.01%, outperforming the strongest baseline by 5.46 percentage points.

Limitations. Although our method generalizes well to a wide range of unseen generative models, there remains room for improvement on generated images that retain substantial real content, such as those produced by face-swapping and image editing methods. Unlike text-to-image models that synthesize entire images, these methods modify only local regions of real images, so most patches carry traces of real content rather than of the generative model, which dilutes the source-specific fingerprint. Furthermore, face-swapping methods typically share similar processing pipelines, such as face alignment, blending, and paste-back, which may lead to category-level rather than method-specific traces. In the future, we will focus on images that retain substantial real content, such as those from face swapping and image editing, where only local regions are modified and many patches remain real-content-dominated, weakening the source-specific fingerprint.

GENERATIVE AI USE DISCLOSURE

Generative AI tools are used in this work to produce image data. Large language models (LLMs) are further adopted to polish writing, detect textual errors, and enhance the overall readability of the manuscript. The authors hold full responsibility for all content, statements and outputs in this final work.

REFERENCES

Lydia Abady, Jun Wang, Benedetta Tondi, and Mauro Barni. A siamese-based verification system for open-set architecture attribution of synthetic images. *Pattern Recognition Letters*, 180:75–81, 2024.

Giulia Bertazzini, Daniele Baracchi, Dasara Shullani, Isao Echizen, and Alessandro Piva. DRAGON: A large-scale dataset of realistic images generated by diffusion models, 2025. URL `https://arxiv.org/abs/2505.11257`.

Tu Bui, Ning Yu, and John Collomosse. Repmix: Representation mixing for robust attribution of synthesized images. In *European Conference on Computer Vision*, pp. 146–163. Springer, 2022.

Dario Cioni, Christos Tzelepis, Lorenzo Seidenari, and Ioannis Patras. Are clip features all you need for universal synthetic image origin attribution? In *European Conference on Computer Vision*, pp. 363–382. Springer, 2024.

Dario Cioni, Christos Tzelepis, Lorenzo Seidenari, and Ioannis Patras. Are CLIP features all you need for universal synthetic image origin attribution? In *Computer Vision – ECCV 2024 Workshops*, pp. 363–382. Springer, 2025. doi: 10.1007/978-3-031-92648-8_22.

Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: From generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 973–982, 2023. doi: 10.1109/CVPRW59228.2023.00104.

Yunshu Dai, Jianwei Fei, and Fangjun Huang. Idguard: Robust, general, identity-centric poi proactive defense against face editing abuse. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11934–11943. IEEE, 2024.

Yunshu Dai, Jianwei Fei, Fangjun Huang, and Chip Hong Chang. Robust secure swap: Responsible face swap with persons of interest redaction and provenance traceability. In *Forty-second International Conference on Machine Learning*, 2025.

Shengbang Fang, Tai D. Nguyen, and Matthew C. Stamm. Open set synthetic image source attribution. In *Proceedings of the 34th British Machine Vision Conference*. BMVA, 2023.

Jianwei Fei, Zhihua Xia, Benedetta Tondi, and Mauro Barni. Supervised gan watermarking for intellectual property protection. In *2022 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–6. IEEE, 2022.

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

Jianwei Fei, Yunshu Dai, Wenyuan Yang, and Zhihua Xia. Distributor-centric model watermarking for image generative models. *Knowledge-Based Systems*, pp. 114422, 2025b.

Jianwei Fei, Yunshu Dai, Peipeng Yu, Zhe Kong, Jiantao Zhou, and Zhihua Xia. Scalable dual fingerprinting for hierarchical attribution of text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15025–15034, 2025c.

Jianwei Fei, Yunshu Dai, Peipeng Yu, Zhihua Xia, Dasara Shullani, Daniele Baracchi, and Alessandro Piva. One for all: Synthesis-free fingerprint learning for attribution of in-the-wild synthetic images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35419–35427, 2026. doi: 10.1609/aaai.v40i42.40851.

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

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3247–3258. PMLR, 2020.

Xin Jiang, Hao Tang, Junyao Gao, Meiqi Cao, Fei Shen, Dongming Zhang, and Yongdong Zhang. DailyBench: A unified benchmark for AI-generated and manipulated images from modern generative models, 2026. URL https://arxiv.org/abs/2607.24016.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? In *Advances in Neural Information Processing Systems*, 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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

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. doi: 10.1109/MIPR.2019.00103.

Midjourney, Inc. Midjourney, 2023. URL https://www.midjourney.com/home/.

Tingshu Mou, Zhipeng Wei, Chao Gong, Jingjing Chen, and Xingjun Ma. ImageAttributionBench: How far are we from generalizable attribution?, 2026. URL https://arxiv.org/abs/2605.12967.

Tai D. Nguyen, Aref Azizpour, and Matthew C. Stamm. Forensic self-descriptions are all you need for zero-shot detection, open-set source attribution, and clustering of AI-generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2503.21003.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Qwen Team. Qwen3-4b-instruct-2507, 2025. URL https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507.

Jun Wang, Omran Alamayreh, Benedetta Tondi, and Mauro Barni. Open set classification of GAN-based image manipulations via a ViT-based hybrid architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 953–962, 2023.

Tianyun Yang, Ziyao Huang, Juan Cao, Lei Li, and Xirong Li. Deepfake network architecture attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4662–4670, 2022a.

Tianyun Yang, Ziyao Huang, Juan Cao, Lei Li, and Xirong Li. Deepfake network architecture attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4662–4670, 2022b. doi: 10.1609/aaai.v36i4.20391.

Tianyun Yang, Danding Wang, Fan Tang, Xinying Zhao, Juan Cao, and Sheng Tang. Progressive open space expansion for open-set model attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15856–15865, 2023a.

Tianyun Yang, Danding Wang, Fan Tang, Xinying Zhao, Juan Cao, and Sheng Tang. Progressive open space expansion for open-set model attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15856–15865, 2023b.

Tianyun Yang, Juan Cao, Danding Wang, and Chang Xu. Model synthesis for zero-shot model attribution. *IEEE Transactions on Multimedia*, 2025.

Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7556–7566, 2019.