QORA: A SUSTAINABLE FRAMEWORK FOR OPEN-WORLD GENERATIVE MODEL ATTRIBUTION WITH QUASI-ORTHOGONAL REPRESENTATION DISENTANGLEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid emergence of new generative models poses significant challenges to static attribution frameworks, which often confidently misattribute images from unknown sources to known ones and struggle to adapt stably to new models. To address these limitations, we propose Quasi-Orthogonal Representation Attribution (QORA), a unified framework for sustainable open-world generative model attribution. QORA consists of two core modules. The Progressive Orthogonal Learning Module (POLM) employs Stiefel manifold optimization to construct a quasi-orthogonal feature space that reduces redundancy while maintaining a stable attribution subspace for open-world settings. The Fingerprint Disentanglement and Enhancement Module (FDEM) leverages classifier-guided attention and multi-auxiliary contrastive learning to disentangle and amplify model-specific fingerprints. To enable continual learning, QORA integrates exemplar replay with feature-similarity-based classifier initialization, achieving lightweight incremental updates for new models while avoiding catastrophic forgetting. Extensive experiments demonstrate that QORA achieves state-of-the-art closed-set accuracy and strong open-set robustness across GAN and diffusion benchmarks, while maintaining stable performance during incremental learning, highlighting its superior scalability and applicability in evolving environments.

1 INTRODUCTION

Generative AI has made remarkable progress in image quality, diversity, and controllability, with applications spanning from entertainment to production. Yet these capabilities also raise serious security concerns, as maliciously crafted synthetic images are exploited to spread misinformation, fabricate events, and manipulate public opinion, threatening the integrity of the digital ecosystem. To mitigate risks, leading AI companies have pledged to embed watermarks into generated content (Bartz & Hu), but such active solutions lack universality. This has driven research into passive methods that detect AI-generated content (Wang et al., 2020b; 2023b; Ojha et al., 2023), though they generally fail to identify the specific source model—information critical for responsibility tracing and accountability.

To address this, the task of generative model attribution has been developed to passively trace the source generator. Early reconstruction-based methods (Albright et al., 2019) exploited cross-model reconstruction errors but were limited to GANs. Fingerprint-based approaches later demonstrated distinct model-specific traces (Yu et al., 2019b; Marra et al., 2019a), enabling multi-class attribution (Yang et al., 2021; Bui et al., 2022), while MAID (Zhu et al., 2025) extended attribution to diffusion models. These methods, however, are closed-world and often misattribute images from unseen generators to the nearest known model. Open-world attribution addresses this limitation by combining attribution with rejection of unknown classes, using strategies such as patch-based contrastive learning (Yang et al., 2022), rejection-aware classifiers (Wang et al., 2023a), metric learning (Fang et al., 2023b), feature augmentation (Yang et al., 2023), Siamese verification (Abady et al., 2024), forensic self-descriptions (Nguyen et al., 2025), and frequency-domain masking (Zhang et al., 2025). Despite these advances, most methods are trained on limited data, are sensitive to irrelevant con-

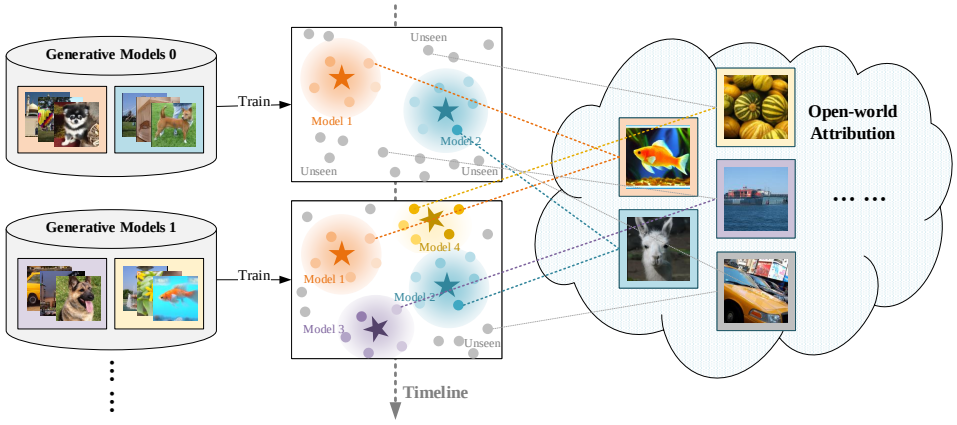


Figure 1: Overview of the SOW-GMA task, which requires the attribution system incrementally incorporates new generative models while maintaining accurate attribution for known sources and reliably rejecting unknown ones, ensuring long-term scalability in dynamic real-world scenarios.

tent or texture, and require full retraining to accommodate new models. As large-scale pretrained Vision–Language Models (VLMs) can produce robust, content-agnostic features, OCC-CLIP (Liu et al., 2024) adapts CLIP for few-shot attribution, while Cioni *et al.* (Cioni et al., 2025) analyze their feature layers for generalization. These approaches, however, typically use embeddings directly and do not optimize VLMs specifically for model attribution task or filter out irrelevant information.

Moreover, the rapid emergence of new generative models further underscores the need for sustainable open-world attribution, as illustrated in Fig. 1. Current solutions suffer from high computational cost, memory overhead, and catastrophic forgetting during incremental updates (Li et al., 2024a). A practical framework must therefore support accurate attribution of known generators, reliable rejection of unknowns, and efficient adaptation to new models without full retraining.

To this end, we propose Quasi-Orthogonal Representation Attribution (QORA), a scalable framework for the Sustainable Open-World Generative Model Attribution (SOW-GMA) task. Built on the CLIP-ViT L/14 backbone, QORA extracts mid-level features containing generative fingerprints and fine-tunes them via LoRA for artifact sensitivity. It introduces the Progressive Orthogonal Learning Module (POLM) to reduce feature redundancy and construct a stable artigact space for the open-world generators, and the Fingerprint Disentanglement and Enhancement Module (FDEM) to isolate and amplify fingerprint-specific signals for closed-set attribution. During incremental learning, QORA freezes most parameters and expands only lightweight classifiers with exemplar replay, enabling efficient adaptation with minimal overhead. The main contributions can be summarized as follows:

- We propose QORA, a practical and scalable framework for SOW-GMA task, which jointly supports accurate closed-set attribution, reliable open-set rejection, and efficient incremental learning for real-world deployment.
- We design a synergistic dual-module architecture, in which POLM construct a stable artigact space for open-world generators, and FDEM decouples and amplifies closed-set model-specific fingerprints.
- We first introduce Stiefel manifold optimization into generative model attribution. By constraining the encoder weights to yield maximally independent feature dimensions that better capture subtle generative fingerprints.

2 Related Works

Artifacts in AI-Generated Images. AI-generated images contain visually subtle but detectable artifacts that differ across architectures and can be exploited for attribution. Early studies emphasized frequency-domain traces, such as irregular mid–high-frequency patterns in GAN outputs (Durall et al., 2020), leading to classifiers based on frequency domains (Frank et al., 2020; Jeong

et al., 2022c). However, these methods generalize poorly to diffusion models, whose artifacts are less frequency-pronounced. Recent works shift focus to spatial-domain cues, leveraging shallow-layer textures (Liu et al., 2020; Zhong et al., 2023), residual modeling (Sinitsa & Fried, 2024), or diffusion-specific reconstruction artifacts (Zhong et al., 2025; Wang et al., 2023b). Pretrained VLMs (Ojha et al., 2023; Sha et al., 2023; Zhu et al., 2023) further improve generalization by extracting robust, content-invariant features. Our approach builds on this line by exploiting mid-level VLM features to extract stable spatial-domain fingerprints.

Generative Model Attribution. Attribution methods aim to identify the source generator of synthetic images. Active approaches embed watermarks but lack generality, while passive approaches exploit model-specific fingerprints. Recent closed-world methods adopt multi-class classification (Yang et al., 2021; Bui et al., 2022) or reconstruction errors (Albright & McCloskey, 2019; Zhu et al., 2025), but fail to generalize to unseen models. Open-world attribution extends to novel classes through strategies such as transformation-pretrained contrastive learning (Yang et al., 2022), Transformer-based localization (Wang et al., 2023a), metric learning (Fang et al., 2023a), feature-space augmentation (Yang et al., 2023), similarity verification (Abady et al., 2024), and spectral masking (Zhang et al., 2025). Despite these advances, most methods require retraining to handle new models and often struggle to suppress irrelevant content. Our work addresses these limitations by introducing quasi-orthogonal projection to suppress redundancy and construct a stable artifact space, while disentangling fingerprints to achieve sustainable attribution in the open world.

Category-Incremental Learning for Attribution. The continual emergence of new generators renders static attribution impractical. Category-Incremental Learning (CIL) (Wang et al., 2024; Ji et al., 2023) expands recognition capacity without full historical data, with prior work exploring contrastive learning (Pan et al., 2023), adapters (Gao et al., 2024), or regeneration-based updates (Li et al., 2024b). For GAN detection, incremental and adapter-based frameworks (Marra et al., 2019b; Tang et al., 2025) alleviate semantic drift. In attribution, however, most solutions remain costly or inflexible. We propose a unified framework that combines open-world rejection with class-incremental expansion, using a compact exemplar memory and feature-similarity-based classifier initialization to achieve scalable, sustainable attribution.

3 Problem Definition

The SOW-GMA task is designed for a realistic and dynamic setting where generative models continuously emerge. The objective is to build an attribution framework supporting open-set recognition and sustainable incremental learning. Training proceeds over sessions $t = 0, 1, \ldots, T$, where the model receives a labeled dataset

$$\mathcal{D}_t^L = \{(x_{t,i}, y_{t,i})\}_{i=1}^{N_t}, \quad y_{t,i} \in \mathcal{C}_t^L, \tag{1}$$

where $x_{t,i}$ is a generated image and $y_{t,i}$ its source model label, together with a memory buffer $\mathcal{D}_t^M \subseteq \bigcup_{i=0}^{t-1} \mathcal{D}_i^L$ that stores exemplars from past sessions. The cumulative known classes are $\mathcal{C}_t^K = \bigcup_{i=0}^{t} \mathcal{C}_i^L$.

In addition to labeled data, a continuously growing unlabeled data pool $\mathcal{D}_t^U = \{x_i\}_{i=1}^m$ is also available, whose classes $\mathcal{C}_t^U$ may include both known $\mathcal{C}_t^K$ and novel unknown classes $\mathcal{C}_t^N \subseteq \mathcal{C}_t^U \backslash \mathcal{C}_t^K$.

The goal of SOW-GMA is to learn a continually adaptive feature extractor $\phi(\cdot)$ that:

1. attributes generated images from known models to $\mathcal{C}_t^K$,
2. rejects generated images from novel unknown models $\mathcal{C}_t^N$ as out-of-distribution,
3. incorporates new model classes through lightweight updates with limited memory $\mathcal{D}_t^M$.

4 Method

To address the SOW-GMA task, we propose QORA (Fig. 2), which uses the CLIP-ViT L/14 encoder pretrained on large-scale image–text data to reduce attribution biases. Features are extracted from the 12-th transformer block, and fine-tuned with LoRA for attribution alignment. These features are further processed by POLM and FDEM, POLM enhances open-world attribution generalization, while FDEM strengthens fingerprint discriminability.

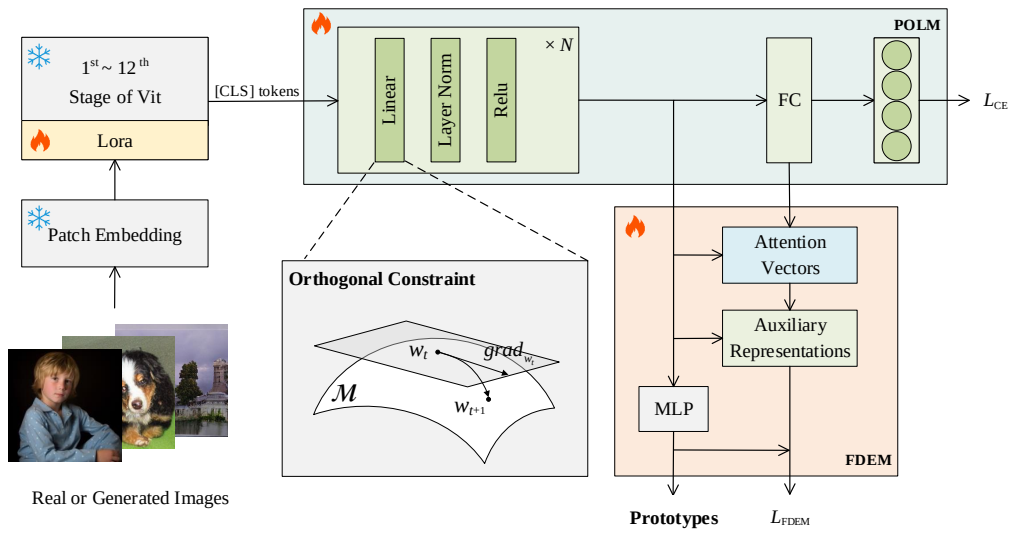


Figure 2: Overview of the proposed QORA framework. CLS tokens are first extracted from a pre-trained CLIP-ViT backbone with LoRA-based fine-tuning. These tokens are transformed by POLM to construct a stable quasi-orthogonal feature space. The FDEM then disentangles and amplifies model-specific fingerprints. After training, class prototypes are obtained by averaging the attribution features produced by FDEM for each category.

4.1 POLM

POLM integrates an orthogonally constrained encoder with a dimension-wise normalized classifier to project pretrained features into a quasi-orthogonal subspace. This space reduces redundancy and amplifies subtle artifact cues, providing a stable foundation for fingerprint disentanglement, enhancement, and sustainable incremental attribution.

Specifically, POLM maps the CLS token $\mathbf{f}_{\text{cls}} \in \mathbb{R}^d$ from the ViT encoder into quasi-orthogonal representations via an N-layer orthogonally-constrained MLP:

$$\mathbf{f}^{(0)} = \mathbf{f}_{\text{cls}}, \quad \mathbf{f}^{(l)} = \text{ReLU}\left(\text{LN}\left(W_{\text{o}}^{(l)}\mathbf{f}^{(l-1)}\right)\right), \quad \mathbf{f}_{\text{o}} = \mathbf{f}^{(N)}, \tag{2}$$

where $l = 1, 2, \ldots, N$, $W_{\text{o}}^{(l)} \in \mathbb{R}^{d\times d}$ denotes the weight of the l-th layer in the encoder, and $\text{LN}(\cdot)$ and $\text{ReLU}(\cdot)$ denote layer normalization and activation. To ensure strict orthogonality, we constrain W_{o} to lie on the Stiefel manifold (Stiefel, 1935):

$$\mathcal{M}_{d,d} = \left\{W_{\text{o}} \in \mathbb{R}^{d\times d} \mid W_{\text{o}} \cdot W_{\text{o}}^{\top} = I_d\right\} \tag{3}$$

Therefore, this constraint can be reformulated as a Riemannian optimization problem:

$$\min_{W_{\text{o}}\in\mathcal{M}_{d,d}} \mathcal{L}(W_{\text{o}}) = \mathcal{L}_{\text{total}} \tag{4}$$

where $\mathcal{L}_{\text{total}}$ denotes the overall loss function of QORA. Meanwhile, to efficiently update W_{o} on the manifold, we compute a skew-symmetric matrix $A = \nabla_{W_{\text{o}}}\mathcal{L}W_{\text{o}}^{\top} - W_{\text{o}}(\nabla_{W_{\text{o}}}\mathcal{L})^{\top}$, where $\nabla_{W_{\text{o}}}\mathcal{L}$ is the gradient of the loss. The weight matrix W_{o} is then updated using the Cayley transform (Li et al., 2020):

$$W_{\text{o}}' = \left(I + \frac{\eta}{2}A\right)^{-1}\left(I - \frac{\eta}{2}A\right)W_{\text{o}} \tag{5}$$

where η is the learning rate. This update guarantees that W_{o}' remains orthogonal and ensures numerical stability throughout training.

In contrast to conventional classifiers that apply class-wise normalization to category vectors, we impose feature-dimension-wise normalization on the classifier weight matrix $W_{fc} \in \mathbb{R}^{C\times d}$:

$$W_{fc}[:, j] \leftarrow \frac{W_{fc}[:, j]}{\|W_{fc}[:, j]\|_2} \quad \forall j \in [1, d] \tag{6}$$

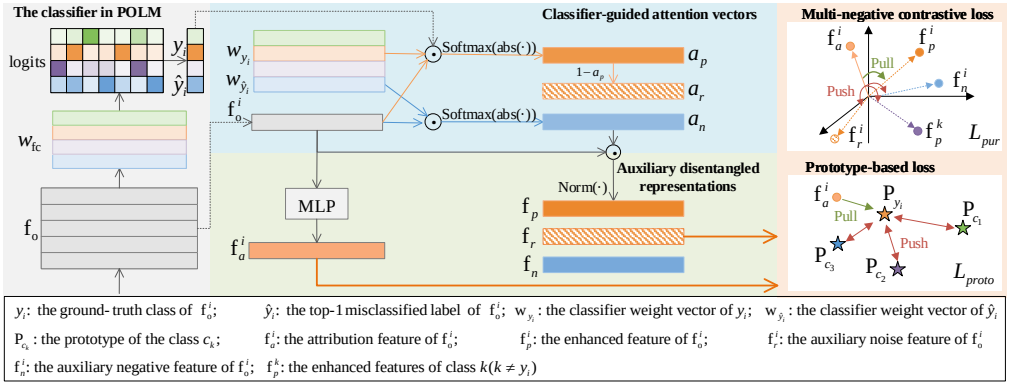


Figure 3: Architecture of FDEM. FDEM disentangles and amplifies generative fingerprints from quasi-orthogonal features produced by POLM. A lightweight MLP projects these features into an attribution space, while classifier weights are used to produce three auxiliary features. Along with class prototypes, these features supervise the attribution learning through contrastive losses.

which balances energy distribution among feature channels, mitigating dominance by high-response channels. This design significantly prevents overfitting to known model categories while enhancing open-set rejection robustness.

4.2 FDEM

FDEM enhances the quasi-orthogonal features from POLM by constructing an attribution subspace that leverages class-specific channel importance to isolate and strengthen generative fingerprints. As shown in Fig. 3, given a sample with feature $\mathbf{f}_o^i$ and label y_i (denoted as $\mathbf{f}_o$ and y), the attention vectors for the ground-truth class y and the top-1 misclassified class $\hat{y}$ are computed as

$$\mathbf{a}_p = \mathrm{Softmax}\left(\frac{|\mathbf{f}_o \odot \mathbf{w}_y|}{\tau}\right), \quad \mathbf{a}_n = \mathrm{Softmax}\left(\frac{|\mathbf{f}_o \odot \mathbf{w}_{\hat{y}}|}{\tau}\right), \tag{7}$$

where $\odot$ denotes element-wise multiplication and τ is a temperature parameter.

These attention maps quantify the channel contributions to correct and confusing predictions. Guided by them, we obtain three disentangled representations:

$$\mathbf{f}_{p/r/n} = \mathrm{Normalize}\big(\mathbf{f}_o \odot \mathbf{a}_{p/r/n}\big), \quad \mathbf{a}_r = 1 - \mathbf{a}_p, \tag{8}$$

where $\mathbf{f}_p$ emphasizes discriminative fingerprints, $\mathbf{f}_r$ suppresses irrelevant noise, and $\mathbf{f}_n$ captures misleading fingerprint artifacts.

Thus, FDEM projects $\mathbf{f}_o$ into an attribution space $\mathbf{f}_a$ using a lightweight MLP and optimizes it with a multi-negative contrastive loss:

$$\mathcal{L}_{\text{pur}} = -\log \frac{\exp(\mathrm{sim}(\mathbf{f}_a, \mathbf{f}_p)/\tau)}{\exp(\mathrm{sim}(\mathbf{f}_a, \mathbf{f}_p)/\tau) + \sum_{f \in (\{\mathbf{f}_r, \mathbf{f}_n\} \cup \{\mathbf{f}_p^j\}_{y_j \neq y_i})} \exp\big(\mathrm{sim}(\mathbf{f}_a, f)/\tau\big)}, \tag{9}$$

where $\mathbf{f}_p$, $\mathbf{f}_r$, and $\mathbf{f}_n$ denote enhanced fingerprints, residuals, and confusing artifacts, respectively, $\{\mathbf{f}_p^j\}_{y_j \neq y_i}$ are fingerprints from other classes, $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity, and τ is a temperature. This formulation aligns $\mathbf{f}_a$ with clean fingerprints while pushing it away from noise, confusions, and unrelated classes.

To further enforce class-level structure, we adopt a prototype-guided loss:

$$\mathcal{L}_{\text{proto}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\mathbf{f}_a^i \cdot \mathbf{p}_{y_i}/\tau\right)}{\sum_{k=1}^{K} \exp\left(\mathbf{f}_a^i \cdot \mathbf{p}_k/\tau\right)} + \frac{1}{K(K-1)} \sum_{j \neq k} \left(\mathbf{p}_j \cdot \mathbf{p}_k\right)^2 \tag{10}$$

where $\mathbf{f}_a^i$ is the normalized attribution feature of sample i, $\mathbf{p}_k$ is the prototype of class k, N is the number of samples, and K is the number of known classes. The first term enforces intra-class

compactness, and the second prevents prototype overlap. Prototypes are updated by exponential moving average:

$$\mathbf{p}_k \leftarrow (1-\lambda)\,\mathbf{p}_k + \lambda\,\overline{\mathbf{f}_a^k} \tag{11}$$

where $\overline{\mathbf{f}_a^k}$ is the batch-wise mean attribution feature of class k, and $\lambda \in (0,1)$ is the momentum factor.

Finally, the total training objective with the classifier cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$ from POLM is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{pur}} + \mathcal{L}_{\mathrm{proto}} \tag{12}$$

During inference, attribution is performed by comparing $\mathbf{f}_a$ to stored prototypes $\mathbf{p}_k$.

4.3 Sustainable Incremental Learning

To integrate new generator classes while preserving performance on previously learned ones, we adopt a memory-efficient incremental learning strategy. In each session t, 20 samples per past class are stored in a replay buffer $\mathcal{D}_{t-1}^M$, covering $\mathcal{C}_{0:t-1}^L$. This buffer is then combined with the current session's labeled data $\mathcal{D}_t^L$ of class set $\mathcal{C}_t^L$ to form the updated buffer $\mathcal{D}_t^M$.

During incremental updates, the CLIP-ViT backbone, LoRA parameters, and the POLM encoder are kept frozen. Only the POLM classifier and the MLP in FDEM are updated. Class-wise mean features are first computed for both previously learned classes $k \in \mathcal{C}_{0:t-1}^L$ in $\mathcal{D}_t^M$, and new classes $n \in \mathcal{C}_t^L$ in $\mathcal{D}_t^L$:

$$\overline{\mathbf{f}_o^k} = \frac{1}{|\mathcal{D}_{t,k}^M|} \sum_{\mathbf{f}_{o,i} \in \mathcal{D}_{t,k}^M} \mathbf{f}_{o,i}, \quad \overline{\mathbf{f}_o^n} = \frac{1}{|\mathcal{D}_{t,n}^L|} \sum_{\mathbf{f}_{o,i} \in \mathcal{D}_{t,n}^L} \mathbf{f}_{o,i} \tag{13}$$

where $\mathcal{D}_{t,k}^M$ and $\mathcal{D}_{t,n}^L$ denote the sets of samples belonging to class k and n, respectively.

For each new class n, the nearest previously known class k^* is identified by

$$k^* = \arg\min_{k \in \mathcal{C}_{0:t-1}^L} \|\overline{\mathbf{f}_o^n} - \overline{\mathbf{f}_o^k}\|_2 \tag{14}$$

and the classifier weight for class n is initialized as

$$\mathbf{w}_n \leftarrow \mathbf{w}_{k^*} \tag{15}$$

Following initialization, incremental training is carried out using the same total loss $\mathcal{L}_{\text{total}}$ as in the initial training phase. After training, updated prototypes are retained for future attribution.

5 Experiment

In this section, we provide a comprehensive evaluation of the proposed QORA framework. We first outline the experimental setups. Then we assess static open-world attribution followed by extensive ablation studies. Finally, we evaluate QORA in a five-session incremental learning scenario, demonstrating its ability of sustainable.

5.1 Experimental Setups

Datasets. We evaluate QORA on two static open-world attribution benchmarks: OSMA (Yang et al., 2023), a GAN-based dataset covering 53 GANs with diverse seeds and architectures, and GenImage (Cioni et al., 2025), a diffusion-based benchmark with ImageNet classes and models from eight diffusion generators. To simulate the continuous emergence of generators, we construct a sustainable open-world benchmark from these datasets. As detailed in in Appendix A, it includes one real-image class and 23 generator classes split into five sessions, each introducing four new seen classes while the unseen set comprises remaining models.

Evaluation Metrics. Following established protocols (Yang et al., 2023; Cioni et al., 2025), we evaluate QORA with three metrics: classification accuracy (Acc.) for closed-set attribution of seen

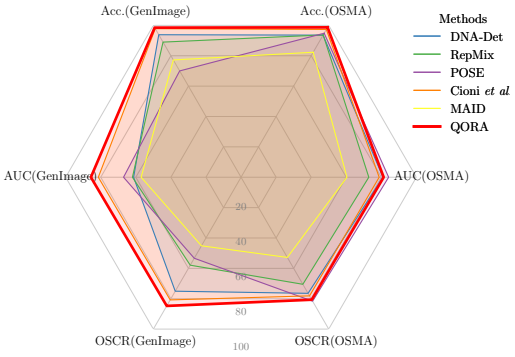


Figure 4: QORA outperforms baselines on OSMA and GenImage in both closed- and open-set metrics, highlighting strong generalization.

Table 1: Performance comparison on the diffusion-based model attribution benchmark GenImage. Results are averaged over five splits. The best performance is shown in bold, and the second-best is underlined.

Method	Acc. (%)	AUC (%)	OSCR (%)
DNA-Det	93.83	61.27	75.08
RepMix	88.98	61.93	57.92
POSE	70.00	67.00	53.35
Cioni *et al.*	<u>97.82</u>	<u>81.39</u>	<u>80.78</u>
MAID	77.23	56.98	45.16
QORA	**98.51**	**85.63**	**84.74**

Table 2: Performance comparison on OSMA. Results are averaged over five splits. The highest score for each metric is shown in bold, and the second-best score is underlined.

Method	Acc.(%)	Unseen Seed		Unseen Arch.		Unseen Data	
		AUC(%)	OSCR(%)	AUC(%)	OSCR(%)	AUC(%)	OSCR(%)
PRNU	55.27	**69.20**	49.16	70.02	49.49	67.68	**48.57**
Yu *et al.*	85.71	53.14	50.99	69.04	64.17	78.79	72.20
DCT-CNN	86.16	55.46	52.68	72.56	67.43	72.87	67.57
DNA-Det	93.56	61.46	59.34	80.93	76.45	<u>66.14</u>	63.27
RepMix	93.69	54.70	53.26	72.86	70.49	78.69	76.02
POSE	94.81	<u>68.15</u>	**67.25**	**84.17**	**81.62**	88.24	85.64
Cioni *et al.*	<u>97.29</u>	54.15	54.00	78.78	78.12	90.60	89.52
MAID	82.30	51.06	46.02	60.40	52.81	**59.04**	<u>52.01</u>
QORA	**98.68**	62.56	<u>62.23</u>	<u>81.34</u>	<u>80.66</u>	80.68	80.08

generators, AUC for open-set detection of unseen generators, and OSCR for jointly assessing attribution accuracy and rejection quality in open-world conditions.

Baseline Methods. We compare QORA against representative attribution baselines spanning both closed- and open-world settings, including PRNU (Marra et al., 2019a), Yu *et al.* (Yu et al., 2019a), DCT-CNN (Frank et al., 2020), DNA-Det (Yang et al., 2022), RepMix (Bui et al., 2022), POSE (Yang et al., 2023), Cioni *et al.* (Cioni et al., 2025), and MAID (Zhu et al., 2025).

Implementation Details. We fine-tune CLIP-ViT L/14 with LoRA with a rank of 16 per adapter, use a one-layer MLP as the POLM encoder, and update FDEM prototypes with a momentum coefficient λ of 0.995. Models are trained for 30 epochs on one-quarter of the training data per class using Adam with cosine annealing, where the initial learning rate is set to 1×10^{-4}. All experiments are implemented in PyTorch 2.0 and run on an NVIDIA RTX 3090.

5.2 Evaluation of Open-Set Model Attribution

We evaluate QORA against baselines on OSMA and GenImage benchmarks, with overall results summarized in Fig. 4. QORA consistently surpasses prior methods in both closed-set and open-set performance, showing strong generalization across architectures.

Comparison with SOTA on GAN-generated images. OSMA evaluates three settings: unseen seeds, unseen architectures, and unseen training data. Strong performance on the first two indicates sensitivity to model-intrinsic fingerprints, while lower performance on unseen data suggests reduced reliance on content semantics. MAID's results were obtained by retraining its open-source implementation under the standard protocol, whereas other baselines are reported from the official OSMA benchmark (Yang et al., 2023). As shown in Table 2, QORA achieves a closed-set attribution accuracy of 98.68%, surpassing the previous best by 1.39%. For open-set evaluation, QORA

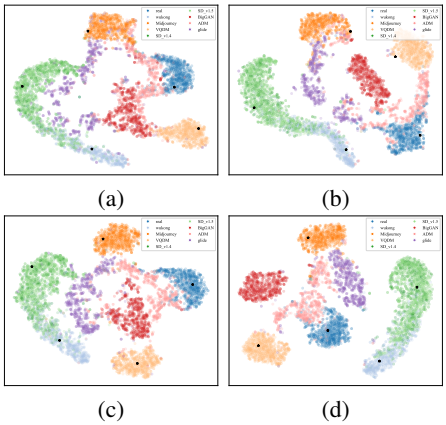


Figure 5: t-SNE under ablations: (a) w/o orthogonality or normalization, (b) orthogonality only, (c) normalization only, (d) full setup.

Table 3: Ablation of loss functions over five GenImage splits. The first three rows measure attribution accuracy using attribution representations, while the last row reports classifier performance in POLM with pure cross-entropy loss $\mathcal{L}_{\text{CE}}$. The best performance is shown in bold.

Losses	Acc. (%)	AUC (%)	OSCR (%)
full model	**98.51**	**85.63**	**84.74**
w/o $\mathcal{L}_{\text{pur}}$	98.27	80.33	79.22
w/o $\mathcal{L}_{\text{proto}}$	96.90	82.29	80.19
w/o $\mathcal{L}_{\text{pur}}$ and $\mathcal{L}_{\text{proto}}$	97.45	81.77	80.18

ranks second on unseen architectures and seeds, trailing POSE, but attains 80.68% AUC and 80.08% OSCR on unseen data, below POSE, which highlights its stronger emphasis on model-intrinsic fingerprints rather than semantic variations.

Comparison with SOTA on diffusion-generated images. As shown in Table 1, on GenImage, closed-set GAN-specific baselines are excluded due to their limited generalization capability. QORA achieves the highest closed-set accuracy of 98.51%, surpassing Cioni *et al.* by 0.69%. For open-set recognition, it achieves 85.63% AUC and 84.74% OSCR, yielding absolute gains of 4.24% and 3.96% over the previous best. These results demonstrate QORA's effectiveness in capturing discriminative fingerprints of diffusion models.

5.3 Ablation Studies

Ablation Study on POLM. We assess the contributions of the orthogonality constraint in the encoder and the dimension-wise normalization in the classifier on split-1 of GenImage. Four configurations are compared: (a) neither constraint, (b) orthogonality only, (c) normalization only, and (d) both constraints (full QORA). t-SNE visualizations of the attribution features (Fig. 5) show that (a) produces scattered distributions for seen categories and shows clear confusion between seen and unseen features, (b) improves inter-class separation for seen categories, (c) reduces overlap between seen and unseen samples, and (d) achieves well-separated clusters and distinct dispersion of unseen samples, demonstrating enhanced open-set rejection and a stable feature space.

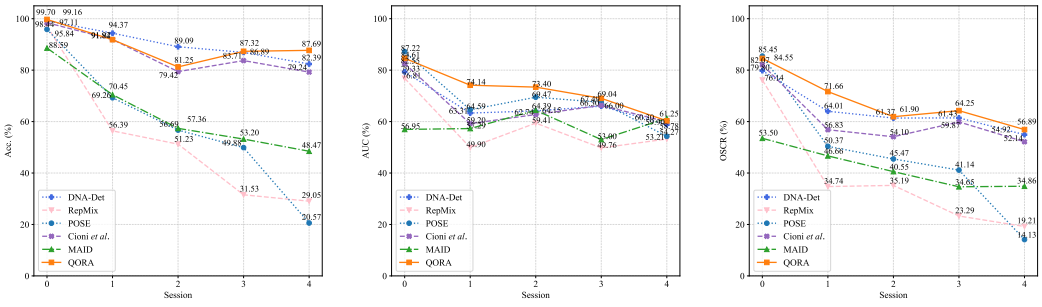
Ablation on Loss Components in FDEM. Table 3 evaluates the prototype-guided loss $\mathcal{L}_{\text{proto}}$ and purification contrastive loss $\mathcal{L}_{\text{pur}}$ across all five GenImage splits. Removing either loss degrades performance: without $\mathcal{L}_{\text{proto}}$, closed-set accuracy drops from 98.51% to 96.90%, highlighting its role in aligning features with class prototypes; without $\mathcal{L}_{\text{pur}}$, open-set AUC and OSCR decrease by 5.30% and 5.52%, showing its importance in purifying model-specific fingerprints. When both losses are removed, reliance on the POLM classifier alone leads to further degradation. These results confirm that FDEM is crucial for learning discriminative, generalizable representations and robust open-world attribution under dynamic conditions.

5.4 Evaluation on Sustainable Open-World Attribution

We evaluate QORA on the five-session SOW-GMA benchmark to assess scalability and adaptability under realistic open-world conditions, comparing it with five baselines: DNA-Det (Yang et al., 2022), RepMix (Bui et al., 2022), POSE (Yang et al., 2023), Cioni *et al.* (Cioni et al., 2025), and MAID (Zhu et al., 2025). All models are trained with official implementations, initializing each incremental session from the previous checkpoint.

Table 4: Performance comparison between Session 0 and Session 4 for different methods. The best performance is shown in bold, and the second-best is underlined. Red arrows and text indicate the increase (↑) or decrease (↓) from Session 0 to Session 4.

Method	Acc. (%)		AUC (%)		OSCR (%)	
	Session 0	Session 4	Session 0	Session 4	Session 0	Session 4
DNA-Det	<u>99.16</u>	<u>82.39</u> ↓ 16.77	79.33	58.78 ↓ 20.55	79.80	<u>54.92</u> ↓ 24.88
RepMix	97.11	29.05 ↓ 68.06	76.81	53.21 ↓ 23.60	76.14	19.21 ↓ 56.93
POSE	95.84	20.57 ↓ 75.27	**87.22**	54.27 ↓ 32.95	**85.45**	14.13 ↓ 71.32
Cioni *et al.*	98.44	79.24 ↓ 19.20	82.55	59.90 ↓ 22.65	82.07	52.14 ↓ 29.93
MAID	88.59	48.47 ↓ 40.12	56.95	**61.25** ↑ 4.30	53.50	34.86 ↓ 18.64
QORA	**99.70**	**87.69** ↓ 12.01	<u>84.61</u>	<u>60.30</u> ↓ 24.31	<u>84.55</u>	**56.89** ↓ 27.66



(a) (b) (c)

Figure 6: Comparison of four attribution methods over five incremental sessions shows that QORA consistently outperforms others in (a) closed-set accuracy, (b) open-set AUC, and (c) open-set OSCR, demonstrating its superior scalability and stability in open-world incremental learning.

Table 4 reports initial and final session performance. QORA achieves 99.70% closed-set accuracy initially and maintains 87.69% in Session 4, representing the smallest decline with 12.01% among all methods. In contrast, POSE, RepMix, and MAID show sharp degradation of 75.27%, 68.06%, and 40.12%. DNA-Det and Cioni drop to 82.39% and 79.24%, remaining 5–8% below QORA. For open-set detection, QORA's initial AUC and OSCR are slightly lower than POSE's but surpass it by Session 4, with gains of 6.03% in AUC and 42.76% in OSCR. MAID shows large AUC fluctuations, as shown in Fig. 6 (b), whereas QORA consistently keeps AUC above 60%, while other baselines decline or fluctuate. Fig. 6 shows metric trends across sessions. QORA maintains balanced, robust performance in closed- and open-set, effectively integrating new classes while preserving prior knowledge and rejecting unseen generators, demonstrating practical suitability for real-world incremental attribution.

6 Conclusion

In this paper, we present QORA, a sustainable framework for open-world generative model attribution. Unlike prior methods hindered by emerging models, QORA integrates accurate closed-set attribution, robust open-set rejection, and efficient class-incremental learning with low memory overhead. POLM leverages Stiefel manifold optimization to construct a quasi-orthogonal space that suppresses redundancy and enhances generalization, while FDEM disentangles and strengthens model-specific fingerprints via classifier-guided attention and contrastive learning. A lightweight incremental strategy further supports rapid adaptation without full retraining. Experiments on GAN- and diffusion-based benchmarks show that QORA achieves state-of-the-art attribution accuracy and preserves strong open-set robustness across sessions, highlighting its scalability and real-world applicability.

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. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2024.03.002. URL https://www.sciencedirect.com/science/article/pii/S0167865524000709.

Michael Albright and Scott McCloskey. Source generator attribution via inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Michael Albright, Scott McCloskey, and ACST Honeywell. Source generator attribution via inversion. In *CVPR workshops*, volume 8, pp. 3, 2019.

Diane Bartz and Krystal Hu. Openai, google, others pledge to watermark ai content for safety, white house says. URL https://www.reuters.com/technology/openai-google-others-pledge-watermark-ai-content-safety-white-house-2023-07-21/.

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

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.

Mo Chen, Jessica Fridrich, Miroslav Goljan, and Jan Lukas. Determining image origin and integrity using sensor noise. *IEEE Transactions on Information Forensics and Security*, 3(1):74–90, 2008. doi: 10.1109/TIFS.2007.916285.

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, 2025.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Shengbang Fang, Tai D Nguyen, and Matthew C Stamm. Open set synthetic image source attribution. *arXiv preprint arXiv:2308.11557*, 2023a.

Shengbang Fang, Tai D Nguyen, and Matthew C Stamm. Open set synthetic image source attribution. *arXiv preprint arXiv:2308.11557*, 2023b.

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*, ICML'20. JMLR.org, 2020.

Caili Gao, Qisheng Xu, Peng Qiao, Kele Xu, Xifu Qian, and Yong Dou. Adapter-based incremental learning for face forgery detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4690–4694. IEEE, 2024.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

Yonghyun Jeong, Doyeon Kim, Seungjai Min, Seongho Joe, Youngjune Gwon, and Jongwon Choi. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 48–57, 2022a.

Yonghyun Jeong, Doyeon Kim, Youngmin Ro, and Jongwon Choi. Frepgan: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1060–1068, 2022b.

Yonghyun Jeong, Doyeon Kim, Youngmin Ro, Pyounggeon Kim, and Jongwon Choi. Fingerprintnet: Synthesized fingerprints for generated image detection. In *European Conference on Computer Vision*, pp. 76–94. Springer, 2022c.

Zhong Ji, Zhishen Hou, Xiyao Liu, Yanwei Pang, and Xuelong Li. Memorizing complementation network for few-shot class-incremental learning. *IEEE Transactions on Image Processing*, 32: 937–948, 2023.

Jun Li, Li Fuxin, and Sinisa Todorovic. Efficient riemannian optimization on the stiefel manifold via the cayley transform. *arXiv preprint arXiv:2002.01113*, 2020.

Meiling Li, Zhenxing Qian, and Xinpeng Zhang. Regeneration based training-free attribution of fake images generated by text-to-image generative models. *ArXiv*, abs/2403.01489, 2024a. URL https://api.semanticscholar.org/CorpusID:268247911.

Meiling Li, Zhenxing Qian, and Xinpeng Zhang. Regeneration based training-free attribution of fake images generated by text-to-image generative models. *arXiv preprint arXiv:2403.01489*, 2024b.

Fengyuan Liu, Haochen Luo, Yiming Li, Philip Torr, and Jindong Gu. Which model generated this image? a model-agnostic approach for origin attribution. In *European Conference on Computer Vision*, pp. 282–301. Springer, 2024.

Zhengzhe Liu, Xiaojuan Qi, and Philip H.S. Torr. Global texture enhancement for fake face detection in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8057–8066, 2020. doi: 10.1109/CVPR42600.2020.00808.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pp. 506–511. IEEE, 2019a.

Francesco Marra, Cristiano Saltori, Giulia Boato, and Luisa Verdoliva. Incremental learning for the detection and classification of gan-generated images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–6. IEEE, 2019b.

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 Computer Vision and Pattern Recognition Conference*, pp. 3040–3050, 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*, pp. 24480–24489, 2023.

Kun Pan, Yifang Yin, Yao Wei, Feng Lin, Zhongjie Ba, Zhenguang Liu, Zhibo Wang, Lorenzo Cavallaro, and Kui Ren. Dfil: Deepfake incremental learning by exploiting domain-invariant forgery clues. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8035–8046, 2023.

Ekta Prashnani, Michael Goebel, and BS Manjunath. Generalizable deepfake detection with phase-based motion analysis. *IEEE Transactions on Image Processing*, 2024.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pp. 86–103. Springer, 2020.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, CCS '23, pp. 3418–3432, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700507. doi: 10.1145/3576915.3616588. URL https://doi.org/10.1145/3576915.3616588.

Sergey Sinitsa and Ohad Fried. Deep image fingerprint: Towards low budget synthetic image detection and model lineage analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4067–4076, January 2024.

Eduard Stiefel. *Richtungsfelder und Fernparallelismus in n-dimensionalen Mannigfaltigkeiten*. PhD thesis, ETH Zurich, 1935.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12105–12114, 2023. doi: 10.1109/CVPR52729.2023.01165.

Shuai Tang, Peisong He, Haoliang Li, Wei Wang, Xinghao Jiang, and Yao Zhao. Towards extensible detection of ai-generated images via content-agnostic adapter-based category-aware incremental learning. *IEEE Transactions on Information Forensics and Security*, 2025.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 24–25, 2020a.

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*, pp. 953–962, 2023a.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695–8704, 2020b.

Xuan Wang, Zhong Ji, Yunlong Yu, Yanwei Pang, and Jungong Han. Model attention expansion for few-shot class-incremental learning. *IEEE Transactions on Image Processing*, 2024.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023b.

Qiang Xu, Shan Jia, Xinghao Jiang, Tanfeng Sun, Zhe Wang, and Hong Yan. Mdtl-net: Computer-generated image detection based on multi-scale deep texture learning. *Expert Systems with Applications*, 248:123368, 2024. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2024.123368. URL https://www.sciencedirect.com/science/article/pii/S0957417424002331.

Tianyun Yang, Juan Cao, Qiang Sheng, Lei Li, Jiaqi Ji, Xirong Li, and Sheng Tang. Learning to disentangle gan fingerprint for fake image attribution. *arXiv preprint arXiv:2106.08749*, 2021.

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, 2022.

Tianyun Yang, Danding Wang, Fan Tang, Xinying Zhao, Juan Cao, and Sheng Tang. Progressive open space expansion for open-set model attribution. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15856–15865, 2023. URL https://api.semanticscholar.org/CorpusID:257496280.

Ning Yu, Larry Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7555–7565, 2019a. doi: 10.1109/ICCV.2019.00765.

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, 2019b.

Junbin Zhang, Yixiao Wang, Hamid Reza Tohidypour, and Panos Nasiopoulos. An efficient frequency domain based attribution and detection network. *IEEE Access*, 13:19909–19921, 2025. doi: 10.1109/ACCESS.2025.3534829.

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

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

Nan Zhong, Haoyu Chen, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang. Beyond generation: A diffusion-based low-level feature extractor for detecting ai-generated images. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 8258–8268, June 2025.

Luyu Zhu, Kai Ye, Jiayu Yao, Chenxi Li, Luwen Zhao, Yuxin Cao, Derui Wang, and Jie Hao. Maid: Model attribution via inverse diffusion. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Mingjian Zhu, Hanting Chen, Mouxiao Huang, Wei Li, Hailin Hu, Jie Hu, and Yunhe Wang. Gendet: Towards good generalizations for ai-generated image detection. *ArXiv*, abs/2312.08880, 2023. URL `https://api.semanticscholar.org/CorpusID:266210505`.

A Details of datasets

We evaluate QORA on two **static open-world attribution** benchmarks:

- OSMA Yang et al. (2023): A GAN-based benchmark built on seven real-image datasets, each paired with two GANs for training. Its unseen set includes 53 GANs held out under three conditions: same architecture/dataset with different seeds, novel architectures, and novel training datasets.
- GenImage Cioni et al. (2025): A diffusion-based attribution dataset. Its known classes comprise real ImageNet images and outputs from four diffusion models. Its unseen set consists of samples generated by four additional diffusion models not used during training.

Both benchmarks are evaluated using five train/test splits, with each split varying the composition of seen and unseen generative models to ensure robust generalization testing.

To simulate real-world conditions where generative models continually emerge, we construct a **sustainable open-world attribution** benchmark based on the two datasets described above. As detailed in Table 5, the benchmark includes the real-image class and 20 generative model classes, chronologically divided into five incremental sessions from 2018 to 2022. Session 0 serves as the initial training pshase for the SOW-GMA task. In each session, four newly introduced generative models serve as the session-specific *seen* classes for training. Meanwhile, the *unseen* set comprises all generative models not yet encountered in the current or any previous session, along with three fixed unseen models, SNGAN, S3GAN, and Wav2Lip, that are consistently included in the open-set across all sessions.

As shown in Table 6, the training and testing protocol for each session t is defined as follows:

Table 5: Chronological split of seen and unseen generative models for SOW-GMA task.

Session	Year	Seen Models	Unseen Models
0	2018	Real, StarGAN, ProGAN, MMDGAN, BigGAN	SNGAN, S3GAN, Wav2Lip + $\mathrm{Seen}_{1,2,3,4}$
1	2019	SAGAN, FSGAN, AttGAN, StyleGAN	SNGAN, S3GAN, Wav2Lip + $\mathrm{Seen}_{2,3,4}$
2	2020	FaceSwap, StyleGAN2, ContraGAN, FaceShifter	SNGAN, S3GAN, Wav2Lip + $\mathrm{Seen}_{3,4}$
3	2021	StyleGAN3, InfoMaxGAN, ADM, Glide	SNGAN, S3GAN, Wav2Lip + Seen_{4}
4	2022	Wukong, Midjourney, Stable Diffusion v1.4, VQDM	SNGAN, S3GAN, Wav2Lip

Table 6: Data Split for training and testing process.

		Data Group
Train	**Closed**	Seen_t, Memory_t
Test	**Closed**	Seen_t, Memory_t
	Open	Unseen_t

- Training: 4K samples are used for each newly introduced class in session t, and 20 exemplars are retained for each previously seen class in a memory set denoted as Memory_t.
- Testing: The closed-set includes all classes in $\mathrm{Seen}_t \cup \mathrm{Memory}_t$, while the open-set consists of Unseen_t.