SREM: Semantic-space Reconstruction Evidence Mining for Generalizable AI-Generated Image Detection

Anwei Luo[1] , **Jinsong Hu**[1] , **Yi Tu**[2] , **Chenqi Kong**[3] , **Dan Ma**[4*]
[1]Jiangxi University of Finance and Economics
[2]Shanghai Jiaotong University
[3]National University of Singapore
[4]Xinjiang University
*Corresponding author
luoanwei@jxufe.edu.cn, madan@xju.edu.cn

Abstract

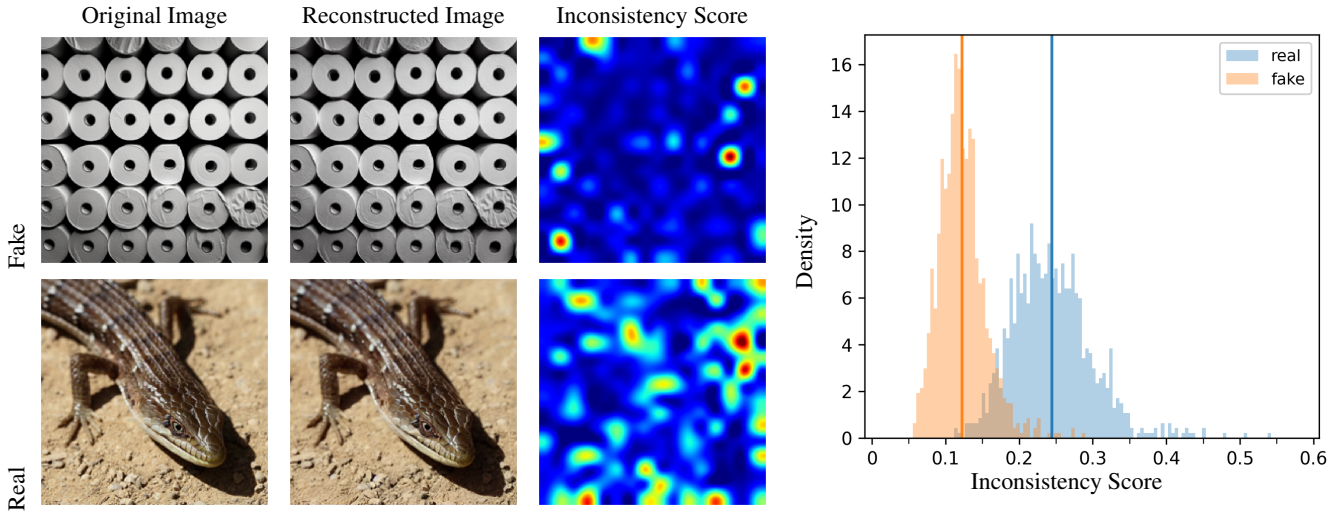
Diffusion-based image synthesis has substantially narrowed the perceptual gap between synthetic and real images, making reliable AI-generated image detection increasingly challenging. Reconstruction-based detectors compare an image with its reconstructed counterpart, but many existing methods characterize this discrepancy primarily through pixel- or latent-space reconstruction errors. These error-centric signals can be sensitive to generator and domain variations, and may not fully capture structured changes induced by reconstruction. In this paper, we propose SREM, a Semantic-space Reconstruction Evidence Mining framework for generalizable AI-generated image detection. Instead of relying solely on low-level reconstruction-error signals, SREM mines reconstruction-induced discrepancies in the CLIP semantic representation space. Specifically, SREM models fine-grained patch-wise inconsistency responses across multiple CLIP layers and characterizes image-level semantic shifts from paired global representations. By jointly exploiting local inconsistency patterns and global semantic shifts, SREM extracts complementary semantic-space reconstruction evidence for detection. Experiments on multiple benchmarks demonstrate the effectiveness and generalization ability of the proposed method.

1 Introduction

Recent advances in diffusion-based image generation have enabled the synthesis of highly realistic visual content, substantially narrowing the perceptual gap between synthetic and real images [Ho *et al.*, 2020; Rombach *et al.*, 2022; Podell *et al.*, 2023]. Meanwhile, the potential misuse of such models for misinformation, image forgery, and other malicious purposes has raised growing concerns about the trustworthiness of visual content. To address this issue, a growing body of work has been devoted to developing detectors for AI-generated images [Corvi *et al.*, 2023; Wang *et al.*, 2020]. Despite substantial progress, existing methods achieve strong performance mainly on images produced by known generators, while their robustness often degrades when applied to unseen generative models.

Recently, reconstruction-based detection has emerged as a promising direction for addressing this challenge. For example, DIRE [Wang *et al.*, 2023] is built on the observation that images generated by diffusion models can often be reconstructed more faithfully than real images using a pre-trained diffusion model. This reconstruction asymmetry provides transferable evidence for detecting synthetic images. Subsequent studies have extended this direction by exploring latent reconstruction errors [Luo *et al.*, 2024], time-step reconstruction errors [Shen *et al.*, 2025], edit-induced reconstruction-error shifts [Jiang *et al.*, 2025], and other reconstruction-based cues. However, the effectiveness of reconstruction-based algorithms generally depends on whether the reconstruction process exposes a sufficiently detectable gap between real and synthetic images. As generative models continue to improve, this gap may become increasingly subtle, making it necessary to develop more effective mechanisms for capturing fine-grained reconstruction-induced differences.

Previous work [Ojha *et al.*, 2023; Liu *et al.*, 2024] suggests that real and generated images become more separable in the CLIP feature space. Meanwhile, recent AI-generated image detection studies show that intermediate CLIP representations also contain complementary fine-grained visual information [Koutlis and Papadopoulos, 2024; Cheng *et al.*, 2025]. Since large-scale pretrained vision-language models encode rich semantic knowledge, we hypothesize that the reconstruction-induced inconsistency gap in such a semantic representation space can serve as an effective forensic cue for detecting synthetic images. We empirically validate this hypothesis in Fig. 1. Specifically, we compute patch-wise similarities between the CLIP representations of each original image and its reconstructed counterpart, and use the normalized complement of the similarity as a patch-wise inconsistency score. As shown in Fig. 1(a), real images exhibit strong reconstruction-induced inconsistency responses across many local regions, whereas synthetic images show high responses only in a few areas. This observation indicates that, after re-



(a) Reconstruction-induced Inconsistency Visualization

(b) Reconstruction-induced Inconsistency Distribution

Figure 1: Reconstruction-induced inconsistency analysis in the CLIP feature space. We compute patch-wise similarities between the CLIP representations of the original image and its reconstructed counterpart, and use the normalized complement of the similarity as the inconsistency score. Images are sampled from the GenImage [Zhu *et al.*, 2023]. (a) Visualization of patch-wise inconsistency scores. (b) Statistical distribution of patch-wise inconsistency scores over 2,000 randomly selected images.

construction, synthetic images tend to preserve their local semantic representations more consistently, while real images undergo more evident local semantic changes. The statistical distribution in Fig. 1(b) further confirms this gap, showing that the patch-wise inconsistency scores of real and synthetic images are clearly distinguishable.

Building on these findings, we propose a **S**emantic-space **R**econstruction **E**vidence **M**ining (SREM) framework for generalizable AI-generated image detection. Given an input image, SREM first constructs an original–reconstruction pair through a diffusion-based reconstruction process and maps both images into a shared semantic representation space using a frozen CLIP encoder. To exploit reconstruction-induced inconsistency in this space, SREM introduces a **Multi-level Patch Inconsistency Modeling** (MPIM) module. MPIM estimates patch-wise inconsistency responses across multiple CLIP layers, which are further aggregated to mine local forensic evidence reflecting fine-grained reconstruction-induced semantic changes. SREM further designs a **Global Semantic Shift Modeling** (GSSM) module to characterize image-level reconstruction behavior. Specifically, GSSM computes the feature difference between the paired global representations, using high-level semantic shifts as global forensic evidence. Finally, these complementary local and global cues are fused to produce the final prediction. Extensive experiments and analyses on multiple benchmarks demonstrate the effectiveness of the proposed method. The main contributions are summarized as follows:

- We reveal a reconstruction-induced inconsistency gap in the CLIP semantic representation space. Through qualitative visualization and statistical analysis, we show that real and synthetic images exhibit distinct patch-wise inconsistency patterns after reconstruction, motivating semantic-space modeling of reconstruction evidence for AI-generated image detection.
- A Semantic-space Reconstruction Evidence Mining framework is proposed to jointly model local reconstruction-induced inconsistency and global semantic shifts in the CLIP feature space. The proposed framework integrates these complementary cues for final prediction.
- Extensive experiments on GenImage, DRCT-2M, and Chameleon validate the effectiveness and generalization capability of SREM across diverse generators and evaluation settings.

2 Related Work

2.1 General Detectors for AI-Generated Images

Early related works primarily focused on detecting GAN-generated images [Frank *et al.*, 2020; Wang *et al.*, 2020]. However, detectors trained to capture GAN-specific artifacts often transfer poorly to diffusion-generated images, as GANs and diffusion models differ substantially in their synthesis mechanisms and resulting image statistics [Corvi *et al.*, 2023]. This limitation has motivated following studies to explore more transferable detection cues beyond generator-specific artifacts. Recently, some works suggest that semantic priors from large-scale pre-trained models can improve cross-generator generalization. UnivFD [Ojha *et al.*, 2023] demonstrated that frozen CLIP features can generalize well across different generative model families with only a simple linear classifier. Sha et al. [Sha *et al.*, 2023] further incorporated BLIP-generated captions [Li *et al.*, 2022] into a CLIP-based multimodal detector, while Wu et al. [Wu *et al.*, 2025] explored language-guided contrastive learning for synthetic image detection. Instead of relying on generic CLIP semantics, FakeInversion [Cazenavette *et al.*, 2024] leverages

Stable Diffusion inversion features to detect images from unseen text-to-image models. These methods indicate that pretrained semantic representations provide more transferable evidence than generator-specific low-level artifacts. Nevertheless, relying only on image-level semantic representations may be insufficient, as realistic AI-generated images can still contain subtle local discrepancies that are not fully captured by global semantics.

To capture such fine-grained evidence, recent studies have explored local patch, intermediate-feature, pixel-relation, and frequency-domain cues. Xi et al. [Xi *et al.*, 2023] combined RGB features with residual cues through a dual-stream cross-attention architecture, while GLFF [Ju *et al.*, 2023] fused multi-scale global representations with informative local patches for AI-synthesized image detection. RINE [Koutlis and Papadopoulos, 2024] further showed that intermediate CLIP encoder blocks contain useful low-level visual information that is complementary to high-level semantic representations. From a local statistics perspective, Zhong et al. [Zhong *et al.*, 2023] modeled inter-pixel correlation differences across texture regions, and NPR [Tan *et al.*, 2024] exploited neighboring pixel relationships induced by up-sampling operations to capture generalized structural artifacts. In addition, frequency-aware methods such as FatFormer [Liu *et al.*, 2024] have incorporated frequency-adaptive representations to expose subtle forgery traces that are difficult to observe directly in RGB space. Overall, these methods indicate that fine-grained local and frequency cues can complement image-level semantic representations. However, many of these cues remain closely tied to low-level statistics and may still be sensitive to post-processing and distribution shifts.

2.2 Reconstruction-Based Detectors for AI-Generated Images

Reconstruction-based detection has become an important direction for detecting AI-generated images. Instead of relying solely on artifacts directly observed in the input image, these methods compare an image with its reconstructed counterpart produced by a pre-trained generative model and use their discrepancy as detection evidence. DIRE [Wang *et al.*, 2023] first formulated diffusion reconstruction error as a discriminative cue, based on the observation that diffusion-generated images are often easier to reconstruct than real images. Following this idea, LaRE [Luo *et al.*, 2024] measures reconstruction error in the latent space, FIRE [Chu *et al.*, 2025] introduces frequency-guided error modeling, and TRE [Shen *et al.*, 2025] exploits reconstruction errors across diffusion time steps. EIRES [Jiang *et al.*, 2025] further converts reconstruction behavior into an edit-induced error shift, while SARE [Kang *et al.*, 2025] measures semantic shifts between an image and its caption-guided reconstruction. In addition, MRCL [Zhuang *et al.*, 2025] and DRCT [Chen *et al.*, 2024] use reconstruction residuals or reconstructed hard samples to improve detector training. These methods demonstrate the effectiveness of reconstruction behavior for AI-generated image detection. However, they mainly focus on constructing or selecting reconstruction-derived signals, such as pixel/latent errors, frequency responses, temporal error sequences, semantic shifts, or hard reconstructed samples.

Different from these signal-level extensions, SREM focuses on model-level residual evidence modeling. Rather than designing another explicit reconstruction error signal, SREM enhances the detector's ability to compare paired original and reconstructed images in the CLIP feature space. It represents reconstruction-induced discrepancies as global semantic residuals and local patch-level discrepancy patterns, enabling residual evidence to be modeled beyond raw error magnitude and used at both semantic and fine-grained levels. Compared with SARE [Kang *et al.*, 2025], which uses caption-guided image-space residuals, SREM compares paired features in the CLIP space, whereas MPIM differs from RINE [Koutlis and Papadopoulos, 2024] by modeling original–reconstruction patch discrepancies rather than single-image intermediate CLIP features.

3 Proposed Method

This section first presents an overview of the proposed SREM framework in section 3.1. We then describe the Multi-level Patch Inconsistency Modeling module in section 3.2 and the Global Semantic Shift Modeling module in section 3.3, which respectively mine local and global reconstruction-induced forensic evidence in the semantic space. Finally, we introduce the training objective in section 3.4.

3.1 Overview

Fig. 2 illustrates the overall architecture of SREM. Given an input image x_{ori}, we obtain its reconstructed counterpart x_{rec} using a pretrained diffusion model, following the reconstruction protocol of DRCT [Chen *et al.*, 2024]. The original and reconstructed images are treated as paired inputs and fed into a frozen CLIP encoder [Radford *et al.*, 2021], which maps them into a shared semantic representation space. From the paired representations, we extract global class-token features and patch features from multiple CLIP layers.

SREM mines complementary reconstruction evidence from both local and global perspectives. In the local branch, the Multi-level Patch Inconsistency Modeling (MPIM) module estimates and aggregates patch-wise inconsistencies across multiple CLIP layers to capture fine-grained local cues. In parallel, the global branch employs the Global Semantic Shift Modeling (GSSM) module to compare paired global representations and characterize image-level semantic shifts induced by reconstruction. Finally, a lightweight fusion head combines the local and global evidence to produce the final prediction.

3.2 Multi-level Patch Inconsistency Modeling

The reconstruction-induced inconsistency patterns observed in Fig. 1 are mainly distributed across local image regions. To explicitly model such fine-grained evidence, we introduce a Multi-level Patch Inconsistency Modeling (MPIM) module, which estimates patch-wise inconsistency responses across multiple CLIP layers and aggregates them to capture local reconstruction-induced semantic changes.

Let $\mathcal{L}$ denote the set of selected CLIP layers for multi-level feature extraction. For each layer $l \in \mathcal{L}$, we denote the paired patch features extracted from the original and reconstructed

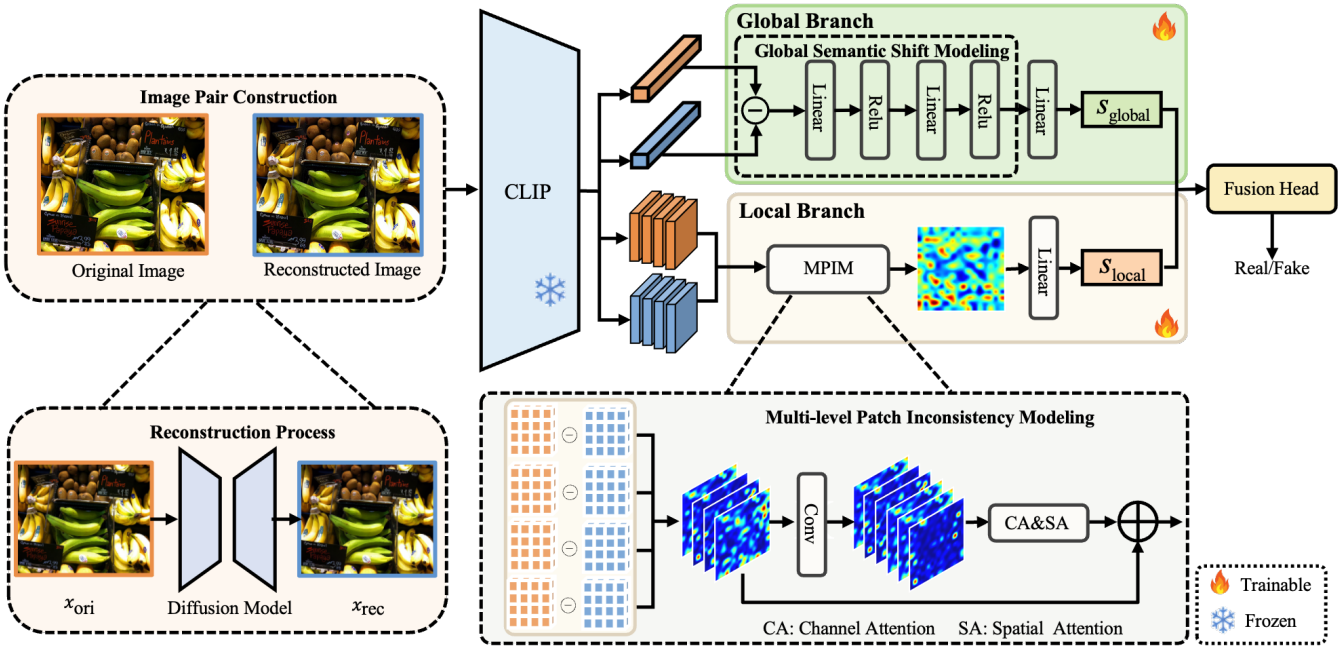


Figure 2: Overview of **SREM**. Given an input image and its diffusion-reconstructed counterpart, a frozen CLIP ViT-L/14 encoder extracts global class-token features and multi-level patch features. GSSM models the global semantic shift between the paired class-token representations, while MPIM aggregates patch-wise inconsistencies across selected CLIP layers. The resulting local and global branch logits are combined by a lightweight learnable fusion head for real/fake prediction.

images as F^l_{ori} and F^l_{rec}, respectively. Both features are represented as $F^l_{\text{ori}}, F^l_{\text{rec}} \in \mathbb{R}^{N\times D}$, where D is the feature dimension of each patch token and $N = H \times W$ is the number of patch tokens arranged on an $H \times W$ spatial grid. For the n-th patch pair at layer l, where $n \in \{1, \ldots, N\}$, the patch-wise inconsistency is computed as

$$d^l_n = \frac{1}{2}\left(1 - \frac{\left\langle F^l_{\text{ori},n}, F^l_{\text{rec},n}\right\rangle}{\left\|F^l_{\text{ori},n}\right\|_2 \left\|F^l_{\text{rec},n}\right\|_2}\right), \tag{1}$$

where a larger d^l_n indicates a stronger inconsistency between the paired patch representations.

For each selected layer, the patch-wise inconsistency scores $\{d^l_n\}_{n=1}^N$ are rearranged according to their spatial locations to form the inconsistency map $I^l \in \mathbb{R}^{H\times W}$. We then concatenate the inconsistency maps from all selected CLIP layers along the channel dimension and feed them into a 3×3 convolutional layer:

$$U = \text{Conv}_{3\times 3}\left(\text{Concat}_{l\in\mathcal{L}}\left(I^l\right)\right), \tag{2}$$

where $\text{Conv}_{3\times 3}(\cdot)$ integrates multi-level inconsistency cues and captures local spatial context.

The fused representation is further refined by sequential channel and spatial attention:

$$\widehat{U} = \mathcal{A}_s\left(\mathcal{A}_c(U)\right), \tag{3}$$

where $\mathcal{A}_c(\cdot)$ and $\mathcal{A}_s(\cdot)$ denote the channel and spatial attention operations, respectively.

To preserve the explicit patch-wise inconsistency cues, we average the response maps over the selected CLIP layers and flatten the result into a parameter-free base representation:

$$B_{\text{patch}} = \text{Flatten}\left(\frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}} I^l\right). \tag{4}$$

The final local representation is obtained by adding a learned refinement to this base representation:

$$M_{\text{local}} = B_{\text{patch}} + \beta\ \text{Flatten}\left(\phi_{\text{out}}(\widehat{U})\right), \tag{5}$$

where $\phi_{\text{out}}(\cdot)$ maps the refined feature $\widehat{U}$ to a single-channel refinement map, and β controls the contribution of the learned refinement. MPIM therefore combines multi-level patch inconsistency aggregation with context-aware refinement, enabling the model to capture local reconstruction-induced forensic evidence.

3.3 Global Semantic Shift Modeling

Although MPIM captures fine-grained cues from local reconstruction-induced inconsistencies, patch-level evidence alone may be insufficient to characterize image-level reconstruction behavior. We therefore introduce a Global Semantic Shift Modeling (GSSM) module to model complementary shifts in the global semantic representation.

Let v_{ori} and v_{rec} denote the global class-token features extracted from x_{ori} and x_{rec}, respectively, using the frozen CLIP encoder. To explicitly capture the reconstruction-induced semantic shift, GSSM first computes the difference between the paired global representations:

$$\Delta v = v_{\text{rec}} - v_{\text{ori}}. \tag{6}$$

This difference operation suppresses the shared semantic content between the original and reconstructed images, while highlighting the global semantic changes caused by reconstruction. The resulting shift feature is then passed through a lightweight MLP to obtain the global semantic shift representation:

$$r_{\text{global}} = \text{MLP}_{\text{g}}\left(\Delta v\right), \tag{7}$$

where $\text{MLP}_{\text{g}}(\cdot)$ is implemented with linear layers and ReLU activations, and r_{global} denotes the resulting global semantic shift representation. This lightweight MLP refines the shift feature into task-relevant image-level evidence, complementing the fine-grained inconsistency cues captured by MPIM.

3.4 Training Objective

MPIM and GSSM capture complementary reconstruction-induced evidence at the local and global levels, respectively. Their representations are first mapped to the corresponding branch logits:

$$s_{\text{local}} = h_{\text{local}}\left(M_{\text{local}}\right), \qquad s_{\text{global}} = h_{\text{global}}\left(r_{\text{global}}\right), \tag{8}$$

where $h_{\text{local}}(\cdot)$ and $h_{\text{global}}(\cdot)$ denote the local and global prediction heads, respectively.

The two branch logits are concatenated and processed by a lightweight learnable fusion head:

$$s_{\text{final}} = h_{\text{fuse}}\left(\text{Concat}\left(s_{\text{local}}, s_{\text{global}}\right)\right), \tag{9}$$

where $h_{\text{fuse}}(\cdot)$ is implemented as a lightweight linear layer.

The final prediction is supervised using the binary cross-entropy classification loss:

$$\mathcal{L}_{\text{ce}} = -\left[y \log(y') + (1-y)\log(1-y')\right], \tag{10}$$

where $y \in \{0, 1\}$ denotes the ground-truth authenticity label and y' is the predicted probability.

In addition, we impose a supervised contrastive metric loss $\mathcal{L}_{\text{con}}$. Therefore, the overall training objective is formulated as

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{\text{ce}} + \alpha\mathcal{L}_{\text{con}}, \tag{11}$$

where α balances the classification and contrastive metric objectives. The contrastive objective is applied only during training and introduces no additional computational cost during inference. In our experiments, α is set to 0.3 by default.

4 Experiments

We first present the implemental details, and then report and discuss the results of our proposed SREM framework and the compared methods. After that, we provide more detailed analysis to verify its effectiveness.

4.1 Experimental Setup

Implementation Details: We use the SDv1-based diffusion reconstruction protocol of DRCT [Chen *et al.*, 2024] to generate reconstructed images. During training, images are randomly cropped to 224×224, while center cropping is applied during testing. For each original image and its reconstruction, we apply the same crop parameters to avoid introducing discrepancies when computing patch-wise inconsistency cues. We adopt CLIP ViT-L/14 as the visual backbone and freeze all its parameters during training. Patch features are extracted from four selected CLIP layers, with $\mathcal{L} = \{6, 12, 18, 24\}$. The trainable modules are optimized using AdamW [Loshchilov and Hutter, 2017] with a learning rate of 2×10^{-5} and a weight decay of 4×10^{-5}. The learning rate uses a linear warm-up schedule followed by cosine annealing, and the total epoch is set to 5. All experiments are implemented in PyTorch [Paszke *et al.*, 2019] and conducted on NVIDIA RTX 4090D and RTX 5090 GPUs.

Evaluation Metrics and Benchmarks: Following previous works [Wang *et al.*, 2023; Chen *et al.*, 2024; Yan *et al.*, 2024], accuracy (ACC) with threshold of 0.5 is adopted as the primary metric for evaluating detection performance. Evaluation experiments are conducted on three representative benchmarks: GenImage [Zhu *et al.*, 2023], DRCT-2M [Chen *et al.*, 2024], and Chameleon [Yan *et al.*, 2024]. For each benchmark, we follow the official training and evaluation protocols. **GenImage** is a large-scale ImageNet-based benchmark that pairs real images with synthetic images generated by diverse models, including Stable Diffusion v1.4 [Rombach *et al.*, 2022], Stable Diffusion v1.5 [Rombach *et al.*, 2022], GLIDE [Nichol *et al.*, 2021], VQDM [Gu *et al.*, 2022b], Wukong [Gu *et al.*, 2022a], BigGAN [Brock *et al.*, 2018], ADM [Dhariwal and Nichol, 2021], and Midjourney [Midjourney, Inc., 2023]. **DRCT-2M** provides a broader evaluation setting with 16 generated-image categories, each containing 120k images. It covers both text-to-image synthesis and image-to-image generation: the text-to-image subset is built from MSCOCO-derived prompts [Lin *et al.*, 2014] and includes 10 Stable Diffusion variants, while the image-to-image subset contains ControlNet-based and diffusion-reconstruction settings conditioned on prompts, Canny edges, masks, or real images. **Chameleon** is designed to assess detector robustness in more realistic scenarios. It contains high-fidelity AI-generated images from diverse scene categories, and only synthetic images misclassified as real by human annotators are retained, making it particularly suitable for evaluating real-world generalization.

4.2 Comparisons Experiments

Performance comparisons are conducted on GenImage, DRCT-2M, and Chameleon benchmarks, respectively.

Comparison on GenImage: Table 1 reports the results on GenImage. Most input-driven detectors achieve high accuracy on SD-related subsets, such as SDv1.4, SDv1.5, and Wukong, but their performance often drops substantially on generators with larger domain shifts, including ADM, GLIDE, VQDM, and BigGAN. For example, UnivFD benefits from CLIP representations and shows better generalization than many conventional detectors, but it still suffers clear performance degradation on several subsets, especially ADM and BigGAN. This suggests that directly using CLIP features is not sufficient to consistently handle diverse generative sources. Reconstruction-driven methods generally improve cross-generator performance by exploiting reconstruction behavior, but their gains remain uneven across generators. Among them, DRCT achieves the second-best average accuracy of 89.49%, showing the effectiveness of reconstruction-based hard sample learning. In comparison, the proposed SREM achieves the highest average accu-

Table 1: Performance comparison on GenImage. All methods are trained on the GenImage/SDv1.4 subset and evaluated on different subsets. Results are reported as accuracy (%), and Avg. denotes the mean accuracy across all testing subsets. The 'Type' column indicates whether each method follows an input-driven (Input) or reconstruction-driven (Recon.) detection paradigm.

Method	Type	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg.
CNNSpot [Wang *et al.*, 2020]	Input	84.92	99.88	99.76	53.48	53.80	99.68	55.50	49.93	74.62
F3Net [Qian *et al.*, 2020]		77.85	98.99	99.08	51.20	54.87	97.92	58.99	49.21	73.51
CLIP/RN50 [Radford *et al.*, 2021]		83.30	99.97	99.89	54.55	57.37	99.52	57.90	50.00	75.31
GramNet [Liu *et al.*, 2020]		73.68	98.85	98.79	51.52	55.38	95.38	55.15	49.41	72.27
De-fake [Sha *et al.*, 2023]		79.88	98.65	98.62	71.57	78.05	98.42	78.31	74.37	84.73
Conv-B [Liu *et al.*, 2022]		83.55	99.99	99.92	51.75	56.27	99.92	58.41	50.00	74.98
UnivFD [Ojha *et al.*, 2023]		91.46	96.41	96.14	58.07	73.40	94.53	67.83	57.72	79.45
AIDE [Yan *et al.*, 2024]		79.38	99.74	99.76	78.54	91.82	98.65	80.26	66.89	86.88
DIRE [Wang *et al.*, 2023]	Recon.	50.40	99.99	**99.92**	52.32	67.23	99.98	50.10	49.99	71.24
LARE [Luo *et al.*, 2024]		74.00	**100.00**	99.90	61.70	88.50	**100.00**	97.20	68.70	86.20
TRE [Shen *et al.*, 2025]		75.20	**100.00**	99.90	62.30	89.60	**100.00**	**98.50**	70.20	87.00
EIRES [Jiang *et al.*, 2025]		72.00	97.20	97.40	69.90	**95.90**	93.50	97.20	**99.80**	87.60
FIRE [Chu *et al.*, 2025]		73.47	98.31	98.38	53.16	62.82	97.32	56.40	50.61	73.93
DRCT [Chen *et al.*, 2024]		**91.50**	95.01	94.41	79.42	89.18	94.67	90.03	81.67	89.49
Ours	Recon.	86.58	97.47	97.41	**81.23**	81.53	97.43	92.17	93.14	**90.87**

Table 2: Performance comparison on DRCT-2M. All methods are trained on the DRCT-2M/SDv1.4 subset and evaluated on different subsets. Results are reported as accuracy (%), and Avg. denotes the mean accuracy across all testing subsets.

Method	SD Variants						Turbo Variants		LCM Variants		ControlNet Variants			DR Variants			Avg.
	LDM	SDv1.4	SDv1.5	SDv2	SDXL	SDXL-Refiner	SD-Turbo	SDXL-Turbo	LCM-SDv1.5	LCM-SDXL	SDv1-Ctrl	SDv2-Ctrl	SDXL-Ctrl	SDv1-DR	SDv2-DR	SDXL-DR	
CNNSpot [Wang *et al.*, 2020]	99.87	99.91	99.90	97.55	66.25	86.55	86.15	72.42	98.26	61.72	97.96	85.89	82.84	60.93	51.41	50.28	81.12
F3Net [Qian *et al.*, 2020]	99.85	99.78	99.79	88.66	55.85	87.37	68.29	63.66	97.39	54.98	97.98	72.39	81.99	65.42	50.39	50.27	77.13
CLIP/RN50 [Radford *et al.*, 2021]	99.00	99.99	99.96	94.61	62.08	91.43	83.57	64.40	98.97	57.43	99.74	80.69	82.03	65.83	50.67	50.47	80.05
GramNet [Liu *et al.*, 2020]	99.40	99.01	98.84	95.30	62.63	80.68	71.19	69.32	93.05	57.02	89.97	75.55	82.68	51.23	50.01	50.08	76.62
De-fake [Sha *et al.*, 2023]	92.10	99.53	99.51	89.65	64.02	69.24	92.00	93.93	99.13	70.89	58.98	62.34	66.66	50.12	50.16	50.00	75.52
Conv-B [Liu *et al.*, 2022]	**99.97**	**100.00**	**99.97**	95.84	64.44	82.00	80.82	60.75	99.27	62.33	**99.80**	83.40	73.28	61.65	51.79	50.41	79.11
UnivFD [Ojha *et al.*, 2023]	98.30	96.22	96.33	93.83	91.01	93.91	86.38	85.92	90.44	88.99	90.41	81.06	89.06	51.96	51.03	50.46	83.46
DIRE [Wang *et al.*, 2023]	98.19	99.94	99.96	68.16	53.84	71.93	58.87	54.35	**99.78**	59.73	99.65	64.20	59.13	51.99	50.04	49.97	71.23
DRCT [Chen *et al.*, 2024]	96.74	96.26	96.33	94.89	96.24	93.46	93.43	92.94	91.17	95.01	95.60	92.68	91.95	**94.10**	**69.55**	**57.43**	90.49
MRCL [Zhuang *et al.*, 2025]	99.92	99.87	99.84	**99.62**	97.28	93.96	98.08	94.81	99.46	97.67	99.72	**98.96**	98.18	85.90	67.36	57.35	93.00
Ours	98.69	98.73	98.74	98.61	**98.73**	**98.50**	**98.72**	**98.68**	98.69	**98.43**	98.72	98.64	**98.62**	93.57	59.79	52.39	**93.02**

racy of 90.87%, outperforming DRCT by 1.38%. Notably, SREM improves the accuracy on BigGAN from 81.67% to 93.14%, yielding 11.47% gain over DRCT. This result indicates that mining original-reconstruction inconsistency in the CLIP feature space can provide complementary evidence beyond directly using CLIP representations or reconstruction-based training, and can better support generalization across different generative paradigms.

Comparison on DRCT-2M: Table 2 reports the results on DRCT-2M, which contains multiple diffusion-model variants. Among them, the DR subsets are synthetic images reconstructed from real images. Most existing detectors achieve high accuracy on SD variants, especially those closely related to the training subset, but their performance drops on other variants, with a more evident degradation on DR subsets. In contrast, SREM maintains accuracy above 98% on most non-DR subsets, covering SD, Turbo, LCM, and ControlNet variants, and only shows a noticeable decrease on the more challenging DR variants. However, similar degradation can also be observed in other competing methods, suggesting that synthetic images reconstructed from real images are more difficult for current detectors to identify. Despite this challenge, SREM achieves the best overall performance with an average accuracy of 93.02%, outperforming existing methods. Although MRCL obtains a very close average accuracy of 93.00%, it relies on multiple reconstruction processes during training, which leads to higher computational and training costs. Overall, SREM demonstrates both effectiveness in mining semantic reconstruction evidence across diverse diffusion variants.

Comparison on Chameleon: Table 3 reports the results on the Chameleon benchmark. Compared with GenImage and DRCT-2M, Chameleon is more challenging because its synthetic images are visually deceptive to humans and therefore more difficult to distinguish from real images. When trained on SD v1.4, which follows the same training setting as the previous benchmarks, most existing detectors achieve only moderate accuracy, generally below 66%. In particular, the recent SOTA method OMAT, which improves generalization by breaking latent prior bias, obtains 66.05% accuracy, showing that Chameleon remains difficult even for strong diffusion-oriented detectors. In contrast, SREM achieves the best accuracy of 67.34%, demonstrating that semantic reconstruction evidence is effective for detecting visually deceptive synthetic images. A similar trend can be observed under the GenImage training setting. It is worth noting that we do not use the full GenImage training set. Instead, we sample 20,000 images from each generator type to construct a reduced training subset. Even under this setting, SREM achieves 67.15% accuracy and outperforms the strongest baseline AIDE by

Table 3: Performance comparison on Chameleon. Results are reported as accuracy (%). Baselines are cited from OMAT [Zhou *et al.*, 2026].

Training Dataset	CNNSpot	FreDect	Fusing	GramNet	LNP	UnivFD	DIRE	PatchCraft	NPR	AIDE	OMAT	Ours
ProGAN	56.94	55.62	56.98	58.94	57.11	57.22	58.19	53.76	57.29	58.37	–	**60.73**
SD v1.4	60.11	56.86	57.07	60.95	55.63	55.62	59.71	56.32	58.13	62.60	66.05	**67.34**
All GenImage	60.89	57.22	57.09	59.81	58.52	60.42	57.83	55.70	57.81	65.77	–	**67.15**

Table 4: Analysis of reconstruction steps in the proposed SREM. Train Steps and Test Steps denote the reconstruction steps used during training and testing, respectively. Results are reported as accuracy (%), and Avg. denotes the mean accuracy across all testing subsets.

Train Steps	Test Steps	SD Variants						Turbo Variants		LCM Variants		ControlNet Variants			DR Variants			Avg.
		LDM	SDv1.4	SDv1.5	SDv2	SDXL	SDXL-Refiner	SD-Turbo	SDXL-Turbo	LCM-SDv1.5	LCM-SDXL	SDv1-Ctrl	SDv2-Ctrl	SDXL-Ctrl	SDv1-DR	SDv2-DR	SDXL-DR	
1	1	96.43	97.62	97.56	97.82	97.85	97.56	96.78	96.80	96.07	97.61	96.51	96.27	96.20	59.93	58.15	57.50	89.79
	10	96.60	96.73	96.78	96.76	96.88	96.52	96.87	96.81	96.70	96.69	96.67	96.54	96.62	59.53	62.13	59.51	89.90
	50	96.68	96.86	96.88	96.82	96.94	96.57	96.97	96.92	96.73	96.72	96.74	96.60	96.70	59.73	62.76	59.55	**90.01**
10	1	97.72	98.05	97.70	98.10	98.72	98.35	98.60	97.60	95.84	97.90	97.49	98.78	97.02	56.57	51.09	50.52	89.38
	10	99.17	99.77	99.75	99.48	99.55	99.00	99.73	99.65	99.34	98.77	99.44	99.43	99.22	53.20	51.05	50.07	90.41
	50	99.30	99.76	99.75	99.47	99.57	99.08	99.74	99.68	99.40	98.94	99.46	99.43	99.31	54.31	51.32	50.13	**90.54**
50	1	98.22	98.62	98.50	98.18	99.00	98.77	98.83	98.80	98.04	98.69	98.38	98.92	98.03	76.47	58.06	53.79	91.83
	10	98.92	99.06	99.04	98.93	99.04	98.76	99.03	98.98	98.97	98.67	99.01	98.92	98.91	91.42	58.21	51.89	92.99
	50	98.69	98.73	98.74	98.61	98.73	98.50	98.72	98.68	98.69	98.43	98.72	98.64	98.62	93.57	59.79	52.39	**93.02**

1.38%, indicating that the proposed method exhibits competitive performance while maintaining good data efficiency. More importantly, when trained on ProGAN, where the synthetic images are generated by a non-diffusion paradigm, SREM still obtains the highest accuracy of 60.73%. This result suggests that even when trained on synthetic images from a non-diffusion paradigm, SREM can still benefit from SD-based reconstruction to mine effective semantic discrepancy cues.

4.3 Ablation Study on Reconstruction Steps

Reconstruction-based detectors may depend on the quality of the reconstructed image. Higher-fidelity reconstruction usually requires more diffusion steps, which also increases inference cost. To study this effect, we vary the number of reconstruction steps used for training and testing on DRCT-2M. Table 4 analyzes the effect of diffusion reconstruction steps in SREM. Overall, using more reconstruction steps during training leads to better performance. When the number of training reconstruction steps is increased from 1 to 50, the best average accuracy improves from 90.01% to 93.02%. This suggests that higher-quality reconstructions are more beneficial for training, likely because they provide more reliable reconstruction counterparts and allow the model to capture subtle semantic shifts between the original and reconstructed images. In contrast, the number of test-time reconstruction steps has a relatively smaller impact. In particular, when the model is trained with 50-step reconstructions, it still achieves 92.99% under 10-step evaluation, which is very close to the 93.02% obtained with 50-step evaluation. Even with 1-step reconstruction, the accuracy maintains 91.83%. These results indicate that high-quality reconstructions are important for learning effective semantic discrepancy patterns during training, while faster few-step reconstruction can be adopted at inference with only minor performance degradation. This provides a practical trade-off between detection accuracy and deployment efficiency.

Table 5: Ablation study of the proposed components in SREM. All variants follow the same training protocol as the corresponding benchmark. Results are reported as average accuracy (%).

Model Components		Avg. ACC (%)	
GSSM	MPIM	GenImage	DRCT-2M
×	×	79.15	88.03
✓	×	79.22	88.96
✓	✓	**90.87**	**93.02**

4.4 Ablation Study on Proposed Components

In this ablation study, we focus on the impact of each proposed component. Table 5 reports the ablation study of the proposed components. When both GSSM and MPIM are removed, the model directly uses the original and reconstructed images as paired inputs for training, without explicitly modeling local patch inconsistencies or global semantic shifts. This baseline achieves 79.15% and 88.03% average accuracy on GenImage and DRCT-2M, respectively. By introducing GSSM alone, the performance improves to 79.22% on GenImage and 88.96% on DRCT-2M, indicating that image-level semantic shifts provide useful reconstruction-induced evidence. When MPIM is further incorporated, the performance increases substantially to 90.87% and 93.02% on the two benchmarks. These results show that local patch-wise inconsistency modeling plays a critical role in capturing fine-grained reconstruction-induced cues, while GSSM provides complementary global semantic evidence. The consistent improvements verify the effectiveness of jointly modeling local and global reconstruction evidence in the semantic space.

5 Conclusion

In this paper, a Semantic-space Reconstruction Evidence Mining (SREM) framework is proposed for generalizable AI-generated image detection. Instead of directly relying on low-level reconstruction-error signals, SREM captures reconstruction stability in the CLIP space through MPIM, which aggregates multi-level patch inconsistencies, and GSSM, which characterizes global semantic shifts from paired repre-

sentations.These complementary cues are integrated to form discriminative forensic evidence for final prediction. Extensive experiments on multiple benchmarks demonstrate the effectiveness of the proposed SREM. Nevertheless, SREM still relies on an external diffusion reconstruction process, which introduces additional inference cost and may be affected by severe reconstructor-generator mismatch. Future work will explore more efficient and robust reconstruction strategies.

References

[Brock *et al.*, 2018] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[Cazenavette *et al.*, 2024] George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. FakeInversion: Learning to detect images from unseen text-to-image models by inverting Stable Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10759–10769, 2024.

[Chen *et al.*, 2024] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-First International Conference on Machine Learning*, 2024.

[Cheng *et al.*, 2025] Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-SPY: Combining semantic and pixel features to detect synthetic images by AI. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13455–13465, 2025.

[Chu *et al.*, 2025] Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. FIRE: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12830–12839, 2025.

[Corvi *et al.*, 2023] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[Frank *et al.*, 2020] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, pages 3247–3258. PMLR, 2020.

[Gu *et al.*, 2022a] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale Chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431, 2022.

[Gu *et al.*, 2022b] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[Jiang *et al.*, 2025] Wan Jiang, Jing Yan, Xiaojing Chen, Lin Shen, Chenhao Lin, Yunfeng Diao, and Richang Hong. EIRES: Training-free AI-generated image detection via edit-induced reconstruction error shift. *arXiv preprint arXiv:2510.25141*, 2025.

[Ju *et al.*, 2023] Yan Ju, Shan Jia, Jialing Cai, Haiying Guan, and Siwei Lyu. GLFF: Global and local feature fusion for AI-synthesized image detection. *IEEE Transactions on Multimedia*, 26:4073–4085, 2023.

[Kang *et al.*, 2025] Ju Yeon Kang, Jaehong Park, Semin Kim, Ji Won Yoon, and Nam Soo Kim. Semantic-aware reconstruction error for detecting AI-generated images. *arXiv preprint arXiv:2508.09487*, 2025.

[Koutlis and Papadopoulos, 2024] Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, pages 394–411. Springer, 2024.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[Lin *et al.*, 2014] 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*, pages 740–755. Springer, 2014.

[Liu *et al.*, 2020] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8060–8069, 2020.

[Liu *et al.*, 2022] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[Liu *et al.*, 2024] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10770–10780, 2024.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Luo *et al.*, 2024] Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. LaRE^2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17006–17015, 2024.

[Midjourney, Inc., 2023] Midjourney, Inc. Midjourney. https://www.midjourney.com/, 2023. Accessed: 2026-05-13.

[Nichol *et al.*, 2021] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[Ojha *et al.*, 2023] 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*, pages 24480–24489, 2023.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[Qian *et al.*, 2020] 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*, pages 86–103. Springer, 2020.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[Sha *et al.*, 2023] 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*, pages 3418–3432, 2023.

[Shen *et al.*, 2025] Chengji Shen, Zhenjiang Liu, Kaixuan Chen, Jie Lei, Mingli Song, and Zunlei Feng. Spatial-temporal reconstruction error for AIGC-based forgery image detection. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[Tan *et al.*, 2024] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024.

[Wang *et al.*, 2020] 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*, pages 8695–8704, 2020.

[Wang *et al.*, 2023] 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*, pages 22445–22455, 2023.

[Wu *et al.*, 2025] Haiwei Wu, Jiantao Zhou, and Shile Zhang. Generalizable synthetic image detection via language-guided contrastive learning. *IEEE Transactions on Artificial Intelligence*, 2025.

[Xi *et al.*, 2023] Ziyi Xi, Wenmin Huang, Kangkang Wei, Weiqi Luo, and Peijia Zheng. AI-generated image detection using a cross-attention enhanced dual-stream network. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1463–1470. IEEE, 2023.

[Yan *et al.*, 2024] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024.

[Zhong *et al.*, 2023] Nan Zhong, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang. Rich and poor texture contrast: A simple yet effective approach for AI-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023.

[Zhou *et al.*, 2026] Yue Zhou, Xinan He, KaiQing Lin, Bing Fan, Feng Ding, and Bin Li. Breaking latent prior bias in detectors for generalizable AIGC image detection. *Advances in Neural Information Processing Systems*, 38:30649–30679, 2026.

[Zhu *et al.*, 2023] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A million-scale benchmark for detecting AI-generated image. *Advances in Neural Information Processing Systems*, 36:77771–77782, 2023.

[Zhuang *et al.*, 2025] Wanyi Zhuang, Qi Chu, Tao Gong, Changtao Miao, and Nenghai Yu. Towards good generalizations for diffusion generated image detection using

multiple reconstruction contrastive learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5431–5440, 2025.