Beyond Artifacts: Real-Centric Envelope Modeling for Reliable AI-Generated Image Detection

Ruiqi Liu[1,2*] Yi Han[3*] Zhengbo Zhang[1] Liwei Yao[4] Zhiyuan Yan[5]
Jialiang Shen[6] ZhiJin Chen[1] Manni Cui[8] Boyi Sun[1] Lubin Weng[1]
Jing Dong[1] Yan Wang[7†] Shu Wu[1†]

[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Advanced Interdisciplinary Sciences, UCAS
[3]Southwest University [4]Shanghai Second Polytechnic University
[5]Peking University [6]The University of Sydney [7]Tsinghua University
[8]Huazhong University of Science and Technology

Abstract. The rapid progress of generative models has intensified the need for reliable and robust detection under real-world conditions. However, existing detectors often overfit to generator-specific artifacts and remain highly sensitive to real-world degradations. As generative architectures evolve and images undergo multi-round cross-platform sharing and post-processing (chain degradations), these artifact cues become obsolete and harder to detect. To address this, we propose **Real-centric Envelope Modeling** (REM), a new paradigm that shifts detection from learning generator artifacts to modeling the robust distribution of real images. REM introduces feature-level perturbations in self-reconstruction to generate near-real samples, and employs an envelope estimator with cross-domain consistency to learn a boundary enclosing the real image manifold. We further build ***RealChain***, a comprehensive benchmark covering both open-source and commercial generators with simulated real-world degradation. REM achieves an average improvement of 6.9% over state-of-the-art (SOTA) methods across 12 benchmark evaluations. Notably, it maintains exceptional generalization on the severely degraded RealChain benchmark, outperforming existing methods by 18.4%. These results establish a solid foundation for reliable synthetic image detection under challenging real-world conditions. The code and datasets are available at `https://github.com/handsome-rich/REM`.

Keywords: AIGI Detection · Chain Degradations · Real-centric

1 Introduction

The rapid progress of generative models has greatly enhanced the realism of synthesized images, driving their widespread use in digital content creation [1,3,34,

* Equal contribution.
† Corresponding author.

35]. However, this also raises increasing concerns about visual content authenticity [15,19,29,55,59]. Thus, reliably detecting AI-generated images (AIGI) under real-world conditions has become an urgent challenge [6,9,11,12,26–28,32,54,56].

Existing detectors follow a generative artifact–driven paradigm [13,14,21,37, 44,51–53,57,60]. These detectors simplify the synthetic image space into a finite set, attempting to learn semantic or pixel-level forgery cues (**artifacts**) from older generators, while assuming images maintain uniform quality during training and inference. However, as shown in Fig. 1, their effectiveness drops sharply under real-world conditions, due to two limitations: (i) they overfit to generator-specific patterns and struggle to generalize to unseen forgery types; and (ii) their reliance on subtle cues makes them vulnerable to real-world perturbations.

In practice, the synthetic image distribution continuously evolves as new architectures and sampling strategies emerge, gradually approaching the real distribution (see Fig. 1(a) and Fig. 2). Therefore, detectors relying on past-generator artifacts often fail on future, more realistic images. Moreover, nearly all images undergo multi-round propagation and post-processing across platforms and devices. As shown in Fig. 1(b) and Fig. 3, such chained processing further weakens forgery cues, making real and synthetic images even harder to distinguish.

Based on these observations, we argue that the artifact-driven paradigm faces fundamental limitations. In contrast, the real image distribution is constrained by physical imaging principles and device characteristics, making it inherently stable and learnable [4,18,33,49]. **To achieve robust detection under complex real-world conditions, it is crucial to explicitly model the real distribution itself, thereby ensuring consistent, transferable, and stable decision boundaries across diverse quality domains.**

To this end, we propose **Real-centric Envelope Modeling (REM)**, a new detection paradigm that shifts the focus from learning generator-specific artifacts to modeling the robust distribution of real images. Specifically, REM first produces diverse near-real samples by applying feature-level perturbations to self-reconstruction models. These samples preserve semantic consistency with real images while exhibiting slight statistical deviations, enabling the model to learn a compact, smooth envelope around the real

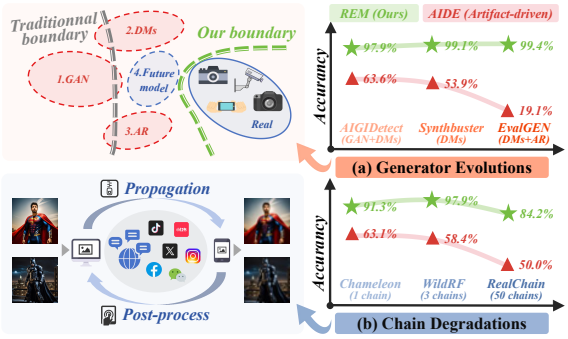

Fig. 1: Challenges of AIGI detection. (a) **Generator Evolution:** Emerging generators (e.g., Diffusion, Autoregressive) increasingly mimic real image manifolds, making old decision boundaries obsolete. (b) **Chain Degradations:** Real-world factors like multi-round compression and post-processing (filters, stickers) destroy detector-relied artifacts.

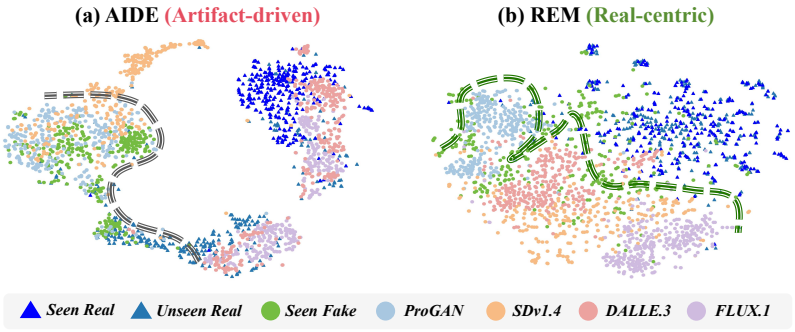

Fig. 2: Challenges of Generator Evolutions. New generators continually emerge with evolving architectures and sampling strategies, causing the feature discrepancy between real and fake images to diminish over time. Detectors trained on obsolete generators overfit to seen artifacts and fail to generalize to new ones, highlighting the need for a real-centric modeling paradigm.

distribution. An Envelope Estimator then explicitly models this boundary in feature space, and a Cross-Domain Consistency constraint ensures its stability under real-world degradations. Together, these components greatly enhance robustness and generalization to unseen generators.

REM performed exceptionally well across all benchmark tests, achieving an average accuracy 6.5% higher than the state-of-the-art (SOTA) model on 6 ideal benchmarks and an average accuracy 5.9% higher on 5 in-the-wild benchmarks.

Furthermore, existing detection benchmarks remain confined to idealized settings and lack coverage of real-world propagation scenarios. To address this limitation, we construct a new dataset named ***RealChain***, which collects images generated by the most advanced and widely used commercial and open-source models, and systematically simulates multi-round cross-platform and cross-device propagation as well as user post-processing (chain degradations). REM still exhibits strong robustness on the *RealChain* benchmark, outperforming SOTA methods by 18.4%.

Our main contributions are summarized as follows:

- We design the Real-centric Envelope Modeling (REM) framework, which relies on real distribution boundary for more reliable discrimination, providing a new perspective for AIGI detection.
- We construct the *RealChain* dataset covering state-of-the-art generators and chain degradations, enabling evaluation of detection performance under realistic conditions.
- Experiments on 12 benchmarks show that REM significantly outperforms existing methods under both degraded and unseen-generator settings.

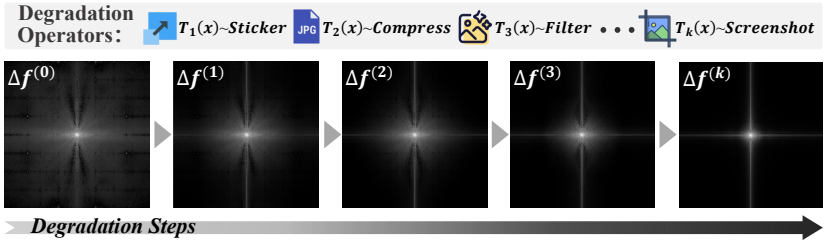

Fig. 3: Challenges of Chain Degradations. As the propagation chain length increases, multi-stage compression and processing progressively suppress generative artifacts, leading to reduced separability between real and fake samples in the frequency domain. This explains the sharp performance drop of artifact-driven detectors (e.g., high-frequency artifacts) in real-world conditions.

2 Related Work

Artifact-driven AIGI Detection. Early methods for AI-generated image (AIGI) detection rely on identifying generator-specific artifacts. CNNSpot [48] first demonstrates that a simple CNN classifier can effectively distinguish synthetic images from known generators but fails to generalize to unseen ones. To enhance cross-generator generalization, UnivFD [36] employs a CLIP-based backbone, leveraging large-scale pretraining for better distinguishing forgery types. FatFormer [31] and C2P-CLIP [45] further optimize CLIP-based detectors by incorporating textual guidance to refine the representation space. NPR [44] focuses on upsampling artifacts introduced by generative models, while SAFE [24] exploits frequency inconsistencies as discriminative cues. AIDE [54] integrates both frequency and semantic signals to improve detection performance.

Although artifact-driven detectors achieve high accuracy under controlled settings, they are easily influenced by non-causal low-level cues such as image format and resolution. As a result, their performance drops significantly in real-world scenarios where such artifacts are suppressed or when facing more realistic generated images. These limitations expose the inherent fragility of the artifact-driven paradigm.

Dataset Alignment AIGI Detection. A recent line of research aims to mitigate evaluation bias by aligning real and generated data. FakeInversion [7] introduces a bias-reduced benchmark by matching real and fake samples in both content and style. SemGIR [58] and DRCT [8] achieve semantic-level alignment through reconstruction or diffusion inversion, while B-Free [17] and Aligned-Forensics [38] use inpainting-based or VAE reconstruction to reduce dataset bias. DDA [10] further enforces format consistency to prevent shortcut learning.

These methods show strong generalization improvements, commonly attributed to better semantic alignment and reduced bias. However, we observe that their success stems from capturing the structure of the real image distribution rather than strict pixel-level alignment. In our experiments (see Appendix), deliberately relaxing image correspondence, either by disabling pixel-level alignment or

introducing mixed augmentations, not only does not harm performance but can even improve it. **This suggests that the advantage of dataset alignment may originate from modeling the real distribution boundary through slight perturbations to the image manifold during the alignment process.** Based on this, we propose the Real-centric Envelope Modeling (REM) framework, which explicitly models the boundary of the real distribution.

3 Method

3.1 Motivation

Challenges of Generator Evolution. The generative landscape itself is open and evolving. Existing methods assume that synthetic samples from different generators share common discriminative forgery cues. However, as new architectures and sampling strategies continually emerge, the learned artifacts quickly become obsolete, while detectors tend to overfit generator-specific biases. Let $\phi(x_r)$ and $\phi(x_f)$ denote feature embeddings of real and synthetic images, respectively. For a given generator g, the feature discrepancy is defined as:

$$d_g = \|\mathbb{E}_{x_r}[\phi(x_r)] - \mathbb{E}_{x_f \sim g}[\phi(x_f)]\|_2. \tag{1}$$

During training, detectors rely on large d_g values from known generators $\mathcal{G}_{train}$. However, for an unseen generator $g' \notin \mathcal{G}_{train}$, the discrepancy $d_{g'}$ often decrease, making its outputs nearly indistinguishable from real images (see Fig. 2 (a)).

Challenges of Chain Degradations. Existing artifact-driven detectors are typically designed under high-quality settings, assuming that the input image x has not undergone significant degradation. However, in real environments, images are subject to chain degradations, which can be formalized as: $x^{(k)} = T_k \circ T_{k-1} \circ \cdots \circ T_1(x)$, where $\{T_i\}_{i=1}^{k}$ represent real-world degradation operators such as compression, filters, digital stickers, etc. As the chain length k increases, the low-level statistics of samples are also changed. The separability between real and synthetic images can be quantified by their frequency discrepancy:

$$\Delta f^{(k)} = \|F(x_r^{(k)}) - F(x_f^{(k)})\|_2, \tag{2}$$

where $F(\cdot)$ denotes the frequency spectrum. As k grows, $\Delta f^{(k)}$ gradually decreases, indicating that real and synthetic images become increasingly indistinguishable in the high-frequency domain, as shown in Fig. 3. Consequently, artifact-driven detectors show severe performance degradation when exposed to real-world data.

Artifact-driven detectors have raised concerns about their ability to detect future generators and real-world chain degradations. To achieve robust detection, as shown in Fig. 4, we propose the **Real-centric Envelope Modeling (REM)** framework, which anchors detection to the real distribution through three modules: Manifold Boundary Reconstruction for generating near-real samples via feature-level perturbations, Envelope Estimator for boundary learning, and Cross-Domain Consistency for maintaining stability under degradations.

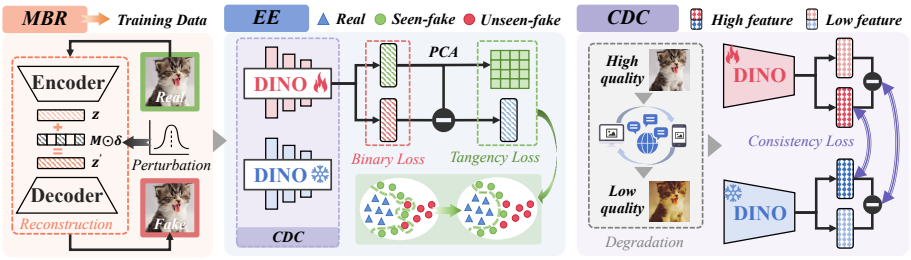

Fig. 4: Overview of the Real-centric Envelope Modeling (REM) framework. REM consists of three key components: (1) Manifold Boundary Reconstruction (MBR) generates diverse near-real samples by applying feature-level perturbations to self-reconstructed inputs; (2) Envelope Estimator (EE) then learns a compact and smooth boundary that tightly encloses the real distribution by separating real samples from the MBR generated near-real ones in feature space; and (3) Cross-domain consistency (CDC) further enforces stability of this boundary under various degradations, ensuring that the learned envelope remains consistent across domains.

3.2 Manifold Boundary Reconstruction (MBR)

The distribution of generated images is open and rapidly evolving, which makes artifact-based detectors trained on limited generators difficult to generalize. In contrast, the distribution of real images is governed by physical imaging laws and device constraints, making it relatively stable and learnable. Building on this observation, we aim to model the boundary of the real manifold rather than relying on generator-specific artifacts. To achieve this, we reconstruct near-real samples around the real distribution and train the model to distinguish them from real samples. This process encourages detector to learn a smooth and discriminative envelope that tightly surrounds the real data distribution.

To ensure that the reconstructed samples provide sufficient coverage of the real manifold, we introduce a feature-level perturbation strategy within a variational autoencoder (VAE). Specifically, we inject controlled Gaussian noise into randomly selected latent dimensions, allowing the decoder to generate samples that remain visually realistic but deviate slightly from the real manifold in multiple directions. Given a real sample x_r, its latent representation is obtained as $z = E(x_r)$. We then apply random noise to a subset of latent dimensions:

$$z' = z + M \odot \delta, \quad \delta \sim \mathcal{N}(0, \epsilon^2 I), \tag{3}$$

where M is a random mask controlling the proportion of perturbed dimensions, and ϵ controls the magnitude of perturbation, ensuring mild deviations that retain semantic fidelity. The perturbed vectors are decoded as $x_f = D(z')$, generating semantically consistent yet slightly shifted near-real samples. By Manifold Boundary Reconstruction, we obtain diverse samples x_f forming a local envelope around the real manifold, providing reliable boundary supervision for the Envelope Estimator.

3.3 Envelope Estimator (EE)

With the near-real samples provided by MBR, we explicitly model the boundary of the real data distribution in the feature space. The core idea is that real samples should reside inside the distribution envelope, while near-real samples are expected to lie closely along its boundary. Formally, the Envelope Estimator aims to learn a smooth and geometrically consistent boundary that encloses the real feature manifold.

Binary Classification. To coarsely separate real and near-real samples, we adopt a binary cross-entropy loss:

$$\mathcal{L}_{\text{bce}} = -\mathbb{E}_{x_r}[\log \sigma(s(h_r))] - \mathbb{E}_{x_f}[\log(1 - \sigma(s(h_f)))], \tag{4}$$

where $h_r = \phi(x_r)$, $h_f = \phi(x_f)$ denote encoded feature embeddings extracted by encoder $\phi(\cdot)$. The function $s(\cdot)$ is a discriminator producing a real-valued score, while $\sigma(\cdot)$ is the sigmoid function converting it into a probability of being real. This loss encourages real samples to have higher confidence (i.e., lie inside the envelope) and pushes near-real samples toward the boundary.

Tangency Regularization. Classification alone may lead to irregular or overfitted boundaries, especially when near-real samples are unevenly distributed. To enhance geometric consistency, we introduce a tangency regularization based on the feature displacement between a real sample and its near-real counterpart, defined as $\Delta h = h_f - h_r$. Let the top-p principal components U_p of real features h_r approximate the local tangent space of the real manifold, and construct the projection matrix $P = U_p U_p^\top$. For each Δh, the term $(I - P)\Delta h$ represents its deviation orthogonal to the tangent space, representing the component moving away from the real manifold surface. We therefore minimize

$$\mathcal{L}_{\text{tan}} = \mathbb{E}_{(h_r, h_f)} \|(I - P)\Delta h\|_2^2, \tag{5}$$

which penalizes off-manifold deviations and encourages near-real features to vary smoothly and consistently along the tangent directions, ultimately yielding a geometry-aligned and continuous envelope boundary.

3.4 Cross-Domain Consistency (CDC)

In real-world scenarios, images inevitably undergo multiple stages of compression, noise injection, and re-encoding. These operations do not alter semantics but introduce significant feature shifts, causing the decision boundary to become unstable across domains. To ensure consistent behavior under such perturbations, we introduce the Cross-Domain Consistency (CDC) module, which enforces stability of the learned envelope boundary.

We adopt DINO [42] as the backbone due to its strong robustness to non-semantic distortions. Through large-scale self-supervised pre-training, DINO produces features that are highly invariant to photometric and compression changes while remaining sensitive to semantic content. This property makes it an ideal

choice for serving as a semantic anchor that preserves high-level meaning under domain degradations. In CDC, a frozen DINO acts as the anchor network $\phi(\cdot)$, providing semantically grounded but non-adaptive representations, while a fine-tuned DINO serves as the learner $f(\cdot)$, adapting to the detection task while maintaining alignment with the anchor space.

Given an image x and its degraded $x' = \textit{Degrade}(x)$, we extract anchor features $h = \phi(x)$ and $h' = \phi(x')$, project them through a linear mapping W to obtain $\hat{h}$ and $\hat{h}'$, and compute learner features $a = f(x)$ and $a' = f(x')$. Two complementary consistency constraints are applied:

Anchor Consistency. This term keeps the learner's features close to the anchor space, preventing semantic drift and maintaining the discriminative structure learned from the frozen backbone:

$$\mathcal{L}_{\text{anc}} = \|a - \hat{h}\|_2^2. \tag{6}$$

Residual Consistency. This constraint ensures that the residuals between the learner and the anchor remain invariant across domains, effectively preserving the geometric shape of the decision envelope under degradations:

$$\mathcal{L}_{\text{res}} = \|(a - \hat{h}) - (a' - \hat{h}')\|_2^2. \tag{7}$$

The low-quality images (**data augmentation**) are generated through simulated chain degradations including compression and perturbation (see Appendix).

3.5 Overall Objective

The overall training objective integrates the envelope modeling and cross-domain consistency terms. Specifically, the total loss is defined as:

$$\mathcal{L} = \underbrace{\left(\mathcal{L}_{\text{bce}} + \lambda_1 \mathcal{L}_{\text{tan}}\right)}_{\mathcal{L}_{\text{EE}}} + \underbrace{\left(\lambda_2 \mathcal{L}_{\text{anc}} + \lambda_3 \mathcal{L}_{\text{res}}\right)}_{\mathcal{L}_{\text{CDC}}}, \tag{8}$$

where the hyperparameters λ_1, λ_2, and λ_3 are empirically determined to balance the importance of each.

4 Dataset Construction

To evaluate AIGI detection robustness under realistic conditions, we build ***RealChain***, a dataset featuring (i) broad coverage of emerging generative models and (ii) realistic simulation of real-world degradation chains (see Fig. 5). *RealChain* reproduces cumulative degradations and distribution shifts in online propagation, providing a comprehensive benchmark for robust AIGI detection.

4.1 Source of Images

Fake Images. To reflect the rapid evolution of generative models, *RealChain* integrates both open-source and commercial sources. Open-source models include *QwenImage* [50], *SDv3.5* [43], and *Flux.1* [23], while commercial ones include *Hunyuan 3.0* [5], *NanoBanana* [16], and *Seedream 4.0* [46]. To capture diverse generation mechanisms, we adopt two primary modes:

(i) Text-to-Image. Images are generated from diverse textual prompts to simulate large-scale content creation. We design 1,000 prompts (500 coarse and 500 fine) and render each using six models, producing 6,000 samples.

(ii) Image-to-Image. Given real references, we perform style transfer and local editing to simulate practical post-editing behaviors. Using 1,000 prompts rendered by *Seedream 4.0* [46], we obtain 1,000 samples.

Real Images. We collect 7,000 real images from MSCOCO [30], OpenImage-v7 [39], Unplash [47], and ImageNet [41], ensuring diversity across domains.

4.2 Chain Degradations Design

To systematically analyze the effects of real-world propagation and post-processing, we define two dataset settings:

No Degradation (ND). This setting includes directly captured real images and cleanly generated synthetic images. It is used to evaluate detector generalization on the most advanced generators and diverse real sources, serving as the upper performance bound.

Chain Degradations (CD). We reproduce realistic cross-platform and device propagation operations that involve compression, re-encoding, and diverse user post-processing. Specifically, we define a library of common and easily reproducible degradation operations observed in real-world scenarios: **(i) Propagation:** Uploading/downloading via PC applications; uploading/downloading via mobile applications; uploading from PC and downloading on mobile; uploading from mobile and downloading on PC. **(ii) Post-processing:** Applying filters; inserting stickers; cropping or changing aspect ratio; taking screenshots. Each sample in the *RealChain* dataset undergoes a random chain of propagation and post-processing operations.

Degradation Chain Construction. Each degradation chain is formed by randomly selecting 1–3 propagation operations and 0–2 post-processing operations. After each operation, the intermediate result is saved and used as the input for the next step. Once all designated operations are completed, the final output of the chain is treated as the fully degraded version used for evaluation. As shown in Fig. 5, *RealChain* includes a total of 50 randomly generated degradation chains. Each chain is applied to 140 real images and 140 synthetic images. Through this design, *RealChain* effectively reproduces the multi-round propagation and user editing processes in real social media environments.

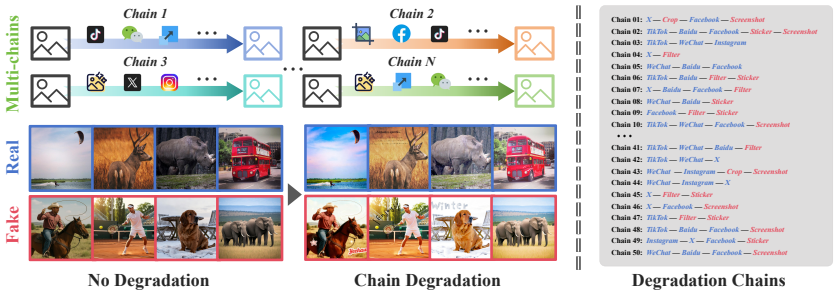

Fig. 5: The *RealChain* dataset contains images with no degradation and images with chain degradations. Each sample undergoes a randomly constructed degradation chain composed of 1–3 propagations and 0–2 post-processing.

Table 1: Overall Comparison on 12 Benchmarks. To ensure fairness and reproducibility, we utilize the official checkpoints released for each method. Following DDA, synthetic images in *GenImage*, *ForenSynths*, and *AIGCDetect* are processed using JPEG compression with a quality factor of 96 to mitigate format bias. Balanced accuracy (B.Acc) is employed across all benchmarks to provide a reliable assessment.

Method	Ideal Benchmarks							In-the-Wild Benchmarks						RealChain		Avg B.Acc
	AIGC-Detect	*Synth-buster*	*Eval-GEN*	*Foren-Synths*	*Gen-Image*	*DRCT-2M*	*Avg*	*Cham-eleon*	*Synth-Wildx*	*WildRF*	*AIGI-Bench*	*Bfree-Online*	*Avg*	ND	CD	
NPR [44]	53.1	50.0	2.9	47.9	51.5	37.3	40.5	59.9	49.8	63.5	54.2	49.5	55.4	74.3	55.7	50.0
UnivFD [36]	72.5	67.8	15.4	77.7	64.1	61.8	59.9	50.7	52.3	55.3	53.0	49.0	52.1	54.2	51.3	55.8
FatFormer [31]	85.0	56.1	45.6	90.0	62.8	52.2	65.3	51.2	52.1	58.9	54.4	50.0	53.3	54.9	51.2	58.8
SAFE [24]	50.3	46.5	1.1	49.7	50.3	59.3	42.9	59.2	49.1	57.2	56.5	50.5	54.5	61.4	49.8	49.3
C2P-CLIP [45]	81.4	68.5	38.9	92.0	74.4	59.2	69.1	51.1	57.1	59.6	53.7	50.0	54.3	51.3	51.3	60.7
AIDE [54]	63.6	53.9	19.1	59.4	61.2	64.6	53.6	63.1	48.8	58.4	55.5	53.1	55.8	60.4	50.0	54.7
DRCT [8]	81.4	84.8	77.8	73.9	84.7	90.5	82.2	56.6	55.1	50.6	72.1	55.7	58.0	69.1	55.4	69.8
Aligned [38]	66.6	77.4	68.0	53.9	79.0	95.5	73.4	71.0	78.8	80.1	72.7	68.5	74.2	68.7	58.0	72.2
DDA [10]	87.8	90.1	97.2	81.4	91.7	98.1	91.1	82.4	90.9	90.3	84.1	95.1	88.6	88.8	65.8	88.0
REM	97.9	99.0	99.2	91.5	98.8	98.9	**97.6**	91.3	95.4	97.8	91.3	96.5	**94.5**	92.4	84.2	**94.9**

Table 2: Performance comparison on the *AIGCDetect* benchmark.

Method	ADM	DALLE2	GLIDE	Midj.	VQDM	Big-GAN	Cycle-GAN	Gau-GAN	Pro-GAN	SDXL	SD1.4	SD1.5	Star-GAN	Style-GAN	Style-GAN2	WFR	Wukong	Avg B.Acc
NPR [44]	43.8	20.0	41.2	53.4	48.4	53.1	76.6	42.2	58.7	59.6	55.1	55.0	67.4	57.9	54.6	58.8	57.4	53.1
UnivFD [36]	62.5	50.0	61.3	55.1	76.9	87.5	96.9	98.8	99.4	58.2	55.6	55.7	95.1	80.0	69.4	69.2	61.1	72.5
FatFormer [31]	80.2	68.5	91.1	54.4	88.0	99.2	99.5	99.1	98.5	71.7	67.5	67.2	99.4	98.0	98.8	88.3	75.6	85.0
SAFE [24]	49.5	49.5	53.0	49.0	50.2	52.2	51.9	50.0	50.0	49.8	49.7	49.8	50.1	50.0	50.0	49.8	50.3	50.3
C2P-CLIP [45]	71.6	52.3	73.5	56.6	73.7	98.4	96.8	98.8	99.3	62.3	77.5	76.9	99.6	93.1	79.4	94.8	79.4	81.4
AIDE [54]	52.9	51.1	60.2	49.8	69.3	70.1	93.6	60.6	89.0	49.6	51.6	51.0	72.1	66.5	59.0	80.6	54.5	63.6
DRCT [8]	79.9	89.2	89.2	85.5	88.6	81.4	91.0	93.8	71.1	88.3	91.4	91.0	53.0	62.7	63.8	73.9	90.8	81.4
Aligned [38]	51.6	52.0	55.6	96.2	72.1	51.2	49.5	50.8	50.7	95.1	99.7	99.6	53.8	52.7	51.6	50.0	99.6	66.6
DDA [10]	89.5	94.6	89.6	95.6	76.6	91.0	72.5	92.7	92.8	99.4	98.7	98.6	72.7	87.8	90.2	52.1	98.8	87.8
REM	98.1	99.3	98.6	96.5	99.3	99.4	98.1	99.8	97.4	99.8	99.3	99.5	90.2	95.1	96.3	98.4	99.5	**97.9**

5 Experiments

We conduct extensive experiments to validate the effectiveness and robustness of our proposed REM under both ideal and real-world conditions.

5.1 Experimental Settings

Datasets. (1) Ideal benchmark. We first evaluate REM on six clean and high-quality datasets, *AIGCDetect* [61], *AIGCDetect* [61], *Synthbuster* [2], *Eval-*

Table 3: Performance comparison on the *Synthbuster* benchmark.

Method	DALLE2	DALLE3	Firefly	GLIDE	Midj.	SD1.3	SD1.4	SD2	SDXL	Avg B.Acc
NPR [44]	51.1	49.3	46.5	48.5	52.8	51.4	51.8	46.0	52.8	50.0
UnivFD [36]	83.5	47.4	89.9	53.3	52.5	70.4	69.9	75.7	68.0	67.8
FatFormer [31]	59.4	39.5	60.3	72.7	44.4	53.7	54.0	52.3	69.1	56.1
SAFE [24]	58.0	9.9	10.3	52.2	56.7	59.4	59.1	53.0	59.5	46.5
C2P-CLIP [45]	55.6	63.2	59.5	86.7	52.9	75.2	76.7	69.2	77.7	68.5
AIDE [54]	34.9	33.7	24.8	65.0	57.5	74.1	73.7	53.2	68.4	53.9
DRCT [8]	77.2	86.6	84.1	82.6	73.7	86.6	86.6	83.2	71.3	84.8
Aligned [38]	50.2	48.9	51.7	53.5	98.7	98.8	98.8	98.6	97.3	77.4
DDA [10]	86.3	90.0	91.9	76.5	93.5	92.9	92.7	93.3	93.5	90.1
REM	99.1	99.9	93.5	98.8	99.9	100.0	100.0	99.9	99.9	**99.0**

Table 4: Performance comparison on the *EvalGEN*.

Method	Flux	GoT	Infinity	NOVA	OmiGen	Avg B.Acc
NPR [44]	0.7	0.2	6.5	4.7	2.2	2.9
UnivFD [36]	4.0	9.2	15.7	8.3	39.6	15.4
FatFormer [31]	9.9	47.9	44.7	98.3	27.3	45.6
SAFE [24]	1.0	0.5	1.9	0.6	1.6	1.1
C2P-CLIP [45]	8.7	49.6	35.3	86.4	14.5	38.9
AIDE [54]	17.9	24.7	3.4	16.3	33.4	19.1
DRCT [8]	72.5	81.4	77.9	84.6	72.5	77.8
Aligned [38]	32.0	72.3	74.0	84.8	77.0	68.0
DDA [10]	89.9	99.5	97.8	99.5	99.5	97.2
REM	97.2	100.0	98.9	99.9	99.9	**99.2**

Table 5: Performance on the *RealChain* with No-Degradation and Chain Degradations.

Method	Flux.1		Hunyuan3		NanoBanana		Qwenlmage		SD3.5		Seedream4		i2i		Avg	
	ND	CD	ND	CD	ND	CD	ND	CD	ND	CD	ND	CD	ND	CD	ND	CD
NPR [44]	84.2	57.6	70.7	55.4	64.7	53.1	53.0	54.4	78.3	56.0	75.4	54.4	93.6	59.0	74.3	55.7
UnivFD [36]	54.2	52.1	55.8	51.7	56.5	51.9	53.2	50.7	52.6	49.6	51.5	51.5	55.4	51.9	54.2	51.3
FatFormer [31]	52.2	48.1	57.3	53.2	59.8	50.5	54.1	50.2	52.5	54.2	54.7	51.9	53.6	50.6	54.9	51.2
SAFE [24]	55.4	49.7	59.6	49.7	66.1	49.5	50.8	48.5	64.6	50.3	69.8	49.8	63.8	50.9	61.4	49.8
C2P-CLIP [45]	51.4	52.3	52.6	48.7	53.1	52.5	50.9	53.2	49.6	51.1	49.2	50.8	52.3	50.4	51.3	51.3
AIDE [54]	69.4	48.6	64.5	51.7	66.3	50.6	57.7	46.3	52.0	57.3	52.6	47.3	60.0	48.4	60.4	50.0
DRCT [8]	54.4	58.3	51.9	55.2	77.5	57.8	63.8	52.6	70.8	53.8	78.0	56.2	83.6	53.7	68.6	55.4
Aligned [38]	71.8	56.0	61.6	54.9	65.8	53.8	62.8	55.6	63.1	64.1	79.3	65.9	76.6	55.7	68.7	58.0
DDA [10]	90.5	67.0	88.0	62.5	83.4	63.6	93.5	70.2	92.1	67.7	93.4	63.5	80.4	65.9	88.8	65.8
REM	96.5	86.7	93.8	85.3	91.9	84.8	94.1	87.3	92.9	84.5	94.2	81.9	83.7	78.8	**92.4**	**84.2**

Table 6: Performance on the *SynthWildx*, *WildRF*, and *AIGIBench*.

Method	SynthWildx			WildRF			AIGIBench		Avg B.Acc
	DALLE3	Firefly	Midj.	Facebook	Reddit	Twitter	SocRF	ComAI	
NPR [44]	43.6	61.3	44.5	78.1	61.0	51.3	53.3	55.1	57.1
UnivFD [36]	45.4	65.3	46.2	49.1	60.2	56.5	54.6	51.4	53.3
FatFormer [31]	46.5	61.6	48.3	54.1	68.1	54.4	56.9	51.9	54.9
SAFE [24]	49.4	48.2	49.6	50.9	74.1	37.5	58.4	54.5	52.7
C2P-CLIP [45]	56.9	61.4	53.0	54.4	68.4	55.8	56.4	51.0	57.3
AIDE [54]	63.4	48.8	51.9	57.8	71.5	45.8	57.3	53.7	57.5
DRCT [8]	58.3	56.4	50.5	46.6	53.1	55.2	67.5	76.7	59.8
Aligned [38]	85.5	58.5	92.2	89.4	69.1	81.8	73.5	71.9	78.2
DDA [10]	92.3	87.3	93.1	93.1	86.4	91.5	79.6	88.6	88.3
REM	96.5	92.8	96.9	97.2	98.1	98.2	92.2	90.4	**94.8**

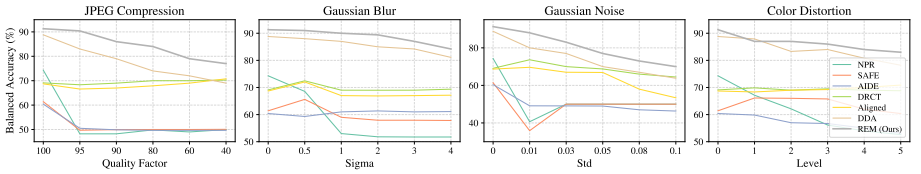

Fig. 6: REM is evaluated under various perturbations, including JPEG compression, Gaussian blur, Gaussian noise, and color distortion.

GEN [10], *ForenSynths* [48], *GenImage* [62], and *DRCT-2M* [8]. These benchmarks include generators with more than 40 different architectures and sampling strategies, as detailed in Table 1. These benchmarks serve as controlled references for assessing the upper bound of generalization performance under ideal conditions. **(2) Real-world degradation benchmark.** We evaluate REM on both public and newly collected real-world datasets, including *Chameleon* [54], *SynthWildx* [11], *WildRF* [6], *AIGIBench* [25], *Bfree-online* [17], and our *RealChain* dataset. These datasets cover diverse propagation environments, black-box post-processing, and multiple image formats generated by both open-source and commercial models. They collectively reflect the complexity of real social media scenarios and serve to assess the model's robustness and generalization under uncontrolled degradations.

Evaluation Protocol. Following works [10,24,31,36,45,54], we use the balanced average accuracy (*B.Acc*) of the real and fake accuracy to evaluate performance. We compare REM with a series of existing methods, including artifact-driven detectors [24, 31, 36, 44, 45, 54], as well as alignment-based approaches [8, 10, 38]. Following the experimental settings of DDA [10], all baseline results are ob-

tained using the official implementations and pretrained weights released by their authors, with JPEG quality alignment applied for the *AIGCDetect* [61],*ForenSynths* [48] and *GenImage* [62]. To further highlight our REM paradigm, we also evaluate the advanced artifact-driven method AIDE [54] and the visual backbone DINOv3 [42]. These models are trained respectively on the diverse generative image dataset GenImage [62] and the alignment-reconstructed DDA dataset [10] (see Appendix).

Implementation Details In all comparative and ablation experiments, we adopt DINOv3 ViT-H+/16 [42] as the backbone network and fine-tune it using the LoRA [20] algorithm. To incorporate generative priors, we employ the pre-trained VAE (FT-MSE version) from Stable Diffusion [40] to project images into the latent space. In the preprocessing stage, images with resolutions larger than 512×512 are resized proportionally so that the shorter side is 512. During training, we use randomly cropped windows of 224×224, while during inference we use center-cropped windows of 224×224. MSCOCO [30] samples are used as the real-image sources, and near-real samples are constructed via the MBR module. We employ the Adam [22] optimizer with a learning rate of 0.0001 and a batch size of 256. Training is performed for only one epoch on four NVIDIA A100 GPUs, taking only 30 minutes to complete.

5.2 Main Results

Overall Comparison on 12 Benchmarks. Table 1 provides a comprehensive evaluation across 12 benchmarks, comprising 6 ideal and 6 real-world datasets, marking one of the most extensive assessments of generator evolution and chain degradations to date. Results show that: (i) REM achieves a new state-of-the-art performance with a 94.9% average balanced accuracy (B.Acc) across over 40 generator subsets, representing a 6.9% improvement over the strongest previous baseline, DDA. (ii) REM demonstrates exceptional stability on in-the-wild benchmarks, with only a slight decrease in detection performance in real-world sampling scenarios. (iii) REM maintains a robust 84.2% accuracy on the severely degraded RealChain benchmark, outperforming existing methods by 18.4% and providing a more reliable decision boundary. These experiments confirm that explicitly modeling the stable real image distribution provides a more robust anchor for detection than learning volatile generator artifacts.

Performance on Ideal Benchmarks. REM achieves state-of-the-art results on datasets including AIGCDetect, Synthbuster, EvalGEN, GenImage, ForenSynths, and DRCT-2M, consistently outperforming artifact-driven methods like SAFE and AIDE, as well as alignment-based approaches such as Aligned and DDA. The detailed performance comparisons in Tables 2, 3, 4 confirm that modeling the real-distribution envelope yields superior cross-generator generalization even for high-fidelity synthetic images.

Performance under Real-world Degradations. Across five in-the-wild benchmarks, including Chameleon, SynthWildx, WildRF, AIGIBench, and Bfree-online, REM show significant improvements (see Tables 1, 6), surpassing DDA by 5.9% in average B.Acc and by 8.9% on Chameleon. On the heavily degraded RealChain

dataset, REM maintains a robust 84.2% B.Acc while most baselines fail, with only DDA exceeding 60% (see Table 5). This performance demonstrates REM's ability to handle image quality degradation in real-world social media networks.

5.3 Analysis and Discussion

Robustness to Perturbations. We further evaluate REM on the *RealChain* (ND) under various perturbations, including JPEG compression, Gaussian blur, Gaussian noise addition, and color distortion. As shown in Fig. 6, artifact-based baselines perform poorly under nearly all perturbations. Both our REM and DDA [10] achieve strong results on clean data, while REM remains more stable under perturbations, further demonstrating the advantage of its cross-quality consistency constraint. Interestingly, the alignment-based methods Aligned [38] and DRCT [8] maintain remarkably steady performance across all perturbations. We speculate that this robustness arises because these methods do not explicitly model format or quality factors during training, which causes them to lose discriminative cues that are beneficial only under high-quality conditions. Their weaker performance on clean data further supports this explanation.

Effect of Real Data Scale. Since REM explicitly models the boundary of the real distribution, we analyze how the scale of real data affects its performance. As shown in Fig. 7 (a), both accuracy and stability consistently improve as the amount of real samples increases. When the number of real samples reaches 100k, REM achieves optimal overall performance across all benchmarks. This trend indicates that a richer real distribution allows the model to learn a more complete and robust envelope boundary. Fortunately, collecting real data is inexpensive, and the computational cost of the MBR module for reconstructing near-real samples is much lower than that of generating diverse fake images from large diffusion models (see Fig. 7 (b)).

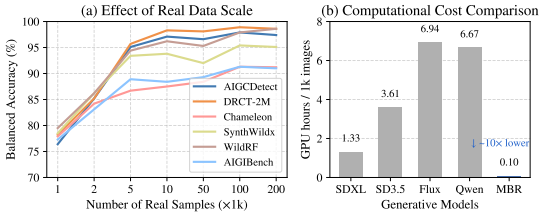

Fig. 7: Efficiency Analysis. (a) **Data Scale:** Accuracy and stability optimize at 100k real samples. (b) **Time Cost:** MBR module requires over **10× lower** GPU cost than diffusion generators.

5.4 Ablation Studies

We conduct a comprehensive ablation study to analyze the contribution of each component in REM (see Table 7 and Fig. 8).

Manifold Boundary Reconstruction (MBR). Replacing the proposed feature-level perturbation with plain VAE reconstruction leads to a 3.6% avg drop in the *B.Acc* on three benchmarks. Without MBR, near-real samples lack diversity and

Table 7: Ablations of modules.

No.	MBR	EE	CDC	*Ideal Bench*	*in-the-wild Bench*	*Real Chain*
1	×	×	×	85.5	80.9	76.4
2	✓	✓	×	93.6	85.8	79.4
3	×	✓	✓	91.5	86.3	83.6
4	✓	×	✓	93.3	89.5	84.3
5	✓	✓	✓	**97.6**	**94.5**	**88.3**

Table 8: Backbone comparison.

Backbone (DINO)	*AIGC-Detect*	*DRCT-2M*	*Chameleon*	*Synth-Wildx*	*WildRF*	*AIGI-Bench*	*RealChain*		**Avg**
							ND	*CD*	
v2-L/14	95.1	97.9	84.6	91.4	95.8	87.2	89.3	78.5	90.0
v2-g/14	96.3	97.8	85.5	91.0	95.8	89.8	89.4	80.7	90.8
v3-L/16	97.0	98.1	89.7	93.2	96.3	89.1	91.1	83.3	92.2
v3-H+/16	97.9	98.9	91.3	95.4	97.9	91.3	92.4	84.2	93.6
v3-7B/16	**98.2**	**99.4**	**93.7**	**96.5**	**98.2**	**93.7**	**93.6**	**86.1**	**94.9**

fail to cover the real boundary. In contrast, feature-level perturbations introduce controlled deviations while preserving semantics, enriching manifold coverage.

Envelope Estimator (EE) When the tangency loss is removed, the model relies solely on the binary classification loss, making the detector more sensitive to image quality degradation, with the *B.Acc* dropping by 2.3%. Feature visualizations further show that the decision surface becomes overfitted and biased toward unseen fake clusters. In contrast, incorporating the tangency constraint enforces local alignment of near-real samples along the real manifold, resulting in a smoother and more stable envelope.

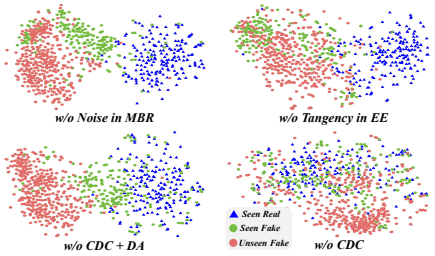

Fig. 8: t-SNE projections on *RealChain*.

Cross-Domain Consistency (CDC) Removing CDC leads to a sharp performance drop under chain degradations (*B.Acc* decreases by 16.4%). Although data augmentation during training can partially mitigate the impact of degradations, it often results in unstable convergence (see Fig. 8). In contrast, CDC explicitly aligns the feature residuals between degraded and clean versions of the same image, thereby maintaining a consistent decision boundary.

Influence of Different Backbones. We further compare REM's performance across different DINO backbones. As shown in Table 8, REM maintains consistently strong performance across all backbones and improves further with increased pretraining quality. This demonstrates the robustness of REM's real-distribution modeling and its ability to efficiently activate and leverage the pre-trained knowledge of DINO-based architectures.

How REM Learns Robust Real Distribution. MBR introduces controllable feature perturbations, generating near-real samples that expand the neighborhood around the real manifold. EE draws a discriminative boundary between real and near-real samples, effectively approximating the envelope of the real distribution. CDC constraint further ensures that representations of the same image remain consistent across different quality levels, keeping the boundary stable under various degradations. Together, these mechanisms enable REM to capture a compact, smooth, and domain-invariant real-centric envelope, resulting in strong robustness to unseen generators and real-world degradations.

6 Conclusion

We introduced REM, a real-centric modeling framework for detecting AI-generated images. REM achieves robust and generalizable performance under real-world degradation conditions, while enabling effective forgery source attribution. Our proposed *RealChain*, featuring emerging generators and chain degradations, supports realistic and challenging real-world evaluation. Extensive experiments across diverse benchmarks further validate the superiority and practical applicability of our approach.

Future work. We will extend REM to video forensics to address the growing challenge of deepfake videos across social media platforms. We also aim to explore more fine-grained modeling of real distributions by incorporating larger-scale datasets and collect more degraded samples during training.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62372454).

References

1. Adobe: Adobe firefly. `https://www.adobe.com/sensei/generative-ai/firefly.html` (2023)
2. Bammey, Q.: Synthbuster: Towards detection of diffusion model generated images. IEEE Open Journal of Signal Processing **5**, 1–9 (2023)
3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. `https://cdn.openai.com/papers/dall-e-3.pdf` (2023)
4. Brooks, T., Mildenhall, B., Xue, T., Chen, J., Sharlet, D., Barron, J.T.: Unprocessing images for learned raw denoising. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11036–11045 (2019)
5. Cao, S., Chen, H., Chen, P., Cheng, Y., Cui, Y., Deng, X., Dong, Y., Gong, K., Gu, T., Gu, X., et al.: Hunyuanimage 3.0 technical report. arXiv preprint arXiv:2509.23951 (2025)
6. Cavia, B.e.a.: Real-time deepfake detection in the real-world. arXiv preprint arXiv:2406.09398 (2024)
7. Cazenavette, G., Sud, A., Leung, T., Usman, B.: 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. pp. 10759–10769 (2024)
8. Chen, B., Zeng, J., Yang, J., Yang, R.: Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In: Forty-first International Conference on Machine Learning (2024)
9. Chen, R., Gao, J., Lin, K., Zhang, K., Zhao, Y., Guan, I., Yao, T., Ding, S.: Task-model alignment: A simple path to generalizable ai-generated image detection. arXiv preprint arXiv:2512.06746 (2025)

10. Chen, R., Xi, J., Yan, Z., Zhang, K.Y., Wu, S., Xie, J., Chen, X., Xu, L., Guan, I., Yao, T., Ding, S.: Dual data alignment makes ai-generated image detector easier generalizable. In: Advances in Neural Information Processing Systems (2025)
11. Cozzolino, D.e.a.: Raising the bar of ai-generated image detection with clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4356–4366 (2024)
12. Dell'Anna, S., Montibeller, A., Boato, G.: Truefake: A real world case dataset of last generation fake images also shared on social networks. arXiv preprint arXiv:2504.20658 (2025)
13. Durall, R.e.a.: 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. pp. 7890–7899 (2020)
14. Frank, J.e.a.: Leveraging frequency analysis for deep fake image recognition. In: International conference on machine learning. pp. 3247–3258. PMLR (2020)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
16. Google DeepMind: Gemini 2.5 flash image (nano banana). `https://aistudio.google.com/models/gemini-2-5-flash-image` (2025)
17. Guillaro, F., Zingarini, G., Usman, B., Sud, A., Cozzolino, D., Verdoliva, L.: A bias-free training paradigm for more general ai-generated image detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18685–18694 (2025)
18. Healey, G.E., Kondepudy, R.: Radiometric ccd camera calibration and noise estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **16**(3), 267–276 (2002)
19. Ho, J.e.a.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
20. Hu, E.e.a.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
21. Ju, Y.e.a.: Fusing global and local features for generalized ai-synthesized image detection. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 3465–3469. IEEE (2022)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Labs, B.F.: Flux. `https://bfl.ai/` (2025)
24. Li, O.e.a.: Improving synthetic image detection towards generalization: An image transformation perspective. arXiv preprint arXiv:2408.06741 (2024)
25. Li, Z., Yan, J., He, Z., Zeng, K., Jiang, W., Xiong, L., Fu, Z.: Is artificial intelligence generated image detection a solved problem? arXiv preprint arXiv:2505.12335 (2025)
26. Lin, K., Lin, Y., Li, W., Yao, T., Li, B.: Standing on the shoulders of giants: Reprogramming visual-language model for general deepfake detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5262–5270 (2025)
27. Lin, K., Yan, Z., Chen, R., Ye, J., Zhang, K.Y., Zhou, Y., Jin, P., Li, B., Yao, T., Ding, S.: Seeing before reasoning: A unified framework for generalizable and explainable fake image detection. arXiv preprint arXiv:2509.25502 (2025)
28. Lin, K., Yan, Z., Chen, R., Zhang, K.Y., Zhou, Y., Piao, C., Li, B., Yao, T., Wang, B., Xiao, Y., et al.: Deep residual injection for full-spectrum forensic signal

perception in multimodal large language models. arXiv preprint arXiv:2606.15880 (2026)
29. Lin, K., Yan, Z., Zhang, K.Y., Hao, L., Zhou, Y., Lin, Y., Li, W., Yao, T., Ding, S., Li, B.: Guard me if you know me: Protecting specific face-identity from deepfakes. arXiv preprint arXiv:2505.19582 (2025)
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
31. Liu, H.e.a.: Forgery-aware adaptive transformer for generalizable synthetic image detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10770–10780 (2024)
32. Liu, R., Cui, M., Qin, Z., Yan, Z., Chen, R., Han, Y., Li, Z., Chen, J., Chen, Z., Lin, K., Shen, J., Weng, L., Dong, J., Wang, Y., Wu, S.: Mirror: Manifold ideal reference reconstructor for generalizable ai-generated image detection. arXiv preprint arXiv:2602.02222 (2026)
33. Lukas, J., Fridrich, J., Goljan, M.: Digital camera identification from sensor pattern noise. IEEE Transactions on Information Forensics and Security **1**(2), 205–214 (2006)
34. Midjourney: Midjourney. `https://www.midjourney.com/home` (2023)
35. Nichol, A.e.a.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
36. Ojha, U.e.a.: 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)
37. Park, J., Owens, A.: Community forensics: Using thousands of generators to train fake image detectors. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8245–8257 (2025)
38. Rajan, A.e.a.: Aligned datasets improve detection of latent diffusion-generated images. arXiv preprint arXiv:2410.11835 (2024)
39. Research, G.: Open images dataset. `https://storage.googleapis.com/openimages/web/index.html` (2017), accessed: 2025-10-27
40. Rombach, R.e.a.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
41. Russakovsky, O.e.a.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
42. Simeoni, O.e.a.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
43. Stability AI: Stable diffusion 3.5. `https://stability.ai/news/introducing-stable-diffusion-3-5` (2024)
44. Tan, C.e.a.: 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. pp. 28130–28139 (2024)
45. Tan, C.e.a.: C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7184–7192 (2025)
46. Team, B.S.: Seedream 4.0. `https://www.volcengine.com/docs/82379/1541523` (2025)
47. Unsplash: Unsplash. `https://unsplash.com` (nd), [Accessed: 2025-10-27]
48. Wang, S.e.a.: 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 (2020)

49. Wang, Y., Xu, T., Zhang, F., Xue, T., Gu, J.: Adaptiveisp: Learning an adaptive image signal processor for object detection. Advances in Neural Information Processing Systems **37**, 112598–112623 (2024)
50. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
51. Yan, J., Li, Z., Wang, F., He, Z., Fu, Z.: Dual frequency branch framework with reconstructed sliding windows attention for ai-generated image detection. IEEE Transactions on Information Forensics and Security (2026)
52. Yan, J., Li, Z., Wang, F., Wang, B., He, Z., Fu, Z.: Dgs-net: Distillation-guided gradient surgery for clip fine-tuning in ai-generated image detection. arXiv preprint arXiv:2511.13108 (2025)
53. Yan, J., Wang, F., Jiang, W., Li, Z., Fu, Z.: Ns-net: Decoupling clip semantic information through null-space for generalizable ai-generated image detection. arXiv preprint arXiv:2508.01248 (2025)
54. Yan, S.e.a.: A sanity check for ai-generated image detection. arXiv preprint arXiv:2406.19435 (2024)
55. Yan, Z.e.a.: Df40: Toward next-generation deepfake detection. Advances in Neural Information Processing Systems **37**, 29387–29434 (2024)
56. Yan, Z.e.a.: Orthogonal subspace decomposition for generalizable ai-generated image detection. arXiv preprint arXiv:2411.15633 (2024)
57. Yang, Y.a.: Dˆ 3: Scaling up deepfake detection by learning from discrepancy. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23850–23859 (2025)
58. Yu, X., Chen, K., Zeng, K., Fang, H., Yang, Z., Shang, X., Qi, Y., Zhang, W., Yu, N.: Semgir: Semantic-guided image regeneration based method for ai-generated image detection and attribution. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 8480–8488 (2024)
59. Zhang, K.Y., Chen, R., Sun, J., Wang, J., Yang, H., Yao, T., Ding, S.: A generalizable face security detection method via unified texture and semantic feature framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 3221–3229 (2025)
60. Zheng, C., Lin, C., Zhao, Z., Wang, H., Guo, X., Liu, S., Shen, C.: Breaking semantic artifacts for generalized ai-generated image detection. Advances in Neural Information Processing Systems **37**, 59570–59596 (2024)
61. Zhong, N.e.a.: Patchcraft: Exploring texture patch for efficient ai-generated image detection. arXiv preprint arXiv:2311.12397 (2023)
62. Zhu, M.e.a.: Genimage: A million-scale benchmark for detecting ai-generated image. Advances in Neural Information Processing Systems **36**, 77771–77782 (2023)