LoRC: Detecting AI-Generated Images via Low-Rank Collapse in Semantic Residuals

Haozhen Yan[1#], Ruoxin Chen[2#], Jiahui Zhan[1], Bo Wang[2], Youchang Xiao[2], Shouhong Ding[2], Liqing Zhang[1], Taiping Yao[2*], and Jianfu Zhang[1*]

[1] Shanghai Jiao Tong University, China
[2] Tencent Youtu Lab, China
{orion810, c.sis}@sjtu.edu.cn, taipingyao@tencent.com

Abstract. Modern generators faithfully model macroscopic semantics, producing synthetic images that appear highly realistic. Consequently, decisive forensic cues reside in subtle non-semantic visual discrepancies. To reveal these cues, we revisit AIGI detection from a geometric perspective and identify an architecture-agnostic signature. Specifically, modern generators exhibit low-rank collapse (*i.e.*, rank degeneracy) in the semantic-residual orthogonal subspace while largely preserving the dominant semantic direction. This structural flattening consistently emerges during the final decoding stage, forming a shared bottleneck across diverse generator architectures. Motivated by this signature, we propose **LoRC**, a framework that decouples semantic dominance to capture the collapsed residual geometry induced by the generative decoding bottleneck. Our method improves accuracy by an average of 7.0% across multiple benchmarks and achieves 97.0% accuracy on 39 unseen generators. These results demonstrate strong cross-model generalization and robustness, making LoRC a reliable approach for AIGI detection in complex real-world environments.

Keywords: AI-Generated Image Detection · Low-Rank Collapse · Orthogonal Projection

1 Introduction

The widespread availability of AI-generated images (AIGIs) [16, 21] raises growing concerns about authenticity and trust in digital content, driving the need for reliable detectors that distinguish AIGIs from real images. The rapid evolution of generative models further complicates detection by exacerbating cross-model generalization challenges, particularly in zero-shot settings involving unseen generators. From a manifold learning perspective, natural images reside on

# These authors contributed equally.
* Corresponding authors.

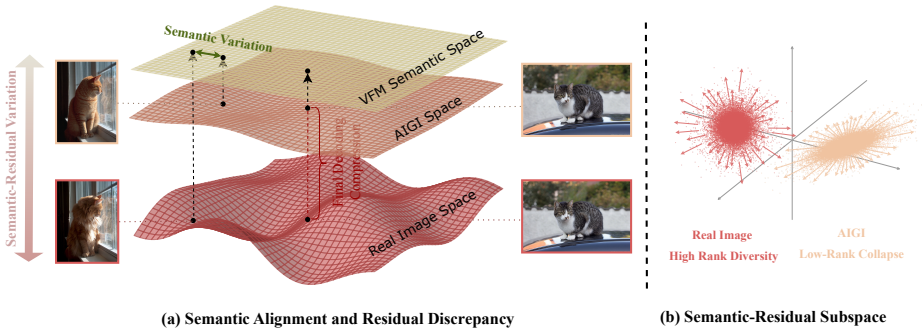


Fig. 1: Geometric perspective of AI-Generated Image (AIGI) detection. We illustrate the structural relationship between real images, AIGIs, and the semantic space of vision foundation models (VFMs). While real images reside on an intricate, high-dimensional manifold (bottom), modern generative models are increasingly aligning with them in the dominant semantic space (top), causing the *Semantic Variation* to shrink significantly. However, in the orthogonal semantic-residual dimension (vertical direction), the discrepancy between AIGIs and real images becomes correspondingly more pronounced. Our empirical analysis reveals that the final decoding stage of generative models acts as a severe information bottleneck, inducing a systematic **low-rank collapse** that forces the complex real manifold to flatten into a constrained AIGI space (middle). The resulting low-rank effect can serve as a universal forensic signature.

a high-dimensional and complex real manifold that modern generators attempt to approximate. With recent advances, generated images have become increasingly *semantically* faithful, often matching real images in object identity and global layout (Fig. 1). This is consistent with the coarse-to-fine nature of many generation pipelines [24, 43], where global structure is stabilized early and later decoding mainly refines fine details. Empirical evidence further supports this observation. As shown in Fig. 2a, generated images from diverse architectures consistently exhibit stronger semantic alignment but weaker semantic-residual energy compared with real images. As a result, the semantic dimension offers diminishing discriminative power for AIGI detection, especially under cross-model shifts and in zero-shot settings. Consequently, the remaining evidence is pushed into the semantic-residual dimensions, where subtle non-semantic imperfections persist (*e.g.*, texture naturalness and local physical consistency).

This observation naturally raises a critical question: **does a universally exploitable and systematic discrepancy actually exist within these semantic-residual dimensions?** The answer is affirmative. To reveal this discrepancy, we analyze deep visual representations by geometrically decoupling them into a primary semantic direction and its orthogonal residual subspace. As illustrated in Fig. 2b, controlled reconstructions of real images induce highly directional feature shifts that consistently reduce residual magnitude while leaving the semantic direction largely unchanged. Furthermore, the variance spectrum of these shifts (Fig. 3) reveals a striking phenomenon within the residual space:

a systematic **Low-Rank Collapse**, where the discrepancy concentrates on only a few dominant dimensions. Specifically, the reconstruction of real images induces feature shifts that are severely compressed and strictly directional in the semantic-residual dimensions, while introducing no significant variations along the semantic direction. Crucially, our empirical analysis in Fig. 4 reveals two key properties of the low-rank collapse, highlighting it as a generalizable and robust generative signature. **First, it is cross-architectural.** This phenomenon appears at the *final decoding stage* across fundamentally different paradigms, including diffusion and autoregressive models, indicating that the final generative mapping acts as a shared information bottleneck. **Second, it is inherently robust.** Because the deviation arises from the geometric structure of the feature manifold, it is largely insensitive to superficial perturbations such as JPEG compression. Together, these properties suggest that low-rank collapse transcends model-specific artifacts and provides a reliable signal for real-world forgery detection.

Motivated by these geometric insights, we propose LoRC (Leveraging Low-Rank Collapse), a framework that shifts AIGI detection from fragile spatial artifacts to robust manifold geometry. Specifically, LoRC isolates and captures this generative signature through a three-stage pipeline: First, to neutralize dominant semantic alignment, we introduce **Semantic Decomposition**. Using a frozen vision foundation model (VFM), we treat the global `[CLS]` token as a semantic anchor and project patch tokens onto its orthogonal complement. This isolates the semantic-residual subspace. Second, because VFM embeddings are high-dimensional, the subtle low-rank collapse may be obscured by irrelevant feature variations. To address this, we introduce a **Low-Rank Attention Block**. This module restricts residual features to a rank-r bottleneck before attention computation, encouraging the network to focus on the collapsed semantic-residual dimensions while suppressing noise. Finally, to explicitly shape the manifold geometry, we introduce an auxiliary **Subspace Separation Loss (SSL)**. By minimizing the inner product between normalized covariance descriptors of real and fake residuals, this loss enlarges the geometric distance between the high-rank natural manifold and the collapsed synthetic subspace. By modeling the architectural limits of modern generators rather than superficial pixel cues, our framework achieves strong robustness and transferability. Extensive evaluations validate this geometric advantage: LoRC improves performance by an average of 7.0% across multiple benchmark datasets. When evaluated on recent unseen generators spanning different architectural paradigms, our method achieves an accuracy of 97.0%. These results demonstrate that LoRC provides a robust and reliable solution for AIGI detection in unconstrained real-world scenarios.

2 Related Works

2.1 Early Low-level Detection

Early AIGI detectors primarily exploit low-level artifacts introduced by the generation pipeline. Zhang [52] shows that GAN images can leave recogniz-

able frequency-domain traces (e.g., spectral duplication) that serve as detection cues. However, later studies [13, 23] report that such frequency patterns are substantially weaker for diffusion-generated images. Building on this line, subsequent works [12, 23, 40] leverage frequency-domain artifacts as auxiliary signals to strengthen discrimination. NPR [41] further observes that generators can introduce upsampling artifacts. However, these low-level cues often exploit spurious, non-causal factors (*e.g.*, image format) and are fragile under common degradations, which limits cross-model generalization and reduces detection accuracy.

2.2 Dataset Bias and Alignment

Recent studies [18] show that detectors can inadvertently exploit non-causal dataset biases (e.g., file format, resolution, compression settings, or acquisition pipelines) as spurious cues, leading to poor cross-domain generalization. A common remedy is to align real and synthetic data in both content and statistical distributions. Diffusion-reconstruction approaches map an input image to a reconstruction generated by a pretrained diffusion model and treat the reconstruction residual as forensic evidence [44], or compare real–real and fake–fake reconstruction pairs to improve robustness [8]. Other works incorporate semantic-guided regeneration [50] or frequency-guided reconstruction errors [12] to improve alignment quality and reduce reconstruction bias. Distribution alignment has also been explored through inpainting and VAE-based reconstruction, as demonstrated in B-Free [19], AlignedForensics [36], REM [31] and M2EA [30]. Beyond pixel-space alignment, DDA [9] observes that alignment may introduce additional mismatch in the frequency domain, and therefore proposes dual alignment in both pixel and frequency spaces to suppress high-frequency bias. Overall, these works highlight the importance of bias mitigation and distribution alignment. However, they primarily operate at the data level and leave unanswered whether a more structural, cross-architectural forensic signature exists.

2.3 Semantic Decoupling

With the rise of foundation vision and vision-language models, frozen generic representations have become strong priors for in-the-wild detection. UnivFD [33] demonstrates that frozen features from CLIP-like models paired with a simple classifier substantially improve generalization to unseen generators. Simplicity Prevails [54] further argues that modern VFM representations may already encode useful forensic cues. For AIGI detection, the key forensic cues can differ from the "what is depicted" semantic cues that VFMs are trained to capture. When real and generated images are highly aligned in semantics, dominant semantic components can mask or suppress subtle but critical forensic signals. Consequently, recent works investigate semantic-forensic decoupling and efficient model adaptation. NS-Net [47] builds a semantic null-space and projects visual

features into it to eliminate semantic interference. VIB-Net [51] imposes an information bottleneck to compress and discard prior semantic knowledge, encouraging minimal sufficient representations. From an architectural perspective, Liu et al. [32] introduce a taxonomy of image generators based on their final architectural components and observe that the final component can imprint transferable traces useful for cross-generator generalization. Effort [49] performs SVD on the parameter space and fine-tunes only the smallest components to learn forensic cues efficiently. While these studies attempt to isolate forensic cues within specific parameter subspaces or architectural traces, their geometric entanglement with semantic structures remains systematically unexplored, motivating explicit subspace decoupling and geometric analysis.

3 Methodology

3.1 Motivation

Semantic Alignment and Residual Discrepancy. Recent advances in photorealistic image synthesis have significantly improved instruction adherence and perceptual realism. Generated images now closely align with real photographs in high-level structure, object identity, and relationships. However, this evolution is asymmetric and non-synchronous: high-level semantics are easier to learn and strongly constrained by high-dimensional conditioning signals (*e.g.*, text), whereas low-level image formation cues remain more entangled and harder to fit. Consequently, this uneven progress pushes the remaining real–synthetic gap into subtle **semantic-residual** dimensions beyond content and composition. Although modern vision foundation models such as CLIP [35] and DINO [6, 38] provide powerful priors, their embeddings are optimized for semantic recognition. They capture what an image depicts, rather than the subtle process-dependent cues that distinguish real images from synthetic ones. As the remaining discrepancy becomes confined to semantic-residual dimensions, naive reliance on global semantic embeddings may suppress these weak forensic signals and introduce semantic interference.

This suggests that decisive forensic evidence does not reside in global semantic embeddings, but in process-dependent residuals left by the generator. We therefore revisit the underlying generation mechanism. A recent study [32] suggests that diverse generators share a final decoding or rendering stage that maps intermediate representations to the final image (*e.g.*, a VAE decoder, a super-resolution diffusion head, or a VQ de-tokenizer). This stage may consistently imprint transferable synthetic traces across model families. Building on this insight, our geometric analysis reveals that fundamentally different paradigms, including diffusion and autoregressive models, exhibit the same structural flaw. Their final decoding stage acts as a severe information bottleneck that induces systematic **low-rank collapse** within the semantic-residual subspace. This intrinsic geometric degradation provides a universal, architecture-agnostic signal for generalized AIGI detection.

Geometric Decoupling in Feature Space. Given an input image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, we extract deep features from a frozen VFM f_ϕ, obtaining a global token and patch tokens $\{\mathbf{c}, \mathbf{X}\} = f_\phi(\mathcal{I})$, where $\mathbf{c} \in \mathbb{R}^D$ denotes the `[CLS]` token and $\mathbf{X} \in \mathbb{R}^{N\times D}$ contains N spatial patch embeddings. We normalize the frozen `[CLS]` token as $\hat{\mathbf{c}} = \mathbf{c}/\|\mathbf{c}\|_2$, which defines the principal semantic direction in feature space. We then perform an orthogonal decomposition of the patch feature map:

$$\mathbf{X}_{sem} = \mathbf{X}(\hat{\mathbf{c}}\hat{\mathbf{c}}^\top), \quad \mathbf{X}_{res} = \mathbf{X}(\mathbf{I} - \hat{\mathbf{c}}\hat{\mathbf{c}}^\top), \tag{1}$$

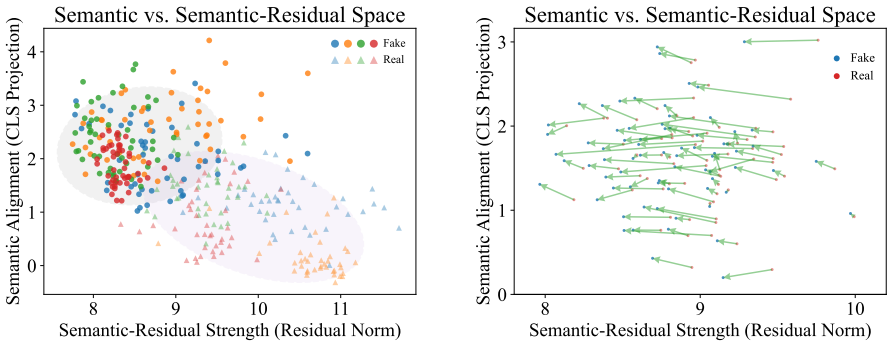
where $\mathbf{I} \in \mathbb{R}^{D\times D}$ is the identity matrix. Here, $\mathbf{X}_{sem}$ captures the dominant global semantic alignment, while $\mathbf{X}_{res}$ isolates the semantic-residual subspace that defines the remaining discrepancy between real and synthetic images.

Systematic Semantic–Residual Shift. Using this geometric formulation, we analyze feature distributions across diverse generative architectures using GenImage [55], a benchmark containing real images and their synthetic counterparts from eight generators. For each image k, we summarize the decomposed feature map using two scalar statistics that measure semantic alignment and residual magnitude:

$$Y^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i^{(k)\top}\hat{\mathbf{c}}^{(k)}, \qquad X^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \left\|\mathbf{x}_{res,i}^{(k)}\right\|_2, \tag{2}$$

where $\mathbf{x}_i^{(k)}$ and $\mathbf{x}_{res,i}^{(k)}$ denote the i-th patch tokens of $\mathbf{X}^{(k)}$ and $\mathbf{X}_{res}^{(k)}$. As shown in Fig. 2a, where colors indicate image categories, generated images (circles) from eight architectures consistently cluster in the upper-left region, exhibiting higher semantic alignment and smaller residual norms. In contrast, real images (triangles) spread toward regions with larger residual norms. This observation reveals two key findings: (i) modern generators tend to over-align with the dominant semantic direction, likely because macroscopic structures are easier to model, and (ii) semantic-residual energy is systematically attenuated in generated images.

Semantic-Controlled Reconstruction. While the previous analysis reveals a universal trend, the real and generated samples share only categorical labels. Moreover, modern image generation pipelines are complex, which may obscure the origin of the observed compression. To strictly align the semantic content and isolate the bottleneck effect, we sample MS-COCO [28] images and reconstruct them exclusively via the Stable Diffusion V1.5 [37] VAE. Using these semantically matched pairs, we examine whether the terminal decoding stage alone can induce systematic attenuation of spatial residuals. As shown in Fig. 2b, we visualize the geometric transitions of these paired samples, where each directed edge represents the feature shift from a pristine real image to its VAE-reconstructed counterpart, denoted as the **feature delta (Δ)**. The directed shifts reveal two consistent patterns: (1) The displacement vectors are predominantly oriented horizontally toward the left, indicating a strong attenuation of the semantic-residual norm. (2) Along the vertical semantic axis, most shifts show a slight upward tendency toward stronger semantic alignment. By controlling the visual



(a) Feature distributions in the decoupled semantic-residual space on the GenImage. Colors denote categories. The consistent upper-left clustering of fake samples exposes a systemic generative bottleneck: semantic over-alignment coupled with attenuated residual energy.

(b) Feature trajectories of paired real and reconstructed samples. Directed edges show systematic leftward and slightly upward shifts, demonstrating that the terminal decoding stage induces residual attenuation and semantic over-alignment.

Fig. 2: Decoupled semantic–residual feature space on GenImage (a) and SD v1.5 VAE reconstructions (b), showing a consistent generative bottleneck: semantic over-alignment with attenuated residual energy.

content, these paired trajectories provide strong evidence that the final decoding component alone can introduce a systematic structural deviation in feature space.

Low-Rank Collapse. To further analyze this structural deviation, we perform PCA on the feature deltas (Δ). The resulting variance spectra (Fig. 3) reveal a clear low-rank structure largely confined to the semantic-residual subspace. The first two principal components account for 29.1% and 16.1% of the variance, indicating that the structural discrepancy concentrates on a few dominant directions. In contrast, the semantic subspace exhibits a more dispersed variance distribution (PC1: 10.7%, PC2: 7.4%), suggesting that macroscopic semantic representations are largely unaffected by this geometric compression.

Universality and Robustness. Having established the low-rank collapse as a definitive structural signature of the SD1.5 VAE bottleneck, two critical questions naturally arise regarding its viability for real-world forgery detection:

- **Cross-Architectural Universality:** Is this pronounced geometric deformation an isolated artifact unique to the SD1.5 decoding stage, or a widespread flaw shared across diverse generative paradigms?
- **Robustness against Degradations:** Can this intrinsic manifold deviation withstand superficial pixel-level perturbations, such as severe image compression?

Regarding cross-architectural universality, we replicate this exact orthogonal decoupling analysis on mainstream architectures representing fundamentally different generative paradigms, including Stable Diffusion V1.5 [37], DiT-XL-2-

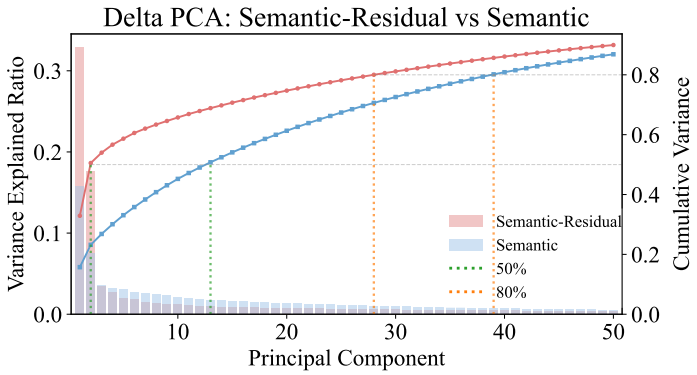


Fig. 3: PCA variance spectra of the reconstructed transition deltas (Δ). The semantic-residual subspace exhibits a stronger low-rank tendency.

512 [34], FLUX.1-dev [2], PixelFlow [10] and JanusPro-7B [11]. As shown in Fig. 4, the empirical results are remarkably consistent: regardless of their distinct underlying mechanisms, every model exhibits a similarly pronounced low-rank tendency within the semantic-residual subspace.

Similarly, we apply JPEG compression (quality factor 96) to the reconstructed images as a preprocessing step, and then repeat the same analysis pipeline. The core conclusions remain unchanged (Fig. 4): the pronounced low-rank tendency in the semantic-residual subspace persists under this lossy perturbation. This indicates that the observed structural deviation is not a trivial artifact of high-frequency reconstruction details, but rather reflects an intrinsic, stable geometric property of the decoding process.

3.2 Method

The above analyses consistently reveal a pronounced low-rank structure in the semantic-residual subspace, which remains stable across architectures and robust under lossy perturbations such as JPEG compression. This suggests that the phenomenon is intrinsic to the generative mechanisms rather than an artifact of specific models. Motivated by these insights, we propose **LoRC** (leveraging **Lo**w-**R**ank **C**ollapse AIGI Detector), a framework designed to explicitly capture and isolate the low-rank structural artifacts intrinsic to generative models (Fig. 5).

Semantic Decomposition. We instantiate the geometric operator in Eq. (1) using a **frozen** DINOv3 encoder. Trained on large-scale natural images, DINOv3 provides a semantically well-aligned feature space, injecting strong semantic priors without task-specific supervision. Within this architecture, the `[CLS]` token aggregates global context from all spatial patches and thus provides an effective reference direction for content alignment in the feature space. To preserve the pretrained natural-manifold geometry, we keep the encoder **frozen** and normal-

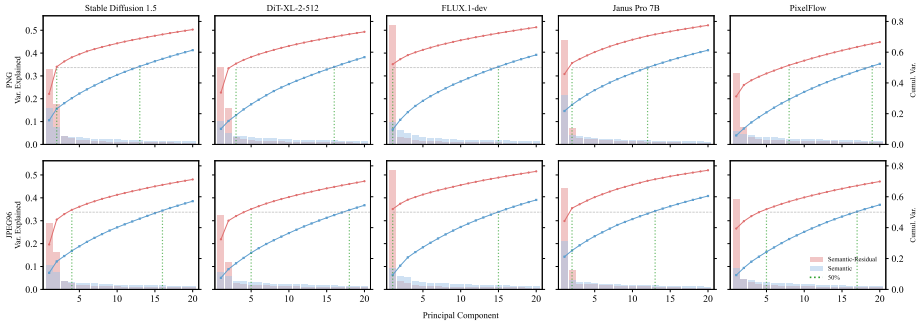


Fig. 4: PCA variance spectra of counterpart deltas (Δ). Comparison of uncompressed PNG (**top**) and JPEG (Q=96) (**bottom**) reconstructions across diverse pipelines. The consistent eigenvalue distributions verify the universality and robustness of the semantic-residual structural deviation.

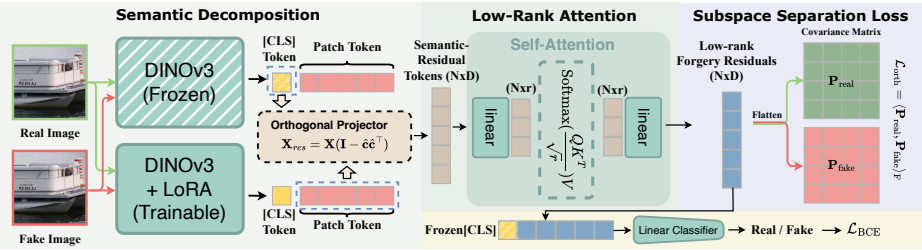


Fig. 5: LoRC Architecture. DINOv3 patch features are decomposed using the normalized frozen `[CLS]` token as a semantic anchor, projecting tokens onto its orthogonal complement to obtain semantic-residuals. A Low-Rank Attention Block performs self-attention in a rank-r bottleneck to amplify structured low-rank forensic cues. The pooled residual feature, concatenated with the frozen `[CLS]` token, is fed to a linear classifier. An auxiliary Subspace Separation Loss further encourages decorrelation between real and fake residual subspaces, reinforcing their geometric distinction.

ize the `[CLS]` token as a sample-specific *semantic anchor* for geometric decoupling. We then project patch tokens onto the orthogonal complement of this anchor to attenuate dominant semantic components, thereby shifting real/fake discrimination from a generic semantic feature space into a dynamic forgery subspace conditioned on the underlying semantics.

Low-Rank Forensic Signal Modeling. The low-rank collapse induced by generative decoders implies that residual discrepancies concentrate in a structured low-dimensional subspace. However, modern VFM embeddings remain inherently high-dimensional (D), which can obscure these structured low-rank patterns among irrelevant feature fluctuations. To enhance sensitivity to low-rank residual cues, we introduce a **Low-Rank Attention Block** that performs self-attention in a rank-r bottleneck space ($r \ll D$). Specifically, we project $\mathbf{X}_{res}$

into $\mathbb{R}^r$ to form the Query, Key, and Value embeddings:

$$\mathbf{Q} = \mathbf{X}_{res}\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{X}_{res}\mathbf{W}^K, \quad \mathbf{V} = \mathbf{X}_{res}\mathbf{W}^V, \tag{3}$$

where $\mathbf{W}^{\{Q,K,V\}} \in \mathbb{R}^{D\times r}$. Self-attention is then performed within this compressed rank-r manifold to capture spatial coherence without noise interference:

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{r}}\right)\mathbf{V}. \tag{4}$$

Finally, the aggregated low-rank features $\mathbf{A} \in \mathbb{R}^{N\times r}$ are projected back to the original dimension via an output linear transformation $\mathbf{W}^O \in \mathbb{R}^{r\times D}$:

$$\mathbf{Y} = \mathbf{A}\mathbf{W}^O. \tag{5}$$

This design ensures that the model focuses exclusively on the dominant low-rank components where the real/fake discrepancy is most pronounced. Meanwhile, we retain the *frozen* `[CLS]` token as a global context anchor to condition the decision and concatenate it with the pooled low-rank residual feature before a linear classifier.

Subspace Separation Loss (SSL). Binary classification alone does not explicitly enforce the geometric separation we hypothesize: real residuals should span a high-dimensional subspace while fake residuals collapse onto a low-dimensional one. We therefore introduce an auxiliary loss that directly operates on the subspace geometry. Given a mini-batch, we partition $\mathbf{X}_{res}$ into real and fake subsets according to the labels. For each subset, all patch residuals are flattened into a matrix $\mathbf{R} \in \mathbb{R}^{BN\times D}$. We then compute a normalized covariance descriptor:

$$\mathbf{P} = \frac{\mathbf{R}^\top\mathbf{R}}{\|\mathbf{R}^\top\mathbf{R}\|_F}. \tag{6}$$

Let $\mathbf{P}_{\text{real}}$ and $\mathbf{P}_{\text{fake}}$ denote the descriptors of real and fake residuals. To encourage their subspaces to diverge, we minimize their Frobenius inner product:

$$\mathcal{L}_{\text{SS}} = \langle \mathbf{P}_{\text{real}}, \mathbf{P}_{\text{fake}} \rangle_F. \tag{7}$$

The total training objective is $\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda_{\text{SS}}\,\mathcal{L}_{\text{SS}}$.

4 Experiments

4.1 Implementation

We employ DINOv3 ViT-H+/16 [38] as the backbone, fine-tuned via LoRA [22] with rank 16 and scaling factor $\alpha = 16$. The model is trained on the DDA-Training-Set, where fake samples are generated via SD2.1 [37] VAE reconstruction of MSCOCO [28] source images. This setup aligns semantic content, compelling the model to focus specifically on capturing low-rank collapse effects. Images are randomly cropped to 224×224 for training and center-cropped for inference, with padding applied for insufficient dimensions. We use the Adam [25] optimizer ($lr = 10^{-4}$) with a batch size of 64 and 4 gradient accumulation steps. The Low-Rank Attention Block rank is set to $r = 32$, and $\lambda_{\text{SS}} = 0.1$.

Table 1: Overall comparison across 7 benchmarks. To ensure fairness and reproducibility, we use official checkpoints released by each method. JPEG compression with a quality factor of 96 is applied to the synthetic images in GenImage and AIGCDetectionBenchmark to mitigate format bias. Bold numbers indicate the best performance.

Method	Standard				In-the-Wild		Recent Generators	Avg.
	GenImage	DRCT-2M	Synthbuster	AIGCDetection Benchmark	Chameleon	WildRF	T2I-CoReBench	
NPR (CVPR'24) [41]	51.5	37.3	50.0	53.1	59.9	63.5	60.0	53.6
UnivFD (CVPR'23) [33]	64.1	61.8	67.8	72.5	50.7	55.3	10.8	54.7
FatFormer (CVPR'24) [29]	62.8	52.2	56.1	85.0	51.2	58.9	7.7	53.4
SAFE (KDD'25) [26]	50.3	59.3	46.5	50.3	59.2	57.2	1.0	46.3
C2P-CLIP (AAAI'25) [39]	74.4	59.2	68.5	81.4	51.1	59.6	70.3	66.4
AIDE (ICLR'25) [48]	61.2	64.6	53.9	63.6	63.1	58.4	9.5	53.5
DRCT (ICML'24) [8]	84.7	90.5	81.3	81.4	56.6	50.6	48.6	71.0
AlignedForensics (ICLR'25) [36]	79.0	95.5	77.4	66.6	71.0	80.1	42.6	73.2
DDA (NIPS'25) [9]	91.7	98.1	90.1	87.8	82.4	90.3	91.1	90.2
LoRC (ours)	**97.9**	**99.3**	**99.9**	**96.3**	**92.6**	**97.6**	**97.0**	**97.2**

4.2 Evaluation Protocol and Benchmark

We evaluate the out-of-distribution (OOD) generalization capabilities across a series of benchmarks. Results for all baseline methods are derived from their official implementations and pre-trained weights. All testing follows the protocol of DDA [9]. PNG images are recompressed using JPEG (quality factor 96) for consistency. Balanced Accuracy (B.Acc) is used as the primary evaluation metric.

We categorize the evaluation benchmarks into three groups: (1) **Standard Benchmarks.** GenImage [55], AIGCDetectionBenchmark [53], Synthbuster [1], DRCT-2M [8]. (2) **In-the-wild Benchmarks.** To evaluate performance against unpredictable real-world degradations and unconstrained generation pipelines, we test on complex in-the-wild datasets, such as Chameleon [48] and WildRF [7]. (3) **Recent SOTA Generators.** To confront the rapid evolution of generative technologies, we perform extensive zero-shot evaluation on **T2I-CoReBench** [27]. This benchmark comprises images generated from 1,080 carefully curated challenging prompts, deliberately designed to cover complex compositional structures and reasoning-intensive scenarios under real-world conditions. It includes **39** cutting-edge text-to-image (T2I) models spanning diverse architectures, including **Diffusion Models** (e.g., SD-3.5-Large [15], FLUX.2-klein-9B [3], Z-Image-Turbo [42]), **Autoregressive Models** (e.g., Infinity-8B [20], GoT-R1-7B [14]), **Unified Models** (e.g., Show-o2-7B [46], OmniGen2-7B [45], Hunyuan-Image-3.0 [5]), as well as **closed-source commercial models** (e.g., Seedream 4.5 [4] and Nano-Banana-Pro [17]). Each model contributes a substantial volume of **4,320** generated images. Overall, T2I-CoReBench supports large-scale benchmarking for AIGI detection across diverse generated model families.

4.3 Overall Results

As summarized in Tab. 1, LoRC performs consistently well across all seven datasets under three evaluation settings, achieving an overall accuracy of **97.2%**

Table 2: Comparison of balanced accuracy on Synthbuster.

Method	DALL·E 2	DALL·E 3	Firefly	GLIDE	Midjourney	SD 1.3	SD 1.4	SD 2	SDXL	Avg.
NPR (CVPR'24) [41]	51.1	49.3	46.5	48.5	52.8	51.4	51.8	46.0	52.8	50.0
UnivFD (CVPR'23) [33]	83.5	47.4	89.9	53.3	52.5	70.4	69.9	75.7	68.0	67.8
FatFormer (CVPR'24) [29]	59.4	39.5	60.3	72.7	44.4	53.7	54.0	52.3	69.1	56.1
SAFE (KDD'25) [26]	58.0	9.9	10.3	52.2	56.7	59.4	59.1	53.0	59.5	46.5
C2P-CLIP (AAAI'25) [39]	55.6	63.2	59.5	86.7	52.9	75.2	76.7	69.2	77.7	68.5
AIDE (ICLR'25) [48]	34.9	33.7	24.8	65.0	57.5	74.1	73.7	53.2	68.4	53.9
DRCT (ICML'24) [8]	77.2	86.6	84.1	82.6	73.7	86.6	86.6	83.2	71.3	81.3
AlignedForensics (ICLR'25) [36]	50.2	48.9	51.7	53.5	98.7	98.8	98.8	98.6	97.3	77.4
DDA (NIPS'25) [9]	86.3	90.0	91.9	76.5	93.5	92.9	92.7	93.3	93.5	90.1
LoRC (ours)	**99.6**	**99.9**	**99.6**	**99.8**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**99.9**

Table 3: Ablation study of the LoRC where modules are added cumulatively.

Setting	**Standard**	**In-the-wild**	**T2I-CoReBench**	**Avg.**
Baseline	97.9	89.4	94.0	93.8
+ Semantic Decomposition	97.2	86.9	**98.1**	94.1
+ Low-Rank Attention	96.9	90.2	96.0	94.4
+ Subspace Separation Loss	**98.1**	**95.2**	97.0	**96.8**

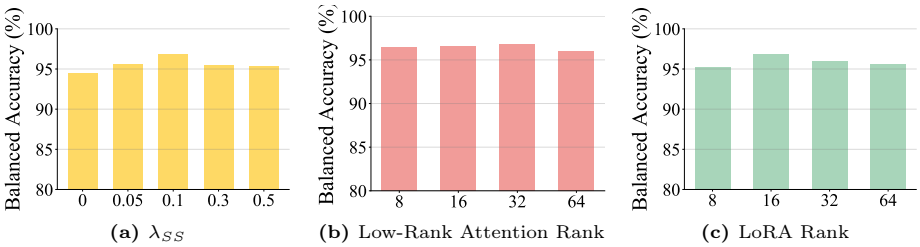
and surpassing the second-best method by **7.0%**. This consistent advantage indicates that LoRC generalizes well across diverse data distributions. On standard benchmarks, LoRC delivers clear and consistent improvements over prior methods on GenImage, DRCT-2M, Synthbuster and AIGCDetectionBenchmark. In particular, it attains a near-perfect mean accuracy on Synthbuster (Tab. 2), suggesting strong generalization beyond source-specific artifacts. **On in-the-wild datasets**, LoRC achieves 92.6% accuracy on Chameleon and 97.6% on WildRF. Notably, on the challenging WildRF benchmark, LoRC outperforms the competitive DDA by **7.3%**, highlighting its effectiveness in highly diverse and uncurated real-world scenarios. Detailed results for each benchmark are provided in the Supplementary Material.

4.4 Zero-Shot Comparison on SOTA Generators.

As generative models evolve at an unprecedented pace, generalizing to unseen architectures remains the most formidable challenge in AIGI detection. We evaluate this zero-shot capability on T2I-CoReBench (Tab. 4). Interestingly, while most baseline methods suffer severe performance degradation, only DDA maintains moderate robustness by aligning frequency-domain statistics. In contrast, by explicitly capturing the universal low-rank collapse injected by the final decoding stage, our proposed LoRC achieves a state-of-the-art average accuracy of **97.0%**, yielding an impressive improvement of **5.9%** over existing approaches. These compelling results thoroughly validate the unparalleled robustness of our method, highlighting its immense potential for defending against rapidly emerging generative models.

Table 4: Comparison of fake accuracy on T2I-CoReBench. C2P., Fat., and Align. are abbreviations for C2P-CLIP, FatFormer, and AlignedForensics, respectively. All test images are compressed using JPEG with a quality factor of 96.

Generator	NPR	UnivFD	C2P.	Fat.	SAFE	AIDE	DRCT	Align.	DDA	LoRC
Diffusion Models										
SD-3-Medium	51.6	5.0	71.0	3.8	1.6	19.2	59.4	60.5	96.8	**99.6**
SD-3.5-Medium	61.6	9.9	66.7	12.9	1.6	7.2	42.4	68.5	97.0	**99.7**
SD-3.5-Large	66.1	12.7	53.0	7.6	1.4	4.2	57.0	58.9	97.0	**99.8**
FLUX.1-schnell	51.6	4.5	59.9	2.5	0.7	16.8	36.0	27.1	94.8	**98.0**
FLUX.1-dev	58.7	3.5	73.5	0.8	0.4	18.7	33.5	21.0	90.6	**96.2**
FLUX.1-Krea-dev	44.9	4.6	**86.4**	1.0	2.6	15.1	22.6	4.2	30.7	81.2
FLUX.2-dev	65.9	3.0	75.9	1.7	0.6	4.0	15.8	2.0	82.7	**91.2**
FLUX.2-klein-4B	52.7	3.2	75.0	1.3	0.5	3.6	12.3	4.1	90.5	**96.6**
FLUX.2-klein-9B	57.0	1.7	67.9	0.8	0.5	2.1	4.6	1.7	91.0	**94.9**
PixArt-α	68.2	16.0	72.9	16.3	1.5	20.8	79.8	96.2	99.5	**99.8**
PixArt-Σ	68.5	15.1	67.6	19.3	1.9	23.3	70.3	89.5	99.5	**99.9**
HiDream-I1	55.4	2.6	60.0	0.3	0.4	9.1	54.9	33.5	83.6	**97.7**
Qwen-Image	86.2	5.2	93.6	1.8	0.3	10.8	35.7	21.7	97.8	**98.6**
Qwen-Image-2512	93.6	9.2	94.4	3.5	0.1	1.8	27.6	86.2	98.7	**99.6**
LongCat-Image	69.1	2.9	65.1	1.8	2.0	2.7	14.6	2.1	95.0	**99.7**
Z-Image	65.6	6.1	74.9	4.5	0.6	9.5	54.2	13.7	96.9	**98.7**
Z-Image-Turbo	43.3	3.5	72.9	0.4	0.4	10.4	56.9	3.2	76.4	**92.8**
Autoregressive Models										
Infinity-8B	67.5	26.9	80.4	24.3	0.7	4.3	78.7	80.0	99.8	**100.0**
GoT-R1-7B	7.7	16.0	72.3	12.8	0.4	6.0	42.1	38.4	100.0	**100.0**
Unified Models										
BAGEL	56.4	8.9	63.1	2.6	0.7	14.6	55.3	17.1	95.4	**99.8**
BAGEL w/Think	62.3	7.1	74.9	2.9	0.6	19.2	57.0	16.3	96.9	**100.0**
show-o2-1.5B	60.5	16.5	73.2	15.1	0.3	20.1	83.1	81.3	99.4	**100.0**
show-o2-7B	9.9	10.0	53.8	15.4	0.1	3.0	61.6	31.9	99.7	**100.0**
Janus-Pro-1B	11.3	44.9	71.6	38.8	0.7	10.0	52.1	54.6	99.4	**99.9**
Janus-Pro-7B	12.9	25.2	68.3	16.1	0.4	7.2	52.3	59.0	99.7	**100.0**
BLIP3o-4B	52.9	19.9	65.6	29.8	2.2	6.0	67.5	75.5	99.7	**100.0**
BLIP3o-8B	49.8	24.2	56.4	25.3	2.1	8.9	63.7	78.5	99.6	**99.9**
OmniGen2-7B	58.3	8.6	77.7	7.9	1.3	28.5	55.1	37.9	97.0	**99.6**
HunyuanImage-3.0	91.8	9.5	77.5	3.4	3.6	3.5	72.3	99.0	99.3	**99.9**
Closed-Source Models										
Seedream 3.0	56.0	4.1	80.6	0.2	0.2	20.2	52.8	19.1	95.8	**99.5**
Seedream 4.0	90.4	22.8	58.4	5.5	0.7	0.0	77.5	81.0	95.4	**99.9**
Seedream 4.5	92.6	26.2	67.0	4.2	0.3	0.0	45.5	80.6	95.7	**99.6**
Gemini 2.0 Flash	49.9	15.7	74.8	4.1	2.3	23.8	73.3	23.8	94.2	**98.5**
Nano Banana	67.5	3.8	54.8	2.4	0.7	1.6	50.9	43.4	70.2	**94.8**
Nano Banana Pro	79.5	4.4	86.0	2.2	0.5	2.9	80.7	80.7	93.2	**94.8**
Imagen 4	57.4	2.9	74.2	1.6	0.1	6.0	45.5	27.3	**90.7**	85.5
Imagen 4 Ultra	73.9	3.4	75.1	1.7	0.2	5.3	48.0	34.9	**90.4**	88.2
GPT-Image 1.5	82.9	5.4	75.4	1.4	4.1	0.3	2.4	6.9	57.5	**91.1**
GPT-4o	87.9	4.2	30.3	2.7	1.2	1.0	0.9	0.1	66.5	**88.8**
Average	60.0	10.8	70.3	7.7	1.0	9.5	48.6	42.6	<u>91.1</u>	**97.0**



(a) λ_{SS} **(b)** Low-Rank Attention Rank **(c)** LoRA Rank

Fig. 6: Hyperparameter sensitivity. (a) The impact of the weight parameter λ_{SS} on Subspace Separation Loss. (b) The impact of the rank in the Low-Rank Attention Block. (c) The effect of varying the LoRA rank during fine-tuning.

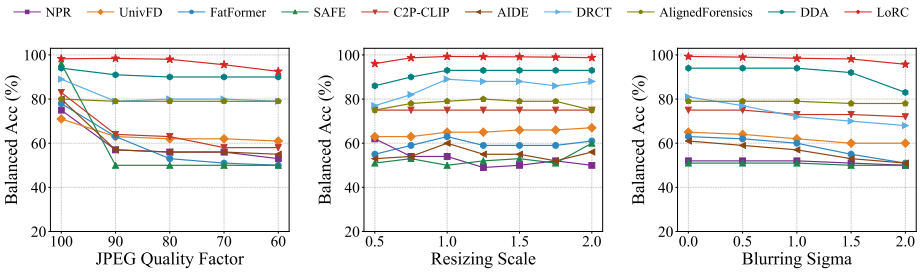


Fig. 7: Robustness analysis on GenImage.

4.5 Ablation Study

Table 3 summarizes an ablation study where we cumulatively add each component of LoRC. Semantic Decomposition notably improves T2I-CoReBench [27], indicating that suppressing dominant semantics is crucial for modern generators with strong semantic alignment. Low-Rank Attention boosts in-the-wild performance by better modeling the collapsed low-rank residual cues. Finally, Subspace Separation Loss enlarges the real/fake subspace gap, yielding the best overall results.

4.6 Hyperparameter Sensitivity

Fig. 6 shows the ablation results of three key hyperparameters. The subspace separation loss improves performance when $\lambda_{SS} > 0$, with the best balanced accuracy at $\lambda_{SS} = 0.1$. The optimal feature rank for the Low-Rank Attention Block is 32, indicating a suitable trade-off between compression and information preservation. For LoRA, rank 16 gives the best result by balancing adaptation capacity and parameter efficiency.

4.7 Robustness

In real-world applications, images frequently undergo various types of degradation. To evaluate practical reliability, we further investigate the robustness

of LoRC. Figure 7 illustrates the model performance under three common perturbations: JPEG compression, image resizing, and Gaussian blurring. LoRC maintains consistently high balanced accuracy across all test conditions. This consistent stability demonstrates that LoRC effectively captures the underlying low-rank structural effect, which provides inherent robustness against common pixel-level degradations.

5 Conclusion

We revisit AI-generated image detection from a geometric perspective and uncover an architecture-agnostic signature: modern generators consistently induce a pronounced *low-rank collapse* in the semantic-residual orthogonal subspace, while largely preserving the dominant semantic direction. This reveals a stable geometric degeneracy introduced by the final decoding bottleneck, shared across diverse generative paradigms and resilient to common post-processing. Motivated by this observation, we propose **LoRC**, which suppresses semantic dominance via orthogonal decomposition and explicitly models the collapsed residual geometry with a low-rank attention bottleneck, further reinforced by a subspace separation objective. Extensive evaluations demonstrate that LoRC yields a 7.0% absolute average improvement across 7 different benchmarks, reaching a 97.2% overall accuracy. In addition, it exhibits strong zero-shot generalization on rapidly evolving generator suites by achieving 97.0% accuracy across **39** unseen generators, which firmly supports low-rank collapse as a reliable universal signature for real-world AIGI detection.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62302295, 62595733, and 62561160155), the Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZDZX0102). This work was also supported by Tencent Youtu Lab.

References

1. Bammey, Q.: Synthbuster: Towards detection of diffusion model generated images. OJSP **5**, 1–9 (2023)
2. Black Forest Labs: FLUX.1 [dev]. `https://huggingface.co/black-forest-labs/FLUX.1-dev` (2024)
3. Black Forest Labs: FLUX.2 [klein] 9b. `https://huggingface.co/black-forest-labs/FLUX.2-klein-9B` (2026)
4. ByteDance Seed: Seedream 4.5. `https://seed.bytedance.com/en/seedream4_5` (2025)
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. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021)
7. Cavia, B., Horwitz, E., Reiss, T., Hoshen, Y.: Real-time deepfake detection in the real-world. arXiv preprint arXiv:2406.09398 (2024)
8. Chen, B., Zeng, J., Yang, J., Yang, R.: Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In: ICML. pp. 7621–7639 (2024)
9. 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: NIPS (2025)
10. Chen, S., Ge, C., Zhang, S., Sun, P., Luo, P.: Pixelflow: Pixel-space generative models with flow. arXiv preprint arXiv:2504.07963 (2025)
11. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
12. Chu, B., Xu, X., Wang, X., Zhang, Y., You, W., Zhou, L.: Fire: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In: CVPR. pp. 12830–12839 (2025)
13. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L.: On the detection of synthetic images generated by diffusion models. In: ICASSP. pp. 1–5 (2023)
14. Duan, C., Fang, R., Wang, Y., Wang, K., Huang, L., Zeng, X., Li, H., Liu, X.: Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. arXiv preprint arXiv:2505.17022 (2025)
15. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML. pp. 12606–12633 (2024)
16. Goodfellow, I.J., et al.: Generative adversarial nets. In: NIPS. pp. 2672–2680 (2014)
17. Google DeepMind: Gemini 3 pro image model card. `https://deepmind.google/models/model-cards/gemini-3-pro-image` (2025)
18. Grommelt, P., Weiss, L., Pfreundt, F.J., Keuper, J.: Fake or jpeg? revealing common biases in generated image detection datasets. In: ECCVW. pp. 80–95 (2024)
19. 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: CVPR. pp. 18685–18694 (2025)
20. Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., Liu, X.: Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In: CVPR. pp. 15733–15744 (2025)
21. Ho, J., et al.: Denoising diffusion probabilistic models. In: NIPS. vol. 33, pp. 6840–6851 (2020)
22. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
23. Karageorgiou, D., Papadopoulos, S., Kompatsiaris, I., Gavves, E.: Any-resolution ai-generated image detection by spectral learning. In: CVPR. pp. 18706–18717 (2025)
24. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: ICLR (2018)
25. Kingma, D.P., et al.: Adam: A method for stochastic optimization. In: ICLR (2015)

26. Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Feng, F.: Improving synthetic image detection towards generalization: An image transformation perspective. In: KDD. pp. 2405–2414 (2025)
27. Li, O., Wang, Y., Hu, X., Huang, H., Chen, R., Ou, J., Tao, X., Wan, P., Qi, X., Feng, F.: Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? In: ICLR (2026)
28. Lin, T.Y., et al.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)
29. Liu, H., Tan, Z., Tan, C., Wei, Y., Wang, J., Zhao, Y.: Forgery-aware adaptive transformer for generalizable synthetic image detection. In: CVPR. pp. 10770–10780 (2024)
30. Liu, R., Han, Y., Sun, B., He, H., Zhou, Y., Wang, Y.: M2ea: Multi-vae manifold envelope alignment for challenging ai-generated image detection. In: ICASSP. pp. 9147–9151. IEEE (2026)
31. Liu, R., Han, Y., Zhang, Z., Yao, L., Yan, Z., Shen, J., Chen, Z., Sun, B., Weng, L., Dong, J., et al.: Beyond artifacts: Real-centric envelope modeling for reliable ai-generated image detection. arXiv preprint arXiv:2512.20937 (2025)
32. Liu, Y., Liu, X., Wang, Y., Soumik, M.: Exploiting the final component of generator architectures for ai-generated image detection. arXiv preprint arXiv:2601.20461 (2026)
33. Ojha, U., et al.: Towards universal fake image detectors that generalize across generative models. In: CVPR. pp. 24480–24489 (2023)
34. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
36. Rajan, A.S., Ojha, U., Schloesser, J., Lee, Y.J.: Aligned datasets improve detection of latent diffusion-generated images. In: ICLR (2025)
37. Rombach, R., et al.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
38. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
39. Tan, C., Tao, R., Liu, H., Gu, G., Wu, B., Zhao, Y., Wei, Y.: C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In: AAAI. vol. 39, pp. 7184–7192 (2025)
40. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In: AAAI. vol. 38, pp. 5052–5060 (2024)
41. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In: CVPR. pp. 28130–28139 (2024)
42. Team, Z.I.: Z-image: An efficient image generation foundation model with single-stream diffusion transformer. arXiv preprint arXiv:2511.22699 (2025)
43. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. In: NIPS. vol. 37, pp. 84839–84865 (2024)
44. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: Dire for diffusion-generated image detection. In: ICCV. pp. 22445–22455 (2023)

45. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
46. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. In: NIPS. vol. 38, pp. 47490–47518 (2025)
47. 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)
48. Yan, S., Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Xie, W.: A sanity check for AI-generated image detection. In: ICLR (2025)
49. Yan, Z., Wang, J., Jin, P., Zhang, K.Y., Liu, C., Chen, S., Yao, T., Ding, S., Wu, B., Yuan, L.: Orthogonal subspace decomposition for generalizable AI-generated image detection. In: ICML. pp. 70268–70288 (2025)
50. 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: ACM MM. pp. 8480–8488 (2024)
51. Zhang, H., He, Q., Bi, X., Li, W., Liu, B., Xiao, B.: Towards universal ai-generated image detection by variational information bottleneck network. In: CVPR. pp. 23828–23837 (2025)
52. Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in gan fake images. In: WIFS. pp. 1–6. IEEE (2019)
53. Zhong, N., Xu, Y., Li, S., Qian, Z., Zhang, X.: Patchcraft: Exploring texture patch for efficient ai-generated image detection. arXiv preprint arXiv:2311.12397 (2024)
54. Zhou, Y., He, X., Lin, K., Fan, B., Ding, F., Li, B.: Simplicity prevails: The emergence of generalizable aigi detection in visual foundation models. arXiv preprint arXiv:2602.01738 (2026)
55. Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y.: Genimage: A million-scale benchmark for detecting ai-generated image. In: NIPS. pp. 77771–77782 (2023)