# GLARE: Towards Generalizable Detection of Latent Diffusion Images with Global–Local Reconstruction Error

Jiangtao Yan[1], Jiazhen Ji[1⋆], Zhongyu Zhang[1], Yuge Huang[1], Wenbin Wang[2], and Shouhong Ding[1]

[1] Tencent YouTu Lab, China
{jiangtyan, royji, wilxyzhang, yugehuang, ericshding}@tencent.com
[2] Wuhan University, China
wangwenbin97@whu.edu.cn

**Abstract.** Latent diffusion models (LDMs) have emerged as a leading paradigm for image generation by performing denoising in a compressed latent space, which provides high fidelity and efficiency but also intensifies forensic and security concerns. While most existing detection approaches rely on supervised learning, training-free detectors are appealing for their simplicity and ease of deployment. However, their performance remains limited and degrades markedly on novel generators. Our analysis reveals a more pronounced context dependence in LDM-generated images compared to real ones. Motivated by this, we introduce GLARE, a novel training-free method designed to exploit this dependence as a detection signal. Specifically, GLARE measures the relative difference between full-image and patch-wise reconstruction errors within a shared autoencoder. Dividing the image into patches implicitly removes global context, thereby inducing a characteristic shift in reconstruction error that serves as a discriminative feature. Furthermore, a lightweight semantic complexity calibration is incorporated to compensate for content-induced variation. Extensive experiments across a wide range of generators demonstrate the effectiveness and remarkable generalization capability of GLARE. Our method outperforms state-of-the-art supervised and training-free baselines significantly and shows strong robustness against common post-processing operations.

**Keywords:** AI-generated image detection · latent diffusion models · training-free detection

## 1 Introduction

Image generation has advanced rapidly, spanning generative adversarial networks [22], variational autoencoders [30], normalizing flows [29], and autoregressive models [52]. Recently, diffusion models have achieved extraordinary progress in image synthesis, yielding increasingly photorealistic results [20, 26]. Building

⋆ Corresponding author.

on this progress, latent diffusion models (LDMs) further improve throughput and visual quality by operating in a compressed latent space [47, 51], enabling high-quality synthesis at scale. While these advances unlock compelling applications, they also amplify risks of misuse, underscoring the urgency for reliable and generalizable detectors of AI-generated images [37]. Prior training-based detectors [10, 12, 17, 23, 32, 44, 57] have achieved promising in-distribution accuracy, yet they rely on curated data and supervised training, incurring nontrivial retraining costs as generators evolve and often degrading under distribution shifts induced by novel architectures or unseen scenarios. Training-free alternatives are therefore appealing for their simplicity and immediate deployability.

Previous training-free approaches can be broadly grouped into two categories: 1) Reconstruction-based methods [18, 46] measure how well a pretrained model can reconstruct or encode the input: AEROBLADE [46] leverages autoencoder reconstruction error, while ZED [18] uses a learned lossless codec to estimate coding cost as a proxy for reconstruction fidelity. Both rely on a single-path absolute score whose magnitude is heavily influenced by image content, yielding fragile discriminative boundaries across domains; 2) Perturbation-based methods probe how a pretrained model's response changes under controlled input transformations, *e.g.*, noise injection [25] or data augmentation consistency [60]. They go beyond a single absolute score by comparing responses under different conditions, yet their performance is sensitive to the choice of perturbation type and strength, which often requires dataset-specific tuning and degrades on unseen domains. Moreover, perturbation-based methods do not exploit the structural consequence of the iterative denoising process that distinguishes LDM outputs; instead, they rely on the assumption that pretrained vision models (*e.g.*, DINOv2 [38]) trained on large-scale real data are inherently robust to perturbations of real images.

We approach the problem from a different angle. Rather than measuring how an image responds to a single signal or external perturbations, we start from the LDM generation process itself and focus on the intrinsic structural traces that iterative denoising imprints on the generated image. The key observation is that denoising steps in the LDM latent space propagate conditioning signals across all spatial locations via cross-attention and multi-step refinement [20, 47], binding local patterns to scene-level context far more tightly than in natural images whose dependencies are predominantly local [55]. We refer to this strengthened spatial correlation as **context dependence**. We instantiate this idea in **GLARE** (**G**lobal–**L**ocal difference of **A**utoencoder **R**econstruction **E**rrors), a training-free detector that reconstructs every test image twice through a single pretrained LDM autoencoder—once as a whole and once after partitioning it into independent patches that sever long-range dependencies—and uses the log-ratio of the two reconstruction errors as its detection score. Because this relative change cancels content-dependent baselines shared by both reconstruction paths, it is inherently more robust than any single-path absolute score. A lightweight **S**emantic **C**omplexity **C**alibration **M**odule (SCCM) further normalizes the score

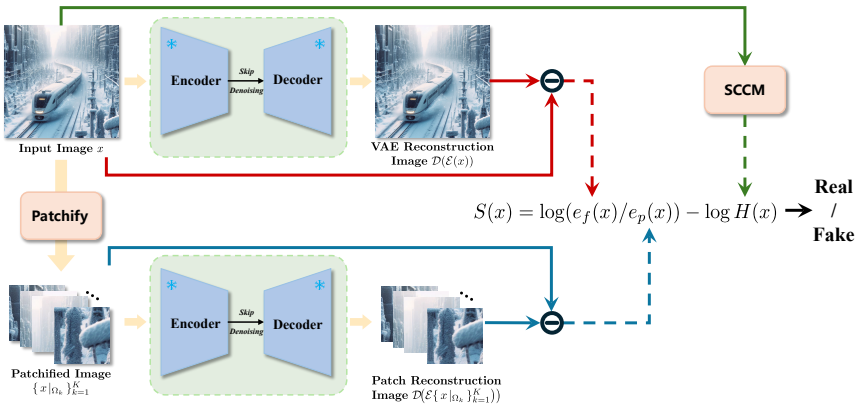



**Fig. 1: Overall framework of GLARE.** Given an input image, we create a patchified variant and pass both through the same pretrained LDM autoencoder, bypassing diffusion denoising to obtain full-image and patch-based reconstructions. The detection score is computed as the logarithmic difference between their reconstruction errors, further calibrated by a semantic complexity term from the SCCM module.

against content complexity. The entire pipeline (Fig. 1) requires no training data and generalizes across LDM generative pipelines.

In summary, our contributions are:

- We reveal that iterative denoising in LDMs induces elevated inter-patch context dependence, and identify its observable proxy, termed *context sensitivity* (the change in AE reconstruction error upon removal of global context), as a new detection cue.
- We propose GLARE, a simple yet effective training-free detector that operationalizes context sensitivity through dual-path reconstruction and logarithmic error differencing, complemented by semantic complexity calibration to suppress content-induced bias.
- Extensive experiments on three benchmarks spanning over 20 generators show that GLARE significantly outperforms all existing training-free methods and most supervised baselines, while maintaining strong robustness under common post-processing perturbations.

## 2 Preliminaries

### 2.1 LDM Image Generation

Latent diffusion models (LDMs) [47] perform generation in a compressed latent space $\mathcal{Z}$ defined by a pretrained autoencoder (AE) $(\mathcal{E}, \mathcal{D})$. Given an image $x$, the encoder produces $z_0 = \mathcal{E}(x)$ and the decoder reconstructs $\hat{x} = \mathcal{D}(z_0)$. AE training minimizes a reconstruction-regularization objective:

$$\min_{\mathcal{E},\mathcal{D}} \ \mathbb{E}\big[\ell\big(x, \mathcal{D}(\mathcal{E}(x))\big)\big] \ + \ \beta\, D_{\mathrm{KL}}\big(q(z \mid x) \,\|\, p(z)\big), \tag{1}$$

where $\ell$ is a perceptual or pixel-level loss and the KL term regularizes the latent distribution toward a standard Gaussian prior $p(z)$. Due to the information bottleneck imposed by the KL regularization, the encoder preferentially retains globally coherent, low-frequency structure while discarding fine-grained details [39], a property we term *semantic compression*.

On top of this latent space, LDMs learn a reverse denoising process. Using the standard $\epsilon$-prediction parameterization, a single sampling step reads

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t, t, c)\right) + \sigma_t \eta, \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, $\sigma_t$ is the noise scale, $\eta \sim \mathcal{N}(0, \mathbf{I})$, and $c$ denotes the conditioning signal (*e.g.*, text). Critically, the denoiser $\epsilon_\theta$ operates on the full spatial extent of $z_t$: cross-attention layers, global feature pooling, and the iterative multi-step refinement jointly propagate conditioning semantics across distant spatial locations, establishing long-range dependencies in the generated latent [20, 47].

### 2.2 Generated Image Detection

**Training-based methods** learn discriminative features from paired real/fake datasets. Representative approaches include NPR [50], which captures upsampling traces via neighboring pixel relationships; FatFormer [34], which adapts a frozen CLIP backbone with forgery-aware adapters; AIDE [57], which fuses semantic and low-level patch features; and ForgeLens [13], which steers a pre-trained backbone toward forgery-relevant cues with lightweight trainable components. More recently, TranX-Adapter [56] bridges low-level artifacts and high-level semantics within multimodal large language models for more robust detection. A growing line of work utilizes AE reconstruction as a feature source: FIRE [16] decomposes reconstruction errors in the frequency domain to isolate bands characteristic of LDM outputs, while LaRE$^2$ [36] and LATTE [53] refine latent-space reconstruction trajectories with learned modules. Despite strong in-distribution results, these methods depend on curated training data and leverage only a single reconstruction error to guide feature extraction.

**Training-free methods** sidestep data dependence by directly leveraging pretrained generative or perceptual models. AEROBLADE [46] thresholds AE reconstruction error, while ZED [18] employs a learned lossless image codec to estimate per-image coding cost. Both rely on a single-path absolute score that is sensitive to content complexity and varies across domains. A family of perturbation-based methods instead probes DINOv2 [38] feature invariance: RIGID [25] under Gaussian noise, WaRPAD [14] under wavelet high-frequency perturbations with patch-wise aggregation, and ConV [60] via augmentation consistency grounded in a manifold orthogonality principle. Manifold Bias [7] analyzes geometric properties of the diffusion score-function manifold. GLARE departs from these paradigms by measuring the differential reconstruction change under context removal, directly targeting the context dependence imposed by the LDM generation process.

# 3 Method

## 3.1 Motivation

Recall from Sec. 2.1 that iterative denoising propagates conditioning semantics across all spatial locations. We now examine why this property creates an exploitable forensic signal.

**Context dependence.** We partition an image $x$ into $K$ non-overlapping patches $\{x_1, \ldots, x_K\}$ and define the context of patch $x_i$ as its complement $x_{\neg i} := \{x_j\}_{j \neq i}$. We quantify the context dependence of a patch by the mutual information

$$C_i \;:=\; I(x_i;\, x_{\neg i}), \qquad \bar{C} \;:=\; \frac{1}{K}\sum_{i=1}^{K} C_i. \tag{3}$$

A larger $C_i$ indicates that the content of patch $x_i$ is more predictable from the rest of the image, reflecting stronger cross-patch correlation.

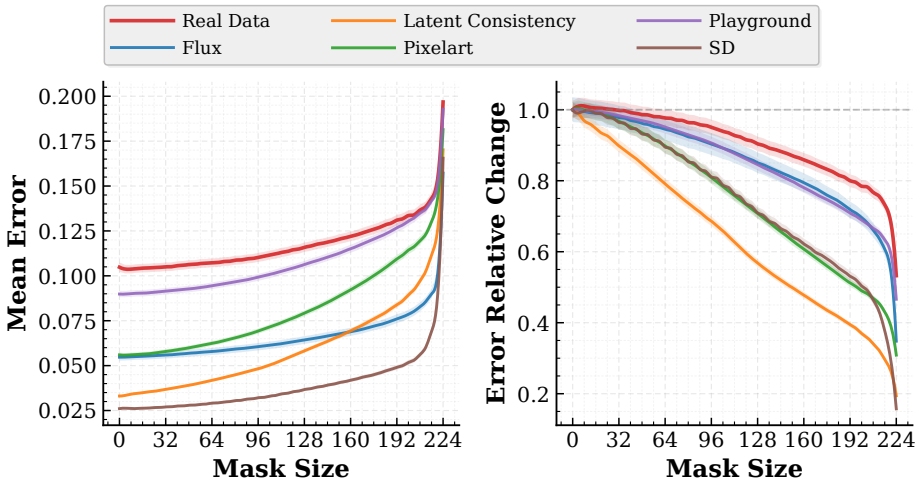
**Spatial correlation in LDM outputs.** As shown in Eq. (2), each denoising step applies $\epsilon_\theta$ to the full spatial extent of $z_t$, where cross-attention and global feature pooling introduce non-local interactions among all spatial positions. Over $T$ iterative steps, these interactions accumulate: correlations between distant patches in $z_t$ propagate and strengthen through the repeated linear-plus-noise updates, so the final generated latent $z_0$ carries strong cross-patch dependencies. Natural images approximately follow the Markov Random Field (MRF) assumption [33, 55], where a pixel's distribution depends primarily on its local neighborhood, yielding moderate $\bar{C}$. By contrast, LDM-generated images exhibit a systematically higher $\bar{C}$ than natural images.

**Reconstruction error as a proxy for context dependence.** The elevated $\bar{C}$ of generated images is not directly observable, but the semantic compression property of the AE (Sec. 2.1) provides an indirect probe: because the bottleneck discards fine-grained details and preserves only globally coherent structure, reconstruction quality becomes sensitive to how much spatial context is available to the encoder. By progressively removing context and monitoring how the AE reconstruction error responds, we can expose this dependence. For an AE with fixed bottleneck capacity, rate–distortion theory [5, 39] shows that the minimum achievable distortion on a region $x_i$ is a monotonically increasing function of the conditional entropy $H(x_i \mid x_{\text{ctx}})$ given the available context (a formal proof is provided in the Supplementary Material):

$$D(x_i \mid x_{\text{ctx}}) \;=\; g\big(H(x_i \mid x_{\text{ctx}})\big), \quad g' > 0. \tag{4}$$

When the full image is provided ($x_{\text{ctx}} = x_{\neg i}$), the encoder exploits cross-region redundancy, and $H(x_i \mid x_{\neg i})$ is low for high $C_i$ patches. As context is progressively masked, $x_{\text{ctx}}$ shrinks, and the conditional entropy increases toward the marginal $H(x_i)$:

$$H(x_i \mid x_{\neg i}) \;\le\; H(x_i \mid x_{\text{ctx}}) \;\le\; H(x_i). \tag{5}$$




(a) Mean Error (b) Error Relative Change

**Fig. 2: Progressive border masking analysis.** (a) Mean VAE reconstruction error on the central patch. (b) Ratio of current reconstruction error to the original full-image (Mask Size = 0) reconstruction error. As the masked area grows, generated samples drop markedly faster than real images, reflecting their stronger dependence on global context.

Because $g$ is monotonically increasing, the *spread* between the two extremes of the entropy chain—$g(H(x_i))$ versus $g(H(x_i \mid x_{\neg i}))$—is controlled by $C_i = H(x_i) - H(x_i \mid x_{\neg i})$: for images with high $\bar{C}$ (generated), the spread is substantial; for images with low $\bar{C}$ (real), it is negligible.

**Empirical verification.** To verify this prediction, we sample 500 real images and, for each of several LDM generators, 500 synthetic images at 512×512 resolution. We progressively mask the image periphery and compute the AE reconstruction error on the central 64×64 patch. As shown in Fig. 2, when we plot the ratio of each error to the full-image baseline (Mask Size = 0), all curves decline from 1.0, but generated images drop markedly faster than real ones as context is removed—consistent with the $\bar{C}$-based analysis above. This divergence is consistent across diverse generators, confirming that the sensitivity of reconstruction error to context removal is a robust, generator-agnostic detection cue.

### 3.2 GLARE

Building on this observation, we introduce GLARE, a training-free framework that operationalizes context sensitivity as a detection signal. The analysis above shows that the separation between generated and real images is governed by the spread between the two endpoints of the entropy chain (Eq. (5)): full context ($x_{\text{ctx}} = x_{\neg i}$) and complete isolation ($x_{\text{ctx}} = \varnothing$). Comparing these two extremes maximizes the discriminative gap. GLARE therefore contrasts them via

a *dual-path* reconstruction scheme and applies a lightweight calibration module to suppress content-induced bias. The overall pipeline is illustrated in Fig. 1.

**Full-context path.** Given a pretrained AE $(\mathcal{E}, \mathcal{D})$ from LDM and a test image $x$, we encode and decode the entire image (global reconstruction):

$$\hat{x}_f = \mathcal{D}(\mathcal{E}(x)), \quad e_f(x) = e(x, \hat{x}_f)\,, \tag{6}$$

where $e(\cdot,\cdot)$ is a perceptual distance. The encoder observes all spatial context; by Eq. (4), the per-patch distortion is $g(H(x_i \mid x_{\neg i}))$, corresponding to the lower bound of the distortion range.

**Context-free path.** To realize the opposite extreme, we partition $x$ into $K$ non-overlapping $p \times p$ patches $\{\Omega_k\}_{k=1}^{K}$ and reconstruct each patch in complete isolation (local reconstruction). Each cropped patch $x\,|_{\Omega_k}$ is resized to the AE's native resolution, independently encoded and decoded, then resized back and reassembled at its grid-aligned location:

$$\begin{aligned} \hat{x}_p^{(k)} &= \mathcal{D}\big(\mathcal{E}(x\,|_{\Omega_k})\big), \quad k = 1, \ldots, K, \\ \hat{x}_s\,|_{\Omega_k} &= \hat{x}_p^{(k)}, \quad k = 1, \ldots, K, \end{aligned} \tag{7}$$

yielding the stitched reconstruction $\hat{x}_s$ and the context-free reconstruction error:

$$e_p(x) = e(x, \hat{x}_s)\,. \tag{8}$$

Here every patch is encoded without any surrounding information, so the per-patch distortion is $g(H(x_i))$, the upper bound of the distortion range in Eq. (5).

**Context Sensitivity Scoring.** As established in Sec. 3.1, the gap between the two extremes of the entropy chain is governed by the context dependence $C_i$. The dual-path design captures this gap through the log-ratio of reconstruction errors:

$$\Delta(x) \;=\; \log e_f(x) \;-\; \log e_p(x). \tag{9}$$

We term $\Delta(x)$ the **context sensitivity score**. By Eq. (4), the reconstruction errors of the two paths are governed by different entropy quantities:

$$\Delta(x) \;\approx\; \log \frac{\sum_i g\big(H(x_i \mid x_{\neg i})\big)}{\sum_i g\big(H(x_i)\big)}, \tag{10}$$

where $g$ is the same monotonically increasing distortion–entropy mapping introduced in Eq. (4) (both paths share the same pretrained decoder). In the Supplementary Material we derive a closed-form expression under the Gaussian–quadratic model, showing that $\Delta(x) \approx -\frac{2}{n}\bar{C}$, *i.e.*, a linear function of average context dependence under those assumptions. In the general (non-Gaussian) case, the monotonicity of $g$ still guarantees that higher $\bar{C}$ yields lower conditional entropies in the numerator while the marginal entropies in the denominator remain unchanged, so $\Delta(x)$ is systematically smaller for generated images (high $\bar{C}$) than for real ones (low $\bar{C}$).

**Semantic Complexity Calibration Module.** Reconstruction error correlates with image complexity [15, 46]. To alleviate this bias, we calibrate using the

inverse of an edge-density proxy, so that more complex images receive stronger normalization. Formally, we define

$$\rho_{\text{edge}}(x) = \frac{1}{HWC} \sum_{c=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} \left\| \nabla \mathbf{F}_c(x)[i,j] \right\|_2, \tag{11}$$

where $\mathbf{F}(x) \in \mathbb{R}^{H \times W \times C}$ are shallow features taken from the same feature space used by the error $e(\cdot,\cdot)$ and $\nabla$ is the Sobel operator. We set the calibration factor:

$$\Phi(x) \;=\; \frac{1}{\rho_{\text{edge}}(x) + \epsilon}, \tag{12}$$

with a small $\epsilon > 0$ for numerical stability. The final detection score becomes

$$S(x) = \Delta(x) + \log \Phi(x) \tag{13}$$

In summary, $\Delta(x)$ captures how sensitively the reconstruction error responds to context removal, and the $\log \Phi(x)$ term controls for image complexity. A higher final score $S(x)$ indicates a higher likelihood that the image is real. Ablation studies validating the contribution of each component are presented in Sec. 4.3.

## 4 Experiment

### 4.1 Setup

**Test datasets.** To evaluate the performance of our method, we conduct experiments on three diverse test sets comprising both real and LDM-generated images, covering more than 20 generative models:

- **LDMFakeDetect** [43] is composed of synthetic images generated by a diverse set of state-of-the-art generative models alongside real images. The generative models included are SD [8, 21, 47], Midjourney [2], Kandinsky 2 [45], Playground v2.5 [31], PixArt-$\alpha$ [11], LCM [35], Flux [1], Würstchen [41] and aMUSEd [40]. The real images are sourced from Redcaps [19], WikiArt [3], and LAION-Aesthetics [48].
- **FakeInversion** [9] introduces a refined LDM detection dataset to counteract evaluation biases in synthetic image assessment by ensuring synthetic images align closely with real images in both semantic content and visual style. The benchmark incorporates fake images generated by a diverse set of leading LDM variants, including DALLE-3 [6], Kandinsky 2 [45], Kandinsky 3 [4], Midjourney [2], SDXL [42], Segmind-Vega [24], PixArt-$\alpha$ [11], Playground v2.5 [31], SDXL-DPO [54], SegMoE [58], SSD-1B [24], Stable-Cascade [41], and Würstchen2 [41].
- **Chameleon** [57] presents a more challenging benchmark generated by diverse commercial and open-source models, including Midjourney [2], DALLE-3 [6], and Stable Diffusion [47] with various LoRA adaptations [27]. Its AI-generated images are sourced "in the wild" from real-world AI art communities and validated via human perceptual studies to ensure photorealistic

**Table 1: Detection performance on the LDMFakeDetect dataset.** We report AUROC and AP (%) for each generator, where **bold** and underline denote the best and second-best methods, respectively. AERO and Manifold abbreviate AEROBLADE [46] and Manifold Bias [7].

| Metric | Generator | *Training-Based Methods* | | | | | *Training-Free Methods* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NPR | FatFormer | AIDE | FIRE | ForgeLens | AERO | RIGID | ZED | Manifold | WaRPAD | ConV | Ours |
| AUROC | SD | 84.50 | 70.88 | 92.86 | 85.24 | **99.94** | 98.14 | 61.37 | 65.73 | 69.59 | 77.02 | 67.86 | 99.23 |
| | Midjourney | 78.43 | 82.73 | 81.14 | 84.84 | 91.19 | 99.67 | 73.66 | 70.09 | 52.11 | 75.51 | 84.78 | **99.96** |
| | Kandinsky | 82.70 | 69.10 | **98.21** | 84.85 | 79.63 | 84.33 | 72.50 | 66.06 | 76.88 | 76.46 | 87.35 | 92.76 |
| | Playground | 47.44 | 75.49 | 97.72 | 84.50 | 91.87 | 83.36 | 46.10 | 85.61 | 77.24 | 73.46 | 93.77 | **99.20** |
| | PixArt-$\alpha$ | 65.61 | 81.67 | 97.18 | 82.68 | 73.52 | 97.29 | 56.40 | 83.61 | 74.75 | 66.73 | 90.05 | **99.89** |
| | LCM | 58.90 | 71.26 | 82.10 | 85.88 | 87.21 | 99.53 | 55.04 | 90.30 | 76.27 | 65.90 | 73.53 | **100.00** |
| | Flux | 82.84 | 47.57 | 94.61 | 86.04 | 81.74 | 77.26 | 59.59 | 63.90 | 74.10 | 57.17 | 29.06 | **96.87** |
| | Würstchen | 86.48 | 91.26 | 97.77 | 83.31 | **100.00** | 90.80 | 66.25 | 65.55 | 86.24 | 92.95 | 96.08 | 96.63 |
| | aMUSEd | 68.13 | 98.64 | 97.81 | 86.29 | **99.92** | 93.50 | 91.08 | 84.50 | 89.15 | 96.13 | 96.35 | 98.61 |
| | **Mean** | 72.78 | 76.51 | 93.27 | 84.85 | 89.45 | 91.54 | 64.66 | 75.04 | 75.15 | 75.70 | 79.87 | **98.13** |
| AP | SD | 86.82 | 68.09 | 93.62 | 78.68 | **99.95** | 96.98 | 59.62 | 72.14 | 68.48 | 76.52 | 66.48 | 99.22 |
| | Midjourney | 79.85 | 80.50 | 84.62 | 77.64 | 91.26 | 99.18 | 77.46 | 70.52 | 54.23 | 74.81 | 84.16 | **99.97** |
| | Kandinsky | 83.24 | 64.29 | **98.24** | 77.89 | 79.82 | 82.71 | 71.61 | 69.46 | 74.66 | 75.61 | 86.93 | 92.69 |
| | Playground | 52.66 | 70.45 | 96.86 | 76.32 | 91.91 | 77.53 | 48.76 | 86.16 | 73.31 | 69.02 | 94.19 | **98.70** |
| | PixArt-$\alpha$ | 60.33 | 78.19 | 97.01 | 74.32 | 73.97 | 95.05 | 59.89 | 83.54 | 77.02 | 61.46 | 90.71 | **99.89** |
| | LCM | 60.45 | 71.46 | 79.00 | 79.61 | 87.42 | 98.55 | 50.66 | 92.43 | 72.36 | 57.40 | 71.74 | **100.00** |
| | Flux | 85.29 | 44.73 | 94.56 | 80.52 | 81.72 | 77.87 | 59.65 | 67.41 | 70.53 | 55.41 | 37.68 | **96.35** |
| | Würstchen | 88.30 | 90.28 | 97.60 | 76.19 | **100.00** | 88.55 | 67.22 | 69.86 | 84.33 | 91.36 | 96.35 | 96.19 |
| | aMUSEd | 68.30 | 98.61 | 97.30 | 80.33 | **99.92** | 89.27 | 91.76 | 87.51 | 88.55 | 95.05 | 96.34 | 97.82 |
| | **Mean** | 73.92 | 74.07 | 93.20 | 77.94 | 89.55 | 89.52 | 65.18 | 77.67 | 73.72 | 72.96 | 80.51 | **97.87** |

quality. These characteristics render Chameleon particularly demanding for evaluating detection algorithms.

**Baselines and Evaluation Metrics.** We conduct a comprehensive comparative analysis of GLARE against a diverse set of AI-generated image detection methods, encompassing both training-based and training-free approaches. The training-based methods include NPR [50], FatFormer [34], AIDE [57], FIRE [16], and ForgeLens [13]. For training-free approaches, we compare against AEROBLADE [46], RIGID [25], ZED [18], Manifold Bias [7], WaRPAD [14], and ConV [60]. Following [7, 25, 46], we employ three standard metrics to evaluate detector performance: Area Under the Receiver Operating Characteristic curve (AUROC), Average Precision (AP), and Accuracy (ACC). Due to space constraints, detailed implementation specifications for these baselines and the ACC results are provided in the Supplementary Material.

**Implementation Details.** For patch-based reconstruction, each image is partitioned into square patches with size $p = 128$ pixels. When the image height or width is insufficient, zero-padding is applied prior to reconstruction, and the outputs are subsequently cropped back to the original spatial extent. Reconstruction is performed using the Stable Diffusion 2.0 base model (SD2-base) [47], which employs a variational autoencoder (VAE) [28] with Kullback–Leibler regularization. As the error computation function $e(\cdot, \cdot)$, we adopt the standard LPIPS metric [59] following [46], using the variant based on second-layer features of the

**Table 2: Detection performance across 13 generators on FakeInversion.** We report AUROC and AP (%).

| Metric | Generator | *Training-Based Methods* | | | | | *Training-Free Methods* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NPR | FatFormer | AIDE | FIRE | ForgeLens | AERO | RIGID | ZED | Manifold | WaRPAD | ConV | Ours |
| AUROC | DALLE-3 | 48.97 | 46.49 | 76.36 | 43.12 | 51.23 | 50.18 | 40.61 | 50.60 | 40.35 | 49.53 | 63.47 | **77.31** |
| | Kandinsky2 | 88.37 | 58.46 | 90.93 | 72.86 | **99.78** | 72.03 | 52.66 | 66.61 | 45.37 | 71.22 | 75.12 | 89.15 |
| | Kandinsky3 | 92.34 | 69.56 | 87.67 | 71.00 | **97.00** | 66.14 | 63.73 | 85.77 | 44.04 | 53.54 | 84.25 | 94.04 |
| | Midjourney | 59.36 | 41.52 | 57.43 | 16.79 | 48.94 | **88.07** | 51.81 | 65.02 | 47.72 | 67.52 | 46.48 | 85.34 |
| | SDXL | **94.65** | 69.04 | 88.74 | 74.23 | 91.41 | 63.00 | 62.27 | 75.22 | 41.77 | 70.75 | 76.20 | 91.60 |
| | Vega | 92.83 | 72.16 | 91.64 | 71.34 | 89.84 | 60.18 | 55.62 | 65.23 | 42.53 | 75.76 | 79.43 | **93.92** |
| | PixArt-1024 | 84.82 | 76.18 | 83.80 | 71.22 | 85.36 | 84.10 | 63.19 | 75.42 | 42.93 | 75.59 | 85.00 | **98.40** |
| | Playground-25 | 76.61 | 76.44 | 90.58 | 72.54 | 80.44 | 76.60 | 56.38 | 73.73 | 40.23 | 82.99 | 86.84 | **93.62** |
| | SDXL-DPO | 92.90 | 76.22 | 77.53 | 71.54 | 92.93 | 70.02 | 63.94 | 70.19 | 43.06 | 80.56 | 86.29 | **97.28** |
| | SegMoE | 89.45 | 65.38 | 81.83 | 72.04 | 93.18 | 66.23 | 54.84 | 68.45 | 41.11 | 74.60 | 82.10 | **97.72** |
| | SSD-1B | 79.58 | 74.24 | 87.63 | 71.49 | 82.63 | 65.77 | 59.22 | 70.67 | 38.89 | 76.18 | 83.29 | **92.78** |
| | Stable-Cascade | 97.51 | 74.89 | 95.27 | 67.26 | **99.88** | 75.70 | 70.44 | 79.85 | 41.51 | 80.05 | 87.03 | 97.42 |
| | Würstchen2 | 93.98 | 85.87 | 96.20 | 72.18 | **99.86** | 71.73 | 58.83 | 66.34 | 48.57 | 90.99 | 92.31 | 88.34 |
| | **Mean** | 83.95 | 68.19 | 85.05 | 65.20 | 85.58 | 69.98 | 57.96 | 70.24 | 42.93 | 73.02 | 79.06 | **92.07** |
| AP | DALLE-3 | 49.62 | 46.20 | **76.06** | 46.16 | 52.14 | 47.44 | 44.94 | 54.16 | 42.46 | 49.04 | 64.19 | 69.75 |
| | Kandinsky2 | 87.80 | 54.70 | 91.42 | 63.82 | **99.75** | 65.23 | 53.27 | 68.66 | 46.62 | 65.87 | 77.44 | 82.49 |
| | Kandinsky3 | 92.83 | 64.75 | 87.19 | 61.42 | **96.87** | 58.60 | 64.94 | 86.08 | 45.46 | 50.77 | 85.70 | 89.05 |
| | Midjourney | 53.12 | 42.60 | 55.90 | 33.57 | 50.72 | 78.82 | 54.70 | 69.88 | 64.88 | 62.76 | 48.38 | **79.36** |
| | SDXL | **94.89** | 63.32 | 87.98 | 65.61 | 90.82 | 56.07 | 64.25 | 77.24 | 44.87 | 70.22 | 77.42 | 88.16 |
| | Vega | **93.75** | 71.00 | 90.40 | 61.69 | 89.82 | 52.38 | 56.72 | 69.16 | 44.81 | 72.58 | 80.95 | 89.63 |
| | PixArt-1024 | 83.45 | 72.83 | 85.40 | 62.27 | 85.11 | 76.83 | 63.51 | 76.00 | 44.74 | 71.75 | 86.23 | **98.00** |
| | Playground-25 | 80.06 | 73.29 | **90.57** | 62.82 | 80.13 | 68.06 | 56.71 | 77.01 | 42.85 | 80.95 | 88.12 | 89.61 |
| | SDXL-DPO | 93.13 | 73.02 | 79.65 | 62.86 | 92.74 | 60.86 | 65.35 | 69.52 | 45.28 | 78.74 | 87.71 | **96.48** |
| | SegMoE | 90.73 | 63.44 | 80.46 | 62.68 | 93.00 | 56.94 | 57.66 | 74.55 | 43.51 | 69.75 | 84.16 | **96.22** |
| | SSD-1B | 82.84 | 70.83 | 85.44 | 61.82 | 82.51 | 56.50 | 60.52 | 74.30 | 42.64 | 72.44 | 84.69 | **88.21** |
| | Stable-Cascade | 97.19 | 72.07 | 95.61 | 58.76 | **99.85** | 66.65 | 71.42 | 81.78 | 43.53 | 76.55 | 88.35 | 95.59 |
| | Würstchen2 | 94.25 | 85.73 | 96.06 | 63.69 | **99.83** | 62.71 | 59.02 | 67.42 | 48.40 | 89.15 | 93.63 | 82.90 |
| | **Mean** | 84.13 | 65.68 | 84.78 | 59.01 | 85.64 | 62.08 | 59.46 | 72.75 | 46.16 | 70.04 | 80.54 | **88.11** |

VGG network [49]. For consistency, the shallow feature maps used to compute the edge-density proxy are also taken from the same VGG second layer.

### 4.2 Main Results

**Performance on LDMFakeDetect Dataset.** Table 1 reports AUROC and AP for a range of training-based and training-free methods evaluated across nine generators on LDMFakeDetect [43]. Overall, our method attains the highest mean AUROC of 98.13% and AP of 97.87%, surpassing prior training-free approaches and exceeding the latest supervised methods. On average, our method outperforms the second-best training-free method, AEROBLADE, by 6.59% in AUROC score. While AEROBLADE performs competitively where the generative model matches the inspected autoencoder, it fails to generalize to more recent models such as Flux [1]. Notably, for the autoregressive latent space generator aMUSEd [40], our performance is comparable to that of ForgeLens [13], suggesting that the context-sensitivity signal extends beyond diffusion-based samplers to other latent-space generation paradigms.

**Performance on FakeInversion Dataset.** Table 2 summarizes AUROC and AP across the FakeInversion subsets, where synthetic images are tightly aligned

**Table 3: Detection performance on the Chameleon dataset.** We report AUROC and AP (%).

| Metric | *Training-Based Methods* | | | | | *Training-Free Methods* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NPR | FatFormer | AIDE | FIRE | ForgeLens | AERO | RIGID | ZED | Manifold | WaRPAD | ConV | Ours |
| **AUROC** | 58.36 | 58.60 | <u>71.62</u> | 51.12 | 52.14 | 67.78 | 55.39 | 53.10 | 16.55 | 54.69 | 46.15 | **82.62** |
| **AP** | 52.94 | 57.79 | 66.02 | 42.80 | 52.19 | 52.10 | <u>66.03</u> | 55.50 | 28.61 | 55.16 | 51.51 | **76.97** |

with real images in both semantic content and visual style. Our method attains the highest mean AUROC of 92.07% and AP of 88.11%, on average surpassing the best training-based method by 6.49% and the best training-free baseline by 13.01% in AUROC. Beyond the aggregate gains, the per-generator results indicate strong cross-model generalization: we secure clear wins on several recent, challenging generators, while remaining competitive on Kandinsky3 and Stable-Cascade. These trends suggest that the proposed LDM context-sensitivity cue transfers well across diverse generator architectures and training recipes. Although ForgeLens achieves the top AUROC on a subset of generators (*e.g.*, Kandinsky2/3, Stable-Cascade, Würstchen2), likely reflecting its emphasis on generator-specific artifacts, its performance varies more across models, whereas GLARE delivers consistently strong results and the best overall mean without supervised training.

**Performance on Chameleon Benchmark.** Unlike traditional curated laboratory datasets, Chameleon collects images "in the wild," where content varies drastically and images may have undergone unknown post-processing (*e.g.*, social-media compression, rescaling). This setting is particularly challenging for detectors that rely on fragile pixel-level artifacts. As reported in Tab. 3, GLARE surpasses the state-of-the-art by +11.00% AUROC and +10.94% AP. We attribute this margin to the fact that context dependence—the core cue exploited by GLARE—originates from the iterative latent sampling process and is therefore largely preserved even after common post-processing, whereas the shallow, low-level cues that other detectors depend on are more fragile. Beyond accuracy, GLARE relies on a single lightweight VAE and never invokes a diffusion U-Net, making it markedly more efficient and deployable than perturbation-based detectors; a detailed runtime and memory analysis is provided in the Supplementary Material.

### 4.3 Ablation Study

To verify the effectiveness of each component, we perform a systematic ablation over module configurations and parameter settings, and report the resulting detection performance. Additional results, including inference time, reconstruction metric selection, failure case analysis, and analysis of non-LDM generators, are provided in the Supplementary Material.

**Component Ablation Study.** We show the impact of each component in GLARE in Tab. 4. Following the design in Sec. 3, we systematically evaluate

**Table 4: Component Ablation Study.** '✓' and '✗' denote that the component is used and not used, respectively.

| Configuration | $e_f(x)$ | $e_p(x)$ | $\Phi(x)$ | AP | AUROC |
|---|---|---|---|---|---|
| Global only ($e_f$) | ✓ | ✗ | ✗ | 89.52 | 91.54 |
| Local only ($e_p$) | ✗ | ✓ | ✗ | 70.38 | 70.11 |
| $\Delta(x)$ | ✓ | ✓ | ✗ | 96.66 | 96.82 |
| $S(x)$ (Full GLARE) | ✓ | ✓ | ✓ | **97.87** | **98.13** |

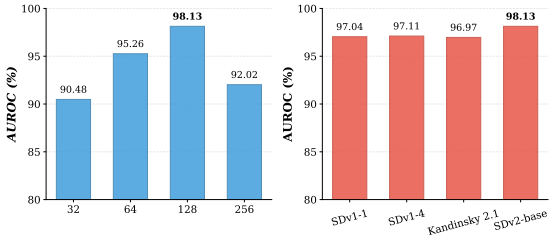
different module configurations and observe their contributions to the final score $S(x)$. Using only full-image reconstruction or only patch-level reconstruction results in suboptimal performance, partly because absolute reconstruction errors vary substantially with image content for both real and generated data; either signal alone is therefore an unreliable discriminator. By comparison, the combination of the two errors lifts AP/AUROC to 96.66%/96.82%. This improvement arises from measuring the relative error change under context suppression. Finally, incorporating the complexity calibration $\Phi(x)$ further raises AP/AUROC to 97.87%/98.13%, highlighting that the calibration effectively mitigates content-specific bias.

**Choice of Patch Size.** The selection of patch size is critical for detection: overly large patches fail to sufficiently suppress global context, while overly small patches may amplify boundary artifacts. As shown in Fig. 3 (a), we evaluate $p \in \{32, 64, 128, 256\}$. Very small patches (*e.g.*, $p$=32) intensify stitching artifacts and influence patch-level reconstruction error $e_p$ for both real and generated images, thereby weakening separability, where the stitching artifacts can be seen in Fig. 5 (c). Conversely, large patches (*e.g.*, $p$=256) preserve considerable global information, reducing the contrast between $e_f$ and $e_p$ and diluting the context-sensitivity signal. A moderate patch size ($p$=128) strikes the best balance between suppressing global context and maintaining reconstruction fidelity, maximizing the discriminative power of GLARE.

**Choice of Autoencoder.** We compare four different autoencoder backbones for reconstruction: SDv1-1 [47], SDv1-4 [47], Kandinsky 2.1 [45], and SDv2-base [47]. As reported in Fig. 3 (b), AUROC scores remain consistently high across different backbones. SDv2-base yields a small gain—likely attributable to stronger semantic compression and reconstruction fidelity—yet the effect size is minor. These results support that our detector primarily leverages a model-agnostic signal rooted in the LDM generative process, rather than specific artifacts of a particular AE.

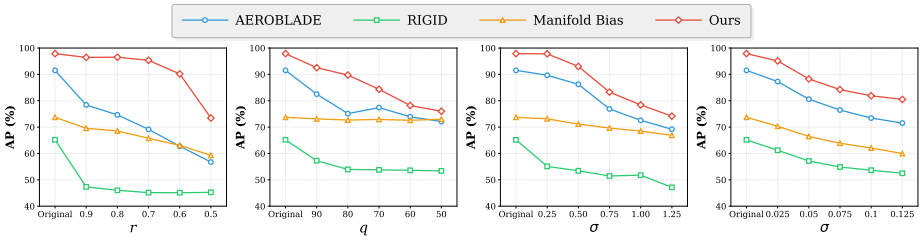
### 4.4 Robustness to Unseen Perturbations

Real-world images frequently encounter various perturbations that challenge the stability of synthetic content detection systems. To rigorously evaluate GLARE's robustness, following previous works [16, 46], we conduct evaluations under four




(a) Impact of Patch Size (b) Impact of VAE

**Fig. 3: Hyperparameters Ablation Studies.** (a) Ablation on patch size $p$; (b) VAE backbone used for image reconstruction.




(a) Center Crop (b) JPEG Compression (c) Gaussian Blur (d) Gaussian Noise

**Fig. 4: Robustness analysis.** AP (%) comparison under common perturbations: (a) center crop, (b) JPEG compression, (c) Gaussian blur, and (d) Gaussian noise. Our method consistently attains higher AP across perturbation levels, demonstrating strong robustness to these transformations.

common perturbation scenarios: center cropping followed by resizing to the original resolution, JPEG compression, Gaussian blur, and Gaussian noise.

As illustrated in Fig. 4, GLARE achieves superior performance over the baselines in every corruption consistently, and degrades more gracefully as severity increases. When subjected to slight perturbations, previous training-free methods such as AEROBLADE and RIGID exhibit marked deterioration in AP scores, whereas our approach consistently maintains stronger resilience. These results highlight the robustness of our approach. However, under extreme degradations, our performance also declines noticeably as these perturbations suppress the global context and high-frequency cues that GLARE relies on. Cropping physically reduces the effective context available to the encoder, thereby diminishing the observable context sensitivity, while blur attenuates fine structures, weakening reconstruction contrast and SCCM calibration, thus reducing discriminability.

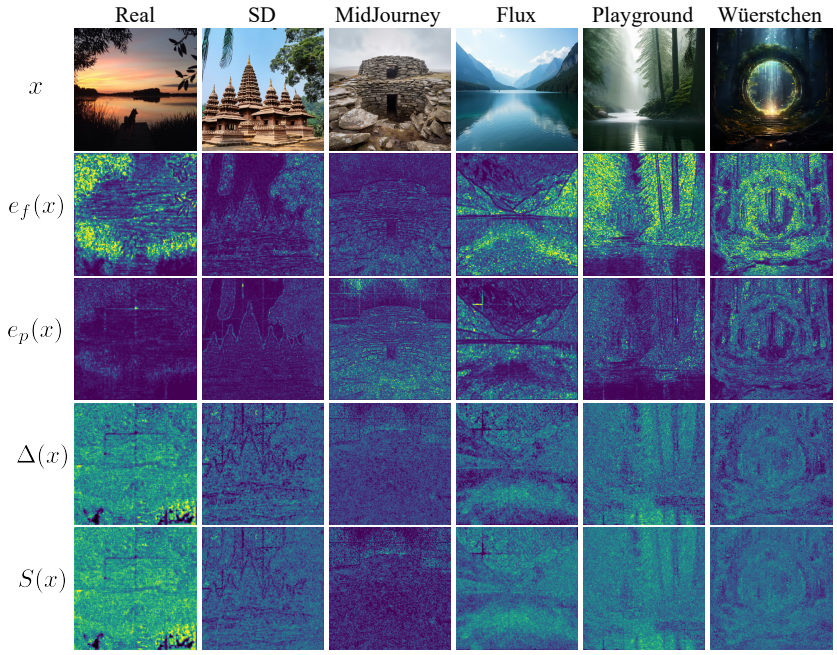



**Fig. 5: Visualization.** Reconstruction error maps on real and LDM-generated images from LDMFakeDetect [43]. Real images exhibit a higher $e_f/e_p$ ratio, while LDM-generated images show a markedly lower ratio, reflecting their stronger context dependence.

### 4.5 Visualization

To elucidate the core operating principle of our model, we visualize real and generated images together with three corresponding reconstruction error maps. As shown in Fig. 5, the patch-based reconstruction suppresses global context and compels the AE to emphasize fine and localized structure. As illustrated in the third row, the change in reconstruction error between the two paths differs systematically across the two classes: real images exhibit a higher $e_f/e_p$ ratio than LDM-generated images, in line with the weaker context dependence of natural images. Notably, for pipelines equipped with stronger VAEs (*e.g.*, Flux, Würstchen), the absolute full-image reconstruction error $e_f(x)$ can be large for both real and synthetic images—potentially undermining detectors that rely solely on full reconstruction. In these cases, $\Delta(x)$ remains discriminative because it captures how reconstruction error changes when global context is removed, providing more reliable separation than any single error map in isolation. Furthermore, the introduction of $\Phi(x)$ partially compensates for bias induced by texture complexity: it slightly amplifies the score for texturally simple images (*e.g.*, the first and fourth columns) while attenuating it for texturally rich ones

(*e.g.*, the third column). Additional visualizations are provided in the Supplementary Material.

## 5 Conclusion

This paper establishes that images generated by Latent Diffusion Models (LDMs) demonstrate stronger coupling between local patterns and global context compared with real images. By contrasting full-image and patch-based reconstructions within a fixed autoencoder, GLARE captures generator-agnostic forensic cues inherent to the LDM sampling process. Further, we introduce a semantic complexity calibration to stabilize the final score. Extensive experiments across diverse datasets and generators show superior performance and strong generalization, while maintaining robustness under various real-world perturbations.

Although GLARE effectively captures context dependence, partitioning images into patches may introduce boundary artifacts that affect patch-level reconstruction errors. Additionally, our current focus on latent space pipelines means that generators with substantially different synthesis mechanisms or heavy post-processing can weaken the discriminative signal. Future work will investigate suitable pretext tasks and incorporate complementary synthesis cues to broaden coverage and improve the robustness of our framework.

## References


1. Flux.1. https://www.blackforestlabs.com, accessed 2026-06-30
2. Midjourney. https://docs.midjourney.com/docs/about, accessed 2026-06-30
3. Wikiart. https://www.wikiart.org/, accessed 2026-06-30
4. Arkhipkin, V., Filatov, A., Vasilev, V., Maltseva, A., Azizov, S., Pavlov, I., Agafonova, J., Kuznetsov, A., Dimitrov, D.: Kandinsky 3.0 technical report. arXiv preprint arXiv:2312.03511 (2023)
5. Berger, T.: Rate-distortion theory. Wiley Encyclopedia of Telecommunications (2003)
6. 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. OpenAI Technical Report **2**(3), 8 (2023)
7. Brokman, J., Giloni, A., Hofman, O., Vainshtein, R., Kojima, H., Gilboa, G.: Manifold induced biases for zero-shot and few-shot detection of generated images. In: International Conference on Learning Representations (2025)
8. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
9. 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)
10. Chen, B., Zeng, J., Yang, J., Yang, R.: Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In: International Conference on Machine Learning (2024)
11. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: PixArt-$\alpha$: Fast training of Diffusion Transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
12. Chen, R., Xi, J., Yan, Z., Zhang, K.Y., Wu, S., Xie, J., Chen, X., Xu, L., Guan, I., Yao, T., et al.: Dual data alignment makes ai-generated image detector easier generalizable. arXiv preprint arXiv:2505.14359 (2025)
13. Chen, Y., Zhang, L., Niu, Y.: Forgelens: Data-efficient forgery focus for generalizable forgery image detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16270–16280 (2025)
14. Choi, S., Lee, H., Lee, M.: Training-free detection of AI-generated images via cropping robustness. In: Advances in Neural Information Processing Systems (2025)
15. Choi, S., Park, S., Lee, J., Kim, S., Choi, S.J., Lee, M.: Hfi: A unified framework for training-free detection and implicit watermarking of latent diffusion model generated images. arXiv preprint arXiv:2412.20704 (2024)
16. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12830–12839 (2025)
17. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L.: On the detection of synthetic images generated by diffusion models. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1–5. IEEE (2023)
18. Cozzolino, D., Poggi, G., Nießner, M., Verdoliva, L.: Zero-shot detection of ai-generated images. In: European Conference on Computer Vision. pp. 54–72. Springer (2024)

19. Desai, K., Kaul, G., Aysola, Z., Johnson, J.: Redcaps: Web-curated image-text data created by the people, for the people. arXiv preprint arXiv:2111.11431 (2021)
20. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Advances in Neural Information Processing Systems. vol. 34, pp. 8780–8794 (2021)
21. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In: Advances in Neural Information Processing Systems (2023)
22. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. vol. 27 (2014)
23. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18685–18694 (2025)
24. Gupta, Y., Jaddipal, V.V., Prabhala, H., Paul, S., Von Platen, P.: Progressive knowledge distillation of stable diffusion xl using layer level loss. arXiv preprint arXiv:2401.02677 (2024)
25. He, Z., Chen, P.Y., Ho, T.Y.: Rigid: A training-free and model-agnostic framework for robust ai-generated image detection. arXiv preprint arXiv:2405.20112 (2024)
26. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
27. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
28. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (2014)
29. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. In: Advances in Neural Information Processing Systems. vol. 31 (2018)
30. Kingma, D.P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., Welling, M.: Improved variational inference with inverse autoregressive flow. In: Advances in Neural Information Processing Systems. vol. 29 (2016)
31. Li, D., Kamko, A., Akhgari, E., Sabet, A., Xu, L., Doshi, S.: Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245 (2024)
32. Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Feng, F.: Improving synthetic image detection towards generalization: An image transformation perspective. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 2405–2414 (2025)
33. Liang, S., Liu, J., Renzhang, C., Guan, Q.: Ferretnet: Efficient synthetic image detection via local pixel dependencies. In: Advances in Neural Information Processing Systems (2025)
34. Liu, H., Tan, Z., Tan, C., Wei, Y., Wang, J., Zhao, Y.: 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)
35. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. In: International Conference on Learning Representations (2024)
36. Luo, Y., et al.: LaRE$^2$: Latent reconstruction error based method for diffusion-generated image detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)

37. Mahara, A., Rishe, N.: Methods and trends in detecting generated images: A comprehensive review. arXiv preprint arXiv:2502.15176 (2025)
38. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
39. Park, S., Adosoglou, G., Pardalos, P.M.: Interpreting rate-distortion of variational autoencoder and using model uncertainty for anomaly detection. Annals of Mathematics and Artificial Intelligence **90**(7), 735–752 (2022)
40. Patil, S., Berman, W., Rombach, R., von Platen, P.: amused: An open muse reproduction. arXiv preprint arXiv:2401.01808 (2024)
41. Pernias, P., Rampas, D., Richter, M.L., Pal, C.J., Aubreville, M.: Würstchen: An efficient architecture for large-scale text-to-image diffusion models. arXiv preprint arXiv:2306.00637 (2023)
42. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: International Conference on Learning Representations (2024)
43. Rajan, A.S., Lee, Y.J.: Stay-positive: A case for ignoring real image features in fake image detection. In: International Conference on Machine Learning (2025)
44. Rajan, A.S., Ojha, U., Schloesser, J., Lee, Y.J.: On the effectiveness of dataset alignment for fake image detection. In: International Conference on Learning Representations (2025)
45. Razzhigaev, A., Shakhmatov, A., Maltseva, A., Arkhipkin, V., Pavlov, I., Ryabov, I., Kuts, A., Panchenko, A., Kuznetsov, A., Dimitrov, D.: Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. arXiv preprint arXiv:2310.03502 (2023)
46. Ricker, J., Lukovnikov, D., Fischer, A.: Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9130–9140 (2024)
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
48. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
49. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
50. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28130–28139 (2024)
51. Vahdat, A., Kreis, K., Kautz, J.: Score-based generative modeling in latent space. Advances in Neural Information Processing Systems **34**, 11287–11302 (2021)
52. Van Den Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: International conference on machine learning. pp. 1747–1756. PMLR (2016)
53. Vasilcoiu, A., et al.: LATTE: Latent trajectory embedding for diffusion-generated image detection. arXiv preprint arXiv:2507.03054 (2025)

54. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2024)
55. Wang, C., Komodakis, N., Paragios, N.: Markov random field modeling, inference & learning in computer vision & image understanding: A survey. Computer Vision and Image Understanding **117**(11), 1610–1627 (2013)
56. Wang, W., Huang, Y., Xu, J., Yu, Y., Yan, J., Ding, S., Zhou, P., Luo, Y.: Tranx-adapter: Bridging artifacts and semantics within mllms for robust ai-generated image detection. In: International Conference on Machine Learning (2026)
57. Yan, S., Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Xie, W.: A sanity check for ai-generated image detection. In: International Conference on Learning Representations (2025)
58. Yatharth Gupta, Vishnu V Jaddipal, H.P.: Segmoe: Segmind mixture of diffusion experts. https://github.com/segmind/segmoe, accessed 2026-06-30 (2024)
59. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
60. Zhang, Y., Nie, J., Tian, X., Gong, M., Zhang, K., Han, B.: Detecting generated images by fitting natural image distributions. In: Advances in Neural Information Processing Systems (2025)