RED: RECONSTRUCTION EVOLUTION DYNAMICS FOR GENERALIZABLE AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid evolution of image generators calls for forensic cues that generalize beyond known generation mechanisms. Existing detectors often rely on static image representations or endpoint reconstruction discrepancies, leaving the evolution of intermediate reconstruction stages underexplored. We observe that the relative token predictability of real and generated images can reverse across reconstruction scales, suggesting that intermediate stages may expose forensic evidence overlooked by endpoint comparisons. Motivated by this observation, we propose RED (Reconstruction Evolution Dynamics), a framework that captures transferable forensic cues from coarse-to-fine reconstruction evolution. To our knowledge, RED is the first framework to use scale-wise token predictability to guide forensic evidence aggregation across intermediate reconstruction states. It represents the reconstruction trajectory produced by a frozen multiscale VQ-VAE in the shared feature space of a frozen CLIP encoder. To connect the observed predictability variations with visual evidence, RED learns image-adaptive stage weights from scale-wise token negative log-likelihoods provided by a frozen VAR model. A cross-stage evidence aggregation module then jointly models the original-image representation and the weighted reconstruction features, capturing complementary forensic cues through interactions along the reconstruction trajectory. Experiments on six diverse benchmarks demonstrate that RED achieves the highest average accuracy of 92.5% and average precision of 97.5% among the evaluated methods. Further evaluations show strong robustness to common image degradations, supporting the value of reconstruction evolution for generalizable AI-generated image detection.

1 INTRODUCTION

Advances in GANs Goodfellow et al. (2014), diffusion models Ho et al. (2020), and visual autoregressive models Tian et al. (2024) have made realistic image synthesis widely accessible, bringing growing challenges to content authenticity and public trust. The continuous emergence of new generators and model versions requires detectors to remain reliable beyond the generation mechanisms represented in their training data. Generalizable AI-generated image detection therefore depends on identifying forensic cues that transfer across diverse and previously unseen generators.

Pretrained foundation models offer a strong starting point for this transfer. Detectors built on vision encoders and multimodal large language models (MLLMs) exploit representations learned from large-scale data to recognize generated images beyond known generators Ojha et al. (2023); Mu et al. (2025); Li et al. (2026b); Tan et al. (2026b). However, separating generation artifacts from semantic correlations in the detection training data remains a challenge Guillaro et al. (2025); Chen et al. (2025). Reconstruction-based detection provides a complementary perspective by examining how an input image responds to a pretrained generative model. The resulting discrepancies provide an additional source of forensic evidence, but comparisons between the input and its final reconstruction emphasize the outcome of reconstruction, leaving the evolution of intermediate states underexplored.

The interpretation of these endpoint discrepancies varies across reconstruction settings. DIRE Wang et al. (2023) reports lower reconstruction errors for diffusion-generated images than for real images, whereas DID Qi et al. (2026) observes more pronounced residuals for generated images in challenging settings. Error accumulation during reconstruction Luo et al. (2024) and content- and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

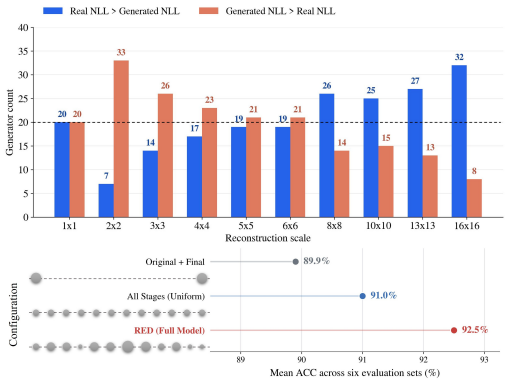


Figure 1: Scale-dependent token predictability and detection performance of RED.

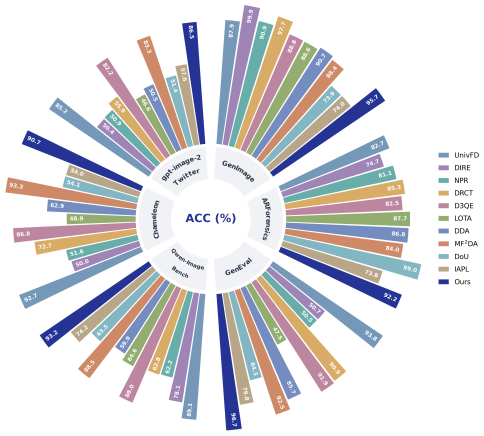


Figure 2: Detection accuracy comparison across six benchmarks (%).

frequency-dependent reconstruction difficulty Chu et al. (2025) further highlight the factors that shape these discrepancies. Together, these findings show that the magnitude of reconstruction error has no single, consistent interpretation across generators and reconstruction settings. An endpoint comparison summarizes the resulting difference, but does not describe how that difference develops as reconstruction progresses.

This leads to our central question: **What can the reconstruction process reveal that its endpoint cannot?** We investigate this question by analyzing real images and samples from 40 generators, analyzed with a pretrained visual autoregressive model (VAR) and its associated multiscale VQ-VAE. The VQ-VAE progressively incorporates quantized information across scales to construct a coarse-to-fine sequence of intermediate reconstructions. VAR evaluates the input-derived tokens through scale-wise negative log-likelihood (NLL), providing a measure of conditional token predictability at each stage of this hierarchy.

As shown in the upper panel of Fig. 1, generated images have higher mean token NLL than real images for 33 of the 40 generators at the 2×2 scale, but for only 8 at the 16×16 scale. This reversal reveals a scale-dependent relationship between image authenticity and token predictability under the pretrained model that a final-scale score alone cannot describe. Further analysis of these scale-dependent NLL profiles are provided in Appendix C. The lower panel further shows that incorporating intermediate reconstruction states improves mean accuracy from 89.9% for the original-plus-final baseline to 91.0% with uniform weights, while adapting stage contributions through NLL-guided weighting in RED further improves accuracy to 92.5%.

Building on this insight, we propose RED (**R**econstruction **E**volution **D**ynamics), a unified framework for capturing transferable forensic cues from coarse-to-fine reconstruction evolution. RED first establishes a *reconstruction trajectory representation* by mapping the original image and successive reconstruction states into an ordered sequence in a shared visual feature space. *NLL-guided stage weighting* translates scale-wise token predictability into image-specific stage contributions, learning how likelihood information should guide the allocation of visual evidence. *Cross-stage evidence aggregation* then jointly contextualizes the original-image and weighted reconstruction representations to capture complementary evidence across the trajectory. The same stage weights guide both input feature modulation and output evidence pooling, coupling evidence allocation with its final readout. The resulting trajectory summary is combined with a dedicated, contextualized original-image readout for detection.

Experiments on GenImage Zhu et al. (2023) and five out-of-domain benchmarks demonstrate RED's generalization across diverse generators and generation paradigms without target-domain adaptation, as shown in Fig. 2, RED achieves strong detection performance across all evaluated benchmarks, supporting its generalization across diverse generators and generation paradigms without target-domain adaptation.

Our contributions are summarized as follows:

- **New perspective.** To our knowledge, we are the first to reveal a reversal in the dominant real–generated token-NLL ordering across reconstruction scales. Building on this finding, we introduce a forensic perspective that shifts attention from endpoint discrepancies to evidence evolving throughout coarse-to-fine reconstruction.
- **Unified framework.** We propose RED, which unifies NLL-guided adaptive stage weighting with cross-stage evidence aggregation. By using token predictability to adapt stage contributions and jointly modeling original-image and intermediate reconstruction representations, RED captures complementary forensic cues through image-specific weighting and cross-stage interactions.
- **Improved generalization.** Experiments on six benchmarks, including five out-of-domain benchmarks, demonstrate RED's generalization across diverse generators and generation paradigms without target-domain adaptation. RED achieves the highest average accuracy of 92.5% and average precision of 97.5% among the evaluated methods, with further evaluations demonstrating robustness to common image degradations.

2 Related Work

AI-Generated Image Detection. Early work established the feasibility of cross-generator detection. CNNSpot Wang et al. (2020) showed that suitable data augmentation enables a detector trained on a single CNN-based generator to generalize to other generators. Subsequent research expanded the representational basis of detection. UnivFD Ojha et al. (2023) leveraged pretrained CLIP features to transfer visual knowledge from large-scale pretraining to image forensics. Beyond the choice of backbone, image-adaptive prompting Li et al. (2026b), complementary feature fusion Mu et al. (2026), and diversity-oriented representation learning He et al. (2026) investigate how visual evidence can be adapted and combined for detection. MLLMs further extend this direction toward semantic anomaly analysis Tan et al. (2026a) and interpretable authenticity reasoning Tan et al. (2026b). Reconstruction provides a further means of probing an image through a pretrained generative model. Its role extends from supplying detection cues, as in the diffusion reconstruction errors used by DIRE Wang et al. (2023), to constructing training examples, as in the reconstructed hard samples used for contrastive learning by DRCT Chen et al. (2024).

AI-Generated Image Detection Benchmarks. Generalization evaluation must account for differences in generation mechanisms, image content, and data sources Xu et al. (2025). GenImage provides a large-scale setting covering GANs and diffusion models across ImageNet Deng et al. (2009) categories. ARForensics Zhang et al. (2025) complements this coverage with visual autoregressive generators that employ different token prediction architectures, while EvalGEN Chen et al. (2025) includes more recent text-to-image generators spanning diffusion and autoregressive models. These benchmarks support evaluating transfer across diverse generator families and generation paradigms. Broader image distributions introduce additional challenges beyond architectural variation. Qwen-Image-Bench Li et al. (2026a), originally developed for text-to-image evaluation, provides outputs from open-source and proprietary systems across varied creative scenarios. Chameleon Yan et al. (2025) includes high-quality real and synthetic images collected online, with generated samples curated to challenge human perception. GPT-Image-2-Twitter Zewde et al. (2026) captures a complementary setting through user-shared images attributed to GPT-Image-2, spanning diverse subjects and visual styles.

3 Method

As illustrated in Fig. 3, RED captures forensic evidence from coarse-to-fine reconstruction evolution through three components.

How to Capture Evidence Throughout Reconstruction? RED constructs a reconstruction trajectory by progressively accumulating and decoding multiscale quantized representations with a frozen VQ-VAE. A frozen CLIP encoder maps the original image and successive reconstruction states into a shared visual feature space (Sec. 3.1).

How to Weight Stages for Each Image? RED maps scale-wise token NLLs from the frozen VAR model to image-specific stage weights, learning how token predictability should guide the contribution of visual evidence across reconstruction stages (Sec. 3.2).

How to Combine Evidence Across Stages? RED modulates the projected reconstruction features with the learned stage weights and jointly processes them with the unweighted original-image representation through cross-stage attention. Learnable stage embeddings retain token identity. The same weights are reused to pool the contextualized reconstruction features, and this summary is concatenated with the contextualized original-image representation for detection (Sec. 3.3).

3.1 Reconstruction Trajectory Representation

We formulate reconstruction evolution as an ordered trajectory of cumulative reconstruction states, making intermediate visual evidence available for joint forensic analysis. Each state incorporates quantized information up to a particular scale, providing a structured representation of how image content develops throughout reconstruction. To construct this trajectory, we use the frozen multi-scale VQ-VAE associated with VAR to extract a hierarchy of discrete token maps from an input image x:

$$(\mathbf{z}_1, \dots, \mathbf{z}_K) = \mathcal{Q}_{\mathrm{ms}}\big(\mathcal{E}_{\mathrm{VQ}}(x)\big), \qquad \mathbf{z}_k \in \{1, \dots, V\}^{n_k \times n_k}, \tag{1}$$

where $\mathcal{E}_{\mathrm{VQ}}$ is the image encoder, $\mathcal{Q}_{\mathrm{ms}}$ denotes multiscale residual quantization, and V is the codebook size. We use $K = 10$ scales, with $(n_1, \dots, n_K) = (1, 2, 3, 4, 5, 6, 8, 10, 13, 16)$.

We organize the quantized residual contributions into cumulative latent states and decode each state to obtain the reconstruction trajectory:

$$\begin{aligned} \widehat{f}_k &= \sum_{j=1}^{k} \mathcal{U}_j\big(e(\mathbf{z}_j)\big), \qquad x_k = \mathcal{D}_{\mathrm{VQ}}(\widehat{f}_k), \\ \mathcal{T}(x) &= (x_1, \dots, x_K), \qquad k = 1, \dots, K, \end{aligned} \tag{2}$$

where e denotes codebook lookup, $\mathcal{D}_{\mathrm{VQ}}$ is the shared decoder, and $\mathcal{U}_j$ resizes the token embeddings to the final latent resolution and applies the pretrained residual transformation. All states have the same output resolution, while successive states incorporate additional quantized residual information. The stage index therefore identifies the extent of accumulated reconstruction information, giving the trajectory a consistent coarse-to-fine organization.

A frozen CLIP ViT-L/14 encoder ϕ embeds the original image and all reconstruction states into a shared feature space:

$$F_k = \phi(x_k) \in \mathbb{R}^d, \qquad \mathcal{F}(x) = (F_0, \dots, F_K), \qquad x_0 = x, \tag{3}$$

where $k = 0, \dots, K$. The ordered sequence $\mathcal{F}(x)$ retains the original-image representation alongside the visual evidence available at successive reconstruction stages. The same token hierarchy that constructs these states also supplies the likelihood evaluation in Sec. 3.2, connecting visual trajectory representation and adaptive stage weighting through a common reconstruction process. Complete reconstruction sequences for real and generated images are visualized in Appendix D.

3.2 NLL-Guided Stage Weighting

We translate the observed scale dependence of token predictability into a learnable rule for allocating reconstruction evidence. The central design is to use scale-wise NLLs to regulate the contributions of visual representations, with their influence learned through the detection objective.

Let $\mathbf{z}_{<k} = (\mathbf{z}_1, \dots, \mathbf{z}_{k-1})$ denote the token maps from preceding scales. Under teacher forcing, the frozen VAR-d16 model evaluates the tokens at scale k conditioned on these input-derived token maps and the unconditional conditioning input $c_\varnothing$. The token-averaged negative log-likelihood is

$$E_k(x) = -\frac{1}{N_k} \sum_{i=1}^{N_k} \log p_\vartheta\big(z_{k,i} \mid \mathbf{z}_{<k}, c_\varnothing\big), \qquad N_k = n_k^2, \tag{4}$$

where ϑ denotes the frozen VAR parameters and $z_{k,i}$ is the i-th token at scale k. Token averaging removes the direct dependence of the sum on token count. Larger E_k indicates lower conditional token predictability under VAR.

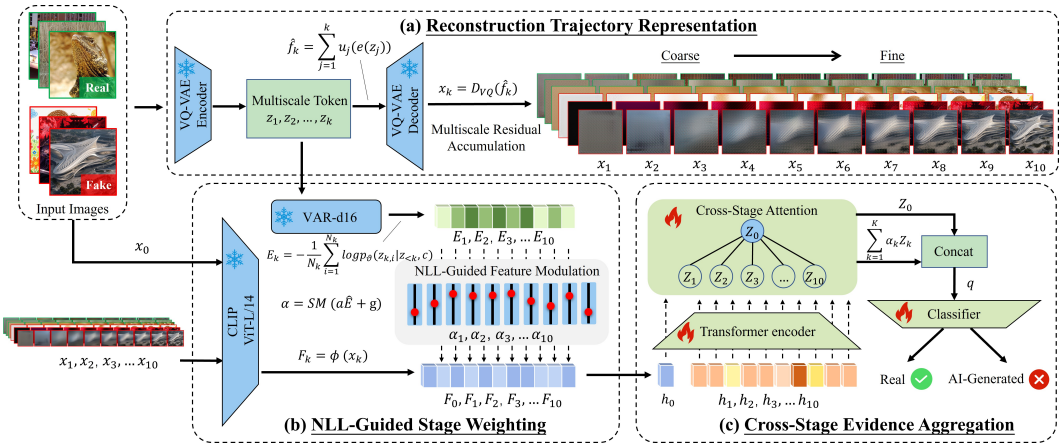


Figure 3: Overview of RED for generalizable AI-generated image detection. (a) Coarse-to-fine reconstruction states form a trajectory of visual evidence. (b) Scale-wise token NLLs guide image-specific weighting of reconstruction features. (c) Cross-stage evidence aggregation jointly models original-image and reconstruction representations, then combines a dedicated original-image readout with a trajectory summary pooled using the same stage weights.

To account for differences in the typical NLL ranges across scales, we standardize each score using the mean μ_k and standard deviation σ_k estimated exclusively from the source training split. These statistics remain fixed during validation and testing. We then map the standardized scores to stage weights:

$$\widehat{E}_k(x) = \frac{E_k(x) - \mu_k}{\sigma_k}, \qquad \widehat{\mathbf{E}}(x) = [\widehat{E}_1(x), \ldots, \widehat{E}_K(x)]^\top, \\ \boldsymbol{\alpha}(x) = \mathrm{softmax}\big(a\widehat{\mathbf{E}}(x) + \mathbf{g}\big). \tag{5}$$

This parameterization combines image-dependent deviations in token predictability with stage preferences learned across the source training set. The shared scalar a controls the direction and strength of the standardized NLL contribution, while the stage biases $\mathbf{g} \in \mathbb{R}^K$ capture preferences shared across images. The softmax normalizes these contributions over reconstruction stages, yielding $\alpha_k > 0$ and $\sum_{k=1}^{K} \alpha_k = 1$. Both a and $\mathbf{g}$ are optimized for detection, allowing supervision to determine how the predictability signal should shape the allocation of visual evidence.

3.3 Cross-Stage Evidence Aggregation

We introduce cross-stage evidence aggregation to couple image-adaptive stage weighting with joint modeling of original-image and reconstruction representations. The aggregation design uses the same stage weights during both token construction and evidence readout, while retaining a dedicated original-image component in the final representation.

We first construct evidence tokens by modulating the projected reconstruction features with their stage weights and adding learnable stage embeddings:

$$h_k = \begin{cases} WF_0 + \mathbf{p}_0, & k = 0, \\ \alpha_k W F_k + \mathbf{p}_k, & 1 \le k \le K, \end{cases} \tag{6}$$

where $W \in \mathbb{R}^{d_h \times d}$ is a shared projection and $\mathbf{p}_k \in \mathbb{R}^{d_h}$ encodes token identity within the sequence, distinguishing the original image and individual reconstruction stages. Stage weights modulate the visual content entering cross-stage interaction, while the embeddings retain stage identity. The original-image token h_0 enters this interaction without NLL-based weighting.

Within the aggregation module, a Transformer encoder (Vaswani et al., 2017), $\mathcal{A}_\eta$, jointly contextualizes the evidence tokens:

$$(Z_0, \ldots, Z_K) = \mathcal{A}_\eta(h_0, \ldots, h_K). \tag{7}$$

Self-attention enables each token to incorporate information from the original image and other reconstruction stages. Consequently, the resulting representations encode relationships across the trajectory in addition to the evidence contained in individual states.

For evidence readout, we reuse the stage weights to summarize the contextualized reconstruction tokens and concatenate this summary with the contextualized original-image token:

$$q = \mathrm{Concat}\left(Z_0, \sum_{k=1}^{K} \alpha_k Z_k\right), \qquad \ell(x) = f_\omega(q), \qquad s(x) = \sigma\big(\ell(x)\big), \tag{8}$$

where f_ω is the classification head, $\ell(x)$ is its logit, and σ is the sigmoid function. The score $s(x) \in (0, 1)$ is used for ranking and thresholded detection, with $y = 1$ denoting an AI-generated image.

Reusing $\boldsymbol{\alpha}$ connects the allocation of reconstruction evidence at the input to its contribution in the final readout: the same image-specific weights modulate the features presented to attention and pool its reconstruction outputs. The separate Z_0 component retains a dedicated original-image readout alongside the trajectory summary, with both informed by cross-stage interactions.

We jointly optimize stage weighting and evidence aggregation using binary cross-entropy with logits. For a minibatch $\mathcal{B}$, the objective is

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} \big[\log\big(1 + \exp(\ell(x))\big) - y\ell(x)\big]. \tag{9}$$

The VQ-VAE, VAR, and CLIP remain frozen. Training updates the stage-weighting parameters, shared projection, stage embeddings, Transformer encoder, and classification head, learning how reconstruction evidence is allocated, contextualized, and combined for detection.

4 Experiments

4.1 Experimental Settings

Baselines and Data. We compare RED with the 13 baselines in Table 1, all trained or fine-tuned on the same GenImage subset covering all training categories. Evaluation uses generated images from GenImage and five external collections listed in Table 1. All real images come from ImageNet, with matched real and generated sample counts within each evaluation split, an additional evaluation with alternative real-image sources is reported in Appendix B (Table 9). We report accuracy (ACC.) and average precision (A.P.), with generated images as the positive class.

Model Configuration. RED uses frozen multiscale VQ-VAE, VAR-d16, and CLIP ViT-L/14 backbones, with $K = 10$ reconstruction states at 256×256 resolution. The detection Transformer has two layers, a hidden dimension of 128, four attention heads, and a feed-forward dimension of 256. Transformer dropout and scale dropout are 0.1 and 0.15, respectively. VAR scores input-derived tokens under teacher forcing with unconditional conditioning $c = c_\varnothing$ and conditional dropout disabled. Token-averaged NLLs are standardized independently at each scale using source-training statistics, which remain fixed during validation and testing.

Training and Evaluation. We optimize the stage-weighting and detection modules using binary cross-entropy with logits and AdamW, with learning rate and weight decay both set to 10^{-4}, batch size 64, and gradient clipping at a global norm of 1.0. Training runs for at most 50 epochs, with early-stopping patience of eight epochs. Checkpoints are selected by source-validation ACC at threshold 0.5; the released checkpoint corresponds to epoch 16. The same threshold is used for all RED evaluations, without target-domain training or calibration. Results are averaged equally across generators within each collection and across the six evaluation sets.

4.2 Main Results

Table 1 summarizes performance across the six evaluation sets. RED achieves the highest average ACC of 92.5% and AP of 97.5%, exceeding the strongest baseline averages by 3.9 and 2.0 per-

Table 1: Comparison with state-of-the-art detectors on in-domain and out-of-domain benchmarks (%). The best, second-best, and third-best results in each column are indicated by **bold**, solid underline, and dashed underline, respectively. † denotes MLLM-based detectors.

Method	Venue	In Domain		Out of Domain										Avg.	
		GenImage		ARForensics		EvalGEN		QIB		Chameleon		G2-T			
		ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.
CNNSpot Wang et al. (2020)	CVPR	61.2	66.9	61.9	66.3	75.8	84.8	54.8	58.1	56.8	61.3	55.4	57.5	61.0	65.8
UnivFD Ojha et al. (2023)	CVPR	87.9	94.9	82.7	90.3	93.8	98.9	89.1	95.4	92.7	98.2	85.2	93.1	88.6	95.1
DIRE Wang et al. (2023)	ICCV	**99.9**	**99.9**	74.7	82.3	50.7	77.9	78.1	86.5	50.0	42.7	50.4	57.6	67.3	74.5
NPR Tan et al. (2024)	CVPR	90.9	97.6	81.1	91.8	50.5	42.5	62.2	70.6	51.6	43.3	50.9	47.9	64.5	65.6
DRCT Chen et al. (2024)	ICML	97.7	98.5	85.3	98.3	90.9	98.7	62.8	89.9	72.7	96.9	55.9	89.9	77.5	95.4
D^{3}QE Zhang et al. (2025)	ICCV	88.8	93.9	82.5	89.5	91.9	97.2	88.0	94.1	86.9	92.2	82.3	90.0	86.7	92.8
LOTA Wang et al. (2025)	ICCV	88.6	96.9	87.7	96.5	47.5	37.9	64.6	61.3	48.9	39.9	46.7	37.5	64.0	61.7
DDA Chen et al. (2025)	NeurIPS	90.7	97.4	86.8	95.8	85.7	95.7	59.9	75.0	62.9	82.8	50.5	56.1	72.7	83.8
MF^{2}DA Mu et al. (2026)	TIFS	88.4	95.4	84.0	92.2	92.5	98.1	88.5	95.8	93.3	**99.2**	83.3	92.1	88.3	95.5
DoU He et al. (2026)	CVPR	73.9	81.6	**99.0**	**99.9**	64.3	88.9	63.5	77.4	54.2	71.0	51.4	61.7	67.7	80.1
IAPL Li et al. (2026b)	CVPR	74.0	81.7	73.8	81.6	79.8	94.3	74.2	85.5	54.1	66.2	57.7	74.9	68.9	80.7
AReason† Tan et al. (2026a)	ICLR	65.6	60.9	77.8	71.6	86.9	79.7	82.6	75.8	84.5	78.4	83.0	77.4	80.1	74.0
Veritas† Tan et al. (2026b)	ICLR	79.8	79.5	77.2	76.9	**99.6**	**99.4**	**95.9**	95.6	**95.5**	95.5	83.5	82.6	88.6	88.2
Ours	/	95.7	98.9	92.2	97.4	96.7	99.2	93.2	**98.0**	90.7	97.1	**86.5**	**94.2**	**92.5**	**97.5**

Note. QIB: Qwen-Image-Bench; G2-T: GPT-Image-2-Twitter;

Table 2: Comparison of detection accuracy (ACC) on ARForensics and EvalGEN (%).

Method	ARForensics (ACC.)								EvalGEN (ACC.)					
	Inf	JP	LG	OMV2	RAR	Switti	VAR	Avg.	Flux	GoT	Inf	NOVA	OG	Avg.
CNNSpot	75.9	74.3	59.0	49.6	54.0	72.2	48.4	61.9	72.4	79.9	67.4	79.1	80.2	75.8
UnivFD	94.7	94.1	73.5	67.1	91.9	92.0	65.7	82.7	92.6	94.0	94.1	94.5	93.7	93.8
DIRE	73.5	74.5	76.0	73.3	76.7	75.0	74.3	74.7	50.6	51.0	50.5	50.7	50.6	50.7
NPR	89.6	96.6	82.8	62.3	57.1	92.8	86.3	81.1	48.5	49.6	51.7	50.5	52.3	50.5
DRCT	98.0	97.9	88.9	74.4	66.8	97.6	73.6	85.3	87.1	97.3	86.1	87.7	96.1	90.9
D^{3}QE	93.2	92.4	74.4	68.0	90.7	90.2	68.8	82.5	90.7	92.2	92.1	92.1	92.2	91.9
LOTA	89.0	87.7	86.7	88.2	87.7	90.3	84.4	87.7	47.6	46.3	48.1	47.7	47.6	47.5
DDA	75.5	97.6	95.2	71.0	82.0	98.2	88.2	86.8	62.5	92.2	80.9	96.4	96.4	85.7
MF^{2}DA	93.0	93.4	72.7	72.7	92.0	91.3	72.9	84.0	91.4	94.0	92.9	92.5	91.9	92.5
DoU	98.9	**99.1**	**99.1**	**98.9**	**99.2**	**99.0**	**99.0**	**99.0**	62.8	67.1	58.3	69.2	63.9	64.3
IAPL	74.1	74.3	73.7	72.4	74.6	73.5	73.9	73.8	70.7	78.0	81.0	90.4	78.9	79.8
AReason†	89.5	81.0	80.0	70.0	77.5	80.5	66.0	77.8	86.5	83.5	89.5	86.5	88.5	86.9
Veritas†	**99.9**	92.0	71.0	58.5	63.5	98.5	57.0	77.2	**99.9**	**99.5**	**99.0**	**99.5**	**99.9**	**99.6**
Ours	96.7	96.3	85.6	91.3	96.3	96.3	82.7	92.2	96.3	96.7	96.6	96.8	96.9	96.7

Note. Inf: Infinity; JP: Janus-Pro; LG: LlamaGen; OMV2: Open-MAGVIT2; OG: OmniGen;

centage points, respectively. Its ACC remains at least 86.5% and AP at least 94.2% across all sets, indicating consistently strong performance across the evaluated generated-image sources.

The comparison with DIRE and DRCT highlights the distinction between source-domain performance and generalization. On GenImage, these methods achieve 99.9% and 97.7% ACC, respectively, exceeding RED's 95.7%. Across the five external evaluation sets, however, their average ACC falls to 60.8% and 73.5%, while RED reaches 91.9%. Individual baselines retain advantages on specific sets, including Veritas and MF^{2}DA on Chameleon, but RED achieves the strongest overall results.

Table 2 further demonstrates RED's balanced performance across generation mechanisms. RED ranks second on both ARForensics and EvalGEN, with ACC of 92.2% and 96.7%, respectively. In comparison, DoU leads on ARForensics with 99.0% but obtains 64.3% on EvalGEN, while Veritas achieves 99.6% on EvalGEN and 77.2% on ARForensics. RED also maintains ACC between 96.3% and 96.9% across all five EvalGEN generators, showing limited variation within this collection.

On Qwen-Image-Bench, RED achieves the highest average AP of 98.0% across 18 generators, ranking first on 10 and maintaining AP between 96.0% and 99.4% (Table 3). Veritas achieves higher ACC on this set (95.9% versus 93.2%), whereas RED achieves higher AP (98.0% versus 95.6%), distinguishing threshold-dependent classification performance from score-ranking performance. On GPT-Image-2-Twitter, RED achieves the best ACC of 86.5% and AP of 94.2%, exceeding UnivFD, the strongest baseline on this set, by 1.3 and 1.1 percentage points, respectively. Appendix B reports the remaining per-generator metrics (Tables 6, 7, and 8).

Table 3: Comparison of average precision (AP) on Qwen-Image-Bench (%).

Method	Qwen-Image-Bench (A.P.)																		
	F2-P	F2-M	GLM	G1	G1.5	HY3	I4	I4-U	QI	Q2-P	Q2512	S4	S4.5	S5	G2	K2.1	NB2	NB-P	Avg.
CNNSpot	57.7	51.7	67.0	62.1	58.4	55.4	63.0	61.6	68.9	53.1	60.6	56.4	56.6	52.9	55.9	62.4	51.4	51.5	58.1
UnivFD	96.0	94.4	97.0	95.8	93.9	97.8	96.1	96.1	96.5	96.4	96.0	95.4	92.6	95.6	95.2	96.1	94.2	92.0	95.4
DIRE	90.7	64.6	81.0	**99.9**	**99.9**	92.6	95.3	95.0	82.1	**99.9**	66.0	75.9	77.0	85.7	**99.9**	79.2	72.2	**99.9**	86.5
NPR	53.8	59.3	45.2	95.1	83.3	80.8	82.3	83.0	69.8	93.7	43.8	59.2	70.4	42.5	86.1	59.6	74.1	89.3	70.6
DRCT	77.6	70.0	**98.5**	95.8	93.1	98.5	97.4	**98.8**	95.4	94.4	94.2	87.6	80.3	80.0	86.9	94.5	83.3	92.3	89.9
D^3QE	95.0	91.5	94.4	94.3	92.1	96.6	94.8	95.0	95.0	95.5	95.2	94.4	92.0	94.7	93.6	94.8	92.9	92.3	94.1
LOTA	37.6	48.7	41.3	96.2	98.5	77.7	37.7	38.1	51.9	99.2	37.1	49.3	47.7	41.2	98.6	50.4	53.2	98.9	61.3
DDA	80.2	64.4	86.7	92.3	67.7	95.6	91.5	90.0	76.4	76.0	69.2	58.6	63.4	79.8	47.4	79.1	64.9	67.5	75.0
MF^2DA	96.3	94.3	96.9	96.8	94.2	98.1	96.2	96.2	96.9	97.1	96.7	95.9	91.7	96.5	95.6	96.8	95.6	92.8	95.8
DoU	56.6	62.0	79.8	95.3	97.5	92.7	76.8	77.4	77.7	94.5	58.1	73.0	76.2	49.6	95.3	68.3	68.7	94.1	77.4
IAPL	59.7	74.4	80.4	99.7	**99.9**	97.5	93.7	94.7	86.4	99.4	60.1	83.4	82.6	66.5	97.7	82.4	81.6	98.5	85.5
AReason†	79.1	74.0	83.6	78.2	64.2	81.5	83.8	77.8	82.2	77.0	74.5	67.3	72.7	77.4	70.0	86.6	68.2	67.0	75.8
Veritas†	95.0	91.5	97.5	99.0	95.0	98.0	97.0	96.0	98.0	96.0	**99.0**	92.5	95.0	95.5	88.0	98.0	95.5	94.0	95.6
Ours	**97.8**	**96.0**	97.6	97.7	96.6	**99.4**	**98.4**	98.5	**99.2**	98.9	98.8	**98.1**	**97.3**	**98.0**	98.5	**98.6**	**98.1**	97.4	**98.0**

Note. F2-P: FLUX.2-pro; F2-M: FLUX.2_max; GLM: GLM-Image; G1: GPT-Image-1; G1.5: GPT-Image-1.5; HY3: HunyuanImage-3.0; I4: Imagen-4.0; I4-U: Imagen-4.0-Ultra; QI: Qwen-Image; Q2-P: Qwen-Image-2.0-pro; Q2512: Qwen-Image-2512; S4: Seedream-4.0; S4.5: Seedream-4.5; S5: Seedream-5.0; G2: gpt-image-2; K2.1: kling_v2_1; NB2: nano-banana-2.0; NB-P: nano-banana-pro;

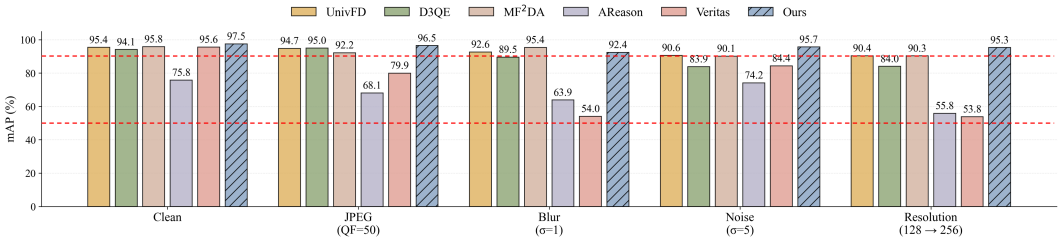


Figure 4: Robustness to common image degradations on Qwen-Image-Bench. Dashed lines mark the 90% and 50% mAP reference levels.

4.3 Robustness to Common Image Degradations

We evaluate robustness on Qwen-Image-Bench, covering 18 generators. Four degradations are tested: JPEG compression with quality factor 50, Gaussian blur with $\sigma = 1$, Gaussian noise with $\sigma = 5$, and downsampling to 128×128 followed by upsampling to 256×256. All models are evaluated without additional fine-tuning, using mean average precision (mAP) across generators.

As shown in Fig. 4, RED achieves the highest mAP under JPEG compression, noise, and resolution reduction, reaching 96.5%, 95.7%, and 95.3%, respectively. Averaged across all four degradations, RED achieves 95.0% mAP, compared with 92.1% for UnivFD and 92.0% for MF^2DA. Blur remains the most challenging degradation for RED: its mAP decreases from 98.0% on clean images to 92.4%, below MF^2DA's 95.4%. This sensitivity reflects the attenuation of fine-scale structures and textures by blurring, which can weaken the forensic evidence captured along the reconstruction trajectory.

The evaluated baselines exhibit markedly different sensitivities to image degradation. UnivFD and MF^2DA maintain at least 90.4% and 90.1% mAP, respectively, across all four conditions. In contrast, the evaluated MLLM-based detectors suffer substantial performance losses, revealing a pronounced gap between clean-image performance and degradation robustness. Most notably, Veritas achieves 95.6% mAP on clean images but drops to 54.0% under blur and 53.8% under resolution reduction, corresponding to declines of 41.6 and 41.8 percentage points. AReason also decreases from 75.8% to 55.8% under resolution reduction, a loss of 20.0 points.

4.4 Ablation Study

Table 4 evaluates the three components of RED: reconstruction trajectory representation, NLL-guided stage weighting, and cross-stage evidence aggregation. The ablations examine which reconstruction states provide useful evidence, how token NLLs should inform detection, and how evidence should be integrated across stages.

Table 4: Component-wise ablation of RED across six evaluation sets (ACC, %).

Variant	Modification / Setting	In Domain	Out of Domain					Avg.
		GenImage	ARForensics	EvalGEN	QIB	Chameleon	G2-T	
Reconstruction Trajectory Representation								
Original image only	Remove all reconstruction inputs	87.9	82.7	93.8	89.1	**92.7**	85.2	88.6
Final reconstruction only	Use the final reconstruction alone	88.0	79.1	93.2	79.5	85.8	72.7	83.0
Original + final reconstruction	Exclude intermediate reconstruction states	95.5	91.8	96.4	90.3	84.1	81.3	89.9
All stages, uniform weights	Use all reconstruction stages with equal weights	95.4	92.0	96.3	92.0	87.8	82.2	91.0
NLL-Guided Stage Weighting								
NLL-only detection	Use NLL scores without visual features	60.9	56.5	68.4	52.5	53.0	51.6	57.2
Visual–NLL concatenation	Concatenate NLL scores with visual features	94.7	92.0	95.3	90.5	86.4	81.9	90.1
Learned static weights	Fix $a = 0$; learn stage biases $\mathbf{g}$	**97.8**	91.3	94.9	88.1	87.6	82.9	90.4
Cross-Stage Evidence Aggregation								
Mean pooling	Replace the aggregation design with mean pooling	86.6	80.5	93.5	75.7	82.5	67.8	81.1
NLL-weighted averaging	Replace the aggregation design with weighted averaging	95.3	**92.4**	96.5	91.6	87.3	84.4	91.2
MLP aggregation	Replace the aggregation design with an MLP	95.6	92.0	96.2	92.1	89.8	83.1	91.5
Full Model								
RED	Complete proposed configuration	95.7	92.2	**96.7**	**93.2**	90.7	**86.5**	**92.5**

QIB: Qwen-Image-Bench; G2-T: GPT-Image-2-Twitter.

Reconstruction Trajectory Representation. This ablation examines the value of intermediate reconstruction states beyond the original image and reconstruction endpoint. Original-image-only and final-reconstruction-only detection achieve average ACCs of 88.6% and 83.0%, respectively. The lower reconstruction-only accuracy may reflect the loss of useful forensic cues during reconstruction. Combining both representations improves ACC to 89.9%, suggesting that reconstruction provides complementary evidence while the original image preserves cues lost in the process. Including all reconstruction stages with uniform weights further raises ACC to 91.0%.

NLL-Guided Stage Weighting. This ablation compares direct NLL features with NLL-guided stage weighting. NLL-only detection achieves 57.2% average ACC, indicating limited standalone discrimination. Visual–NLL concatenation reaches 90.1%, below uniform weighting at 91.0%, suggesting that directly appending NLLs does not effectively exploit their information in this setting. To isolate image-dependent guidance from learned stage preferences, we fix $a = 0$ while retaining learnable biases $\mathbf{g}$. The resulting static weights achieve 90.4% average ACC, also below uniform weighting. Although static weighting outperforms RED on GenImage (97.8% versus 95.7%), RED performs better on all five out-of-domain sets and reaches 92.5% overall, a gain of 2.1 percentage points. This pattern is consistent with fixed stage preferences being more suited to the source distribution and transferring less effectively.

Cross-Stage Evidence Aggregation. This ablation compares the proposed aggregation design with pooling and MLP-based alternatives. Mean pooling achieves 81.1% average ACC, while NLL-weighted averaging and MLP aggregation reach 91.2% and 91.5%, respectively. The weak mean-pooling result suggests that indiscriminate fusion may dilute useful stage-specific evidence. Full RED achieves 92.5%, exceeding weighted averaging and MLP aggregation by 1.3 and 1.0 percentage points, respectively. These gains suggest additional value in jointly contextualizing stage representations within the complete aggregation design. Compared with weighted averaging, RED improves ACC by 3.4 points on Chameleon and 2.1 points on GPT-Image-2-Twitter, but decreases slightly by 0.2 points on ARForensics, indicating dataset-dependent benefits.

5 Conclusion

We presented **RED** (**Reconstruction Evolution Dynamics**), a framework for generalizable AI-generated image detection that models visual representations across coarse-to-fine reconstruction stages. RED integrates multiscale VQ-VAE reconstruction, VAR-based NLL-guided stage weighting, and cross-stage evidence aggregation to combine original-image features with intermediate reconstruction evidence. Experiments on six diverse benchmarks show that RED achieves the highest average accuracy and average precision among the evaluated detectors, with strong generalization to unseen generators without target-domain adaptation. Further evaluations support its robustness to common image degradations. These findings highlight reconstruction evolution as a valuable source of complementary forensic evidence for generalizable AI-generated image detection.

AI USE STATEMENT

Generative AI tools were used to assist with experimental code development and language editing of the manuscript. The authors reviewed and verified all AI-assisted content and take full responsibility for the final work.

REFERENCES

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion-generated images. In *Proc. Int. Conf. Mach. Learn. (ICML)*, volume 235, pp. 7621–7639, July 2024.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes AI-generated image detector easier generalizable. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 38, pp. 106475–106500, December 2025.

Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. FIRE: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 12830–12839, June 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 248–255, June 2009.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 27, pp. 2672–2680, December 2014.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general AI-generated image detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 18685–18694, June 2025.

Qinghui He, Haifeng Zhang, Qiao Qin, Bo Liu, Xiuli Bi, and Bin Xiao. Diversity over uniformity: Rethinking representation in generated image detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 40407–40417, June 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 33, pp. 6840–6851, December 2020.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, Ziyi He, Wei Wang, Dalin Li, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiao Xu, Xiaoyue Chen, Yuxiang Chen, Yan Shu, Yanran Zhang, Yilei Chen, Yixian Xu, Zekai Zhang, Zhendong Wang, Zihao Liu, Zikai Zhou, Hongzhu Shi, Yi Wang, Bing Zhao, Hu Wei, Lin Qu, and Chenfei Wu. Qwen-Image-Bench: From generation to creation in text-to-image evaluation. *arXiv preprint arXiv:2605.28091*, May 2026a. doi: 10.48550/arXiv.2605.28091.

Yiheng Li, Zichang Tan, Guoqing Xu, Zhen Lei, Xu Zhou, and Yang Yang. Towards generalizable AI-generated image detection via image-adaptive prompt learning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 21262–21272, June 2026b.

Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. LaRE^2: Latent reconstruction error based method for diffusion-generated image detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 17006–17015, June 2024.

Wenpeng Mu, Zheng Li, Qiang Xu, Xinghao Jiang, and Tanfeng Sun. Exda: Towards universal detection and plug-and-play attribution of ai-generated ex-regulatory images. In *Proc. ACM MM*, pp. 11512–11521, 2025.

Wenpeng Mu, Qiang Xu, Yining Zhang, Xiao Zhang, Huiyuan Wang, Yuefan Li, Xinghao Jiang, and Tanfeng Sun. MF2DA: Multi-level feature fusion for robust detection and attribution of universal AI-generated images. *IEEE Trans. Inf. Forensics Security*, 2026.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 24480–24489, June 2023.

Xinyi Qi, Kai Ye, Chengchun Shi, Ying Yang, Jin Zhu, and Hongyi Zhou. A difference-in-difference approach to detecting ai-generated images. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 42964–42975, June 2026.

Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 28130–28139, June 2024.

Chuangchuang Tan, Xiang Ming, Jinglu Wang, Renshuai Tao, Bin Li, Yunchao Wei, Yao Zhao, and Yan Lu. Semantic visual anomaly detection and reasoning in AI-generated images. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, volume 2026, pp. 111907–111942, 2026a.

Hao Tan, Jun Lan, Zichang Tan, Senyuan Shi, Ajian Liu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas: Generalizable deepfake detection via pattern-aware reasoning. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, volume 2026, pp. 66420–66477, 2026b.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 37, pp. 84839–84865, December 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 30, pp. 5998–6008, December 2017.

Hongsong Wang, Renxi Cheng, Yang Zhang, Chaolei Han, and Jie Gui. LOTA: Bit-planes guided AI-generated image detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 17246–17255, October 2025.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 8695–8704, June 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 22445–22455, October 2023.

Qiang Xu, Wenpeng Mu, Jianing Li, Tanfeng Sun, and Xinghao Jiang. Advancements in ai-generated content forensics: A systematic literature review. *ACM Comput. Surv.*, 58(3):1–36, 2025.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, volume 2025, pp. 70702–70720, April 2025.

Kidus Zewde, Simiao Ren, Xingyu Shen, Jiaqi Wu, Yuchen Zhou, Tommy Duong, Zikang Zhang, Ethan Traister, and Kewen Xie. GPT-Image-2 in the wild: A twitter dataset of self-reported AI-generated images from the first week of deployment. *arXiv preprint arXiv:2604.25370*, April 2026. doi: 10.48550/arXiv.2604.25370.

Yanran Zhang, Bingyao Yu, Yu Zheng, Wenzhao Zheng, Yueqi Duan, Lei Chen, Jie Zhou, and Jiwen Lu. D^3QE: Learning discrete distribution discrepancy-aware quantization error for autoregressive-generated image detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 16292–16301, October 2025.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A million-scale benchmark for detecting AI-generated image. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 36, pp. 77771–77782, December 2023.

A Dataset Construction and Reproducibility Details

Source Data and Splits. Using sampling seed 2026, we select 1,000 images from each of eight generators in the GenImage training split: ADM, BigGAN, GLIDE, Midjourney, Stable Diffusion v1.4, Stable Diffusion v1.5, VQDM, and Wukong. Together with 8,000 ImageNet training images, these form a source dataset of 16,000 samples. A fixed 70/30 split, stratified across the eight generator groups and the real-image group, yields 11,200 training and 4,800 validation samples, with equal numbers of real and generated images in each split. Each generator contributes 700 training and 300 validation images. No ImageNet validation images are included in either source split. GenImage testing uses held-out samples from the same eight generators.

Evaluation Data and Metric Aggregation. The main evaluation covers 40 generator/dataset groups across six collections, comprising 37,500 generated-image records (Table 5). A fixed pool of 1,000 ImageNet validation images serves as the real-image reference. EvalGEN uses a fixed 500-image subset of this pool and 500 generated images per group; all other collections use the full reference pool and 1,000 generated images per group. The same real-image subset is reused across groups within each collection. We first macro-average metrics over groups within each collection and then compute the overall results as the unweighted mean of the six collection-level scores.

Table 5: Per-group evaluation sample counts. Real images are drawn from a fixed pool of 1,000 ImageNet validation images; EvalGEN uses a fixed 500-image subset.

Evaluation collection	Groups	Real/group	Generated/group
GenImage	8	1,000	1,000
ARForensics	7	1,000	1,000
EvalGEN	5	500	500
Qwen-Image-Bench	18	1,000	1,000
Chameleon	1	1,000	1,000
GPT-Image-2-Twitter	1	1,000	1,000

Preprocessing and Feature Extraction. Following VAR's validation preprocessing, images are decoded as RGB, resized using Lanczos interpolation so that the shorter side is 288 pixels, and center-cropped to 256×256. They are then converted to tensors in $[0, 1]$, without online random augmentation. The frozen multiscale VQ-VAE decodes ten cumulative reconstruction states at 256×256, using token-grid scales $(1, 2, 3, 4, 5, 6, 8, 10, 13, 16)$. We use the same frozen CLIP ViT-L/14 encoder as UnivFD to map the original image and each reconstruction to a 768-dimensional feature vector. CLIP preprocessing applies a 224×224 center crop and channel-wise normalization with means $(0.48145466, 0.45782750, 0.40821073)$ and standard deviations $(0.26862954, 0.26130258, 0.27577711)$.

NLL Computation and Model Selection. The official frozen VAR-d16 model evaluates input-derived tokens under teacher forcing, using the unconditional input $c = c_\varnothing$ with conditional dropout disabled. Tokens at the current scale serve only as prediction targets, while preceding scales provide the token conditioning. Token-averaged NLLs are standardized independently at each scale using statistics estimated exclusively from the source-training split and held fixed during validation and testing. RED contains 381,548 trainable parameters, while the VQ-VAE, VAR, and CLIP remain frozen. Training follows Section 4. Checkpoints are selected using source-validation ACC at a fixed decision threshold of 0.5. No target-domain checkpoint selection, threshold calibration, or hyperparameter tuning is performed.

Real-Image Source Control. To examine the effect of the real-image source, we pair a fixed set of 1,000 Chameleon generated images separately with 1,000 real images from ImageNet validation, Chameleon, LSUN, or MS COCO. The detector checkpoint, preprocessing, decision threshold, and NLL standardization statistics remain fixed, so that only the real-image source varies. Samples are selected without consulting detector scores or predictions. Real and generated images are not matched by content category because the sources do not share a common set of category annotations.

B Additional Experimental Results

We complement the main evaluation with detailed per-generator results, reporting ACC and AP on GenImage, AP on ARForensics and EvalGEN, and ACC on Qwen-Image-Bench.

Detailed Results on GenImage. Table 6 shows that RED achieves an average ACC of 95.7% and AP of 98.9%. Although DIRE and DRCT obtain higher average ACC, RED maintains ACC between 93.2% and 97.7% across all eight generators. Its AP ranges from 98.0% to 99.8%, yielding the second-highest average AP on this benchmark.

Table 6: Per-generator accuracy (ACC) and average precision (AP) on GenImage. All results are reported in percentages, and Avg. denotes the mean across eight generators.

Method	GenImage																	
	ADM		BigGAN		Glide		Midjourney		SD1.4		SD1.5		VQDM		Wukong		Avg.	
	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.
CNNSpot	53.6	56.1	62.4	63.8	63.6	69.0	59.0	64.7	63.1	71.9	65.1	73.8	58.7	62.8	63.9	73.1	61.2	66.9
UnivFD	75.6	86.3	90.9	97.3	91.2	97.2	83.4	92.4	91.1	96.9	92.1	97.2	89.9	96.3	89.2	95.7	87.9	94.9
DIRE	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**	**99.8**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**
NPR	93.2	98.6	91.4	97.7	96.9	99.5	80.2	92.9	92.0	98.1	92.0	98.1	92.1	98.3	89.4	97.8	90.9	97.6
DRCT	92.1	97.7	99.2	98.7	97.9	98.8	97.6	98.7	99.0	98.1	99.4	99.3	98.0	98.3	99.0	98.6	97.7	98.5
D^3QE	83.4	89.9	92.3	96.1	90.5	94.6	84.1	91.4	91.0	95.7	90.2	94.7	90.2	95.3	88.9	93.9	88.8	93.9
LOTA	86.3	96.0	87.5	96.8	93.3	99.0	93.0	98.8	88.6	97.0	88.4	96.8	85.8	95.5	86.3	95.6	88.6	96.9
DDA	89.6	97.4	89.8	97.9	88.2	97.0	88.9	97.3	98.3	**99.9**	98.3	**99.9**	74.1	89.4	98.2	**99.9**	90.7	97.4
MF^2DA	81.7	91.4	92.7	97.9	90.9	96.4	86.0	94.4	89.2	95.8	88.6	95.5	92.7	98.6	85.4	93.2	88.4	95.4
DoU	75.2	81.6	74.2	81.2	73.9	82.0	72.7	81.6	74.5	82.9	72.8	80.4	74.7	82.2	73.1	81.1	73.9	81.6
IAPL	72.8	80.2	75.2	83.1	76.0	83.5	72.0	80.6	73.1	80.7	74.3	82.2	75.2	82.4	73.7	81.0	74.0	81.7
AReason*	60.5	56.6	61.5	57.3	68.5	64.2	66.0	61.5	67.5	62.7	65.0	59.9	68.5	63.2	67.0	61.9	65.6	60.9
Veritas*	63.5	63.0	67.5	67.0	71.0	70.5	88.5	88.0	94.5	94.5	97.5	97.5	63.0	63.0	92.5	92.5	79.8	79.5
Ours	94.8	98.6	97.1	99.4	97.7	99.8	93.2	98.0	96.6	99.2	96.6	99.2	95.2	98.7	95.0	98.5	95.7	98.9

Average Precision on ARForensics and EvalGEN. As shown in Table 7, RED achieves 97.4% average AP on ARForensics, ranking behind DoU and DRCT. On EvalGEN, it achieves the second-highest average AP of 99.2%, following Veritas at 99.4%. RED's AP remains between 98.9% and 99.5% across all five EvalGEN generators, demonstrating consistently strong performance on this benchmark.

Table 7: Per-generator average precision (AP, %) on ARForensics and EvalGEN. Avg. denotes the mean across generators within each benchmark.

Method	ARForensics (A.P.)								EvalGEN (A.P.)					
	Inf.	JP	LG	OMV2	RAR	Switti	VAR	Avg.	Flux	GoT	Inf.	NOVA	OG	Avg.
CNNSpot	84.9	82.6	62.5	50.7	55.4	80.2	47.8	66.3	80.9	90.3	74.8	87.2	90.5	84.8
UnivFD	98.9	99.0	84.5	79.1	97.4	97.6	75.8	90.3	98.2	**99.6**	**99.1**	98.8	98.8	98.9
DIRE	81.2	81.7	83.1	81.3	84.0	82.9	81.8	82.3	77.9	80.4	71.1	80.7	79.5	77.9
NPR	97.7	99.5	95.1	83.1	72.0	98.5	96.4	91.8	35.5	35.2	49.0	44.7	48.2	42.5
DRCT	99.5	98.5	98.5	97.8	98.0	99.1	97.0	98.3	97.9	98.4	98.4	**99.7**	99.1	98.7
D^3QE	96.9	97.5	83.1	78.8	96.2	95.2	78.8	89.5	95.0	96.8	97.8	98.2	98.0	97.2
LOTA	96.6	96.7	96.2	96.8	96.6	97.7	95.2	96.5	37.6	36.5	38.9	38.3	38.4	37.9
DDA	92.9	99.8	99.1	86.5	95.2	**99.9**	97.0	95.8	85.8	98.8	95.1	99.5	99.4	95.7
MF^2DA	98.7	98.7	83.6	84.7	97.7	97.5	84.4	92.2	97.7	99.4	98.5	97.3	97.7	98.1
DoU	**99.9**	**99.9**	**99.8**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**	85.3	92.4	83.2	94.4	89.2	88.9
IAPL	81.1	82.3	81.5	81.5	82.3	81.5	81.1	81.6	90.9	93.6	95.4	97.9	93.8	94.3
AReason*	82.6	74.4	73.8	64.1	71.5	73.9	60.7	71.6	78.9	75.2	84.3	78.9	81.3	79.7
Veritas*	**99.9**	91.0	70.5	58.5	62.6	98.5	57.0	76.9	**99.9**	99.0	99.0	99.0	**99.9**	**99.4**
Ours	99.8	99.6	93.9	97.1	99.5	99.4	92.6	97.4	98.9	99.3	**99.1**	99.4	99.5	99.2

Note. Inf.: Infinity; JP: Janus-Pro; LG: LlamaGen; OMV2: Open-MAGVIT2; OG: OmniGen;

Accuracy on Qwen-Image-Bench. Table 8 reports performance at the fixed decision threshold. RED achieves the second-highest average ACC of 93.2%, behind Veritas at 95.9%. Its ACC ranges from 88.9% to 95.8% across 18 generators and exceeds 90% on 17 of them. These results complement the AP evaluation in the main text, showing that RED also maintains strong detection performance at its fixed operating threshold.

Table 8: Per-generator accuracy (ACC, %) on Qwen-Image-Bench. Avg. denotes the mean across 18 generators.

Method	Qwen-Image-Bench (ACC.)																		
	F2-P	F2-M	GLM	G1	G1.5	HY3	I4	I4-U	QI	Q2-P	Q2512	S4	S4.5	S5	G2	K2.1	NB2	NB-P	Avg.
CNNSpot	54.9	50.5	61.9	58.4	54.9	53.2	58.6	57.1	62.1	51.3	56.5	53.6	53.5	51.1	53.5	57.3	48.9	49.6	54.8
UnivFD	89.4	87.6	91.8	89.2	86.4	92.9	90.7	90.4	91.2	90.9	90.2	89.4	84.5	90.1	89.0	90.0	87.3	83.9	89.1
DIRE	88.1	61.9	62.4	**99.9**	**99.8**	88.2	88.6	88.2	62.5	**99.9**	50.8	62.2	62.6	63.2	**99.9**	62.3	65.2	**99.9**	78.1
NPR	52.5	54.4	51.1	84.9	65.1	65.0	70.3	71.1	58.5	81.7	50.1	52.4	55.2	49.3	71.2	55.5	58.5	73.9	62.2
DRCT	51.6	51.0	84.7	68.5	56.2	90.9	81.3	83.1	62.0	60.4	57.6	53.4	51.9	52.4	55.1	62.3	52.6	56.1	62.8
D^3QE	88.6	85.1	88.6	89.5	85.9	92.3	88.3	88.8	89.5	90.3	89.3	88.3	84.8	87.3	87.8	89.4	86.4	84.7	88.0
LOTA	46.5	55.9	49.1	87.7	91.8	75.7	46.5	47.0	58.7	94.9	46.2	55.3	54.2	48.4	93.1	57.7	61.0	93.8	64.6
DDA	61.6	53.5	66.0	74.2	53.3	82.0	72.3	71.9	54.9	57.6	52.7	51.3	52.1	58.7	50.2	59.2	52.6	54.6	59.9
MF^2DA	88.9	86.4	90.0	90.2	86.4	91.9	88.6	89.0	90.1	90.1	89.7	88.5	82.7	89.5	88.2	89.7	88.2	84.4	88.5
DoU	50.7	55.5	56.8	78.3	85.0	86.9	57.3	56.8	61.4	73.8	50.3	57.1	56.1	49.5	79.9	61.1	54.0	73.1	63.5
IAPL	50.8	58.6	59.9	98.3	99.0	91.3	84.9	85.4	66.7	96.5	50.7	64.0	60.5	52.4	92.2	63.9	66.8	93.9	74.2
AReason*	86.5	81.5	90.0	85.5	71.0	88.5	90.0	85.5	89.0	84.0	82.0	74.0	79.5	84.5	76.0	91.5	75.0	73.5	82.6
Veritas*	**95.0**	**92.0**	**98.0**	99.0	95.5	**99.0**	**97.5**	**96.5**	**98.0**	96.5	**99.0**	92.5	**95.5**	**95.5**	88.5	**98.0**	**96.5**	94.5	**95.9**
Ours	92.7	88.9	92.4	92.5	90.2	95.8	93.6	94.0	95.5	95.2	94.7	**93.2**	91.8	92.5	94.6	94.0	93.5	92.0	93.2

Note. F2-P: FLUX.2-pro; F2-M: FLUX.2_max; GLM: GLM-Image; G1: GPT-Image-1; G1.5: GPT-Image-1.5; HY3: HunyuanImage-3.0; I4: Imagen-4.0; I4-U: Imagen-4.0-Ultra; QI: Qwen-Image; Q2-P: Qwen-Image-2.0-pro; Q2512: Qwen-Image-2512; S4: Seedream-4.0; S4.5: Seedream-4.5; S5: Seedream-5.0; G2: gpt-image-2; K2.1: kling_v2_1; NB2: nano-banana-2.0; NB-P: nano-banana-pro;

Table 9: Comparison across real-image sources with a fixed generated-image source. ACC and AP are reported in percentages.

Method	**ImageNet-val**		**Chameleon**		**LSUN**		**MS COCO**		**Avg.**	
	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.
UnivFD	92.7	98.2	60.2	75.3	88.2	92.2	82.7	88.6	80.95	88.58
D^3QE	86.9	92.2	62.9	67.6	88.3	94.1	87.3	94.2	81.35	87.03
MF^2DA	93.3	99.2	69.0	85.4	90.0	93.7	89.9	93.9	85.55	93.05
Ours	90.7	97.1	88.5	95.5	89.6	96.3	90.1	96.9	**89.73**	**96.45**

Generalization across real-image sources. Table 9 compares RED with three baselines across four real-image sources while keeping the Chameleon-generated images fixed. RED achieves the highest average ACC of 89.73% and AP of 96.45%, exceeding the strongest baseline, MF^2DA, by 4.18 and 3.40 percentage points, respectively. Its ACC ranges from 88.5% to 90.7% and AP from 95.5% to 97.1%, showing limited variation across sources. Replacing ImageNet-val real images with Chameleon real images reduces RED's ACC by 2.2 points, compared with drops of 24.0–32.5 points for the baselines. On the Chameleon real-image source, RED exceeds MF^2DA by 19.5 points in ACC and 10.1 points in AP. Although MF^2DA performs better on ImageNet-val and achieves higher ACC on LSUN, RED maintains stronger overall performance and greater consistency. These results support improved robustness to real-image source changes under this evaluation protocol.

C Scale-Wise NLL Analysis

Fig. 5 presents scale-wise profiles of standardized token-NLL differences relative to the ImageNet validation reference. For each image group, we compute the difference in mean standardized token NLL at each scale and subtract its average across scales. This row-centering highlights the shape of each profile: warm and cool colors indicate differences above and below the corresponding row mean, respectively.

White circles mark the scale with the largest signed difference in each profile, and the upper panel summarizes these peak locations across the 40 generated-image groups. Five, sixteen, and thirteen groups peak at the 1×1, 2×2, and 3×3 scales, respectively, while the remaining six peak at 16×16. These final-scale peaks correspond to VQDM, BigGAN, ADM, RAR, VAR, and Open-MAGVIT2. The differing peak locations and profile shapes demonstrate that scale-wise NLL variation depends on the generator. These peaks characterize likelihood variation and do not directly identify the most discriminative reconstruction stages.

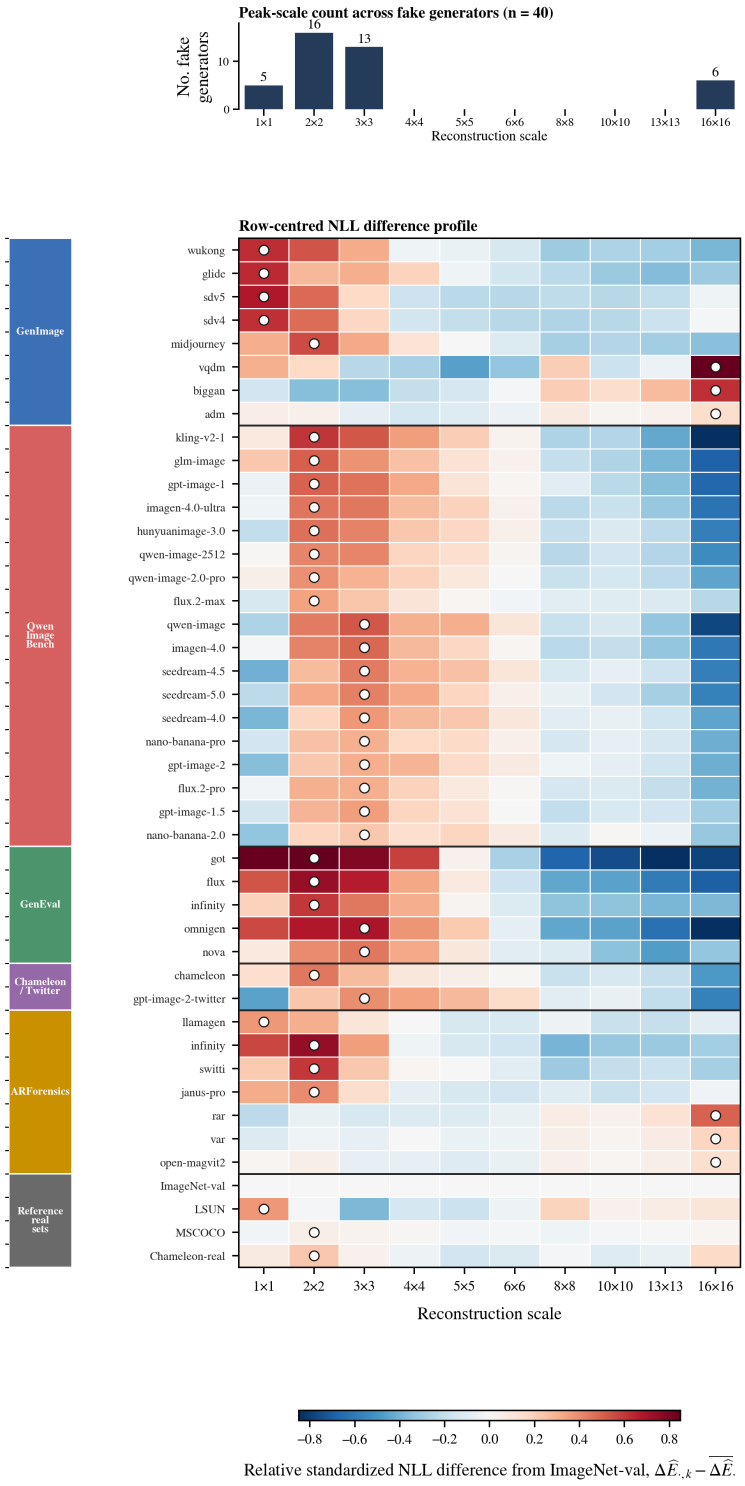


Figure 5: Scale-wise NLL difference profiles relative to ImageNet validation.

The additional real-image profiles illustrate variation across real-image sources, with the ImageNet validation profile equal to zero by construction. Together, these observations motivate examining likelihood information throughout reconstruction and learning how it should guide stage contributions, rather than relying on a single endpoint value or a fixed preferred scale.

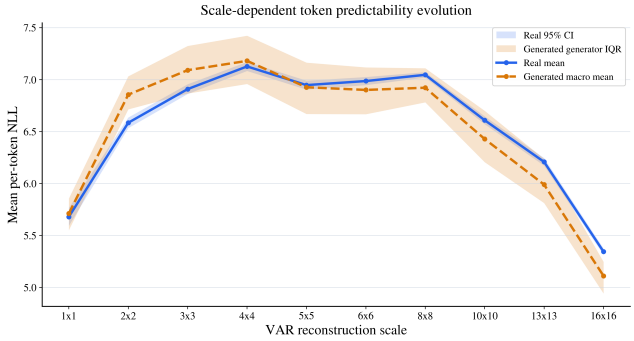


Figure 6: Scale-wise token-NLL profiles under the frozen VAR model.

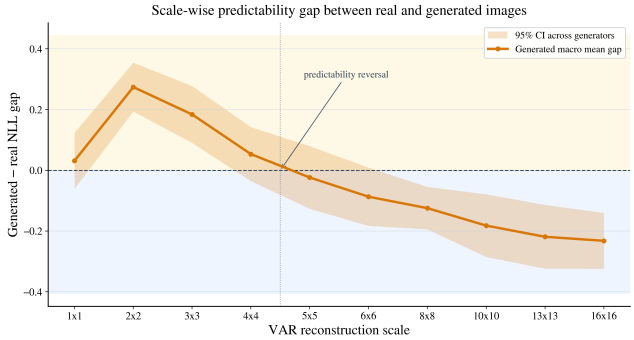


Figure 7: Scale-wise generated-minus-real token-NLL gaps.

Figures 6 and 7 examine token predictability across reconstruction scales under the frozen VAR model. We compare the real-image mean NLL with the macro-average of generator-group mean NLLs. The scale-wise gap is defined as $\Delta_k = \bar{E}_k^{\text{gen}} - \bar{E}_k^{\text{real}}$: positive values indicate lower conditional token predictability for generated images, while negative values indicate the reverse.

Scale-dependent predictability reversal. The macro-mean gap is positive at coarse scales, particularly 2×2 and 3×3, changes sign between 4×4 and 5×5, and becomes negative at finer scales. This reversal shows that the relationship between authenticity and token predictability depends on the reconstruction stage. A final-scale score therefore omits the opposite ordering observed at earlier stages. The generated-image interquartile range also reveals variation across generator groups, indicating that the aggregate trend does not imply identical behavior for every generator.

Connection to RED. The NLL profile and reconstruction trajectory arise from the same token hierarchy: E_k measures the predictability of newly introduced tokens conditioned on preceding scales, while x_k incorporates their contribution into the cumulative reconstruction. This correspondence provides stage-aligned predictability information for the visual representations. RED standardizes NLLs using source-training statistics and learns image-dependent weights through $\boldsymbol{\alpha}(x) = \mathrm{softmax}(a\widehat{\mathbf{E}}(x) + \mathbf{g})$. These weights modulate reconstruction features before cross-stage attention and pool the contextualized outputs, connecting evidence allocation with the final readout. The plotted population gaps motivate considering the full scale profile; they do not directly determine stage importance or prescribe a fixed weighting rule.

Complementary evidence from ablations. Table 4 connects these observations to detection performance. Adding intermediate states with uniform weights improves average ACC from 89.9% to 91.0%, supporting the value of intermediate visual evidence. Replacing learned static weights with image-dependent NLL guidance further improves ACC from 90.4% to 92.5%, with gains on all five out-of-domain sets. Together, these results support combining multistage visual evidence with adaptive predictability guidance, while the aggregate NLL profiles alone do not establish which stages are most discriminative for individual images.

D Qualitative Visualization of Reconstruction Trajectories

We visualize coarse-to-fine reconstruction trajectories for representative samples from all four real-image sources and six generated-image evaluation collections used in this work. For each sample, the frozen multiscale VQ-VAE used in RED produces ten successive reconstruction states by progressively accumulating and decoding quantized residual information. All states share a common output resolution, while successive stages incorporate additional information, progressing from coarse image structure toward finer visual detail. These examples illustrate reconstruction evolution across diverse image sources and provide an intuitive view of the intermediate states underlying RED's trajectory representation, complementing the quantitative scale-wise NLL analysis.

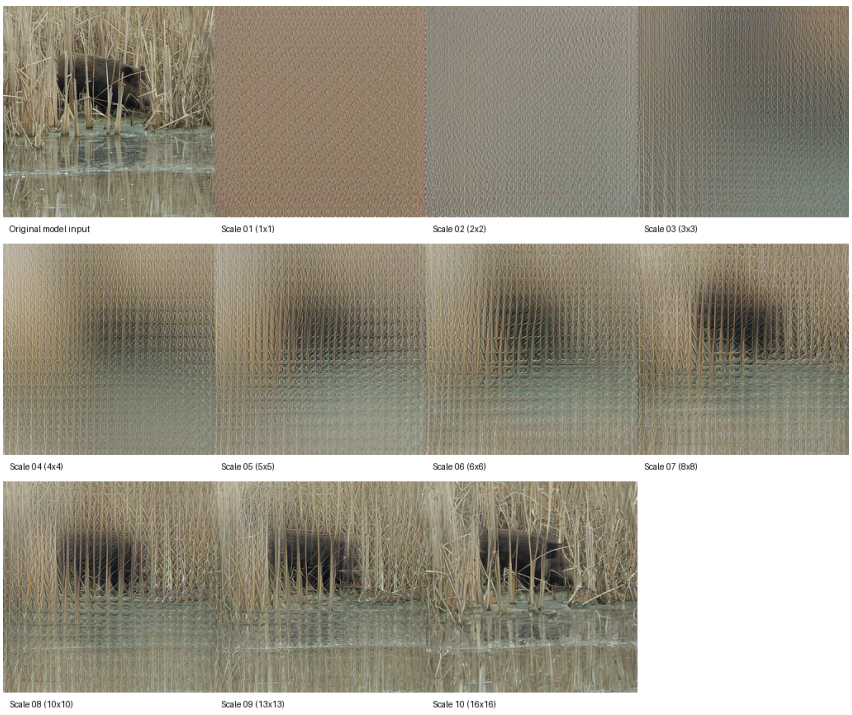


Figure 8: Coarse-to-fine reconstruction trajectory of a real image from ImageNet.

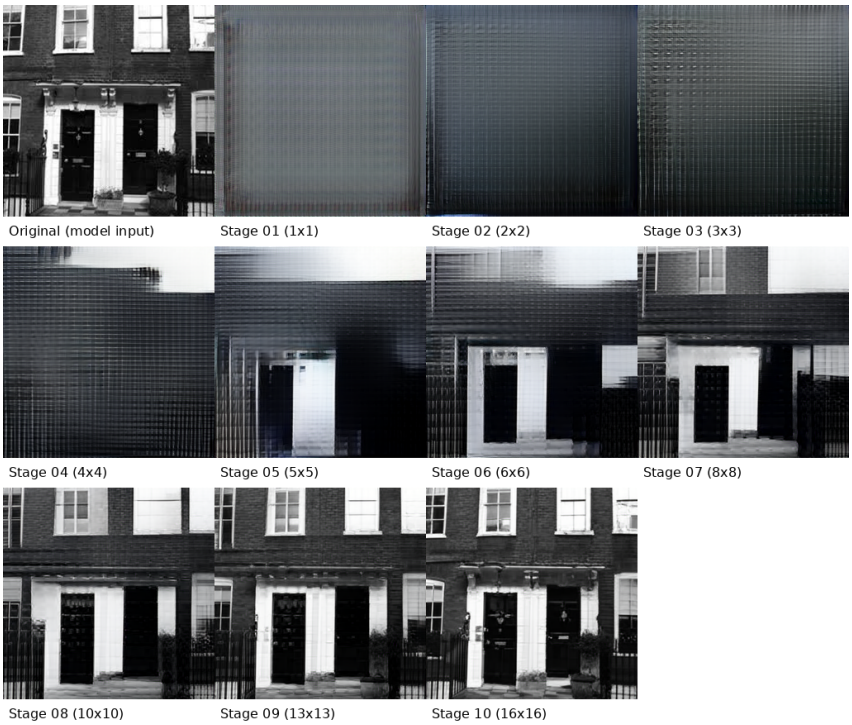


Figure 9: Coarse-to-fine reconstruction trajectory of a real image from LSUN.

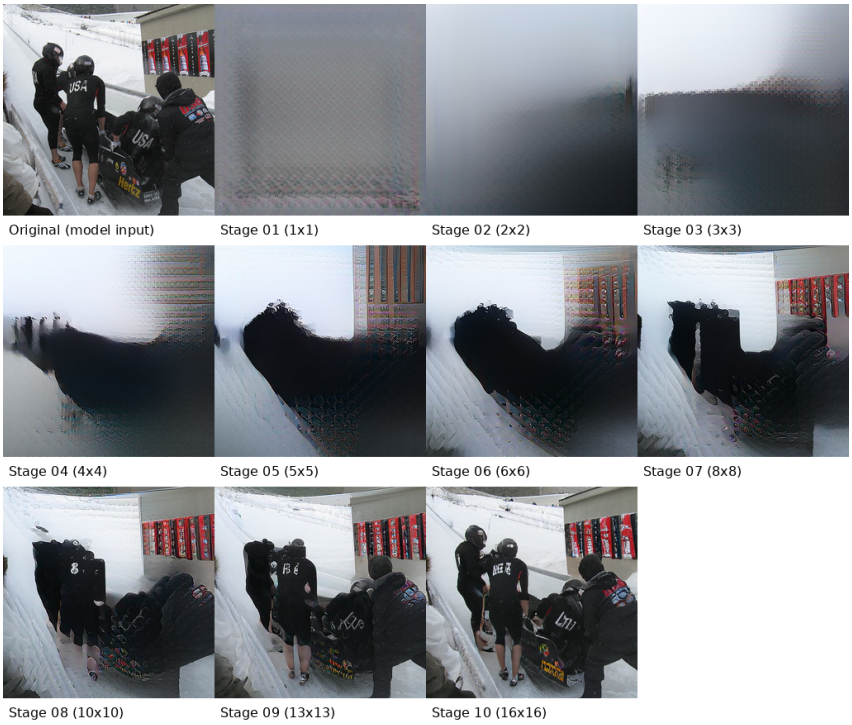


Figure 10: Coarse-to-fine reconstruction trajectory of a real image from MS COCO.



Figure 11: Coarse-to-fine reconstruction trajectory of a real image from Chameleon.

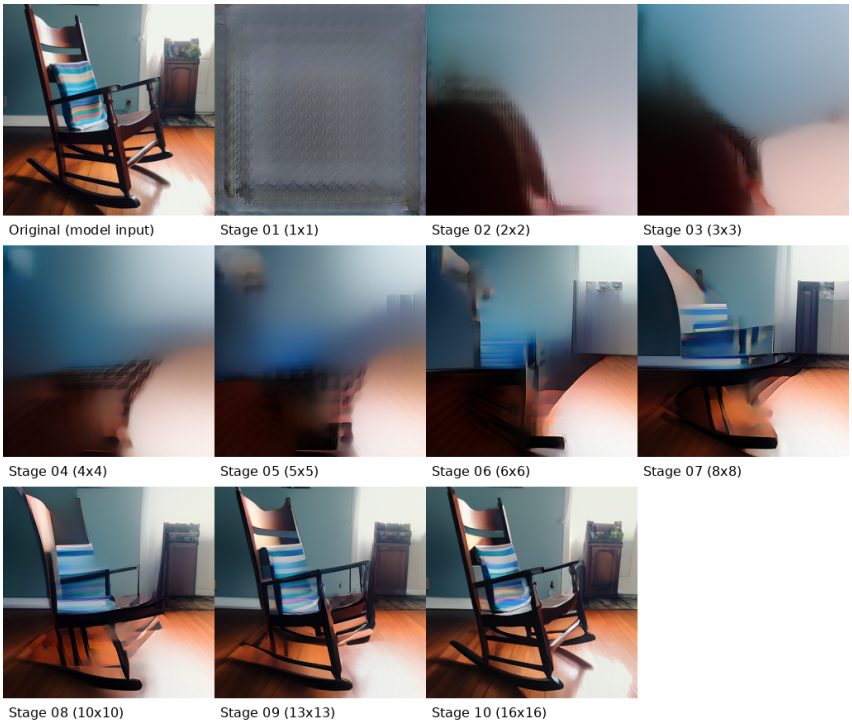


Figure 12: Coarse-to-fine reconstruction trajectory of an AI-generated image from GenImage.



Figure 13: Coarse-to-fine reconstruction trajectory of an AI-generated image from ARForensics.

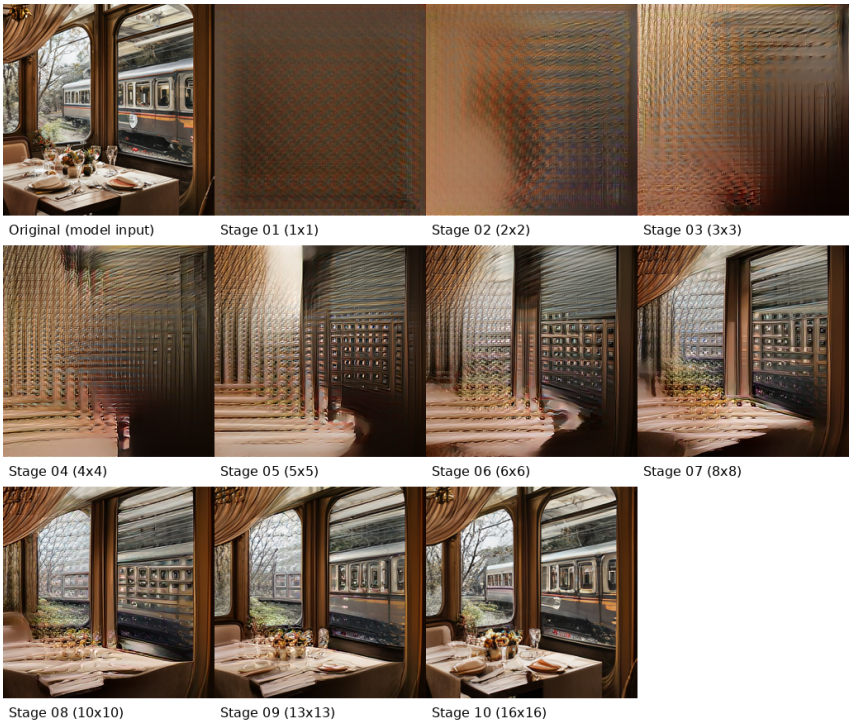


Figure 14: Coarse-to-fine reconstruction trajectory of an AI-generated image from EvalGEN.

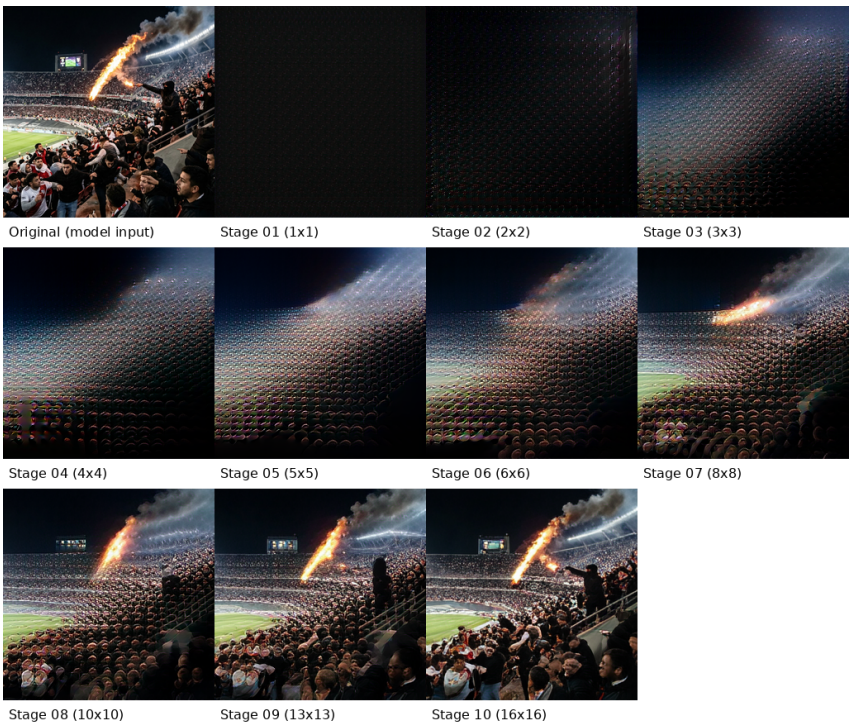


Figure 15: Coarse-to-fine reconstruction trajectory of an AI-generated image from Qwen-Image-Bench.



Figure 16: Coarse-to-fine reconstruction trajectory of an AI-generated image from Chameleon.

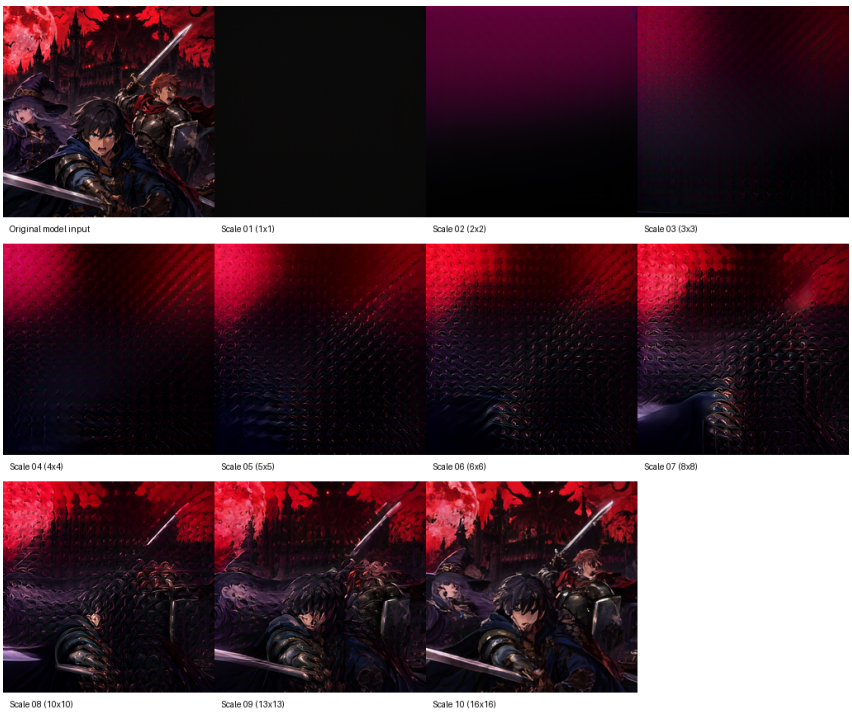


Figure 17: Coarse-to-fine reconstruction trajectory of an AI-generated image from GPT-Image-2-Twitter.