RESCOPE: LEARNING RECONSTRUCTION-CONSISTENT REPRESENTATIONS FOR GENERALIZABLE AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Detectors trained on a single generator often rely on generator-specific cues that transfer poorly to unseen generators. We use consecutive reconstruction to investigate whether label-preserving reconstruction variations can provide a training signal for improving cross-generator generalization. Since consecutive reconstructions share the same fake label, their feature discrepancy reflects the encoder's sensitivity to an additional reconstruction rather than a change in real/fake labels. We use this signal to reduce reconstruction sensitivity while preserving real/fake discrimination, encouraging the encoder to learn more stable forensic cues that transfer across generators. We propose ReSCOPE, a reconstruction-consistent adaptation framework that aligns encoder representations of consecutive fake–fake reconstruction pairs while retaining source-domain real/fake supervision. Training alternates between (i) joint optimization of source-domain classification and reconstruction consistency and (ii) consistency-only adaptation. In the adapted representations, distances between consecutive reconstructions decrease while classification margins for images from unseen generators increase. Reconstruction is used only for offline training-pair construction and is not required at inference. With single-generator training, ReSCOPE achieves 97.57% mean accuracy on GenImage and 88.93% overall accuracy on Chameleon. It also improves generalization on additional curated and in-the-wild benchmarks.

1 INTRODUCTION

AI-generated image detectors often achieve high accuracy on the training generator but exhibit a substantial generalization gap on unseen generators. Pretrained visual representations support cross-generator detection (Ojha et al., 2023), yet detectors remain vulnerable to content, compression, and generator-specific shortcuts (Grommelt et al., 2024; Zhou et al., 2025). Single-source supervision does not distinguish transferable detection cues from generator-specific cues that also predict the training labels (D'Amour et al., 2022). Drawing on prior distinctions between content, method-specific, and common forgery features (Yan et al., 2023), we adopt the following conceptual view of an encoded image:

$$Z(x) = C(x) + U_{\text{forensic}}(x) + S_{\text{src}}(x) + \varepsilon(x). \tag{1}$$

Here, C denotes content and semantics, U_{forensic} transferable forensic cues, S_{src} generator-specific cues, and ε residual variation and noise. In this conceptual decomposition, both U_{forensic} and S_{src} can support source-domain classification, so classification supervision alone does not ensure that the encoder favors transferable forensic cues over generator-specific shortcuts (D'Amour et al., 2022). A detector can therefore achieve strong source-domain performance by relying on generator-specific cues that transfer poorly to unseen generators (Zhou et al., 2025). We seek to reduce reliance on S_{src} while retaining real/fake discrimination.

We investigate this problem through consecutive reconstruction. Building on reconstruction-based image detection (Rajan et al., 2025; Qi et al., 2026), we apply a fixed reconstruction operator T twice to a real image A, obtaining $A_1 = T(A)$ and $A_2 = T(A_1)$. Unlike the original–reconstruction transition, the transition from A_1 to A_2 is between two fake images: the additional reconstruction

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.

changes the image but preserves its real/fake label. Their feature difference thus reflects the encoder's sensitivity to reconstruction rather than a change in class label. We define this difference as the second-pass reconstruction response, $R_{12} = F(A_2) - F(A_1)$, where F is the image encoder.

Our analysis shows that these responses concentrate in local low-dimensional subspaces in both CLIP and SigLIP2 (Section 3), but removing the entire response subspace reduces detection accuracy, and the responses are difficult to predict accurately. We therefore adapt the encoder directly using the observed reconstruction pairs to reduce the representation discrepancy between consecutive reconstructions. By reducing this label-preserving response, we aim to reduce the encoder's reliance on generator-specific shortcuts.

We propose ReSCOPE, a reconstruction-consistent adaptation framework that uses LoRA to align the normalized representations of A_1 and A_2 while retaining source-domain real/fake supervision (Fig. 3). Joint phases optimize classification and pairwise consistency together, whereas consistency-only phases keep the classification head fixed and focus encoder updates on reducing representation discrepancies between reconstruction pairs. The schedule alternates these focused consistency updates with joint classification and consistency training. In the adapted detectors, reconstruction-pair distances contract while classification margins increase for images from the unseen BigGAN generator (Fig. 2). Reconstruction is required only for offline training-pair construction. At inference, the adapted encoder and classification head process the input image directly, without reconstruction.

On GenImage, with SD1.4 as the sole training generator, ReSCOPE improves mean accuracy over frozen-encoder linear baselines by 15.81 percentage points for CLIP and 12.33 percentage points for SigLIP2, bringing SigLIP2's mean accuracy to 97.57%. The same detector also demonstrates strong cross-dataset generalization on seven additional benchmarks, achieving 88.93% overall accuracy on Chameleon. Our contributions are:

- **Reconstruction-response analysis and research data.** We identify the representation changes induced by label-preserving consecutive reconstructions and characterize their concentrated local low-dimensional structure. We construct approximately 320K aligned triplets of original images and their two consecutive reconstructions to support research on image reconstruction.
- **Reconstruction-consistent adaptation.** We introduce ReSCOPE, which learns transferable representations by constraining the encoder's response to label-preserving reconstruction variation. It combines pairwise representation consistency with real/fake classification supervision through alternating adaptation, encouraging the encoder to learn more transferable forensic cues and thereby improve cross-generator generalization. Inference requires no reconstruction.
- **Cross-generator and cross-dataset generalization.** We validate the effectiveness of ReSCOPE on both CLIP and SigLIP2 backbones and demonstrate generalization across multiple generation benchmarks and in-the-wild datasets using the same trained detector.

2 Related Work

Generalizable AI-generated image detection. Generalizable detection has been studied through both transferable visual representations and the design of training data. Universal fake-image detection uses pretrained CLIP features to transfer across generative models (Ojha et al., 2023), while Community Forensics expands training-generator diversity (Park & Owens, 2025). Other studies identify compression and size biases (Grommelt et al., 2024) and generator-specific latent-prior cues (Zhou et al., 2025) that can limit cross-generator performance. DDA addresses training-data alignment through JPEG quality matching and mixup (Chen et al., 2025). Our focus is single-source adaptation: using reconstruction-induced variation to guide representation learning without adding supervised examples from other generators.

Reconstruction-based detection and training. Reconstruction serves both as a detection signal and as a tool for building training data. DIRE uses diffusion reconstruction errors (Wang et al., 2023), FakeInversion combines inversion and reconstruction signals (Cazenavette et al., 2024), and AEROBLADE detects generated images through autoencoder reconstruction errors without detector

training (Ricker et al., 2024). DID also applies consecutive reconstructions, but uses differences of pixel residuals as detector inputs (Qi et al., 2026). For training, Aligned Datasets and B-Free construct content-aligned real–fake data through reconstruction (Rajan et al., 2025; Guillaro et al., 2025), while DRCT combines reconstruction-augmented data with same-label contrastive learning (Chen et al., 2024). In contrast, ReSCOPE focuses on a relation within reconstructed data: the consecutive fake–fake pair $(T(A), T^2(A))$ from the same real image. It directly minimizes the distance between their normalized features, without a contrastive negative term, and requires neither residual inputs nor reconstruction at inference.

Representation adaptation for generalization. Adaptation methods regulate how pretrained representations change during detector training. EFFORT constrains updates through orthogonal weight subspaces (Yan et al., 2025b), CausalCLIP studies feature filtering and disentanglement (Liu et al., 2026), and CLIP-ADA and SDD develop artifact and semantic-aware adaptation designs (Deng et al., 2026; Wang et al., 2025). The ARA preprint uses anchor-classifier constraints for reconstruction-based adaptation (Choi et al., 2026a). Consistency regularization can also guide learning through teacher-based targets (Tarvainen & Valpola, 2017). ReSCOPE places its consistency constraint directly on the normalized features of consecutive fake–fake reconstructions, rather than classifier outputs or teacher targets. Coupled with source-domain classification, this constraint trains the representation to remain stable under repeated reconstruction.

3 MOTIVATION: RECONSTRUCTION RESPONSE ANALYSIS

D'Amour et al. (2022) describe underspecification as a problem in which a learning pipeline produces predictors with similar in-distribution performance but different behavior under distribution shifts. In single-source image detection, both transferable forensic cues U_{forensic} and generator-specific shortcuts S_{src} can support accurate classification. Source-domain accuracy alone therefore cannot tell us which cues the detector relies on. We investigate how representation changes induced by consecutive reconstructions relate to S_{src}, while the fake label remains unchanged. We aim to reduce the encoder's sensitivity to these variations and its reliance on generator-specific shortcuts, encouraging it to use transferable forensic cues U_{forensic}.

3.1 LABEL-PRESERVING RECONSTRUCTION RESPONSES AND GENERATOR-SPECIFIC SHORTCUTS

For a real image A, we apply a fixed reconstruction operator T twice to obtain the first reconstruction A_1 and its subsequent reconstruction A_2:

$$A_1 = T(A), \qquad A_2 = T(A_1), \qquad R_{12} = F(A_2) - F(A_1), \tag{2}$$

where F is a frozen image encoder and R_{12} denotes the feature response to the second reconstruction. Both A_1 and A_2 are generated by the same reconstruction operator and share the fake label. The response R_{12} therefore captures the encoder's sensitivity to changes introduced by an additional application of this operator. These changes motivate a connection to generator-specific shortcuts: a detector may rely on features tied to a particular generation process even when those features vary within the fake class. We investigate whether reducing this reconstruction sensitivity can weaken the detector's reliance on S_{src} and encourage the use of transferable forensic cues U_{forensic}.

We analyze 160K consecutive-reconstruction triplets using features extracted by frozen CLIP ViT-L/14 (Radford et al., 2021) and SigLIP2-Giant (Tschannen et al., 2025). We examine the structure of R_{12} and test whether it can be predicted or removed to improve cross-generator generalization.

3.2 LOCAL STRUCTURE AND EXPLICIT RESPONSE REMOVAL

We first test whether similar images can help estimate the second-pass response. Using frozen encoder features, we retrieve K similar triplets from a reference subset of the data in Sec. 3.1 and estimate d local directions from their responses. We normalize each response to unit length as $r_{12} = R_{12}/\|R_{12}\|_2$ and measure coverage as the fraction of its energy captured by these directions, comparing local directions with global directions and random feature axes at the same dimension.

Fig. 1 shows clear local low-dimensional structure. For CLIP with $K = 512$, 8 local directions capture 1.79 times as much response energy as global directions. With 64 local directions, just 6.25% of the feature dimensions capture 39.47% of the response energy, compared with 28.10% for global directions. Local coverage exceeds global coverage in all 22 valid configurations across both backbones.

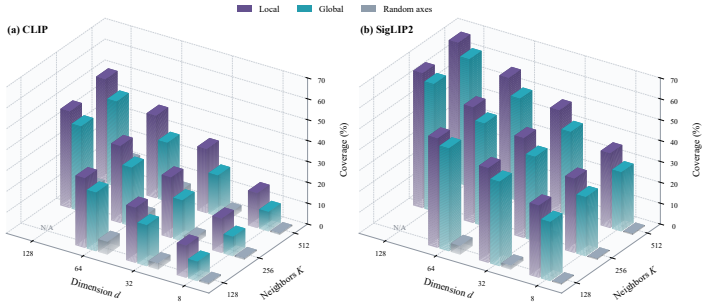

Figure 1: **Local structure of second-pass reconstruction responses.** Unit-response energy coverage (%) of local directions, global directions, and random feature axes for CLIP and SigLIP2. K denotes the number of reference neighbors and d the subspace dimension. N/A indicates exclusion by the rank limit. Detailed results are provided in Appendix A.

However, this structure does not directly yield accurate response predictions: both linear and MLP predictors fall well below the directional alignment attainable within the local subspace. Projecting out the response subspace also fails to produce consistent generalization gains, with full removal reducing detection performance (see Appendices A and A.3). We therefore train the encoder directly on observed fake–fake pairs to reduce reconstruction-induced discrepancies, rather than estimating and removing the response.

3.3 Reconstruction Consistency and Cross-Generator Discrimination

Reconstruction consistency encourages the encoder to produce similar normalized representations for the first and second reconstructions, A_1 and A_2. Fig. 2 connects this approach to the discrimination learned under single-source supervision. In Fig. 2(a), gray, orange, green, and blue denote real images, first reconstructions A_1, subsequent reconstructions A_2, and unseen BigGAN images, respectively. The Fake arrow illustrates the learned discrimination direction, with a horizontal source-specific component S_{src} and a vertical transferable component U_{forensic}. Because both components support source-domain classification, the learned direction may rely on both. The source-biased boundary perpendicular to this direction passes through the BigGAN cluster, causing some unseen fake images to be classified as real.

The red R_{12} arrow connects A_1 and A_2, showing the representation change induced by another reconstruction within the fake class. The schematic places this response along the axis of source-related variation. We test whether suppressing such reconstruction variation while retaining real/fake discrimination encourages the encoder to use more transferable forensic information.

ReSCOPE combines representation consistency between consecutive fake reconstructions with source-domain real/fake supervision. Fig. 2(b) illustrates the training objective: bring A_1 and A_2 closer while maintaining separation between real and fake images, encouraging a decision boundary that depends less on horizontal source-related variation. Panels (c–f) in Fig. 2 examine whether this behavior appears in the trained detectors.

We compare a CLIP detector trained with real/fake classification supervision alone with ReSCOPE, using each model's encoder features and classification head. In Fig. 2(c–d), the horizontal axis measures position along a fixed reconstruction-sensitive direction, while the vertical axis comes from the actual classification head. The zero line therefore corresponds directly to the detector's decision boundary. After ReSCOPE adaptation, the distributions of A_1 and A_2 move closer together, while more BigGAN images fall on the fake side of the boundary.

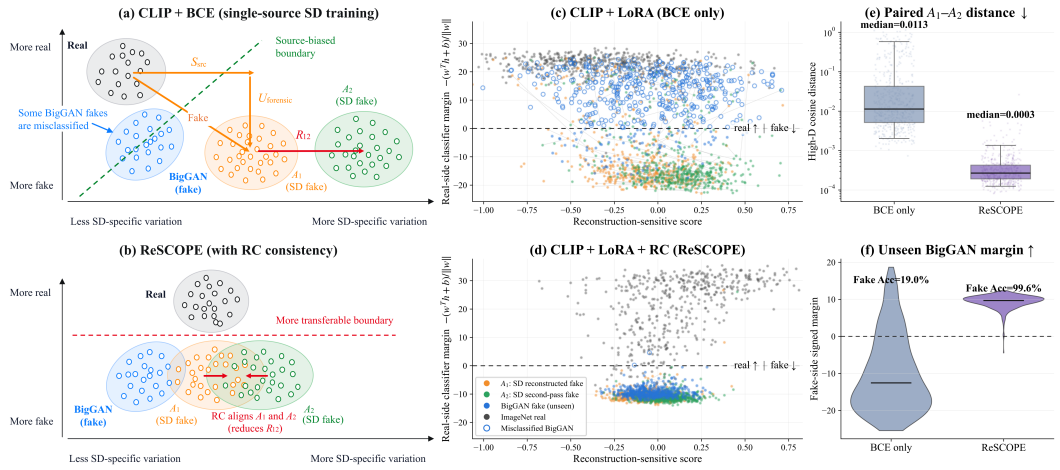

Figure 2: **Reconstruction consistency and unseen-generator discrimination.** (a–b) Conceptual illustrations of single-source discrimination and reconstruction-consistency adaptation. (c–d) Features from the classification-only and ReSCOPE detectors; zero marks each classifier's actual decision boundary. Hollow blue points denote misclassified BigGAN images. (e–f) Reconstruction-pair distances contract while BigGAN fake-side margins improve. Projection and sampling details are provided in Appendix A.4.

Panels (e–f) in Fig. 2 quantify this behavior in the original high-dimensional feature space. On the samples used for this analysis, the median cosine distance between reconstruction pairs decreases from 0.0113 to 0.000269, approximately a 42-fold reduction. Meanwhile, the median fake-side margin for BigGAN images increases from -12.56 to $+9.68$, and fake accuracy rises from 19.0% to 99.6%, with real-image accuracy at 95.0%.

4 METHODOLOGY: ReSCOPE

ReSCOPE consists of three components: pair construction, reconstruction-consistency adaptation, and source-domain classification. As shown in Fig. 3, consecutive reconstruction pairs are constructed offline and provide a consistency constraint on the shared encoder, while labeled source images provide real/fake supervision. Alternating adaptation combines these two training signals. We denote the encoder's pooled feature by $h_\theta(x)$ and the linear classification head by g_ϕ. We train only rank-r LoRA modules inserted into the attention projections (Hu et al., 2022), with parameters θ, together with the head parameters ϕ; pretrained backbone weights remain frozen.

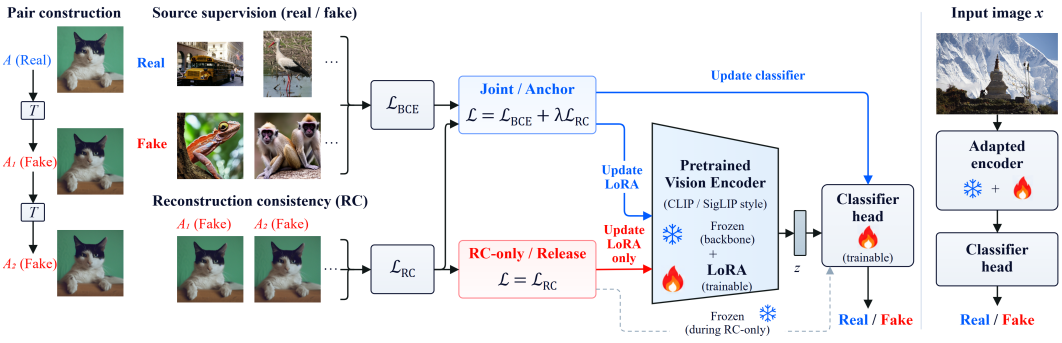

Figure 3: Overview of ReSCOPE. Consecutive reconstructions form the fake–fake pair used by RC. Joint stages update LoRA adapters and the classification head using BCE+RC; RC-only stages update only the adapters. Original backbone weights remain frozen throughout. At inference, only the input image, encoder, and head are used.

4.1 Reconstruction Operator and Pair Construction

Inspired by the caption-conditioned DDIM inversion used in FakeInversion (Cazenavette et al., 2024), we implement a reconstruction operator using frozen Stable Diffusion v1.4 (Rombach et al., 2022) and apply it twice consecutively. We use BLIP-2 (Li et al., 2023) to generate one caption c_A for each preprocessed real image A. The same caption is reused in both reconstruction passes, and the frozen SD1.4 text encoder maps it to the conditioning embedding e_A used during both inversion and denoising. We define the reconstruction operator and consecutive reconstructions as

$$T(x; c_A) = Q_8\left[D_{\mathrm{VAE}}\left(\frac{\mathcal{D}_{e_A}(\mathcal{I}_{e_A}(sE_{\mathrm{VAE}}(x)))}{s}\right)\right], \quad A_1 = T(A; c_A), A_2 = T(A_1; c_A). \tag{3}$$

where E_{VAE} uses the posterior mode, D_{VAE} is the decoder, and s is the latent scaling factor. $\mathcal{I}_{e_A}$ and $\mathcal{D}_{e_A}$ denote deterministic DDIM inversion and denoising (Song et al., 2021) under the same text conditioning, traversing the same timestep grid in opposite directions. Q_8 clips and quantizes the decoded image to 8-bit RGB. We save A_1 as a PNG and reload it for the second pass, reusing the same caption. The pairs $\mathcal{D}_{\mathrm{RC}} = \{(A_{1,i}, A_{2,i})\}_i$ are generated offline with the frozen reconstruction model. During training, both images in each pair share the adapted encoder. Detailed reconstruction settings are provided in Appendix B.1.

4.2 Reconstruction-consistent learning

For each nonzero pooled feature, we define its unit-normalized representation as $z_\theta(x) = h_\theta(x)/\|h_\theta(x)\|_2$. The reconstruction-consistency objective is

$$\mathcal{L}_{\mathrm{RC}}(\theta) = \mathbb{E}_{(A_1,A_2)\sim\mathcal{D}_{\mathrm{RC}}}\left[1 - z_\theta(A_1)^\top z_\theta(A_2)\right]. \tag{4}$$

The pair (A_1, A_2) undergoes another reconstruction without changing its fake label, providing a label-consistent target for representation alignment. RC treats the corresponding feature discrepancy as variation to attenuate: the encoder is encouraged to represent the two fake views consistently despite the additional reconstruction. The loss aligns feature directions without requiring equal feature magnitudes and bypasses the classification head; both views contribute gradients through the shared encoder. This objective connects to the response analysis in Section 3 through the normalized response $\widetilde{R}_{12,\theta} = z_\theta(A_2) - z_\theta(A_1)$. For each pair, $1 - z_\theta(A_1)^\top z_\theta(A_2) = \frac{1}{2}\|\widetilde{R}_{12,\theta}\|_2^2$. Thus, minimizing RC directly penalizes the second-pass response in the normalized representation space. Fig. 2(e) measures this same cosine discrepancy in the original detector features.

To retain real/fake discrimination during reconstruction-consistency learning, we apply binary cross-entropy to labeled source images:

$$\mathcal{L}_{\mathrm{BCE}}(\theta, \phi) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{src}}}\left[\mathrm{BCE}(\sigma(s_{\theta,\phi}(x)), y)\right]. \tag{5}$$

Here, $\mathcal{D}_{\mathrm{src}}$ contains real images ($y = 0$) and fake images from a single training generator ($y = 1$), and $s_{\theta,\phi}(x) = g_\phi(h_\theta(x))$ is the fake-image logit. The classifier uses unnormalized encoder features; unit normalization is specific to RC.

4.3 Alternating adaptation

We initialize the detector with source-domain classification, training the LoRA parameters and classification head with BCE. We then alternate between the two update paths in Fig. 3: the blue Joint path updates both the LoRA adapters and classification head, whereas the red RC-only path updates only the adapters, keeping the head fixed. The corresponding objectives are

$$\text{Joint:} \quad \min_{\theta,\phi} \ \mathcal{L}_{\mathrm{BCE}}(\theta, \phi) + \lambda\mathcal{L}_{\mathrm{RC}}(\theta), \tag{6}$$

$$\text{RC-only:} \quad \min_{\theta} \ \lambda\mathcal{L}_{\mathrm{RC}}(\theta), \qquad \phi \text{ fixed}. \tag{7}$$

Here, $\lambda > 0$ controls the consistency weight relative to classification in the Joint objective.

The **Joint phase** anchors representation adaptation to the source-domain real/fake task while introducing reconstruction consistency. BCE maintains real/fake discrimination, while RC reduces reconstruction-induced variation in the normalized representations. Both losses update the LoRA

parameters, and BCE also updates the head, so consistency is learned under continued source supervision. The **RC-only phase** temporarily omits the classification loss and concentrates encoder updates on the discrepancy between A_1 and A_2. Only the LoRA parameters are updated; the head is fixed and bypassed by the loss. Through the normalized-response identity above, these updates directly target sensitivity to repeated reconstruction.

During RC-only adaptation, the head remains fixed but its input representations change without direct classification supervision. Returning to Joint periodically reapplies source-domain classification alongside RC. This schedule combines focused reductions in reconstruction sensitivity with updates that maintain real/fake discrimination.

At inference, an image is classified as fake when $s_{\theta,\phi}(x) \geq 0$, equivalently when its sigmoid probability is at least 0.5. Inference requires only one forward pass through the adapted encoder and classifier, with no reconstruction.

5 Experiments

Implementation details. We use CLIP ViT-L/14 and SigLIP2-Giant as backbones, fine-tuning them with rank-4 LoRA and training a linear classification head. Supervised training uses real and fake images from the SD1.4 training split of GenImage. For reconstruction consistency, we use 10K offline-generated pairs of consecutive reconstructions. The input resolution is 224×224 for CLIP and 384×384 for SigLIP2. We apply random horizontal flips during training, using the same transformation for both images in each pair. All cross-dataset evaluations use the same GenImage-trained SigLIP2 detector with a fixed zero-logit threshold and no target-dataset adaptation. Inference requires no reconstruction. Full training configurations are provided in Appendix B.

5.1 Comparison with Other Methods

Cross-generator comparison on GenImage. Table 1 compares ReSCOPE with other methods on GenImage (Zhu et al., 2023b). ReSCOPE achieves the best performance among the compared methods, with an mAcc of 97.57%. Its accuracy exceeds 95% on all seven unseen generators, including 96.15% on Midjourney and 95.97% on ADM. These results demonstrate strong cross-generator generalization from a single training generator.

Cross-dataset comparison on generation benchmarks. Tables 2, 4–6 report results on DiffusionForensics (Wang et al., 2023), AIGCDetectionBenchmark (Zhong et al., 2024), Synthbuster (Bammey, 2024), EvalGEN (Chen et al., 2025), and DRCT-2M (Chen et al., 2024). We evaluate the same GenImage-trained detector without target-domain adaptation and observe the following: **(1) Strong cross-dataset performance.** ReSCOPE ranks first or second among the compared methods on each benchmark. It achieves 98.15% mAcc on DiffusionForensics, with the best accuracy on six of eight generators, and ranks second on both DRCT-2M and AIGCDetectionBenchmark. **(2) Generalization across diverse generators.** These benchmarks cover diffusion models, GANs, and autoregressive models. ReSCOPE achieves 89.01% mAcc across the 17 generators in AIGCDetectionBenchmark and 99.42% mean fake-image accuracy on EvalGEN, with accuracy above 98% on all five EvalGEN generators. These results demonstrate that reconstruction-consistent adaptation supports transfer across datasets and generation models.

Comparison on in-the-wild datasets. We further evaluate ReSCOPE on Chameleon (Yan et al., 2025a) and the available B-Free Online subset (Guillaro et al., 2025). On Chameleon, ReSCOPE achieves the highest overall accuracy of 88.93% among the compared methods, with 79.27% accuracy on fake images and 96.19% on real images (Table 3). On B-Free Online, the same detector achieves 89.02% balanced accuracy (Table 5). These results demonstrate generalization from a single training generator to in-the-wild images.

5.2 Ablation and Robustness Studies

Ablation studies. Table 7 compares four training strategies. ReSCOPE achieves the highest mAcc of 93.37% with CLIP and 97.57% with SigLIP2. The gains over BCE-only support learning from label-preserving reconstruction pairs. Joint-only achieves 86.51% with CLIP and 92.77% with SigLIP2. Alternating adaptation improves on both results, showing the benefit of alternat-

ing joint optimization with consistency-only updates. **Robustness.** Table 8 reports robustness to JPEG compression, Gaussian blur, and resizing on GenImage. Experimental details are provided in Appendix F. Together, these results demonstrate the effectiveness of ReSCOPE and its robustness to JPEG compression and resizing.

Table 1: **Cross-model accuracy (Acc) on GenImage.** ReSCOPE is trained on the SD1.4 subset of GenImage. Baseline results are taken from the cited papers. mAcc denotes the unweighted mean across generators; MJ denotes Midjourney.

Method	MJ	SD1.4	SD1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	mAcc
CNNSpot (Wang et al., 2020)	52.80	96.30	95.90	50.10	39.80	78.60	53.40	46.80	64.21
DIRE (Wang et al., 2023)	60.20	**99.90**	**99.80**	50.90	55.00	**99.20**	50.10	50.20	70.66
UnivFD (Ojha et al., 2023)	73.20	84.20	84.00	55.20	76.90	75.60	56.90	80.30	73.29
NPR (Tan et al., 2024a)	81.00	98.20	97.90	76.90	89.80	96.90	84.10	84.20	88.60
DRCT/UnivFD (Chen et al., 2024)	91.50	95.01	94.41	79.42	89.18	94.67	90.03	81.67	89.49
AIDE (Yan et al., 2025a)	79.38	99.74	99.76	78.54	91.82	98.65	80.26	66.89	86.88
EFFORT (Yan et al., 2025b)	82.40	99.80	**99.80**	78.70	93.30	97.40	91.70	77.60	91.10
OMAT ($r = 4$) (Zhou et al., 2025)	90.36	97.52	97.46	83.82	97.41	97.62	95.53	97.34	94.63
SAFE (Li et al., 2025)	95.30	99.40	99.30	82.10	96.30	98.20	96.30	97.80	95.60
C2P-CLIP (Tan et al., 2025)	88.20	90.90	97.90	**96.40**	99.00	98.80	96.50	**98.70**	95.80
PPL (Yang et al., 2026b)	94.80	98.50	98.30	94.70	96.10	98.60	**98.50**	98.00	97.20
HiDA-Net (SD1.4) (Mu et al., 2026)	94.20	99.10	99.20	92.70	**99.10**	98.00	96.90	97.80	97.10
ReSCOPE (CLIP)	92.65	97.40	97.34	77.44	97.17	97.22	90.65	97.10	93.37
ReSCOPE (SigLIP2)	**96.15**	98.18	98.21	95.97	98.05	98.38	97.80	97.84	**97.57**

Table 2: **Cross-dataset and cross-model accuracy (Acc) on DiffusionForensics.** All methods are trained on GenImage SD1.4. Baseline results are taken from RippleNet (Yang et al., 2026a).

Method	Train	ADM	DDPM	iDDPM	PNDM	LDM	VQDM	SD-v1	SD-v2	mAcc
F3Net (Qian et al., 2020)	SD1.4	55.00	55.00	46.00	46.40	45.00	44.30	81.10	65.00	54.70
FreqNet (Tan et al., 2024b)	SD1.4	58.40	64.10	50.00	49.80	88.30	88.90	98.70	97.80	74.50
FatFormer (Liu et al., 2024)	SD1.4	74.00	81.30	66.30	71.60	83.30	93.30	99.30	91.90	82.60
AIDE (Yan et al., 2025a)	SD1.4	60.80	67.20	56.80	67.80	84.50	86.50	**99.80**	94.30	77.20
VIBNet (Zhang et al., 2025)	SD1.4	70.40	95.50	68.40	77.00	65.50	95.80	98.60	93.00	83.00
EFFORT (Yan et al., 2025b)	SD1.4	85.90	87.80	78.50	87.30	92.10	86.10	97.80	86.20	87.70
NPR (Tan et al., 2024a)	SD1.4	60.80	81.00	73.60	70.00	**100.00**	97.70	99.40	88.50	83.90
FerretNet (Liang et al., 2025)	SD1.4	71.90	82.50	70.00	75.80	**100.00**	85.30	99.30	86.90	84.00
RippleNet (Yang et al., 2026a)	SD1.4	84.70	88.70	77.00	87.00	93.10	90.00	98.80	92.70	89.00
ReSCOPE (SigLIP2)	SD1.4	**94.75**	**97.34**	**98.00**	**98.50**	99.15	**99.15**	99.15	**99.15**	**98.15**

Table 3: **Cross-dataset accuracy (Acc) on Chameleon.** All detectors are trained on SD1.4. Baseline results are taken from AIDE (Yan et al., 2025a). The first row reports overall accuracy, and the second reports accuracy on fake and real images separately (Fake / Real).

Metric	CNNSpot (2020)	FreDect (2020)	Fusing (2022)	UnivFD (2023)	DIRE (2023)	PatchCraft (2024)	NPR (2024a)	AIDE (2025a)	ReSCOPE (SigLIP2)
Overall Acc	60.11	56.86	57.07	55.62	59.71	56.32	58.13	62.60	**88.93**
Fake / Real	8.86/98.63	1.37/98.57	0.00/99.96	74.97/41.09	11.86/95.67	3.07/96.35	2.43/100.00	20.33/94.38	79.27/96.19

Table 4: **Cross-dataset and cross-model accuracy (Acc) on AIGCDetectionBenchmark.** Baseline results are taken from Table 2 of PatchCraft (Zhong et al., 2024).

Method	ADM	DALLE2	GLIDE	MJ	VQDM	BigGAN	CycleGAN	GauGAN	ProGAN	SDXL	SD1.4	SD1.5	StarGAN	StyleGAN	StyleGAN2	WFR	Wukong	mAcc
CNNSpot (Wang et al., 2020)	60.39	50.45	58.07	51.39	56.46	71.17	87.62	81.42	**100.00**	53.03	50.57	50.53	94.60	90.17	86.91	**91.65**	51.03	69.73
FreDect (Frank et al., 2020)	63.42	34.70	54.13	45.87	77.80	81.97	78.77	80.57	99.36	51.23	38.79	39.21	94.62	78.02	66.19	50.75	40.30	63.28
Fusing (Ju et al., 2022)	49.00	52.80	57.20	52.20	55.10	77.40	87.00	77.00	**100.00**	55.60	51.00	51.40	97.00	85.20	83.30	66.80	51.70	67.63
GramNet (Liu et al., 2020)	58.61	49.25	54.50	50.02	52.86	67.33	86.07	69.35	99.99	64.53	51.70	52.16	95.05	87.05	87.28	86.80	50.76	68.43
LNP (Liu et al., 2022a)	83.91	92.45	83.50	69.55	85.03	88.43	79.07	79.17	99.95	87.75	89.33	88.81	**100.00**	92.64	**93.82**	50.00	86.39	85.28
LGrad (Tan et al., 2023)	67.15	65.45	66.11	65.35	72.99	85.62	86.94	78.46	99.83	71.30	63.02	63.67	99.27	91.08	85.32	55.70	59.55	75.11
DIRE-G (Wang et al., 2023)	75.78	66.48	71.75	58.01	53.68	70.12	74.19	67.79	95.19	55.35	49.74	49.83	95.47	83.03	75.31	58.05	54.46	67.90
DIRE-D (Wang et al., 2023)	**98.25**	92.45	92.42	89.45	91.90	49.70	49.58	51.23	52.75	91.28	91.24	91.63	46.72	51.31	51.72	53.30	90.90	72.70
UnivFD (Ojha et al., 2023)	66.87	50.75	62.46	56.13	85.31	95.08	**98.33**	**99.47**	99.81	50.73	63.66	63.49	95.75	84.93	74.96	86.90	70.93	76.80
PatchCraft (Zhong et al., 2024)	82.17	**96.60**	83.79	90.12	88.91	**95.80**	70.17	71.58	**100.00**	98.43	95.38	95.30	99.97	**92.77**	89.55	85.80	91.07	**89.85**
ReSCOPE (SigLIP2)	91.38	92.90	**98.56**	**97.95**	**98.20**	82.30	81.79	89.17	81.70	**98.80**	**99.56**	**99.64**	86.67	82.19	79.24	53.50	**99.60**	89.01

Table 5: **Cross-dataset accuracy on Synthbuster, EvalGEN, and B-Free Online.** Baseline results are taken from DDA (Chen et al., 2025). We report balanced accuracy (%) on Synthbuster and B-Free Online, and fake-image accuracy (%) on EvalGEN. ‡ReSCOPE is evaluated on the available B-Free Online subset. †The Synthbuster mean for DRCT is recomputed from its nine reported generator-wise results.

Method	Synthbuster										EvalGEN						B-Free Online
	DALL-E 2	DALL-E 3	Firefly	GLIDE	MJ	SD1.3	SD1.4	SD2	SDXL	Mean	Flux	GoT	Infinity	NOVA	OmniGen	Mean	BAcc
NPR (Tan et al., 2024a)	51.10	49.30	46.50	48.50	52.80	51.40	51.80	46.00	52.80	50.00	0.70	0.20	6.50	4.70	2.20	2.90	49.50
UnivFD (Ojha et al., 2023)	83.50	47.40	89.90	53.30	52.50	70.40	69.90	75.70	68.00	67.80	4.00	9.20	15.70	8.30	39.60	15.40	49.00
FatFormer (Liu et al., 2024)	59.40	39.50	60.30	72.70	44.40	53.70	54.00	52.30	69.10	56.10	9.90	47.90	44.70	98.30	27.30	45.60	50.00
SAFE (Li et al., 2025)	58.00	9.90	10.30	52.20	56.70	59.40	59.10	53.00	59.50	46.50	1.00	0.50	1.90	0.60	1.60	1.10	50.50
C2P-CLIP (Tan et al., 2025)	55.60	63.20	59.50	86.70	52.90	75.20	76.70	69.20	77.70	68.50	8.70	49.60	35.30	86.40	14.50	38.90	50.00
AIDE (Yan et al., 2025a)	34.90	33.70	24.80	65.00	57.50	74.10	73.70	53.20	68.40	53.90	17.90	24.70	3.40	16.30	33.40	19.10	53.10
DRCT (Chen et al., 2024)	77.20	86.60	84.10	82.60	73.70	86.60	86.60	83.20	71.30	81.32†	72.50	81.40	77.90	84.60	72.50	77.80	55.70
AlignedForensics (Rajan et al., 2025)	50.20	48.90	51.70	53.50	**98.70**	**98.80**	**98.80**	**98.60**	**97.30**	77.40	32.00	72.30	74.00	84.80	77.00	68.00	68.50
DDA (Chen et al., 2025)	86.30	**90.00**	**91.90**	76.50	93.50	92.90	92.70	93.30	93.50	**90.10**	89.90	99.50	97.80	99.50	99.50	97.20	**95.10**
ReSCOPE (SigLIP2)	**88.90**	88.90	76.95	**88.95**	81.05	89.20	89.20	88.95	87.90	86.67	**98.55**	**99.82**	**98.91**	**99.82**	**100.00**	**99.42**	89.02‡

Table 6: **Cross-model accuracy (Acc) on DRCT-2M.** Baseline results are taken from DRCT (Chen et al., 2024). Baselines are trained on DRCT-2M, while ReSCOPE is trained on GenImage; all use SD1.4 as the training generator. DIRE and DRCT use the SDv1 reconstruction model.

Method	SD variants						Turbo variants		LCM variants		ControlNet variants			DR variants			mAcc
	LDM	SD1.4	SD1.5	SD2	SDXL	SDXL-Refiner	SD	SDXL	SD1.5	SDXL	SD1	SD2	SDXL	SD1	SD2	SDXL	
CNNSpot (Wang et al., 2020)	99.87	99.91	99.90	**97.55**	66.25	86.55	86.15	72.42	98.26	61.72	97.96	85.89	82.84	60.93	51.41	50.28	81.12
F3Net (Qian et al., 2020)	99.85	99.78	99.79	88.66	55.85	87.37	68.29	63.66	97.39	54.98	97.98	72.39	81.99	65.42	50.39	50.27	77.13
CLIP/RN50 (Radford et al., 2021)	99.00	99.99	99.96	94.61	62.08	91.43	83.57	64.40	98.97	57.43	99.74	80.69	82.03	65.83	50.67	50.47	80.05
GramNet (Liu et al., 2020)	99.40	99.01	98.84	95.30	62.63	80.68	71.19	69.32	93.05	57.02	89.97	75.55	82.68	51.23	50.01	50.08	76.62
De-fake (Sha et al., 2023)	92.10	99.53	99.51	89.65	64.02	69.24	92.00	93.93	99.13	70.89	58.98	62.34	66.66	50.12	50.16	50.00	75.52
Conv-B (Liu et al., 2022b)	**99.97**	**100.00**	**99.97**	95.84	64.44	82.00	80.82	60.75	99.27	62.33	99.80	83.40	73.28	61.65	51.79	50.41	79.11
UnivFD (Ojha et al., 2023)	98.30	96.22	96.33	93.83	91.01	93.91	86.38	85.92	90.44	88.99	90.41	81.06	89.06	51.96	51.03	50.46	83.46
DIRE (Wang et al., 2023)	98.19	99.94	99.96	68.16	53.84	71.93	58.87	54.35	**99.78**	59.73	99.65	64.20	59.13	51.99	50.04	49.97	71.23
DRCT/Conv-B (Chen et al., 2024)	99.91	99.90	99.90	96.32	83.87	85.63	91.88	70.04	99.66	78.76	**99.90**	**95.01**	81.21	**99.90**	**95.40**	**75.39**	**90.79**
DRCT/UnivFD (Chen et al., 2024)	96.74	96.26	96.33	94.89	96.24	93.46	93.43	92.94	91.17	95.01	95.60	92.68	**91.95**	94.10	69.55	57.43	90.49
ReSCOPE (SigLIP2)	97.27	97.00	96.97	97.27	**97.23**	**97.27**	**97.27**	**97.27**	97.27	**97.27**	96.53	93.86	87.23	74.06	64.42	60.29	90.53

Table 7: **Training-strategy ablation on GenImage.** mAcc (%) for both backbones.

Backbone	Linear	BCE-only	Joint-only	ReSCOPE
CLIP ViT-L/14	77.56	80.09	86.51	**93.37**
SigLIP2-Giant	85.24	89.75	92.77	**97.57**

Table 8: **Robustness on GenImage.** BAcc/AUROC (%), averaged over eight generators.

Metric	Raw	JPEG Compression					Gaussian Blur			Resizing			
		Q100	Q90	Q80	Q70	Q60	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$0.5\times$	$0.75\times$	$1.25\times$	$1.5\times$
BAcc	97.57	96.75	96.44	95.53	93.03	90.64	80.13	86.90	90.43	88.85	96.36	94.19	93.37
AUROC	99.73	99.55	99.46	99.73	99.70	99.63	95.09	94.95	96.34	97.17	99.57	98.89	99.00

6 CONCLUSION

We present ReSCOPE, a reconstruction-consistent adaptation method for generalizable AI-generated image detection. We find that feature responses from consecutive reconstructions exhibit clear local structure. This reconstruction-sensitive variation is closely related to generator-specific shortcuts. ReSCOPE aligns the representations of consecutive fake–fake reconstructions while preserving real/fake supervision. This weakens reliance on source-specific cues and encourages the encoder to use more transferable forensic cues. It also provides a new way to study generalization through label-preserving reconstruction variation. ReSCOPE requires no reconstruction at inference. Experiments on multiple benchmarks show strong cross-generator and cross-dataset generalization.

Reproducibility and AI assistance

The appendices provide diagnostic protocols, training configurations, and supplementary results. AI tools were used for literature searches and language polishing.

References

Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024.

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

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 7621–7639, 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 *Advances in Neural Information Processing Systems*, 2025.

Hyeongjun Choi, Juhun Lee, Davide Cozzolino, Luisa Verdoliva, and Simon S. Woo. Anchor-regularized adaptation for generalizable AI-generated image detection with DINOv3. *arXiv preprint arXiv:2608.15196*, 2026a.

Sungik Choi, Hankook Lee, Jaehoon Lee, Robin Kim, Stanley Jungkyu Choi, and Moontae Lee. A debiased reconstruction-based framework for training-free detection of AI-generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Jingyi Deng, Chong Zhang, Chenken Xu, Chenhao Lin, Zhengyu Zhao, Shuai Liu, Qian Wang, and Chao Shen. CLIP-ADA: CLIP-guided artifact-invariant generalizable synthetic image detection. *IEEE Transactions on Information Forensics and Security*, 21:3588–3601, 2026.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 3247–3258, 2020.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or JPEG? revealing common biases in generated image detection datasets. *arXiv preprint arXiv:2403.17608*, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18685–18694, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing global and local features for generalized AI-synthesized image detection. In *IEEE International Conference on Image Processing*, 2022.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19730–19742, 2023.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2405–2414, 2025.

Shuqiao Liang, Jian Liu, Renzhang Chen, and Quanlong Guan. FerretNet: Efficient synthetic image detection via local pixel dependencies. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *Computer Vision – ECCV 2022*, pp. 95–110, 2022a.

Bo Liu, Qiao Qin, and Qinghui He. CausalCLIP: Causally-informed feature disentanglement and filtering for generalizable detection of generated images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(9):7069–7077, 2026.

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

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

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

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

Lianrui Mu, Haoji Hu, Xingze Zou, Jianhong Bai, and Jiaqi Hu. No pixel left behind: A detail-preserving architecture for robust high-resolution AI-generated image detection. In *International Conference on Learning Representations*, 2026.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8245–8257, 2025.

Xinyi Qi, Kai Ye, Chengchun Shi, Ying Yang, Jin Zhu, and Hongyi Zhou. A difference-in-difference approach to detecting AI-generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42964–42975, 2026.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pp. 86–103, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. In *International Conference on Learning Representations*, 2025.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. 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.

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

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3418–3432, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for GAN-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12105–12114, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024a.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):5052–5060, 2024b.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2P-CLIP: Injecting category common prompt in CLIP to enhance generalization in deepfake detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):7184–7192, 2025.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ana Vasilcoiu, Ivona Najdenkoska, Zeno Geradts, and Marcel Worring. LATTE: Latent trajectory embedding for diffusion-generated image detection. *arXiv preprint arXiv:2507.03054*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8695–8704, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023.

Ziye Wang, Minghang Yu, Chunyan Xu, and Zhen Cui. Semantic discrepancy-aware detector for image forgery identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18388–18398, 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 *International Conference on Learning Representations*, 2025a.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. UCF: Uncovering common features for generalizable deepfake detection. *arXiv preprint arXiv:2304.13949*, 2023.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 70268–70288, 2025b.

Jiazhen Yang, Ruijin Jin, Junjun Zheng, Xiangheng Kong, Zunlei Feng, and Jie Lei. Structured local differential modeling for AI-generated image detection. *arXiv preprint arXiv:2608.12811*, 2026a.

Zheng Yang, Ruoxin Chen, Zhiyuan Yan, Keyue Zhang, Xinghe Fu, Shuang Wu, Xiujun Shu, Taiping Yao, Shouhong Ding, Zequn Qin, and Xi Li. All patches matter, more patches better: Enhance AI-generated image detection via panoptic patch learning. In *International Conference on Learning Representations*, 2026b.

Haifeng Zhang, Qinghui He, Xiuli Bi, Weisheng Li, Bo Liu, and Bin Xiao. Towards universal AI-generated image detection by variational information bottleneck network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23828–23837, 2025.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. PatchCraft: Exploring texture patch for efficient AI-generated image detection. *arXiv preprint arXiv:2311.12397*, 2024.

Yue Zhou, Xinan He, Kaiqing Lin, Bing Fan, Feng Ding, and Bin Li. Breaking latent prior bias in detectors for generalizable AIGC image detection. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Mingjian Zhu, Hanting Chen, Mouxiao Huang, Wei Li, Hailin Hu, Jie Hu, and Yunhe Wang. GenDet: Towards good generalizations for AI-generated image detection, 2023a.

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 *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023b.

A Additional Analysis of Reconstruction Responses

A.1 Local Structure of Reconstruction Responses

We analyze the response directions of frozen CLIP and SigLIP2 encoders using the same 160K consecutive-reconstruction triplets. For each test sample, we construct a local subspace from the responses of similar images in the reference set. We compare the response energy captured by local subspaces, global subspaces, and random feature axes at the same dimension.

Features and sampling. CLIP uses preprojection CLS features ($D = 1024$), and SigLIP2 uses pooler features ($D = 1536$). The 160K triplets contain 160 triplets from each of 1,000 classes. Both encoders use the same triplet IDs.

All geometry and prediction analyses use the same split within each class: triplets indexed 0–139 form the reference/training set, 140–149 the validation set, and 150–159 the test set. Across 1,000 classes, these splits contain 140K, 10K, and 10K triplets. Both backbones use identical sample IDs. Geometry uses validation only to check implementation and metric definitions; the reported results use the test split after the protocol is frozen. Reconstruction-response predictor checkpoints are selected by validation mean directional cosine. The projection experiment in Section A.3 uses a separate protocol.

Neighbor retrieval and coverage aggregation. For each test triplet, we retrieve neighbors twice, once using $F(A)$ and once using $F(A_1)$. For either input feature, each reference triplet is scored by the maximum of its cosine similarities to $F(A_i)$ and $F(A_{1,i})$. We select the top K distinct triplets, and form U from the top d right singular vectors of their row-stacked unit responses. The test response is not used for neighbor selection or subspace fitting. For each test triplet, we average the two squared projection norms; the reported coverage is

$$C_{\text{local}} = \text{mean}\left[\|U^\top r_{12}\|_2^2\right]. \tag{8}$$

where the mean spans both retrievals and all test triplets. Global SVD uses all 140K reference unit responses. The random-coordinate control selects feature axes with a fixed random seed, without reading the responses or fitting an SVD.

Geometry explores $K \in \{128, 256, 512\}$ and $d \in \{8, 32, 64, 128\}$. The $(128, 128)$ configuration is excluded because it reaches the sample-count rank bound and the neighbor matrices are not uniformly numerically full rank.

Table 9: **Unit-response coverage for all valid configurations in Fig. 1.** Triples report Local/Global/Random coverage (%). Δ CI is the class-bootstrap 95% interval for local minus global coverage, in percentage points.

		CLIP		SigLIP2	
K	d	Local/Global/Random	Δ CI	Local/Global/Random	Δ CI
128	8	15.42/9.32/0.70	[5.95, 6.26]	34.61/28.33/0.48	[6.09, 6.45]
128	32	26.39/19.61/2.95	[6.59, 6.97]	45.38/40.48/1.91	[4.71, 5.08]
128	64	33.39/28.10/6.06	[5.10, 5.48]	52.43/49.29/4.10	[2.96, 3.33]
256	8	16.43/9.32/0.70	[6.96, 7.27]	35.57/28.33/0.48	[7.07, 7.41]
256	32	28.98/19.61/2.95	[9.18, 9.56]	47.82/40.48/1.91	[7.16, 7.51]
256	64	36.86/28.10/6.06	[8.57, 8.93]	55.43/49.29/4.10	[5.96, 6.33]
256	128	46.30/40.75/12.58	[5.37, 5.74]	64.61/61.01/8.76	[3.44, 3.76]
512	8	16.71/9.32/0.70	[7.25, 7.54]	35.78/28.33/0.48	[7.29, 7.60]
512	32	30.66/19.61/2.95	[10.87, 11.22]	49.25/40.48/1.91	[8.61, 8.94]
512	64	39.47/28.10/6.06	[11.19, 11.56]	57.59/49.29/4.10	[8.13, 8.47]
512	128	49.77/40.75/12.58	[8.85, 9.18]	67.29/61.01/8.76	[6.13, 6.43]

Observations. Table 9 shows that local coverage exceeds global coverage in all 22 valid configurations across both backbones. All class-bootstrap 95% confidence intervals for the local-minus-global difference lie above zero. The local structure therefore holds across the tested neighborhood

sizes and subspace dimensions. At a fixed dimension, increasing the number of reference neighbors further improves local coverage. For example, at $d = 64$, increasing K from 128 to 512 raises coverage from 33.39% to 39.47% for CLIP and from 52.43% to 57.59% for SigLIP2.

These results show that reconstruction responses from similar images provide useful local directional information. Subspaces constructed from image neighborhoods describe reconstruction-induced feature variation more effectively than a single global subspace shared by all samples.

A.2 Predicting Reconstruction Response Directions

Capturing response energy in a local subspace does not determine the specific direction of each sample's response. We train a linear map and a two-hidden-layer MLP to predict the unit response from a single image's frozen features. Prediction is evaluated by the signed cosine similarity between the predicted and true directions. A local oracle provides a reference by using the true response to compute the best directional alignment attainable within the retrieved subspace.

Local oracle. Given the true test response, its best attainable directional cosine in a retrieved local subspace is $\|U^\top r_{12}\|_2$. We average this value over the two retrievals and all test triplets. Because the oracle uses the target response, it measures representability rather than deployable prediction accuracy. Response magnitude is not the prediction target.

Predictor objective and optimization. The predictor returns unit vectors $p_A = P(F(A))$ and $p_1 = P(F(A_1))$, with one feature input per forward pass. Training minimizes

$$\mathcal{L}_P = \tfrac{1}{2}[(1 - p_A^\top r_{12}) + (1 - p_1^\top r_{12})] + (1 - p_A^\top p_1). \tag{9}$$

Linear is a $D \to D$ affine map. The MLP is $D \to 2048 \to 2048 \to D$, with LayerNorm, GELU, and dropout 0.1. Both are initialized to the training mean direction and use AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}, ReduceLROnPlateau, and early stopping. Validation mean signed cosine selects the reconstruction-response predictor checkpoint. The principal-direction control's sign is aligned with the training mean unit response.

Table 10: **Mean signed directional cosine on held-out triplets.** Brackets give 95% confidence intervals obtained by bootstrapping the 1,000 classes. Oracles use the true response within the retrieved subspace.

Method	CLIP	SigLIP2
Mean direction	0.1219	0.3287
Principal direction	0.1165	0.3255
Linear	**0.2350** [0.2320, 0.2380]	**0.4446** [0.4392, 0.4495]
Two-hidden-layer MLP	0.2100 [0.2069, 0.2131]	0.4291 [0.4239, 0.4346]
Local oracle, $K = 512, d = 8$	0.3936	0.5718
Local oracle, $K = 512, d = 64$	0.6226 [0.6199, 0.6252]	0.7512 [0.7482, 0.7542]

Observations. In Table 10, the linear predictor achieves mean directional cosines of 0.2350 for CLIP and 0.4446 for SigLIP2. It outperforms the mean direction, principal direction, and tested MLP on both backbones. With $K = 512, d = 64$, the local oracle reaches 0.6226 and 0.7512. This gap shows that the predictors do not fully recover the response direction information available within the local subspace. The nonlinear MLP does not improve direction prediction in these experiments.

Local structure and accurate sample-specific prediction are therefore distinct. This observation supports adapting the encoder directly with observed reconstruction pairs, without making response prediction a prerequisite for adaptation.

A.3 Explicit Response Subspace Removal

We next examine whether directly suppressing feature components within the response subspace improves detection. Given a local subspace basis U, we transform the feature as

$$h' = h - \alpha U U^\top h, \tag{10}$$

where α controls the removal strength.

Experimental setup. This experiment uses a 150K reference database, separate from the 140K reference split used for the geometry and prediction analyses. The basis consists of 64 uncentered SVD directions from 512 neighboring unit responses. Each value of α receives its own classifier: $1536 \to 512 \to 128 \to 1$, with LayerNorm, GELU, and dropout 0.2. Classifier training uses 10K real and 10K fake images. Each projection diagnostic classifier's checkpoint is selected by BAcc on SD1.4 validation with 1,000 images per class. The cross-generator evaluation contains 40K samples. The exploratory soft-projection sweep used strengths from 0.005 to 0.125 and inspected this same cross-generator set. Projection classifiers are separate from the main LoRA detectors; Raw denotes the unprojected representation in this diagnostic.

Results and interpretation. The unprojected representation achieves 92.875% BAcc. Full removal ($\alpha = 1$) reduces BAcc to 89.507%. Partial removal depends on its strength: $\alpha = 0.25$ yields 92.755%, while $\alpha = 0.125$ yields 93.348%. The corresponding AUROC values for Raw, $\alpha = 1$, $\alpha = 0.25$, and $\alpha = 0.125$ are 0.993394, 0.978106, 0.993375, and 0.993117. Thus, weaker projection can preserve detection performance and modestly improve BAcc, whereas removing the entire subspace harms discrimination.

The geometry of the response subspace alone is therefore insufficient to guide wholesale removal of feature components. ReSCOPE instead adapts the encoder through reconstruction-pair consistency while retaining real/fake supervision, allowing adaptation to account for both reconstruction variation and the detection task.

A.4 Reconstruction Consistency and Cross-Generator Discrimination

Models and held-out images. Fig. 2(c–f) uses BCE-only and ReSCOPE CLIP ViT-L/14 detectors at their 15,000-step endpoints. We extract the 1,024-dimensional features consumed by each model's trained linear head. Both models use the same 2,000 held-out images: 500 SD1.4 reconstruction pairs (A_1, A_2), 500 unseen BigGAN-generated images, and 500 ImageNet real images. The reconstruction pairs and BigGAN images cover the same fixed set of 500 ImageNet leaf classes. Orange, green, blue, and gray denote A_1, A_2, BigGAN, and real images. Hollow blue markers identify BigGAN-generated images classified as real; thin gray segments connect a fixed subset of 100 reconstruction pairs.

Actual classifier margins. For each detector, let $w^\top h + b$ be its trained head's logit, with positive values denoting fake. We define

$$m_{\text{fake}}(h) = \frac{w^\top h + b}{\|w\|_2}, \qquad m_{\text{real}}(h) = -m_{\text{fake}}(h). \tag{11}$$

Panels (c–d) use m_{real} as the vertical coordinate, so points above zero are classified as real and points below zero as fake. The horizontal dashed line is therefore each detector's actual boundary, expressed in coordinates aligned with its own trained head. Panel (f) uses m_{fake}, for which positive values indicate correct BigGAN predictions.

Fixed reconstruction-sensitive axis. We construct the horizontal axis using an additional 1,000 reconstruction pairs disjoint from all displayed samples. After projecting out the subspace spanned by the two trained classifier-head directions, we fit a shrinkage-LDA direction distinguishing A_1 from A_2 on the BCE-only calibration features. This direction is then fixed and applied unchanged to both detectors and all four image groups. BigGAN and real images do not participate in fitting the axis. Along this direction, the projected A_1/A_2 centroid gap decreases from approximately 0.191 for BCE-only to 0.094 for ReSCOPE.

Original-space consistency and decision margins on images from unseen generators. Panel (e) reports $d_{\cos}(A_1, A_2) = 1 - \cos(h_\theta(A_1), h_\theta(A_2))$ for all 500 pairs in the original 1,024-dimensional feature space. Points denote individual pairs, boxes show the interquartile range, center lines indicate medians, and whiskers span the 5th–95th percentiles. The vertical axis is logarithmic. The median falls from 0.0113 to 0.000269, approximately $42\times$ lower under ReSCOPE. On the same

Table 11: **DDIM scheduler settings for the 160K SD1.4 reconstruction pool.**

Setting	Value
Inversion scheduler	`DDIMInverseScheduler`
Denoising scheduler	`DDIMScheduler`
Inversion / denoising steps	50 / 50
`num_train_timesteps`	1000
`beta_schedule`	`scaled_linear`
`beta_start` / `beta_end`	0.00085 / 0.012
`set_alpha_to_one`	`false`
`steps_offset`	1
`clip_sample`	`false`
Guidance scale / DDIM η	1.0 / 0.0

500 BigGAN-generated images, the median fake-side margin changes from -12.56 to $+9.68$, and fake accuracy increases from 19.0% to 99.6%. ReSCOPE classifies 95.0% of the 500 displayed real images correctly. These sample-level statistics describe the Fig. 2 analysis; full-benchmark results are reported in Section 5.

These results show that reconstruction-pair contraction accompanies a shift of images from an unseen generator toward the correct side of the decision boundary. Together with the preceding analyses, they support using reconstruction-induced variation to guide encoder adaptation. ReSCOPE reduces sensitivity to this variation through training, without first predicting the response accurately or directly removing the entire response subspace.

B Data Construction and Implementation Details

B.1 Reconstruction Data Construction

We select approximately 320K real images from GenImage to construct consecutive-reconstruction data. For each original image A, we apply the frozen reconstruction model twice to obtain A_1 and A_2, forming an aligned triplet (A, A_1, A_2). The complete dataset contains approximately 320K triplets, with (A_1, A_2) forming a label-preserving fake–fake reconstruction pair. All reconstructions are generated offline, and the reconstruction model is not optimized during detector training.

Different experiments use different subsets of these data. The geometry and response-prediction analyses use 160K triplets, with the splits described in Appendix A. Reconstruction-consistency training uses 10K pairs, sampled as 5K pairs from each of two reconstruction pools. Thus, the complete reconstruction dataset contains 320K triplets, while detector training uses 10K reconstruction pairs.

Reconstruction settings. The 160K SD1.4 pool uses real images from GenImage's SD1.4 `train/nature` split. After EXIF correction and sRGB conversion, we resize the shorter side to 512 pixels while preserving aspect ratio and take a 512×512 center crop. BLIP-2 OPT-2.7B generates one caption c_A per original image with `max_new_tokens=30`, `num_beams=1`, and seed 0. The same caption conditions both passes, using SD1.4's own tokenizer and CLIP text encoder:

$$A_1 = \mathcal{R}(A; c_A), \qquad A_2 = \mathcal{R}(A_1; c_A).$$

Here, $\mathcal{R}$ implements the fixed-caption operator T in the main text.

For each pass, the frozen SD1.4 VAE encodes the input with `latent_dist.mode()` and scales the latent by 0.18215. We perform full 50-step DDIM inversion followed by 50-step denoising on the same timestep grid in reverse, with identical caption conditioning. After inverse latent scaling and VAE decoding, the output is converted to $[0, 1]$ and clipped, multiplied by 255, rounded, and saved as an 8-bit RGB PNG. A_1 is reloaded from PNG before the second pass; c_A is reused without recaptioning. Both passes use float32, seed 0, guidance scale 1.0, and $\eta = 0$, with no stochastic noise, CFG extrapolation, or negative-prompt branch. This full-inversion procedure has no img2img `strength` parameter. Table 11 gives the scheduler settings.

B.2 Detector Training Data and Model Configuration

Detector training combines real/fake classification supervision with reconstruction-consistency supervision. Classification uses the GenImage SD1.4 training set, containing 162,000 real and 161,996 generated images. Consistency training uses the 10K reconstruction pairs described above. Training applies random horizontal flips, synchronized between the two images in each reconstruction pair.

We use CLIP ViT-L/14 and SigLIP2-Giant-opt-patch16-384 with input resolutions of 224×224 and 384×384. CLIP adapts all 96 query, key, value, and output attention projections across its 24 Transformer blocks. SigLIP2 adapts 160 attention projections. Both backbones use LoRA rank 4, alpha 8, and dropout 0.1, with the original pretrained weights frozen.

The linear classification head consumes the raw pooled representation, while the reconstruction-consistency loss uses its L2-normalized version. In the CLIP configuration, reconstruction pairs receive only the consistency objective, without an additional pair fake-classification loss. Table 12 lists the trainable parameter counts, batch sizes, gradient accumulation, and loss weights.

Table 12: Training configurations. Effective batches are across two GPUs per optimizer step. Every named phase lasts 625 steps.

Setting	CLIP	SigLIP2
Input resolution	224	384
Attention projections with LoRA	96	160
LoRA rank	4	4
Joint trainable parameters	787,457	1,967,617
Supervised micro-batch per GPU	16	8
Pair micro-batch per GPU	8	4
Gradient accumulation	1	2
Effective supervised batch	32	32
Effective pair batch	16	16
Joint / RC-only loss weight	0.1 / 0.1	0.1 / 0.1
B0 steps	625	625
Total optimizer steps (including initialization)	15,000	9,375

B.3 Optimization and Alternating Training

Training begins with source-domain real/fake classification initialization, followed by alternating Joint and RC-only phases. Joint phases optimize classification and reconstruction consistency together, updating both LoRA adapters and the classification head. RC-only phases keep the head fixed and update only the adapters through the consistency loss. Each phase contains 625 optimizer steps.

After initialization B0, CLIP completes six Joint phases J1–J6, followed by RC-only E7, and alternates Joint and RC-only phases through E23, for 15,000 total steps. SigLIP2 completes B0 and five Joint phases J1–J5, followed by RC-only P1 and four Joint/RC-only cycles ending at P5, for 9,375 total steps. Both totals include initialization. For SigLIP2, learning rates are 10^{-4} for B0, 10^{-5} for J1–J5, and 10^{-6} for the subsequent alternating phases. Training uses bfloat16 mixed precision.

We report results from CLIP E23 and SigLIP2 P5. For each backbone, evaluations across datasets use the same reported weights without target-dataset adaptation.

B.4 Baseline Implementations

Frozen-feature linear baselines. Each Linear baseline extracts features with the original frozen pretrained encoder and trains only a linear classification head. Training uses 160K ImageNet real training images and 160K generated images sampled from the 161,996-image SD1.4 training fake pool. CLIP uses the 1,024-dimensional post-LayerNorm CLS feature before visual projection, and SigLIP2 uses the 1,536-dimensional pooled output. Both apply L2 normalization before the linear head during training and inference. We use AdamW with learning rate 10^{-3}, weight decay 10^{-4},

2,048 examples per class per batch, and seed 20260908. Both heads are trained for 10 epochs (790 optimizer steps) and evaluated using the final-epoch weights on the GenImage validation split.

BCE-only and Joint-only. Both controls start from a shared 625-step classification initialization and train to a total of 15,000 steps for CLIP or 9,375 for SigLIP2. BCE-only optimizes only real/fake classification. Joint-only continuously optimizes classification and reconstruction consistency throughout adaptation, without RC-only phases. Both controls use the same source data, preprocessing, and LoRA configuration as alternating training; Joint-only uses the same 10K reconstruction-pair pool.

Control training uses seed 0 and AdamW with weight decay 0.01, $\beta = (0.9, 0.999)$, gradient clipping at 1.0, and cosine learning-rate schedules. The initialization learning rate is 10^{-4}; the next 3,125 steps use 10^{-5}, followed by 625-step stages at 10^{-6}. The effective supervised batch contains 32 images per step. Joint-only additionally uses 16 reconstruction pairs and optimizes $\mathcal{L}_{\mathrm{BCE}} + 0.1\mathcal{L}_{\mathrm{RC}}$. Each stage transfers adapter and head weights and creates a new optimizer. The two controls share a newly trained initialization, while ReSCOPE uses the initialization saved in its original run.

C Effect of Adaptation Length

We examine how cross-generator detection performance changes as alternating adaptation progresses. We evaluate SigLIP2 after each RC-only stage, comparing six endpoints from P1 to P6 on a 40K-image subset of GenImage. This subset contains 2,500 real and 2,500 fake images for each of the eight generators. All evaluations use a fixed logit threshold of zero. The subset is drawn from the same official validation split used for the full GenImage evaluation.

P1 follows the first RC-only stage. Each subsequent endpoint completes another cycle of 625 Joint steps and 625 RC-only steps. Evaluating at the same point in each cycle allows us to compare different adaptation lengths.

Table 13: **Effect of adaptation length on GenImage.** BAcc (%) of SigLIP2 evaluated after each RC-only stage on a 40K-image subset. Total steps include classification initialization.

Stage endpoint	Total steps	BAcc (%)
P1	4,375	97.3200
P2	5,625	97.3200
P3	6,875	97.4525
P4	8,125	97.5175
P5	9,375	97.5850
P6	10,625	97.5225

As shown in Table 13, BAcc remains between 97.32% and 97.59% across 4,375–10,625 training steps, with an overall range of 0.265 percentage points. These results show that high detection performance is maintained across the evaluated adaptation lengths.

The training schedules and total step counts for CLIP E23 and SigLIP2 P5 are provided in Appendix B.

D Comparison with Reconstruction-Based Detectors

We compare ReSCOPE with reconstruction-based detectors on GenImage. Table 14 reports accuracy for SD1.4-trained methods and AUROC for both trained and training-free methods. ReSCOPE uses the same SigLIP2 detector as in Table 1.

LaRE^2 and LATTE/Avg results are taken from Table 1 of Vasilcoiu et al. (2025) (arXiv v2), where Avg denotes average pooling. SARE results follow Tables 1–2 of Kang et al. (2025) (arXiv v3). AEROBLADE and RDD results are taken from Table 1 of Choi et al. (2026b), which evaluates both methods using an ensemble of SD1.4, SD2-base, and MiniSD autoencoders. AUROC values are reported as percentages.

Table 14: **Comparison with reconstruction-based detectors on GenImage.** Panel (a) reports Acc (%) for SD1.4-trained detectors. Panel (b) reports AUROC (%), including training-free AEROBLADE and RDD. Bold and underline indicate the best and second-best results within each panel.

Method	MJ	SD1.4	SD1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Mean
(a) Acc (%)									
LaRE^2 (Luo et al., 2024)	74.00	**100.00**	**99.90**	61.70	88.50	**100.00**	<u>97.20</u>	68.70	86.20
LATTE/Avg (Vasilcoiu et al., 2025)	88.80	**100.00**	**99.90**	74.00	<u>95.80</u>	<u>98.90</u>	80.80	89.80	91.00
SARE (Kang et al., 2025)	<u>90.32</u>	97.21	97.04	<u>84.47</u>	93.55	97.05	93.66	<u>92.05</u>	<u>93.17</u>
ReSCOPE (SigLIP2)	**96.15**	<u>98.18</u>	<u>98.21</u>	**95.97**	**98.05**	98.38	**97.80**	**97.84**	**97.57**
(b) AUROC (%)									
AEROBLADE (Ricker et al., 2024)	<u>98.50</u>	97.60	97.80	85.10	98.90	97.80	72.10	97.80	93.20
RDD (Choi et al., 2026b)	98.30	**100.00**	99.90	92.60	<u>99.60</u>	<u>99.90</u>	94.60	**99.70**	98.10
SARE (Kang et al., 2025)	96.83	<u>99.94</u>	<u>99.93</u>	<u>94.87</u>	98.00	99.83	<u>98.31</u>	97.51	<u>98.15</u>
ReSCOPE (SigLIP2)	**99.40**	**100.00**	**100.00**	**99.13**	**99.81**	**100.00**	**99.87**	<u>99.65</u>	**99.73**

ReSCOPE achieves the highest mean accuracy of 97.57% and mean AUROC of 99.73% among the compared methods. These results demonstrate its effectiveness relative to reconstruction-based detectors. ReSCOPE requires no reconstruction at inference.

E Evaluation Details

Metrics. We report mAcc as the unweighted mean of generator-wise Acc on GenImage, DRCT-2M, AIGCDetectionBenchmark, and DiffusionForensics. BAcc is the unweighted mean of real- and fake-class accuracies. Synthbuster and B-Free Online use BAcc in Table 5; EvalGEN uses mean Fake Acc because it contains only generated images. Chameleon uses overall image-level Acc. GenImage and Synthbuster are class-balanced within each generator, so their Acc and BAcc coincide.

Baseline result sources. Baseline results in Table 4 are taken from Table 2 of PatchCraft (Zhong et al., 2024); those in Table 5 are taken from Tables 7, 10, and 2 of DDA (Chen et al., 2025). For B-Free Online, the published baselines use the original benchmark, while our result uses the verified available subset described in Appendix E.1.

Inference settings. All external evaluations use the same SigLIP2 P5 checkpoint (Appendix B), with RGB conversion, bicubic resizing to 384×384, and channel-wise normalization with mean and standard deviation 0.5. Inference uses BF16, no random crop or flip, and a fixed zero-logit threshold (fake probability 0.5), without target adaptation or test-set selection of checkpoint, threshold, or preprocessing.

GenImage. We evaluate on the complete official validation split, containing 100K images across eight generators. SD1.5 contributes 8,000 real and 8,000 fake images, while each other generator contributes 6,000 images per class. We report mAcc as the unweighted mean of the eight generator-wise accuracies.

Sources of GenImage baseline results. For Table 1, CNNSpot's generator-wise values originate from GenImage, Table 3 (Zhu et al., 2023b). DIRE and UnivFD match the SD1.4-trained reimplementations in GenDet, Table 3 (Zhu et al., 2023a). The displayed means for these three methods follow AIDE, Table 4 (Yan et al., 2025a). NPR results are reported in C2P-CLIP, Table 3 (Tan et al., 2025). Results for AIDE, EFFORT, DRCT, and OMAT are taken from their own papers: Table 4 of Yan et al. (2025a), Table 9 of Yan et al. (2025b), Table 2 of Chen et al. (2024), and Table 1 of Zhou et al. (2025). SAFE, C2P-CLIP, PPL, and HiDA-Net results come from Table 4 of Li et al. (2025), Table 3 of Tan et al. (2025), Table 1 of Yang et al. (2026b), and Table 3 of Mu et al. (2026). Published means are retained as reported in these sources.

DRCT-2M. We evaluate 16 diffusion generators with 5,000 fake images per generator and the same 5,000 MSCOCO validation real images. This gives 85K unique images and 160K evaluation entries after real-image reuse. Table 6 reports the generator-wise Acc.

AIGCDetectionBenchMark. Baselines are trained on LSUN real images and ProGAN fake images, except DIRE-D, which is trained on ADM. ReSCOPE is trained on GenImage SD1.4. The complete evaluation contains 152,598 images across 17 balanced domains, including SDXL. To mitigate encoding-related differences between real and generated images, we re-encode both classes as JPEG at quality 80 before bicubic resizing to 384×384. Table 4 reports the generator-wise Acc.

DiffusionForensics. We evaluate the eight LSUN-Bedroom generator groups in Table 2, sharing 1,000 real images across groups. DDPM contains 768 fake images and each other group contains 1,000, giving 8,768 unique images. Means weight the eight generators equally. DDPM's unequal class sizes yield 97.34% Acc and 97.20% BAcc; macro BAcc is 98.13%, with shared Real Acc of 98.30%.

Chameleon. We evaluate all 26,033 test images (14,863 real and 11,170 fake). Table 3 reports overall, fake-, and real-class Acc. BAcc is 87.73%, with Real Acc of 96.19% and Fake Acc of 79.27%.

E.1 Additional cross-benchmark evaluations

Synthbuster + RAISE. Each of nine generators contributes 1,000 fake images, paired with the same 1,000 RAISE-1k photographs (10K unique images). We use released files without additional JPEG re-encoding. Table 5 reports per-generator Acc and the equal-weight mean. Shared Real Acc is 78.40%; the across-generator sample SD of Acc is 4.48 percentage points.

EvalGEN. Following DDA Tables 2/10, we use generation index 0 for each of 553 prompts from five generators: 2,765 fake images in total. This subset is complete. Author-released JPEG Q96 files are used without recompression. Table 5 reports Fake Acc and its equal-weight mean, since no real images are included. The across-generator sample SD is 0.64 percentage points.

B-Free Online. The B-Free Online manifest contains 1,111 images. We evaluate the 799 successfully downloaded, official-MD5-verified images (71.92% coverage): 262 real and 537 fake, retaining 34 source IDs. Table 5 reports this available subset. Verified web files are used without extra JPEG re-encoding. Acc is image-level accuracy; BAcc is 89.02%, with Real Acc of 93.13% and Fake Acc of 84.92%.

F Robustness Evaluation on GenImage

Table 8 evaluates the fixed SigLIP2 P5 detector on the complete 100K-image GenImage validation set. Each transformation is applied to both real and fake images, with no detector adaptation or threshold adjustment. The logit threshold remains zero. We test JPEG quality factors of 100, 90, 80, 70, and 60; Gaussian blur with $\sigma \in \{1, 2, 3\}$; and resizing by factors of 0.5, 0.75, 1.25, and 1.5. JPEG encoding uses Pillow with its default subsampling; blur uses Pillow GaussianBlur with radius set to σ; scaling uses bicubic interpolation. The detector retains its 384×384 input resolution and normalization settings.

Raw results are reused from the full evaluation of the same checkpoint. All 12 transformed conditions cover the complete 100K-image set, giving 1.2 million additional model evaluations with no failures. BAcc and AUROC are averaged equally across the eight generators. Because each generator group is class-balanced, mean BAcc equals mAcc.

BAcc reaches 95.53% under JPEG Q80 and 96.36% at $0.75\times$ resizing. Gaussian blur causes larger, non-monotonic accuracy drops. Under $\sigma = 1$, mean real-image accuracy is 62.17%, while mean fake-image accuracy remains 98.10%, accounting for the larger accuracy decrease in this condition.