RESIDUAL FAMILY AUGMENTATION FOR DETECTING IMAGES FROM HIGH-FIDELITY AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated image (AIGI) detection has advanced rapidly, yet on the newest text-to-image models even robust detectors degrade sharply. In this paper, we diagnose the failure with a Residual Scaling Sweep, which scores a detector as the reconstruction residual is scaled, and identify an overlooked patch-size mismatch. The sweep shows that detectors trained with existing recipes respond only near the magnitude and the sign of their training residual, and they fail to keep up with recent generators. We therefore propose *Residual Family Augmentation*: the residual is added back to the photograph within a bounded range of magnitudes and at both signs, all labeled fake, plus a second reconstruction pass. The detector then no longer depends on the magnitude or the sign of one residual and responds across that range, which includes the weaker residuals of autoencoders never seen in training. We further observe that the autoencoders of recent generators leave a reconstruction residual, the difference between a photograph and its reconstruction, that is weaker and concentrated at finer spatial scales than the residual of earlier autoencoders, scales that the pretrained filters of vision backbones attenuate. Notably, we find that simply halving the patch of the pretrained backbone, which adds no parameter, restores the residual to scales the backbone retains, and the augmentation reaches its full gain only under the resampled patch. Trained on a single, older autoencoder, our detector leads seven public detectors by five points or more in mean balanced accuracy over eleven sets from the newest generators, while remaining competitive with the best baseline across public benchmarks.

1 INTRODUCTION

Existing AI-generated image (AIGI) detectors, trained on real photographs and their reconstructions labeled fake, generalize to a wide range of latent generators (Rajan et al., 2025; Guillaro et al., 2025; Chen et al., 2025). On recent generators, typically built on a latent autoencoder, the accuracy of these detectors drops sharply, and on the newest it falls to chance. To understand this failure, we examine what each detector reads with a *Residual Scaling Sweep* and identify an overlooked patch-size mismatch behind it.

The *Residual Scaling Sweep* rescales the residual of a decoder at test time and reads each detector's response, as illustrated in Figure 1. We add the residual back to the photograph at scale β: $\beta = 0$ gives the photograph, $\beta = 1$ its reconstruction, and $\pm\beta$ change the image by the same amount with opposite sign. On this sweep, the public detectors show two dependencies: they follow the residual magnitude, or they rise only for the signed pattern of the training decoder. Either way, most read the older decoders well and the newest, whose residuals are the faintest and the finest, far worse. Both dependencies come from the training data: the standard recipe offers one fake per photograph, the reconstruction, so the detector learns the magnitude and the signed pattern of that single point first. Indeed, our detector under the standard recipe repeats the same failures, and scaling the residual during training largely removes both, all other settings held fixed. *We therefore train not on one point of the residual, but on the family of images it spans.*

The *family* alone does not suffice: a residual the backbone cannot resolve is not learned at any magnitude. Our observation starts from the reconstruction residuals of multiple autoencoders, from older diffusion models to the newest generators. The residual is the difference between a photograph and its reconstruction, and each decoder leaves its own. Across decoders, the residuals are correlated,

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.

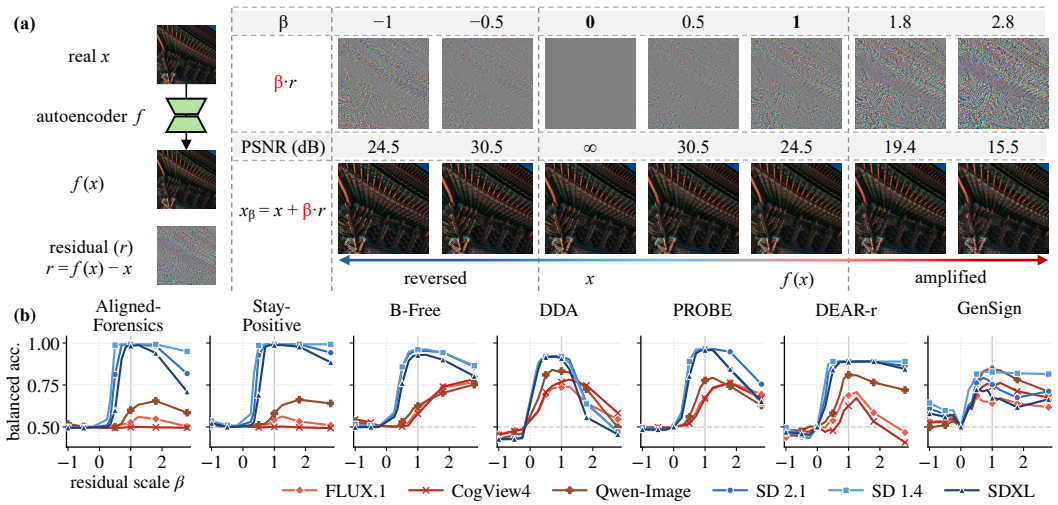

Figure 1: (a) Residual Scaling Sweep: one decoder's residual added back to the photograph at scale β; β and $-\beta$ change the image by the same amount (equal PSNR) and differ in sign. Residuals at $4\times$. (b) Seven public detectors on the sweep: older 4-channel decoders in blue, newer decoders with 16 or more latent channels in red; one line per decoder.

and newer decoders write them weaker and finer. On the newest decoders, a third to a half of the residual lies at scales of a few pixels, which vision transformers are known to attenuate (Park & Kim, 2022; Wang et al., 2022; Bai et al., 2022). Notably, resampling the patch of the pretrained backbone to a finer size restores the residual to scales the backbone retains, while adding no parameter; we call this step *Patch Resampling*. With the resampled patch, however, the public detectors gain little, and our detector under the standard recipe keeps both dependencies of the sweep; the family above is needed on top of it.

Based on the above findings, we propose *Residual Family Augmentation*, a training scheme for the *resampled patch*: Patch Resampling lets the finer residual reach the network, and the family widens the training residuals in magnitude and in sign so that the network reads it across that range. We add the residual back to the photograph within a bounded range of magnitudes and at both signs and label every result fake, and we pass the reconstruction through the autoencoder once more for a fake example the decoder itself produced. Nothing else changes: the same photographs, the same single decoder, the same binary loss. A feature tied to the residual magnitude now sees the same label at every magnitude, and a feature matched to the signed pattern sees the pattern reversed with the same label. The range brackets the weaker residuals of recent decoders, although those decoders never appear in our training data.

- We diagnose the failure with the *Residual Scaling Sweep*, a test-time rescaling of the reconstruction residual: existing detectors depend on the magnitude or on the signed pattern of their training residual, and both dependencies come from training on a single reconstruction per photograph.
- We identify an overlooked patch-size mismatch: the residuals of recent generators concentrate at scales the pretrained filters attenuate, and *Patch Resampling*, adding no parameter, restores them to scales the backbone retains.
- We propose *Residual Family Augmentation* for the resampled patch: residuals added back within a bounded range of magnitudes and at both signs, all labeled fake, plus a second reconstruction pass, so that the detector responds across the range of the newest decoders.
- On 67 generator sets from eight benchmarks, our detector, trained on a single older autoencoder, leads seven public detectors by five points or more on recent generators, and neither the training photographs nor Patch Resampling alone reproduces the full gain.

2 Related Work

Reconstruction-based AIGI detectors. Training on real images and their autoencoder reconstructions removes content bias from the real/fake split (Rajan et al., 2025), and Stay-Positive constrains the classifier to positive evidence on the same data (Rajan & Lee, 2025); B-Free adds self-conditioned reconstructions and content augmentation (Guillaro et al., 2025), DDA aligns pixel and frequency statistics (Chen et al., 2025), and DRCT uses reconstructions as hard contrastive samples (Chen et al., 2024). Others vary the source of the residual: PROBE adds detector-guided hard samples (Cao et al., 2026), DEAR prunes channels located with inpainted photographs (Kim et al., 2026), GenSign reconstructs through randomly varied architectures (Fei et al., 2026), DID classifies first- and second-order reconstruction errors (Qi et al., 2026), and Liu et al. (2026) pass photographs through the final component of the generator. All train at the reconstruction, and DDA also at blends of it with the photograph; we keep one older decoder and extend its residual beyond the reconstruction, to the opposite sign and to a second pass. Section 4 compares with the three that release weights (PROBE, DEAR, GenSign).

Detector backbones and other cues. HiDA-Net feeds full-resolution tiles to a frozen CLIP backbone (Mu et al., 2026) and PPL trains on photographs with reconstructed patches (Yang et al., 2026); both keep the patch size of the pretraining. FlexiViT resamples the patch embedding to train one model at several patch sizes (Beyer et al., 2023); we resample it instead to match the filter to the scale of the residual (Section 3.1). Classifiers read the traces of upsampling (Zhang et al., 2019; Durall et al., 2020; Frank et al., 2020; Tan et al., 2024) or learn the spectrum of real images (Karageorgiou et al., 2025), and Corvi et al. (2023) documented the spectral peaks of diffusion generators in the residual; our diagnostic varies the amplitude and sign of one such residual with its shape held fixed. In shortcut learning, a classifier reads the cheapest predictive cue (Geirhos et al., 2020); JPEG history and resolution are documented shortcuts (Grommelt et al., 2024), benchmarks now pair fakes with photographs (Xiao et al., 2026) or test in the wild (Yan et al., 2025; Li et al., 2025), and frozen CLIP probes (Ojha et al., 2023; Cozzolino et al., 2024; Koutlis & Papadopoulos, 2024), reasoning-aligned distillation (Huang et al., 2026), probabilistic prompts (Wang et al., 2026) and energy scores (Cai et al., 2025) take routes without reconstruction training. We identify two dependencies inside the residual itself and widen the training family over them rather than penalizing them in the loss.

3 Proposed Method

We first resample the patch embedding (Section 3.1), then define the Residual Scaling Sweep (Section 3.2) and Residual Family Augmentation (Section 3.3). Let x be a photograph, $\mathrm{AE}(x) = D(E(x))$ its reconstruction by the encoder E and decoder D of a latent autoencoder, and $f_k(x)$ its k-th pass, $f_1(x) = \mathrm{AE}(x)$. The reconstruction residual is

$$r = \mathrm{AE}(x) - x. \tag{1}$$

The residuals of different autoencoders are correlated: over eight autoencoders and 400 photographs, the first principal component carries 48% of the variance with a positive loading on every autoencoder (Appendix C), so a detector trained on one autoencoder can respond to the residual of another.

3.1 Patch Resampling

Detectors built on pretrained vision transformers read the image through the patch embedding of the pretraining and keep its patch size (Ojha et al., 2023; Koutlis & Papadopoulos, 2024; Guillaro et al., 2025; Chen et al., 2025). This section measures what that layer does to the residual of a recent generator and resamples its filters to half their size, a step that adds no parameters.

Recent residuals are weak twice. Autoencoders differ in the amplitude and the spatial scale of the residual (Figure 3). The Stable Diffusion residuals form patches of several pixels; the residuals of FLUX.1, FLUX.2, CogView4 and Qwen-Image are faint, at .45 to .64 of the SD 2.1 amplitude, and change from one pixel to the next, with .31 to .52 of their energy at periods below four pixels against .21 to .28 for Stable Diffusion (Figure 3a,b). A higher-fidelity autoencoder thus leaves a weaker residual, and a third to a half of it varies from one pixel to the next. A vision transformer reads the image through the patch embedding, a bank of 14×14 filters at stride 14, so that every

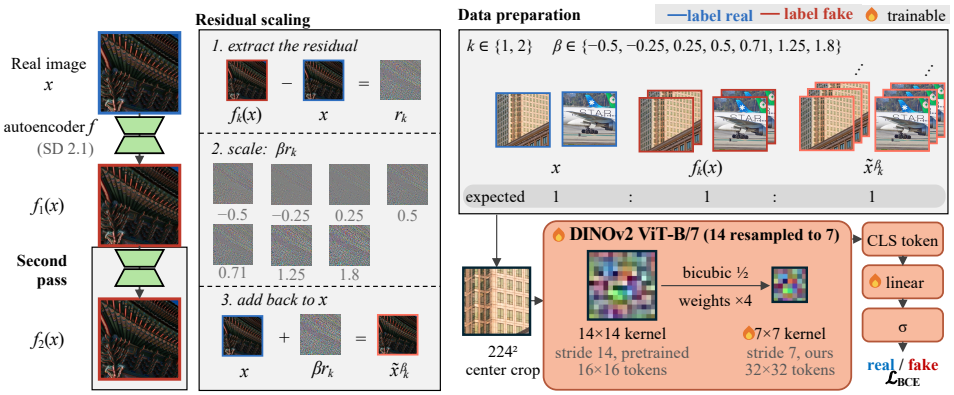

Figure 2: Training of our detector with DINOv2 ViT-B/7, the ViT-B/14 backbone with its patch embedding resampled to 7 pixels (Section 3.1). $f_k(x)$ is the k-th pass of the fixed SD 2.1 autoencoder and $\tilde{x}_k^\beta$ the image with its residual scaled (Section 3.3); residuals are shown at $8\times$.

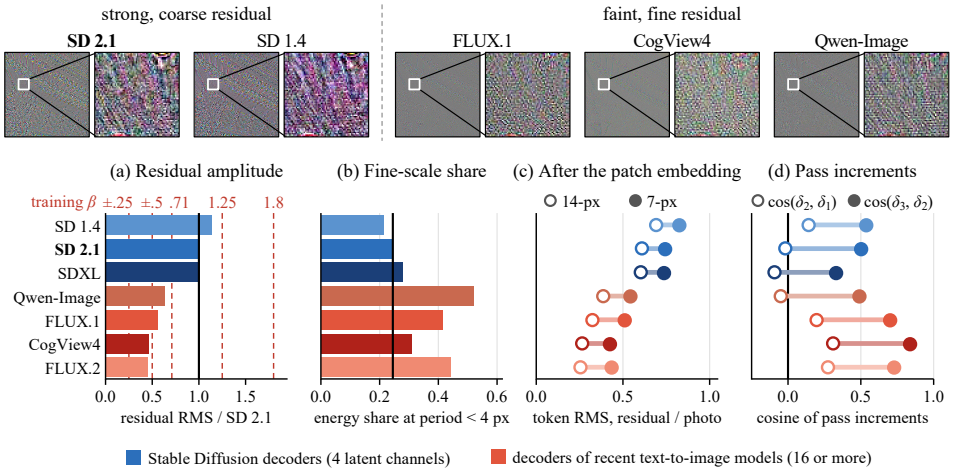

Figure 3: Top: residuals of five decoders at $4\times$, with the flattest 64-px block at $12\times$. Bottom, seven decoders: (a) amplitude relative to SD 2.1, with the training values of β; (b) share of residual energy at periods below four pixels; (c) token RMS of the residual after the DINOv2-B patch embedding, relative to the photograph, with 14-px (open) and 7-px (filled) filters; (d) cosine between increments of successive passes, δ_2 against δ_1 (open), δ_3 against δ_2 (filled). Training decoder in bold.

token is a weighted sum over 196 pixels, $t_k = \langle W_k, P\rangle + b_k$ for the pretrained filter W_k and the patch P. The filters are tuned to natural images, and this sum is what a residual must pass before the first attention block. At the token level, the residual amplitude relative to the photograph is .61 for SD 2.1 and .26 to .39 for the four recent autoencoders (Figure 3c, open; Appendix C). The recent residual is therefore weak twice: in the pixels, and again after the filters that read it.

Patch Resampling. The *resampled patch* is half the size of the pretrained patch: *Patch Resampling* resamples every 14×14 filter of our ViT-B/14 to 7×7 and applies it at stride 7,

$$t_k = \langle \widetilde{W}_k, Q\rangle + b_k, \qquad \widetilde{W}_k = \mathcal{R}_{14\to 7}(W_k), \tag{2}$$

where Q is the 7×7 block under the patch and $\mathcal{R}$ is bicubic resampling with the response to a constant input preserved (cf. Beyer et al., 2023). The crop stays 224×224, the token grid becomes 32×32 with the positional embedding interpolated, and no parameter is added. Each token now covers 49 pixels in place of 196, so every period of the residual doubles relative to the filter, and the token-level amplitude rises to .74 for SD 2.1 and to .43 to .54 for the recent residuals (Figure 3c, filled): 1.4 to 1.7 times more passes. The step applies to any vision transformer, in training and

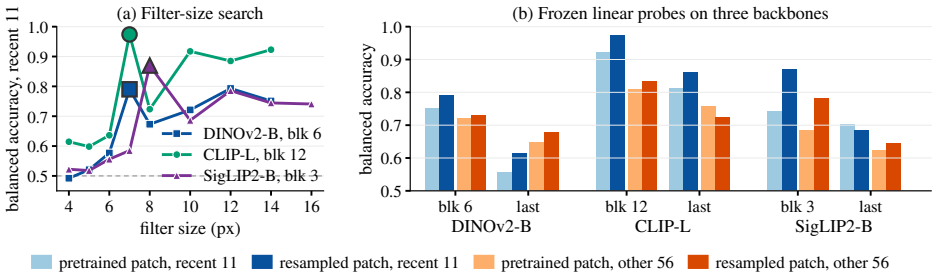

Figure 4: Frozen linear probes on three backbones. (a) Filter-size search at each backbone's best block; the outlined marker is half the pretrained size. (b) Probes at the best intermediate and the last block with the pretrained and resampled patch, on the recent and other 56 sets (Tables 11, 12).

inference; the cost is four times the tokens, 214 GFLOPs per image against 47 (Table 7). We write ViT-B/7 for the resampled backbone.

The factor is a search result (Figure 4a): frozen linear probes on three backbones peak at half the pretrained size (DINOv2 within .003), the smallest size that still computes the pretrained tokens (Table 12; Appendix C). The gain appears without fine-tuning and at the last block of DINOv2 and CLIP (Figure 4b). The gain comes from the pixels under each token, not their number: a 448 crop with the pretrained patch has the same grid and stays at the standard recipe (Table 13). Patch Resampling changes what the detector can read, not what it learns: it lifts the standard recipe from .830 to .921 (Table 2) and three of four released detectors at inference (Section 4.2), yet the response still follows the magnitude and the sign of the training residual, as the next section shows.

3.2 Residual Scaling Sweep

The resampled patch leaves open which property of the residual a detector depends on. The Residual Scaling Sweep measures this dependence by varying the residual of a photograph without changing the shape of the residual. For a residual scale $\beta \in [-1, 2.8]$ the sweep forms

$$x_\beta = x + \beta\, r = x + \beta\,(\mathrm{AE}(x) - x), \tag{3}$$

so that x_0 is the photograph and x_1 is the reconstruction. A value $\beta > 1$ amplifies the residual, and a value $\beta < 0$ adds the residual with the opposite sign, which sharpens the photograph beyond the original. The images x_β and $x_{-\beta}$ differ from x by the same power spectrum and differ from each other only in sign (Figure 1a). We score x_β against x for eight decoders and 200 photographs at thirteen values of β (Appendix B) and read three quantities: the *amplitude response*, the gain in balanced accuracy from $\beta = 0.25$ to $\beta = 1$, which is steep for a detector that depends on the residual magnitude; the *sign reading*, the AUROC of $x_{-0.5}$ against x, which is near 0.5 for a detector matched to the signed pattern of one decoder and below 0.5 for one that reads real-image statistics; and the *decoder spread*, the range of balanced accuracy across decoders at matched damage (27 dB PSNR), which is large for a detector matched to one decoder. Figure 1b applies the sweep to seven public detectors: each follows the magnitude of the residual or the signed pattern of its training decoder, and most read the recent decoders well below the older ones (Appendix D.3).

The sweep under the resampled patch. The sweep separates what *patch resampling* repairs from what it leaves. We train our detector with the standard recipe, one reconstruction per photograph, with the original and the resampled patch. At $\beta = 1$ the resampled patch narrows the gap between the older and the recent decoders from .27 to .09, so it removes most of the decoder spread; the other two readings do not change. On the recent decoders the same model scores .68 at $\beta = 2.8$ and .50 at $\beta = -1$, so the response still depends on the magnitude and sign of the residual. Both remaining dependencies come from the training data, not the backbone: the standard recipe offers one fake per photograph, the reconstruction, at one magnitude and one sign, and a detector trained on this single point has no reason to respond to another magnitude or the opposite sign.

3.3 Residual Family Augmentation

Standard reconstruction training uses two points per photograph: x with label real and $\mathrm{AE}(x)$ with label fake. Residual Family Augmentation replaces the single fake point by a bounded family of images around it: the residual at several magnitudes and at both signs, and the residual of a second pass. The photographs, the single decoder, and the loss stay unchanged.

Residual scaling. Let $f_k(x)$ be the k-th pass of the autoencoder, $f_1(x) = \mathrm{AE}(x)$ and $f_2(x) = \mathrm{AE}(\mathrm{AE}(x))$, with residual $r_k = f_k(x) - x$. With probability $p_\beta = 0.5$ we replace $f_k(x)$ by

$$\tilde{x}_k^\beta = \mathrm{clip}\big(x + \beta\, r_k\big), \qquad \beta \sim \mathcal{U}\{-0.5,\, -0.25,\, 0.25,\, 0.5,\, 0.71,\, 1.25,\, 1.8\}, \tag{4}$$

and we keep the label fake. The scaling is applied at full resolution, before cropping and before any resampling, so that x and $f_k(x)$ are pixel-aligned. The magnitudes span a factor of seven at both signs and bracket the residuals of the recent decoders, which have .45 to .64 of the SD 2.1 amplitude (Figure 3a): the detector sees residuals of that strength in training without seeing those decoders. The negative values label a reversed residual fake, so neither the sign of the residual nor an excess of high-frequency content marks a photograph. Real images are never scaled, and no single range of β reproduces the gain of the full set (Table 3).

Second pass. We pass the reconstruction through the same autoencoder once more and label $f_2(x) = \mathrm{AE}(\mathrm{AE}(x))$ fake. The scaled images $x + \beta r_1$ lie on one line through x and $f_1(x)$; the second pass leaves this line and adds a direction that scaling the first residual cannot produce, whereas a third pass repeats the second at a smaller magnitude (Figure 3d; Appendix D.1). Two passes is therefore the smallest depth that adds a new direction. Each photograph contributes three items, x, $f_1(x)$ and $f_2(x)$, and residual scaling applies to both fake items, so that in expectation a photograph contributes one real item, one reconstruction and one scaled residual (Figure 2); the real item has loss weight 2 (Appendix B). All models of Table 2 further share a set of label-preserving augmentations, JPEG re-encoding, rescaling, and a sharpened-real control that keeps sharpness from becoming a fake cue, defined with their probabilities in Appendix B.

4 Experiments

4.1 Experiment Setup

Training Setup. Training reals are 53,725 SA-1B photographs at native resolution (Kirillov et al., 2023); fakes are the reconstructions of the same photographs by a single 4-channel decoder (SD 2.1), after one and two passes. No image of SDXL or of any decoder with 16 or more latent channels is used in training, so every number on these decoders below is zero-shot. Following B-Free and DDA, our detector fine-tunes a DINOv2 ViT-B/14 (Oquab et al., 2024) with the resampled patch of Section 3.1; all optimization and augmentation settings are in Appendix B.

Evaluation Protocol. Every balanced accuracy uses the fixed threshold 0.5 (logit 0 for B-Free), as B-Free and DDA measure their competitors; AUROC and the false-positive rate on each benchmark's reals are reported alongside. Each image is scored on a single center crop, and no evaluation benchmark is used for checkpoint or threshold selection (Appendix B). We evaluate 67 generator sets: six public benchmarks, X-AIGD (Xiao et al., 2026), GenImage (Zhu et al., 2023), Synthbuster (Bammey, 2024), the two AlignedForensics sets (Rajan et al., 2025) and the GAN sets of ForenSynths (Wang et al., 2020), plus newT2I, six text-to-image models released after the baselines, and twelve reconstruction sets over the X-AIGD test photographs (Appendix A). Each set contributes 500 fake and 500 real images (479 pairs for newT2I), paired where pairs exist. A generator is recent when its latent decoder has 16 or more channels and no baseline saw it in training; eleven sets qualify (Table 1), the other 56 form the second group. Short names follow Table 4; (iPh.) marks the iPhone-style variant of a generator.

Baseline Methods. We compare against seven public detectors that train mainly on photographs and their reconstructions: AlignedForensics (Rajan et al., 2025), Stay-Positive (Rajan & Lee, 2025), B-Free (Guillaro et al., 2025), DDA (Chen et al., 2025), PROBE (Cao et al., 2026), DEAR (Kim et al., 2026) (DEAR-r checkpoint) and GenSign (Fei et al., 2026) (GenImage SD 1.4 checkpoint).

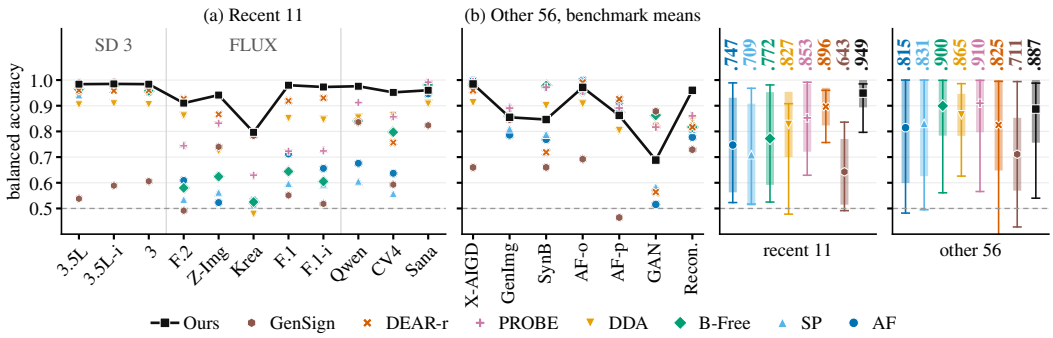

Figure 5: bAcc at 0.5, our detector (line) against seven public detectors (markers): (a) recent 11, (b) other 56, benchmark means; right: mean, ±1 sd, range. Names as in Table 1; GenImg: GenImage; SynB: Synthbuster; AF-o, AF-p: AlignedForensics original and post-processed; Recon.: reconstructions of X-AIGD test photographs.

Table 1: Recent generators: bAcc at 0.5 per generator, mean AUROC and FPR on the real photographs. Backbones in parentheses with the patch size after the slash; best in bold, second best underlined. Our mean carries the standard deviation over three seeds. Gray rows apply the resampled patch at inference to the released weights; †, ‡, § mark input sizes (Appendix B). AF, SP: AlignedForensics, Stay-Positive; F.: FLUX; Z-Img: Z-Image; CV4: CogView4; -i: iPhone-style.

	SD 3 lineage			FLUX lineage										
Model	3.5L	3.5L-i	3	F.2	Z-Img	Krea	F.1	F.1-i	Qwen	CV4	Sana	Mean	AUROC	FPR
Official checkpoints														
AF (R50)	.970	**.989**	.968	.609	.523	.531	.712	.656	.676	.637	.946	.747	.920	.001
SP (R50)	.942	.968	.954	.535	.563	.517	.597	.595	.606	.558	.967	.709	.956	.002
B-Free (DINOv2-B/14)	.964	.976	.958	.580	.624	.525	.644	.605	.841	.797	.981	.772	.827	.035
+ resampled patch†	.886	.895	.890	.648	.779	.604	.799	.803	.852	.753	.898	.801	.887	.206
DDA (DINOv2-L/14)	.904	.908	.905	.861	.723	.478	.851	.845	.853	.863	.907	.827	.900	.180
+ resampled patch‡	.926	.941	.932	.819	.841	.597	.902	.888	.862	.852	.941	.864	.934	.107
PROBE (DINOv2-L/14)	**.990**	**.989**	**.987**	.744	.832	.629	.723	.724	.912	.857	**.992**	.853	.952	.014
+ resampled patch‡	.983	.983	.973	.601	.841	.611	.705	.735	.948	.787	**.992**	.833	.934	.009
DEAR-r (R50)	.960	.958	.960	**.927**	.866	.787	.919	.930	.839	.757	.957	.896	.958	.080
GenSign (CLIP-L/14)	.538	.589	.606	.492	.740	.783	.551	.518	.836	.593	.824	.643	.712	.245
+ resampled patch§	.715	.695	.718	.656	.738	.782	.668	.615	.754	.553	.756	.695	.762	.382
Our method (DINOv2-B/7), by training photographs														
COCO 50k	.950	.948	.947	.913	.838	.805	.937	.921	.878	.819	.939	.900	.959	.095
COCO 25k + LSUN 25k	.941	.940	.944	.896	.890	.788	.935	.922	.921	.909	.927	.910	.974	.117
SA-1B (Ours)	.984	.985	.984	.910	**.942**	**.796**	**.980**	**.973**	**.976**	**.952**	.960	**.949**±.007	**.980**	.027

Every baseline runs from its official checkpoint with its own preprocessing and threshold, without calibration (Appendix B). We additionally retrain AlignedForensics, Stay-Positive and DDA with their official recipes on our SA-1B photographs (Appendix C).

4.2 Experimental Results

Results on recent generators. Table 1 compares our detector with seven public detectors on the eleven recent sets. Our detector attains the best balanced accuracy on six of the eleven sets and is within .006 of the best on the three SD 3 sets; its mean of .949 exceeds DEAR-r, the strongest baseline, by 5.3 points and DDA by 12.2 points. Over the four earlier detectors the largest margins are on the FLUX lineage (FLUX.1 .980 against .851, Z-Image .942 against .723, Krea .796 against .531), where most of them are near chance at the fixed threshold. PROBE and DEAR-r narrow this gap and lead on the SD 3 lineage, Sana and FLUX.2 by at most .032; GenSign, trained on one older generator, is near chance on seven of the eleven sets. The standard deviation over the eleven sets is .055 for our detector and .073 to .199 for the baselines, so the gain is not concentrated in a few sets, and every number of our detector is zero-shot. Krea is our weakest set although it shares the FLUX.1 decoder, which we detect at .980 on FLUX.1 and at .932 on its reconstructions; the deficit therefore does not come from the decoder residual, and we leave the cause open.

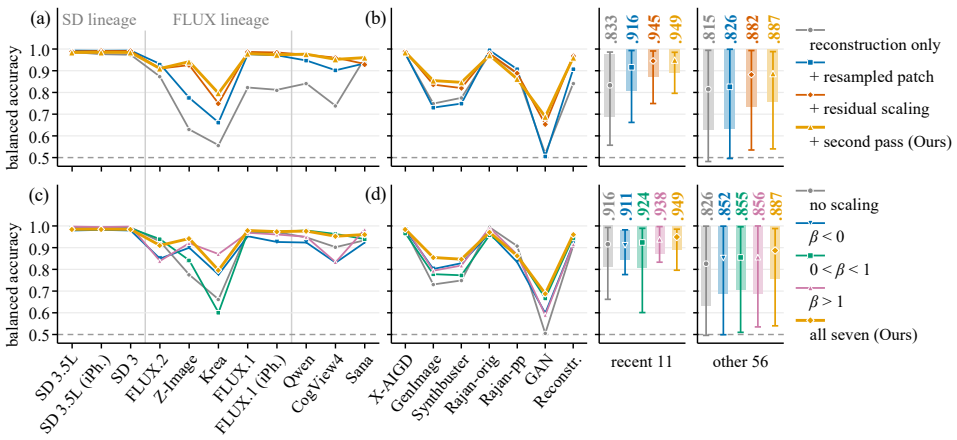

Figure 6: Ablation, first seed. (a) bAcc per recent generator as the components of Table 2 are added one at a time and (b) the same models on the other 56 sets, mean per benchmark; (c, d) the same for one training range of β at a time. Right of each row: mean, ± 1 sd and range.

Results on the other generators. On the other 56 sets our detector reaches a mean of .887, below PROBE (.910) and B-Free (.900) and above the five other baselines (.711 to .865); Table 8 lists every generator and Figure 5b shows the means per benchmark. Our detector leads on the reconstruction sets (.960 against .861) and is within .04 of the best baseline on the older X-AIGD generators, GenImage and the original AlignedForensics set; it trails on the post-processed AlignedForensics set (.862 against .926 for DEAR-r), Synthbuster (.846 against .977 for B-Free) and the GAN sets (.688 against .879 for GenSign). The deficit lies in generators without a learned latent decoder or with an undisclosed pipeline, GauGAN, BigGAN, DALL-E 3 and Firefly (.54 to .62), which are outside the target of this work. Our mean false-positive rate on the reals is .043, below B-Free (.060) and DDA (.165). Each baseline has its own strength, PROBE on the older and more varied generators, DEAR-r on the AlignedForensics sets and GenSign on the GAN sets, and is weak elsewhere. The gap on the recent sets is not a matter of training photographs: retraining AlignedForensics, Stay-Positive and DDA on our SA-1B photographs with their own autoencoders (Table 8) moves them to .822, .806 and .822 on the recent sets, and none reaches our .949; retrained Stay-Positive has a false-positive rate of .367 and falls to .665 on the other 56.

Resampled patch at inference and training photographs. The gray rows of Table 1 apply the resampled patch at inference to the released weights, without retraining. DDA rises from .827 to .864 on the recent sets and its false-positive rate falls from .180 to .107; B-Free and GenSign rise by .03 and .05 at false-positive rates of .206 and .382; PROBE, fine-tuned on 336-pixel tiles, falls by .02. On the other 56 sets only DDA gains, the others being tuned to the pretrained kernel. The substitution points in the direction of Section 3.1; the gain is realized in training, where our detector reaches .949 and .887 on the recent and the other sets. The last block trains our method on the photographs of DDA (COCO 50k) and on photographs matching AlignedForensics (COCO 25k + LSUN 25k): .900 and .910, above every baseline, so the lead does not require SA-1B, which adds a further .04 to .05. With the SD 1.4 autoencoder of AlignedForensics and Stay-Positive in place of SD 2.1, our method reaches .942 (Appendix D.1), so each baseline is exceeded with the autoencoder it was trained on.

Robustness to post-processing. Under JPEG compression, Gaussian blur and downscaling applied to real and generated images alike, our detector and DEAR-r are tied over ten conditions (.826 and .828), each leading in four, with DEAR-r at close to three times our false-positive rate on clean images (.082 against .030). The loss under post-processing comes from missed generated images; our false-positive rate stays below .09 in every condition (Appendix D.4).

4.3 Ablation and Diagnostic Analysis

Ablation studies. Table 2 adds the three components of our method one at a time, every cell the mean over three seeds; Figure 6a, b shows the same models per generator. Each component does

Table 2: Ablation: components added one at a time; every cell is the mean over three training seeds, with the standard deviation in the summary columns. Recent 11: mean, standard deviation and minimum of bAcc at 0.5. Other 56: mean per benchmark and over the 56 sets. Benchmark abbreviations as in Figure 5. FPR on the reals of all benchmarks.

	Recent 11			Other 56								
Training	Mean	Std	Min	X-AIGD	GenImg	SynB	AF-o	AF-p	GAN	Recon.	Mean	FPR
Reconstruction only	.830±.004	.146±.001	.554±.004	.978	.752	.778	.969	.888	.517	.841	.816±.004	.015±.002
+ resampled patch	.921±.006	.102±.006	.670±.024	.978	.730	.747	**.988**	**.900**	.509	.906	.824±.002	**.006±.000**
+ residual scaling	.947±.002	.064±.007	.774±.025	**.989**	.835	.831	.978	.885	.663	**.969**	.884±.002	.039±.004
+ second pass (Ours)	**.949±.007**	**.055±.004**	**.802±.015**	.986	**.849**	**.843**	.973	.870	**.700**	.964	**.889±.007**	.042±.001

Table 3: Range of the residual scaling: each model trains with the listed β values only. bAcc at 0.5, mean and minimum over the recent sets, FPR on their reals, and the five recent sets where the models differ most; first seed.

		Recent 11			Per generator				
Training β		Mean	Min	FPR	Krea	Z-Image	CogView4	FLUX.2	Qwen
none (one pass)		.916	.662	.005	.662	.776	.902	.930	.947
$\beta < 0$	$\{-.5, -.25\}$	.911	.776	.028	.776	.899	.831	.849	.924
$0 < \beta < 1$	$\{.25, .5, .71\}$	.924	.601	.021	.601	.841	**.963**	**.939**	**.978**
$\beta > 1$	$\{1.25, 1.8\}$	.938	**.833**	**.004**	**.871**	.921	.833	.839	.949
all seven (Ours)		**.949**	.796	.027	.796	**.942**	.952	.910	.976

a different job. The resampled patch raises the mean on the recent sets from .830 to .921 and by only $+.008$ on the other 56: it helps where the residual is small and fine. Residual scaling raises the recent mean to .947 and, beyond the mean, lowers the standard deviation from .102 to .064 and lifts the weakest set from .670 to .774, at a false-positive rate of .039. The second pass adds $+.002$ to the mean, within the spread over seeds, but lowers the standard deviation to .055 and raises the weakest set, Krea, to .802 and the GAN sets to .700; we keep it for uniformity, not for the mean. A third pass, another training autoencoder or another pooling change the recent mean by less than .01 (Appendix D.1). The components also depend on each other: without scaling the model responds only near the training amplitude (Section 3.2), whereas with scaling the recent decoders score .958 at $\beta = 1$ and, outside the training band, .980 and .958 at 2.8 and -1; without the resampled patch the full recipe closes only half of the sweep gap and trails under every post-processing (Appendix D.4).

Range of the residual scaling. Table 3 trains the detector with one range of β at a time (Figure 6c, d). No range alone reproduces the gain: the three models reach .911 to .938 against .916 without scaling (one pass) and .949 with the full set, because each range covers different generators. The negative values matter most: only the model trained on them responds at both signs of β on the sweep, so they remove the dependence on the sign of the residual, at the cost of the false-positive rate (.028 against .004 for the values above 1; Appendix D.2).

5 Conclusion

We proposed Residual Family Augmentation for detecting images from recent latent generators, together with the patch-size mismatch it builds on. A detector trained on reconstructions reads the cheapest cue for its label, the amount of damage or the signed pattern of one decoder, and both fail on recent generators, whose residuals are the faintest and the least similar to the training one. The mismatch compounds the failure: the residual of a recent generator concentrates at scales the pretrained filters attenuate, and Patch Resampling, adding no parameter, returns it within reach. On the resampled patch, the augmentation adds the residual back at multiple magnitudes and both signs, all labeled fake, plus a second reconstruction pass, so the detector responds across the residuals of recent decoders. Trained on a single older autoencoder, our detector holds where every baseline falls short, while remaining competitive across public benchmarks. The Residual Scaling Sweep needs no generated images; we recommend it as a standard check for reconstruction-trained detectors. Appendix E discusses the limitations of our method.

AI USE STATEMENT

In this work, we used generative AI tools to polish the English text, correct grammar, and improve readability, and for retrieval and discovery, namely searching for and summarizing related work. We also used them in research ideation and execution: for feedback on experimental design and on the interpretation of results in discussion with the authors, and for writing and debugging the training, evaluation, and analysis code, the figures, and the LaTeX tables. We have not used generative AI tools for generating datasets beyond the images of the generators under study, or for proving mathematical claims, which this work does not make. We have reviewed all AI-assisted work. Every experimental number in this paper was produced by runs the authors launched and logged, AI-assisted code was checked by the authors on held-out data and by inspection of its outputs, and every reference was checked against the original paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies the detection of AI-generated images, which supports the integrity of visual media. All real images come from publicly released datasets under their respective licenses; no new data involving human subjects were collected. Detectors of this kind can produce false accusations, and we report false-positive rates on real photographs throughout so that deployment thresholds can be chosen with that risk in view.

REPRODUCIBILITY STATEMENT

Section 3.1 defines Patch Resampling, Section 3.2 the diagnostic, Section 3.3 the augmentation, and Appendix B every training setting. Data construction (real sources, minimum-size filters, no resampling, leakage checks against evaluation sets by file hash and identifier) is given in Appendix A. Per-image scores for every benchmark, the fixed evaluation manifests, and the training identifier lists will be released with the code.

REFERENCES

Jiawang Bai, Li Yuan, Shu-Tao Xia, Shuicheng Yan, Zhifeng Li, and Wei Liu. Improving vision transformers by revisiting high-frequency components. In *European Conference on Computer Vision (ECCV)*, 2022.

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

Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. FlexiViT: One model for all patch sizes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Qianshu Cai, Chao Wu, Yonggang Zhang, Jun Yu, and Xinmei Tian. Towards generalizable detector for generated image. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zijie Cao, Weijie Tu, Yao Xiao, Weijian Deng, Weiyan Chen, Liang Lin, and Pengxu Wei. Where detectors fail: Probing generative space for generalizable AI-generated image detection. In *International Conference on Machine Learning (ICML)*, 2026.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *International Conference on Machine Learning (ICML)*, 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 (NeurIPS)*, 2025.

Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of AI-generated image detection with CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Jianwei Fei, Yunshu Dai, Xiaoyu Zhou, Zhihua Xia, and Alessandro Piva. Enabling supervised learning of generative signatures for generalized AI-generated images detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning (ICML)*, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or JPEG? revealing common biases in generated image detection datasets. In *European Conference on Computer Vision Workshops (ECCVW)*, 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Qing Huang, Zhipei Xu, Xuanyu Zhang, Xiangyu Yu, and Jian Zhang. ReAlign: Generalizable image forgery detection via reasoning-aligned representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution AI-generated image detection by spectral learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Dahye Kim, Jaehyun Choi, Hyun Seok Seong, Seongho Kim, Donghun Lee, Sungwon Yi, and Jang-Ho Choi. Dissect and prune: Enhancing robustness in AI-generated image detection. In *International Conference on Machine Learning (ICML)*, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision (ECCV)*, 2024.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025.

Yanzhu Liu, Xiao Liu, Yuexuan Wang, et al. Exploiting the final component of generator architectures for AI-generated image detection. *arXiv preprint arXiv:2601.20461*, 2026.

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 (ICLR)*, 2026.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations (ICLR)*, 2022.

Xinyi Qi, Kai Ye, Chengchun Shi, Ying Yang, Jin Zhu, and Hongyi Zhou. A difference-in-difference approach to detecting AI-generated images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Anirudh Sundara Rajan and Yong Jae Lee. Stay-positive: A case for ignoring real image features in fake image detection. In *International Conference on Machine Learning (ICML)*, 2025.

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 (ICLR)*, 2025.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the Fourier domain analysis: From theory to practice. In *International Conference on Learning Representations (ICLR)*, 2022.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Xinyuan Wang, Yingxin Lai, Zhiming Luo, and Zhihui Liu. PPM-CLIP: Probabilistic prompt modeling for generalizable AI-generated image detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yao Xiao, Weiyan Chen, Jiahao Chen, Zijie Cao, Weijian Deng, Binbin Yang, Ziyi Dong, Xiangyang Ji, Wei Ke, Pengxu Wei, and Liang Lin. Unveiling perceptual artifacts: A fine-grained benchmark for interpretable AI-generated image detection. In *International Conference on Learning Representations (ICLR)*, 2026.

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 (ICLR)*, 2025.

Zheng Yang, Ruoxin Chen, Zhiyuan Yan, Ke-Yue 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 (ICLR)*, 2026.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2019.

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 (NeurIPS) Datasets and Benchmarks Track*, 2023.

A DATA CONSTRUCTION

Real images. Our training reals are photographs from SA-1B (Kirillov et al., 2023), drawn from twelve archive shards. Each photograph is kept at its native resolution (short side about 1,500 pixels), and a 512×512 center crop is taken without any resampling; photographs with a short side below 512 are dropped. The pool has 53,725 images, of which 200 are held out for validation. No X-AIGD real image is used for training.

Reconstructions. Each real crop is encoded and decoded once ($\times 1$) and twice ($\times 2$) with the SD 2.1 autoencoder (Rombach et al., 2022), in 32-bit precision, and stored as PNG so that no compression is applied between steps.

The SD 2.1 autoencoder. SD 2.1 is the autoencoder shipped with Stable Diffusion 2.1, distributed as `sd-vae-ft-mse`; it is the autoencoder of the training data of B-Free, DDA and PROBE, so three of the seven baselines share our training decoder.

Generators and autoencoders. Table 4 lists every generator and autoencoder that this paper names, with the name used in the tables and the source of the weights. The six newT2I generators are run with the public pipeline of each model at about one megapixel, with the X-AIGD prompt of the same identifier and a fixed seed; the images are stored as PNG. The X-AIGD generators keep the names of Xiao et al. (2026): SD 3.5L is Stable Diffusion 3.5 Large, HYDiT 1.2 is HunyuanDiT v1.2, the two iPhonePhoto sets apply a LoRA to FLUX.1 and SD 3.5L, and SDXL RE (Realism Engine) and SD 1.5 RV (Realistic Vision) are community fine-tuned checkpoints.

Table 4: Generators and autoencoders named in this paper. Latent: number of latent channels and spatial reduction factor.

Name in tables	Model	Weights	Latent
newT2I generators			
CogView4	CogView4-6B	`THUDM/CogView4-6B`	16, $f8$
Krea	FLUX.1 Krea [dev]	`black-forest-labs/FLUX.1-Krea-dev`	16, $f8$ (FLUX.1 VAE)
FLUX.2	FLUX.2 [klein] 4B, distilled	`black-forest-labs/FLUX.2-klein-4B`	32, $f16$ (FLUX.2 VAE)
Qwen	Qwen-Image	`Qwen/Qwen-Image`	16, $f8$
Sana	Sana 1.5 4.8B, 1024px	`Efficient-Large-Model/SANA1.5_4.8B_1024px`	32, $f32$ (DC-AE)
Z-Image	Z-Image-Turbo	`Tongyi-MAI/Z-Image-Turbo`	16, $f8$ (FLUX.1 VAE)
Reconstruction autoencoders			
SD 2.1 (training)	Stable Diffusion 2.1 VAE	`stabilityai/sd-vae-ft-mse`	4, $f8$
SD 1.4	Stable Diffusion 1.x VAE	`CompVis/ldm-text2im-large-256` (identical)	4, $f8$
SDXL	SDXL VAE	`stabilityai/sdxl-vae`	4, $f8$
FLUX.1	FLUX.1 [dev] VAE	`black-forest-labs/FLUX.1-dev`	16, $f8$
FLUX.2	FLUX.2 VAE	shipped with FLUX.2 [klein] 4B	32, $f16$
CogView4	CogView4 VAE	shipped with CogView4-6B	16, $f8$
Qwen	Qwen-Image VAE	shipped with Qwen-Image	16, $f8$
DC-AE	DC-AE f32c32 (Sana 1.1)	`mit-han-lab/dc-ae-f32c32-sana-1.1`	32, $f32$
VQ-8, VQ-16	LlamaGen tokenizers	`FoundationVision/LlamaGen`	VQ, $f8$ and $f16$
Emu3 (VQ)	Emu3 vision tokenizer	`BAAI/Emu3`	VQ, $f8$
Real-ESRGAN 4×	Real-ESRGAN x4plus (control, no autoencoder)	`RealESRGAN_x4plus`	–

Evaluation sets. Table 5 lists the eight evaluation sets with the source of the generated and of the real images. Two of the sets are ours, newT2I and the reconstruction sets; the six others are public benchmarks used as released. The eleven recent sets of Table 1 come from two sources: five from X-AIGD (SD 3, SD 3.5L, SD 3.5L iPh., FLUX.1, FLUX.1 iPh.) and six from newT2I. Within a set, the scored images are the first 500 in sorted order of the paired identifiers when the real and the generated images are paired, and otherwise the first 500 real and the first 500 generated images in sorted file order; no random draw is involved. The only sets with fewer than 500 pairs are the six newT2I sets (479, see below).

newT2I. We generated the six newT2I sets with the public diffusers pipeline of each model in bfloat16, at the number of steps and the guidance scale recommended by each model card (Table 6), for the first 500 identifiers of the X-AIGD test split in sorted order. The prompt is the `original_prompt` field of the X-AIGD test annotation of that identifier, a descriptive caption in sentence form. The resolution is the size of the X-AIGD image of the same identifier that is closest to 1.04 megapixels, rounded to the grid of the model (about one megapixel; for example 1344×768). The seed is 42 plus the index of the identifier. The images are stored as PNG. The paired real image

Table 5: Evaluation sets. Pairs: the real image is the photograph from which the generated image or the reconstruction was made.

Set	Gen.	Generated images	Real images	Images per set
X-AIGD (Xiao et al., 2026)	13	X-AIGD test split	X-AIGD test photograph of the same identifier	500 pairs
newT2I (ours)	6	generated by us from the X-AIGD prompt of the same identifier	X-AIGD test photograph of the same identifier	479 pairs
Reconstructions (ours)	12	one pass of each of eleven autoencoders over the X-AIGD test photographs, and a Real-ESRGAN control	the same photograph	500 pairs
GenImage (Zhu et al., 2023)	8	GenImage test split	ImageNet photographs of the same generator folder	500 / 500
Synthbuster (Bammey, 2024)	9	Synthbuster	RAISE-1k, shared by the nine generators	500 / 500
AlignedForensics, original and post-processed (Rajan et al., 2025)	6 + 6	the LDM sets of the benchmark	the real images of the benchmark, original and post-processed	500 / 500
ForenSynths GAN (Wang et al., 2020)	7	ForenSynths test split	the real images of each GAN folder	500 / 500

is the X-AIGD test photograph of the same identifier; 479 of the 500 identifiers have a photograph on disk, so every newT2I set has 479 pairs. The images were generated in July and August 2026.

Table 6: Generation settings of the newT2I sets (diffusers pipelines, bfloat16).

Model	Pipeline	Steps / guidance	Resolution grid
CogView4-6B	CogView4Pipeline	50 / 3.5	32
FLUX.2 [klein] 4B	Flux2KleinPipeline	4 / 1.0	32
FLUX.1 Krea [dev]	FluxPipeline	28 / 4.5	16
Qwen-Image	QwenImagePipeline	50 / 4.0 (true CFG)	16
Sana 1.5 4.8B	SanaPipeline	20 / 4.5	32
Z-Image-Turbo	ZImagePipeline	9 / 0	16

Reconstruction sets. The twelve reconstruction sets use the X-AIGD test photographs with a short side of at least 512 pixels (1,605 of the 2,009 test photographs on disk), at a 512×512 center crop without resampling. Each crop passes once through one of eleven autoencoders (SD 2.1, SD 1.4, SDXL, FLUX.1, FLUX.2, CogView4, Qwen-Image, DC-AE, Emu3, VQ-8, VQ-16) and is stored as PNG. The twelfth set is a control with a learned upsampler but no autoencoder, Real-ESRGAN $4\times$: the crop is downscaled by four with a bicubic filter and upscaled by Real-ESRGAN x4plus back to 512 pixels (PSNR 22.1 dB to the crop). Each set is scored on the first 500 photographs in sorted order, paired with the crop itself as the real image. None of these photographs is used in training.

Leakage. Before every training run the identifier set of the training reals is intersected with the identifier and file-hash set of every evaluation benchmark, and the run starts only if the intersection is empty. The residual scaling sweep uses real photographs from the X-AIGD test split, none of which appears in training.

B Implementation details and evaluation protocol

Data. The training reals are 53,725 SA-1B photographs (Kirillov et al., 2023). Each photograph keeps the native resolution, and we take a 512×512 center crop without resampling. We reconstruct each crop once and twice with the SD 2.1 autoencoder in 32-bit precision and store the results as

PNG (Appendix A). We exclude 200 photographs and train on the remaining 53,525, that is 160,575 items. No image produced by a generator is used for training.

Network. Following B-Free and DDA, the detector fine-tunes a DINOv2 vision transformer (Oquab et al., 2024). We use the ViT-B/14 variant with four registers, the backbone of B-Free, and train all of its parameters (86.2M after Patch Resampling). Patch Resampling (Section 3.1) resamples the pretrained 14×14 patch-embedding filters to 7×7 by bicubic interpolation, multiplies the filters by four so that the response to a constant input is unchanged, a simpler variant of the pseudo-inverse resize of FlexiViT (Beyer et al., 2023), and applies them with stride 7; the positional embedding is interpolated to the 32×32 token grid. We write ViT-B/7 for this backbone. The classifier reads the class token of the last block after the final layer normalization and applies one linear layer ($768 \rightarrow 1$) that outputs the logit; the same pooling is used in training and in inference. Averaging the class tokens of blocks 4, 6, 8 and 12 (Koutlis & Papadopoulos, 2024) in place of the final class token changes the mean balanced accuracy by less than .005 (Appendix D.1).

Loss. The label is 0 for a real item and 1 for every fake item: the reconstruction, the second pass, and every scaled image. The network outputs one logit, and the loss is the binary cross-entropy between the sigmoid of the logit and the label. Each photograph contributes one real item and two fake items, so a real item receives a weight of 2 and the two classes have equal mass. The decision threshold is 0.5 on the sigmoid output, that is a logit of 0.

Optimization. We use AdamW with a weight decay of 10^{-4} and layer-wise learning-rate decay with a factor of 0.75 over the 12 blocks: the head has a learning rate of 10^{-4}, the last block 3×10^{-5}, and the patch and positional embeddings 9.5×10^{-7}. The learning rate follows a cosine schedule to zero without warm-up. The batch size is 16. One epoch is a random draw of 60,000 items without replacement (3,750 steps), and training runs for 16 epochs (60,000 steps) with bfloat16 autocast and a fixed seed. The budget is fixed in advance; we apply no early stopping and no validation-based model selection, and the final weights are the weights after the last step. One run takes 2.6 hours on one RTX PRO 5000 GPU (48 GB), with gradient checkpointing; two runs that share a node with two RTX 6000 Ada GPUs take 3.7 hours each.

Augmentation. Augmentations apply in training only, in the following order. (1) Residual scaling (Eq. 4) on fake items, at 512×512 and in floating point; the clipping in Eq. 4 touches 2.8% of the channel values at $\beta = -0.5$ and 1.5% at $\beta = 1.8$, none for $0 < \beta \leq 1$, at pixels the photograph already saturates, and at most 1% of the values lose more than four levels (200 training photographs). (2) The sharpened-real control: with probability 0.3 a real item is replaced by its unsharp-masked version $a + s\,(a - G_\sigma(a))$, with G_σ a Gaussian blur of $\sigma \sim \mathcal{U}(1, 2)$ and $s \sim \mathcal{U}(0.5, 1.5)$, clipped to $[0, 255]$, and keeps the label real, so that sharpness alone does not mark an image as fake. (3) Symmetric downscaling of real and fake items with probability 0.5, by a factor in $[0.5, 0.9]$ and with a random interpolation kernel. (4) Downscaling of real items alone with probability 0.3 and the same factors. (5) A 224×224 center crop; with probability 0.2 the crop is instead taken at side $224/s$, $s \in [1.05, 3]$, and upscaled to 224. (6) JPEG re-encoding of real and fake items with probability 0.3: in half of these events a standard quantization table at quality in $[40, 90]$ with random chroma subsampling, otherwise a quality in $[60, 95]$. (7) A second JPEG pass with probability 0.2 at quality in $[60, 95]$. Every augmentation keeps the label. We apply no flip, rotation, blur, additive noise or color jitter to either class.

Inference. We score a single 224×224 center crop with the resampled patch, as in training; an image smaller than the crop is upscaled to the crop size.

Thresholds. Every balanced accuracy in this paper uses the fixed threshold 0.5 (logit 0 for B-Free) for every model; a deployed detector must commit to a threshold, so an output distribution that does not fit the threshold is a defect of the model, not of the measurement. No evaluation benchmark is used for checkpoint or threshold selection. The design choices of Section 3 (the resampling factor, the set of β values and the number of passes) were compared on the recent sets and are reported as ablations; the other 56 sets serve as the check that these choices are not specific to the recent group (Appendices C and D).

Evaluation. Each generator set is scored with 500 fake images and 500 real images of the same benchmark, 479 pairs for each newT2I set (Appendix A); on X-AIGD, newT2I, and the reconstruction sets each fake image is paired with the real photograph of the same identifier. All detectors, the baselines included, are scored by one pipeline on the same image lists.

Baselines. AlignedForensics, Stay-Positive, B-Free, and DDA run from the official checkpoints with the preprocessing of each method, at the fixed threshold and without calibration. The retrained rows of Table 8 use the public training code and recipe of each method; only the training set changes to our SA-1B photographs and their reconstructions by the autoencoder of each method (Appendix C).

Detectors published in 2026. In Table 1, backbones in parentheses are ResNet-50 (R50), DINOv2 ViT-B/14 and ViT-L/14 (DINOv2-B/L) and CLIP ViT-L/14 (CLIP-L); † marks 252-px windows, ‡ a 336-px crop or tiles and § a 224-px input. With the resampled kernel at inference, the five-window wrapper of B-Free keeps 36×36 tokens per window, so each window covers 252 pixels instead of 504 without resizing; DDA and PROBE keep their 336-pixel crop or tiles (48×48 tokens); GenSign keeps its 224-pixel input (32×32 tokens). PROBE runs from its released DINOv2-L checkpoint with its official protocol (non-overlapping 336 tiles, mean logit, threshold .5); we run it in fp16 and upscale images whose short side is below 336, as for every detector. DEAR runs from the DEAR-r checkpoint (the AlignedForensics ResNet-50 with its pruned channels) and GenSign from its GenImage-SD 1.4 dual-stream checkpoint (CLIP ViT-L/14 with LoRA and an EfficientNet-B0 on the extracted signature), each with its own preprocessing and threshold and with the label convention of its released code.

Compute. Table 7 lists the inference cost of the resampled patch on one RTX 6000 Ada GPU at batch size 32 (FLOPs at batch 1, fp32). The resampled patch slightly reduces the parameter count (86.2M against 86.6M) and quadruples the number of tokens.

Table 7: Compute per image. Throughput and peak memory in fp32 (fp16 in parentheses).

Model	Crop / patch	Tokens	GFLOPs	Images/s	Memory (GB)
DINOv2 ViT-B/14	224 / 14	256	47.1	430 (1,720)	0.73 (0.81)
DINOv2 ViT-B/7 (Ours)	224 / 7	1,024	214.1	91 (339)	1.78 (1.58)
DINOv2 ViT-B/14, 448 crop (Table 13)	448 / 14	1,024	214.8	88 (325)	1.89 (1.69)
DDA, DINOv2 ViT-L/14	336 / 14	576	381.9	53 (194)	2.17 (2.45)

Residual scaling sweep. The sweep uses 200 real photographs from the X-AIGD test split, none of which appears in training, and eight decoders (SD 2.1, SD 1.4, SDXL, FLUX.1, FLUX.2, CogView4, Qwen-Image, and LlamaGen VQ-16), with $\beta \in \{-1, -0.5, -0.25, 0, 0.25, 0.35, 0.5, 0.71, 0.85, 1, 1.25, 1.8, 2.8\}$. The amplitude response is the mean over decoders of $\mathrm{bAcc}(\beta{=}1) - \mathrm{bAcc}(\beta{=}0.25)$; the sign reading is the AUROC of $x_{-0.5}$ against x_0; the decoder spread is the range of bAcc over decoders, each taken at the grid point whose PSNR is closest to 27 dB. For a model trained with residual scaling the sign and spread readings are partly circular, since the model has seen scaled residuals of its own decoder in training; the $\beta = 1$ AUROC on decoders absent from training is not, and it is the number we treat as evidence.

C Additional results

Retraining the baselines on our photographs. AlignedForensics, Stay-Positive and DDA are retrained with their own public code and recipe, changing only the training set: the real images are our SA-1B photographs, and the fake images are the reconstructions of these photographs by the autoencoder of each method (the LDM autoencoder, identical to the SD 1.4 autoencoder, for AlignedForensics and Stay-Positive; SD 2.1 for DDA), produced with our reconstruction protocol (Appendix A); B-Free is not retrained because its fakes come from an inpainting pipeline rather than a reconstruction. Tables 8–10 give bAcc and AUROC for every generator of the eight evaluation sets (six public benchmarks, newT2I, and the reconstruction sets).

Frozen probes by block. Table 11 gives the linear probes at every block that we read, with the original and the resampled patch. The probes are trained on 3,000 SA-1B photographs and the SD 2.1 reconstructions of these photographs at a 224 input, without augmentation, and scored on the eleven recent sets and the other 56 sets at threshold 0.5; an evaluation image smaller than the crop is upscaled to the crop, as in the main evaluation (Appendix B). The resampled patch raises the

Table 8: Per-generator results for every generator of the eight evaluation sets (six public benchmarks, newT2I, and the reconstruction sets): bAcc at 0.5 and AUROC; 500 real and 500 paired fake images per generator, fewer where unavailable; B-Free at its native logit threshold 0, all others at 0.5. Ours: the model of Table 1 (first seed). "SA-1B" rows are the baseline above them retrained with its own code and recipe on our SA-1B photographs and the reconstructions of these photographs by the autoencoder of the method. Best in bold, second best underlined; ties share the formatting.

X-AIGD	**FLUX.1**		**FLUX.1 iPh.**		**HunyuanDiT**		**Infinity**		**Lumina**		**PixArt-α**		**SDXL**	
Method	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC
AlignedForensics	.712	.945	.656	.941	**.994**	**1.00**	**.999**	**1.00**	.997	**1.00**	**.998**	.998	.994	**1.00**
SA-1B	.846	.991	.756	.970	.992	**1.00**	.998	**1.00**	**.998**	**1.00**	.997	**1.00**	**.997**	**1.00**
Stay-Positive	.597	.948	.595	.966	.989	**1.00**	**.999**	**1.00**	.995	.998	.997	**1.00**	**.997**	**1.00**
SA-1B	.816	.983	.817	.985	.819	**1.00**	.819	**1.00**	.818	.998	.819	**1.00**	.819	**1.00**
B-Free	.644	.774	.605	.680	.984	**1.00**	.983	**1.00**	.984	**1.00**	.984	**1.00**	.984	**1.00**
DDA	.851	.932	.845	.929	.911	.995	.913	.998	.912	.997	.912	.996	.912	.998
SA-1B	.767	.881	.693	.852	.959	.998	.959	.998	.961	**1.00**	.961	.999	.960	.999
PROBE	.723	.917	.724	.905	.993	**1.00**	.993	**1.00**	.993	**1.00**	.993	**1.00**	.993	**1.00**
DEAR-r	.919	.974	.930	.982	.959	**1.00**	.960	**1.00**	.960	.999	.960	**1.00**	.960	**1.00**
GenSign	.551	.621	.518	.587	.611	.687	.744	.811	.543	.606	.672	.758	.605	.693
Ours	**.980**	**.998**	**.973**	**.995**	.981	.998	.987	**1.00**	.988	**1.00**	.976	.995	.986	.999

X-AIGD (cont.)	**SDXL RE**		**SD 1.4**		**SD 1.5 RV**		**SD 3**		**SD 3.5L**		**SD 3.5L iPh.**		**Mean**	
Method	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC
AlignedForensics	.997	**1.00**	**.999**	**1.00**	**.999**	**1.00**	.968	.996	.970	.999	.989	.994	.944	.990
SA-1B	.998	**1.00**	.998	**1.00**	.998	**1.00**	**.994**	.999	**.998**	**1.00**	**.992**	.996	.966	.997
Stay-Positive	**.999**	**1.00**	**.999**	**1.00**	**.999**	**1.00**	.954	.999	.942	.998	.968	.995	.925	.993
SA-1B	.819	**1.00**	.819	**1.00**	.819	**1.00**	.819	.999	.819	**1.00**	.817	.996	.818	.997
B-Free	.984	**1.00**	.982	**1.00**	.982	.999	.958	.989	.964	.992	.976	.997	.924	.956
DDA	.912	.998	.910	.991	.912	.997	.905	.987	.904	.985	.908	.994	.901	.984
SA-1B	.961	.999	.959	.996	.958	.997	.948	.992	.954	.994	.957	.997	.923	.977
PROBE	.993	**1.00**	.993	**1.00**	.993	**1.00**	.987	.999	.990	.999	.989	**.999**	.951	.986
DEAR-r	.960	**1.00**	.960	.999	.960	**1.00**	.960	**1.00**	.960	**1.00**	.958	**.999**	.954	.996
GenSign	.572	.652	.808	.878	.723	.790	.606	.680	.538	.611	.589	.656	.622	.695
Ours	.988	**1.00**	.985	.999	.988	**1.00**	.984	.999	.984	.998	.985	**.999**	**.983**	**.999**

newT2I	**CogView4**		**FLUX.2**		**Krea**		**Qwen**		**Sana**		**Z-Image**		**Mean**	
Method	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC
AlignedForensics	.637	.879	.609	.859	.531	.796	.676	.879	.946	.983	.523	.851	.654	.874
SA-1B	.682	.951	.840	.970	.721	.970	.710	.958	.968	.995	.532	.883	.742	.955
Stay-Positive	.558	.909	.535	.925	.517	.926	.606	.925	.967	.999	.563	.922	.624	.934
SA-1B	.733	.842	.814	**.995**	**.813**	**.989**	.800	.951	.812	.990	.811	.957	.797	.954
B-Free	.797	.892	.580	.619	.525	.529	.841	.928	.981	**1.00**	.624	.701	.725	.778
DDA	.863	.951	.861	.944	.478	.425	.853	.943	.907	.997	.723	.817	.781	.846
SA-1B	.928	.980	.832	.925	.480	.487	.909	.975	.955	.999	.616	.771	.787	.856
PROBE	.857	.967	.744	.896	.629	.838	.912	.983	**.992**	**1.00**	.832	.966	.828	.942
DEAR-r	.757	.861	**.927**	.974	.787	.911	.839	.885	.957	.996	.866	.953	.856	.930
GenSign	.593	.671	.492	.509	.783	.870	.836	.909	.824	.915	.740	.808	.711	.780
Ours	**.952**	**.983**	.910	.962	.796	.870	**.976**	**.996**	.960	.989	**.942**	**.987**	**.923**	**.965**

GenImage	**ADM**		**BigGAN**		**GLIDE**		**MJ**	
Method	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC
AlignedForensics	.522	.731	.524	.497	.630	.751	.967	.995
SA-1B	.553	.576	.560	.605	.684	.849	.985	**.999**
Stay-Positive	.517	.850	.556	.722	.595	.862	.981	**.999**
SA-1B	.541	.600	.554	.678	.568	.714	.621	.956
B-Free	.701	.864	.688	.957	.739	.905	.939	.985
DDA	**.835**	**.965**	.672	.952	.818	.967	.954	.994
SA-1B	.832	.943	.865	.952	**.898**	.965	.933	.980
PROBE	.730	.954	.723	.972	.763	**.987**	**.988**	**.999**
DEAR-r	.632	.877	.594	.746	.683	.860	.959	.989
GenSign	.623	.883	**.964**	**.992**	.888	.976	.824	.967
Ours	.709	.881	.549	.818	.799	.951	.973	.998

Table 9: Table 8, continued.

GenImage (cont.)	SD 1.4		SD 1.5		VQDM		Wukong		Mean	
Method	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC
AlignedForensics	.996	**1.00**	.998	**1.00**	.654	.922	**1.00**	**1.00**	.786	.862
SA-1B	.997	**1.00**	.992	**1.00**	.715	.893	.993	**1.00**	.810	.865
Stay-Positive	**.998**	**1.00**	**1.00**	**1.00**	.835	.988	.999	**1.00**	.810	.928
SA-1B	.648	**1.00**	.655	**1.00**	.632	.887	.624	**1.00**	.605	.854
B-Free	.986	.999	.985	.999	.846	.963	.995	**1.00**	.860	.959
DDA	.981	.999	.986	.998	.626	.897	.986	**1.00**	.857	.972
SA-1B	.955	.993	.967	.996	.736	.910	.966	.998	**.894**	.967
PROBE	.995	**1.00**	.997	**1.00**	.931	**.996**	.997	**1.00**	.890	**.988**
DEAR-r	.993	.999	.994	**1.00**	**.958**	.995	.995	**1.00**	.851	.933
GenSign	.913	.976	.923	.974	.727	.945	.905	.985	.846	.962
Ours	.974	.996	.968	.994	.890	.967	.978	.997	.855	.950

Synthbuster	DALL-E 2		DALL-E 3		Firefly		GLIDE		MJ v5	
Method	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC
AlignedForensics	.488	.533	.482	.200	.486	.475	.570	.669	.982	**1.00**
SA-1B	.454	.461	.429	.235	.437	.257	.632	.763	.929	.998
Stay-Positive	.500	.181	.500	.414	.561	.915	.550	.780	**1.00**	**1.00**
SA-1B	.467	.324	.410	.248	.438	.257	.501	.620	.535	.998
B-Free	**.949**	**.998**	.983	**1.00**	.997	**1.00**	.878	**.992**	.997	**1.00**
DDA	.855	.937	.897	.969	.916	.983	.782	.871	.934	.998
SA-1B	.927	.979	.938	.980	.939	.983	**.932**	.979	.974	.999
PROBE	.924	.996	**.996**	**1.00**	**.999**	**1.00**	.835	.963	.999	**1.00**
DEAR-r	.437	.286	.396	.137	.578	.681	.593	.623	.894	.995
GenSign	.469	.469	.620	.630	.675	.723	.703	.769	.594	.599
Ours	.734	.928	.610	.853	.622	.888	.793	.941	.981	**1.00**

Synthbuster (cont.)	SD 1.3		SD 1.4		SD 2		SDXL		Mean	
Method	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC
AlignedForensics	.981	.999	.982	**1.00**	.978	.997	.974	.997	.769	.763
SA-1B	.929	**1.00**	.929	**1.00**	.927	.997	.926	.994	.732	.745
Stay-Positive	**1.00**	**1.00**	**1.00**	**1.00**	.997	.998	.993	.997	.789	.809
SA-1B	.535	**1.00**	.535	**1.00**	.534	.996	.535	.994	.499	.715
B-Free	.999	**1.00**	.997	**1.00**	.997	**1.00**	**.999**	**1.00**	**.977**	**.999**
DDA	.929	.992	.926	.992	.932	.995	.934	.998	.901	.971
SA-1B	.969	.997	.966	.995	.975	.999	.975	**1.00**	.955	.990
PROBE	.999	**1.00**	.999	**1.00**	**.999**	**1.00**	.998	**1.00**	.972	.995
DEAR-r	.894	.976	.894	.976	.891	.974	.892	.996	.719	.738
GenSign	.761	.843	.753	.837	.731	.813	.635	.632	.660	.702
Ours	.960	.995	.959	.994	.980	**1.00**	.978	.999	.846	.955

AF-o	Kandinsky		LCM		MJ		PixArt		Playground		SD		Mean	
Method	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC
AlignedForensics	**.999**	**1.00**	.999	**1.00**	**.997**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**.999**	**1.00**
SA-1B	.996	**1.00**	.996	**1.00**	.995	**1.00**	.996	**1.00**	.996	**1.00**	.996	**1.00**	.996	**1.00**
Stay-Positive	**.999**	**1.00**	**1.00**	**1.00**	.996	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**.999**	**1.00**
SA-1B	.694	**1.00**	.694	**1.00**	.690	.983	.694	**1.00**	.694	**1.00**	.694	**1.00**	.693	.997
B-Free	.955	.999	.958	.999	.958	.999	.958	**1.00**	.958	**1.00**	.958	**1.00**	.958	.999
DDA	.908	.997	.909	.998	.906	.992	.909	.999	.909	.999	.899	.979	.907	.994
SA-1B	.815	.989	.817	.988	.814	.976	.817	.997	.817	.994	.760	.869	.807	.969
PROBE	.958	.997	.958	**1.00**	.952	.994	.958	**1.00**	.958	**1.00**	.955	.998	.956	.998
DEAR-r	.991	**1.00**	.991	**1.00**	.989	**1.00**	.991	**1.00**	.991	**1.00**	.991	**1.00**	.991	**1.00**
GenSign	.726	.788	.775	.817	.648	.708	.598	.703	.689	.738	.717	.772	.692	.754
Ours	.952	.985	.979	**1.00**	.973	.998	.979	**1.00**	.978	.999	.968	.997	.972	.996

Table 10: Table 8, continued.

AF-p	**Kandinsky**		**LCM**		**MJ**		**PixArt**		**Playground**		**SD**		**Mean**	
Method	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC
AlignedForensics	.897	.971	**.966**	.994	.873	.941	.908	.932	.876	.930	**.967**	**.995**	.914	.961
SA-1B	.707	.928	.726	.883	.759	.891	.854	.828	.697	.795	.801	.940	.757	.878
Stay-Positive	.845	.956	.934	.982	.792	.885	.898	.960	.809	.934	.933	.985	.868	.950
SA-1B	.591	.707	.675	.868	.564	.663	.693	.921	.674	.841	.619	.762	.636	.794
B-Free	.913	**.992**	.894	.970	.853	.921	.859	.939	.878	.956	.838	.902	.873	.947
DDA	.849	.966	.845	.965	.794	.894	.796	.906	.847	.972	.690	.788	.803	.915
SA-1B	.825	.949	.805	.932	.747	.843	.778	.906	.811	.952	.536	.573	.750	.859
PROBE	.905	.985	.912	**.998**	.838	.912	.892	**.976**	.908	**.991**	.892	.973	.891	.972
DEAR-r	**.936**	.984	.953	.996	**.897**	**.947**	**.913**	.962	.905	.962	.951	.994	**.926**	**.974**
GenSign	.490	.486	.460	.485	.452	.449	.511	.553	.428	.471	.447	.467	.465	.485
Ours	.788	.873	.927	.988	.824	.906	.891	.942	**.923**	.978	.821	.905	.862	.932

GAN	**BigGAN**		**CycleGAN**		**GauGAN**		**ProGAN**		**StarGAN**		**StyleGAN**		**StyleGAN2**		**Mean**	
Method	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC
AlignedForensics	.513	.376	.487	.296	.507	.431	.528	.846	.520	.794	.554	.620	.500	.866	.516	.604
SA-1B	.497	.291	.472	.131	.502	.344	.541	.671	.501	.274	.503	.618	.502	.520	.503	.407
Stay-Positive	.534	.579	.536	.684	.509	.542	.547	.917	.824	.931	.636	.759	.500	.809	.584	.746
SA-1B	.401	.367	.399	.345	.427	.409	.692	.812	.636	.696	.694	.799	.658	.737	.558	.595
B-Free	**.927**	**.987**	.810	.931	**.962**	.987	**.949**	.988	.840	.919	.784	.902	.772	.930	.863	.949
DDA	.843	.944	.701	.866	.909	.972	.902	.984	.659	.720	.874	.949	.848	.941	.819	.911
SA-1B	.925	.975	.819	.912	.929	.988	.809	**.991**	.864	.940	.883	.960	**.851**	.976	.869	**.963**
PROBE	.846	.983	.681	.919	.910	**.998**	.929	.987	.569	.618	**.952**	**.989**	.829	**.978**	.817	.925
DEAR-r	.569	.549	.534	.591	.539	.486	.631	.936	.648	.930	.521	.582	.505	.822	.564	.699
GenSign	.853	.934	**.944**	**.988**	.865	.984	.871	.890	**.994**	**.999**	.845	.934	.778	.863	**.879**	.942
Ours	.681	.901	.616	.731	.540	.674	.779	.909	.696	.809	.801	.910	.706	.811	.688	.821

Reconstructions	**CogView4**		**Emu3 (VQ)**		**FLUX.1**		**FLUX.2**		**VQ-16**		**VQ-8**	
Method	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC
AlignedForensics	.501	.507	.793	.922	.549	.794	.517	.604	.960	.990	.993	.998
SA-1B	.517	.560	.877	.988	.506	.678	.531	.677	.952	.996	**.995**	.998
Stay-Positive	.495	.563	**.961**	.989	.548	.739	.499	.618	**.989**	**1.00**	.988	.998
SA-1B	.670	.736	.828	**.994**	.677	.735	.702	.789	.830	.998	.830	.999
B-Free	.575	.672	.850	.943	.588	.702	.561	.653	.943	.990	.947	.990
DDA	.777	.861	.831	.918	.746	.823	.740	.813	.883	.975	.897	.993
SA-1B	.864	.944	.905	.969	.705	.845	.726	.858	.953	.990	.970	.998
PROBE	.668	.837	.909	.977	.679	.850	.566	.686	.961	.996	.962	.997
DEAR-r	.614	.660	.892	.982	.700	.764	.614	.647	.903	.999	.902	.999
GenSign	.728	.786	.793	.858	.638	.695	.635	.696	.659	.728	.632	.688
Ours	**.973**	**.995**	.926	.979	**.932**	**.985**	**.950**	**.984**	.981	.999	.987	**1.00**

Reconstructions (cont.)	**Qwen**		**DC-AE**		**SD 1.4**		**SD 2.1**		**SDXL**		**Real-ESRGAN** $4\times$		**Mean**	
Method	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC	bAcc	AUROC
AlignedForensics	.601	.848	.889	.965	.995	**1.00**	**.994**	.998	.974	.994	.564	.660	.777	.857
SA-1B	.551	.788	.898	.983	**.996**	**1.00**	.993	.998	**.990**	.998	.821	.982	.802	.887
Stay-Positive	.560	.829	**.976**	.992	.990	**1.00**	.989	.998	.981	.994	.720	.886	.808	.884
SA-1B	.787	.906	.825	.993	.831	**1.00**	.831	.999	.829	.997	.830	.995	.789	.928
B-Free	.638	.751	.933	.984	.949	.992	.950	.990	.915	.976	.959	**.999**	.817	.887
DDA	.829	.922	.888	.981	.893	.989	.898	.995	.887	.980	.892	.989	.847	.937
SA-1B	.877	.957	.967	**.997**	.967	.996	.972	.998	.961	.995	**.967**	.997	.903	.962
PROBE	.765	.901	.958	**.997**	.967	.998	.966	.999	.963	.997	.965	.998	.861	.936
DEAR-r	.829	.895	.898	.991	.902	.988	.903	.996	.900	.994	.743	.802	.817	.893
GenSign	.834	.914	.795	.872	.825	.906	.745	.805	.651	.712	.811	.877	.729	.795
Ours	**.981**	**.998**	.973	.994	.983	.999	.987	**1.00**	.985	**1.00**	.862	.945	**.960**	**.990**

balanced accuracy on the recent sets at every block of DINOv2 and CLIP and at the intermediate blocks of SigLIP2; on the other 56 sets the change is smaller and is negative at the last two blocks of CLIP. In every backbone the best intermediate block is more accurate than the last block.

Table 11: Linear probes on frozen backbones by block, original against resampled patch. Recent 11: bAcc at 0.5 and AUROC averaged over the eleven recent sets, FPR on the real photographs of these sets. Other 56: bAcc averaged over all 56 other sets; images smaller than 224 pixels are upscaled to 224 before the crop, as in the main evaluation.

		Original patch				Resampled patch			
		Recent 11			Other 56	Recent 11			Other 56
Backbone	Block	bAcc	AUROC	FPR	bAcc	bAcc	AUROC	FPR	bAcc
DINOv2 ViT-B/14	3	.645	.700	.226	.631	.767	.841	.093	.726
	6	.752	.835	.126	.720	.790	.883	.109	.728
	9	.677	.764	.171	.707	.728	.807	.141	.741
	12 (last)	.557	.581	.292	.648	.613	.648	.289	.678
CLIP ViT-L/14	6	.794	.880	.041	.755	.930	.991	.006	.828
	12	.922	.990	.021	.809	.974	.998	.010	.832
	18	.902	.982	.031	.812	.961	.993	.033	.791
	24 (last)	.813	.914	.092	.758	.859	.930	.135	.722
SigLIP2 ViT-B/16	3	.741	.814	.128	.683	.870	.943	.057	.782
	6	.722	.789	.189	.690	.788	.871	.134	.716
	9	.639	.687	.295	.635	.691	.753	.289	.640
	12 (last)	.701	.764	.243	.622	.684	.747	.286	.645

Tables behind Figure 4. Table 12 gives the filter-size search, Table 11 the frozen probes of the three backbones at every block, and Table 13 the two inputs of our detector that share the same 32×32 token grid.

Table 12: Filter-size search on frozen backbones. Linear probes at every block that we read, 224 crop, bAcc at 0.5 on the eleven recent sets (protocol of the main evaluation: images smaller than the crop are upscaled). The pretrained filter is 14 px for DINOv2 and CLIP and 16 px for SigLIP2; bold marks half the pretrained size. Last row: agreement of the resampled DINOv2-B bank with the pretrained tokens (Appendix C, measurements).

Backbone	Block	4	5	6	7	8	10	12	14	16
DINOv2 ViT-B/14	3	.494	.542	.563	**.767**	.629	.682	.682	.646	–
	6	.492	.521	.577	.790	.673	.721	**.793**	.751	–
	9	.486	.518	.560	**.729**	.639	.669	.693	.677	–
	12 (last)	.495	.486	.508	**.614**	.533	.585	.580	.557	–
CLIP ViT-L/14	6	.647	.671	.700	**.930**	.790	.919	.800	.794	–
	12	.614	.598	.636	**.974**	.724	.917	.885	.923	–
	18	.554	.574	.581	**.961**	.598	.853	.886	.901	–
	24 (last)	.529	.543	.547	**.859**	.561	.752	.801	.813	–
SigLIP2 ViT-B/16	3	.522	.518	.556	.585	**.870**	.686	.786	.745	.741
	6	.503	.504	.534	.552	**.788**	.658	.706	.684	.722
	9	.462	.465	.477	.513	**.691**	.602	.638	.656	.639
	12 (last)	.461	.495	.489	.513	.684	.597	.654	**.709**	.700
Token agreement, DINOv2-B		.13	.18	.28	**.87**	.51	.89	.98	1.00	–

Choice of the factor, details. Two variables shape the curve of Figure 4a. The number of photograph pixels under each token helps at every size: with the filters untouched and the crop resized to the 32×32 grid, the DINOv2 probe rises from .730 at 14 pixels to .837 at 4 pixels. Whether the resampled bank still computes the pretrained tokens is not monotone in the size: for DINOv2-B, the cosine between the tokens of the pretrained bank and the tokens of the resampled bank on matched

Table 13: Patch size and input of our detector. The models share every other setting of our method; balanced accuracy at the threshold 0.5.

		Recent 11			Other 56	
Input and patch	Pixels per token	Mean	Std	Min	Mean	FPR
448 crop, patch 14	196	.834	.170	.517	.870	.030
224 crop, patch 14	196	.840	.122	.608	.849	.037
224 crop, patch 7 (Ours)	49	.949	.056	.796	.887	.027

content is .98 at size 12, .89 at size 10, .87 at size 7, .51 at size 8, and below .3 at sizes 6 to 4. Half the pretrained size is the smallest size at which the resampled bank reproduces the pretrained tokens, and among the sizes that keep the bank the half has the fewest pixels under each token. The one exception to the gain at the last block is the attention pooling of SigLIP2, which discards the gain of the intermediate blocks. A frozen CLIP-L backbone with the resampled patch and a linear head reaches .97 on the recent generators, and the same probe is near chance on the post-processed and the high-resolution real photographs (.50 and .48) with a false positive rate of .17 on the other sets, so the patch alone does not make a detector.

Filter-level measurements of Section 3.1. The three measurements use the patch-embedding weights of DINOv2 ViT-B/14 and the weights resampled by Eq. 2, without any training. (i) Grating response: for a period P we form 64 gratings $\cos(2\pi(x\cos\theta + y\sin\theta)/P + \phi)$ of unit amplitude with random orientation θ and phase ϕ on a 224×224 canvas, apply the bank with its own stride, and report the root mean square of the tokens over channels, positions and gratings. The original bank gives .33, .46, .53, .42, .32, .16 and .08 at periods of 2, 3, 4, 6, 8, 16 and 32 pixels; the resampled bank gives .73, .56, .41, .27, .21, .11 and .06. (ii) Residual pass-through: on the 224 centre crop of 400 held-out photographs of the X-AIGD test split, we apply the bank to the normalized photograph x and to the residual r of each autoencoder and report the ratio of the root-mean-square tokens, $\|W * r\|/\|W * x\|$. Original bank: SD 1.4 .69, SD 2.1 .61, LlamaGen VQ-8 .65, Qwen-Image .39, FLUX.1 .32, CogView4 .27, FLUX.2 .26; resampled bank: .82, .74, .80, .54, .51, .43, .44. (iii) Token agreement: for a filter size k, the cosine between the tokens of the original bank on a 448 crop (32×32 tokens) and the tokens of the bank resampled to k on the same crop resized to $32k$ pixels with antialiased bicubic interpolation, averaged over positions and the same 400 photographs (last row of Table 12).

Correlation of the residuals across autoencoders. For the statement of Section 3 that the residuals of different autoencoders are correlated, we take 400 held-out photographs at a 512 center crop, reconstruct each with eight autoencoders (SD 2.1, SD 1.4, SDXL, FLUX.1, FLUX.2, CogView4, Qwen-Image and DC-AE), and convert each residual to gray. For each photograph we compute the 8×8 correlation matrix of the residual values over all pixels, average the matrices over photographs, and take the eigendecomposition. The first component carries 47.9% of the variance (second 15.1%, third 7.9%) with loadings of .78, .73, .74, .62, .52, .68, .77 and .67, positive for every autoencoder; restricted to flat regions, the first component carries 40.1% with loadings from .49 to .72, again all positive. The correlation is an observation about the residuals and not a statement about which cue our detector uses for its decision.

Trade-off of a finer patch. With 3.5 pixels under each token, measured with the earlier implementation of the same operation in which the input crop was resized to the token grid, the mean on the recent generators rises to .972 and the weakest generator to .906, and the detector rates 31% of the high-resolution photographs of Synthbuster as fake and misses half of the post-processed fakes of the AlignedForensics set (.884 to .707), so the mean over the other 56 generator sets falls from .890 to .877. The frozen probes show the same trade-off on the other generator sets: from 14 to 3.5 pixels under each token, DINOv2 (block 6) rises on the reconstruction sets (.765 to .899) and falls on Synthbuster (.768 to .670) and on the post-processed AlignedForensics set (.680 to .528), and CLIP follows the same directions (.857 to .921, .666 to .432). The trade-off is a property of the patch size and not of our training.

Resampled patch at inference. The resampled patch is required at inference as well: with the weights of our detector unchanged and each token covering 14 photograph pixels at inference (the 448 crop downscaled to 224 inside the network, so that the 7×7 filters read 14 photograph pixels per token on the same 32×32 grid), the mean on the recent generators falls from .949 to .782, while the other 56 sets move from .887 to .856. Resampling the fine-tuned 7×7 filters back to 14×14 instead gives .622, but that also alters the weights. The 448 crop of Table 13 is above the standard recipe on the other 56 sets (.870 against .815): the wider view helps where the residual is coarse.

D Additional ablations and analyses

Every model in this appendix starts from the configuration of our detector (Table 2, last row) and changes one setting. All models are trained once with the same seed and scored with the protocol of Appendix B. Differences below .01 in a mean are within the run-to-run variation that we observe and are not interpreted.

D.1 Design choices

Table 14: One setting changed at a time. bAcc at 0.5: mean and minimum over the eleven recent sets, mean over the other 56 sets, the post-processed AlignedForensics set and the GAN sets; FPR on the real images of the recent sets.

	Recent 11		Other 56	Post-proc.	GAN	
Model	Mean	Min	Mean	Mean	Mean	FPR
Ours (resampled patch, final class token, two passes, SD 2.1)	.949	.796	.887	.862	.688	.027
Third pass added	.950	.831	.907	.851	.789	.033
SD 1.4 autoencoder in place of SD 2.1	.942	.741	.893	.810	.864	.014
Mean of the class tokens of blocks 4, 6, 8, 12 in place of the final class token	.952	.816	.888	.866	.690	.024
Input upsampled $\times 2$ with the original patch in place of the resampled patch	.946	.790	.886	.876	.692	.032

Number of passes. A third reconstruction, labeled fake like the first two (the real weight matches the number of fake items: 2 for two passes, 3 for three), raises the mean on the other 56 sets from .887 to .907, the GAN sets from .688 to .789, and the weakest recent set from .796 to .831 (Table 14). The same model loses on the post-processed AlignedForensics set (.862 to .851) and rates more real photographs as fake (FPR .027 to .033); under the post-processing of Appendix D.4 the two depths are within .02 of each other in every condition (Table 16). The third pass therefore moves weight toward older generators at the cost of false positives and of the post-processed set. The second pass adds a component that the first does not contain, and the third pass repeats the second: on 200 SA-1B crops (512×512, from the training pool) the increment of the second pass, $f_2(x) - f_1(x)$, is orthogonal to the residual of the first (cosine $-.02$) while the increment of the third pass has a cosine of $+.50$ with that of the second and a smaller magnitude (RMS 4.5 against 5.4); the block-wise error maps of consecutive passes overlap (correlation .85 and .90), so every pass writes at the same locations; a reconstruction lies close to a fixed point of its autoencoder, the property that reconstruction-error detectors use with the autoencoder (Ricker et al., 2024) or with a diffusion model (Wang et al., 2023). Two passes is the smallest depth that contains the new component. A third pass adds a third to the training items per photograph, and the accuracy it adds on older generators comes with the losses above, so we keep two passes. The pattern holds beyond SD 2.1: over the same 200 training photographs, the cosine between the increments of the second and the first pass lies between $-.09$ and $.31$ for ten decoders, and between the third and the second pass between .33 and .84 for the nine continuous autoencoders; for the VQ-16 tokenizer the third pass adds another new direction (.04). Figure 3d shows seven of the ten decoders.

Training autoencoder. Replacing the SD 2.1 autoencoder by the SD 1.4 autoencoder, the autoencoder that AlignedForensics and Stay-Positive train on, changes the mean on the recent sets from .949 to .942 and on the other 56 sets from .887 to .893. The decrease is concentrated in Krea (.796 to .741) and Z-Image (.942 to .905); FLUX.1, SD 3, SD 3.5L, Qwen, FLUX.2 and Sana change by at most .02. The model trained with SD 1.4 is better on the GAN sets (.688 to .864) and has a lower false positive rate (.027 to .014), and worse on the post-processed set (.862 to .810) and on

Synthbuster (.846 to .797). The method therefore does not depend on the SD 2.1 autoencoder; the choice of autoencoder redistributes accuracy across benchmarks.

Pooling and implementation. Reading the mean of the class tokens of blocks 4, 6, 8 and 12 (Koutlis & Papadopoulos, 2024) in place of the final class token changes every mean by at most .005 (Table 14), so the differences between blocks in the frozen probes of Table 11 disappear after fine-tuning. Replacing the resampled patch by a $\times 2$ nearest-neighbour upsampling of the input with the original patch, which gives each token the same 49 photograph pixels, changes the recent and other-56 means by at most .003. The four combinations of pooling and implementation lie within .006 of each other on the recent sets (.946 to .952).

D.2 Range of the residual scaling

On the other 56 sets the range models reach .852 ($\beta < 0$), .855 ($0 < \beta < 1$) and .856 ($\beta > 1$) against .826 without scaling (one pass) and .887 with the full set, so the full set is the best choice on both groups. Table 3 in Section 4.3 trains the detector with one range of β at a time. Table 15 runs the sweep of Section 3.2 on the same models, averaged over the eight decoders of the sweep. Figure 7 shows the sweep per decoder for the four models of Table 2.

Table 15: Residual scaling sweep of the models of Table 3. bAcc at 0.5 averaged over the eight decoders of the sweep, as a function of the scaling factor β applied at test time.

Training β	-1	$-.5$	$-.25$	.25	.5	.71	1	1.25	1.8	2.8
none (one pass)	.500	.500	.499	.513	.699	.861	.925	.919	.814	.631
$\{-.5, -.25\}$	.969	.955	.834	.500	.679	.828	.914	.939	.959	.946
$\{.25, .5, .71\}$	.499	.499	.498	.807	.957	.976	.965	.917	.710	.532
$\{1.25, 1.8\}$	.537	.503	.501	.503	.633	.806	.911	.950	.981	.986
all seven (Ours)	.966	.916	.737	.775	.937	.965	.971	.972	.976	.982

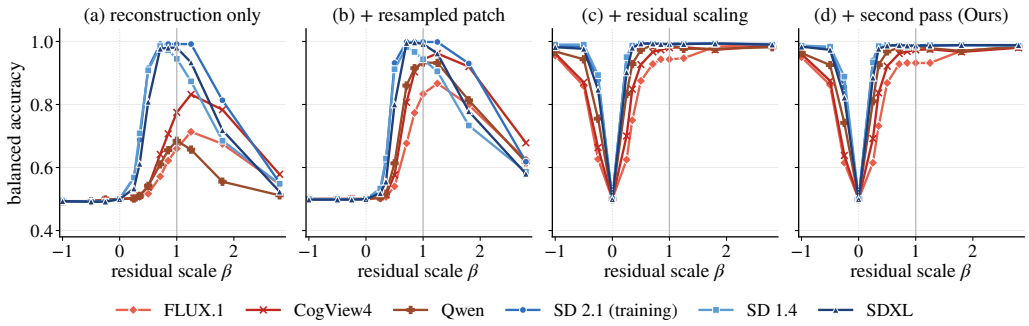

Figure 7: Residual Scaling Sweep of the four models of Table 2: bAcc at 0.5 of x_β against x for six decoders; the three recent decoders are not in training.

Each model responds near the range of β that it was trained on, and not to smaller values or to the opposite sign: the model trained on $\{.25, .5, .71\}$ is at chance at every negative β and at .532 at $\beta = 2.8$, and the model trained on $\{1.25, 1.8\}$ is at chance at every $|\beta| \leq .25$. The model trained on the negative values is the exception: it scores .914 to .959 at every $\beta \geq 1$, so a reversed residual teaches the detector to respond to the magnitude of the residual and not to its sign. No range alone reproduces the response of the full set, in the sweep or on the recent generators (Table 3).

D.3 Public detectors on the sweep

Detectors of 2025. Figure 1b places the four public detectors of 2025 on the sweep. Aligned-Forensics and Stay-Positive have the largest decoder spread: at matched damage the two detectors score .97 to .99 on the 4-channel decoders and .50 to .66 on all four recent decoders (FLUX.1,

CogView4 and Qwen-Image with 16 latent channels, FLUX.2 with 32). Amplifying the residual of a 16-channel decoder to the PSNR of the SD 2.1 reconstruction ($\beta = 1.8$, 27.4 dB) leaves both detectors at .55 to .56, so neither detector misses the recent decoders for lack of damage: both detectors depend on the signed pattern of the training decoder. B-Free depends on the magnitude of the residual. On the recent decoders B-Free scores .60 at $\beta = 1$, .74 at $\beta = 1.8$ and .79 at $\beta = 2.8$, so B-Free misses a residual as small as the residual of a recent decoder. DDA scores an AUROC of .25 at $\beta < 0$: DDA rates a sharpened photograph as more real than the photograph, and amplification lowers the accuracy of DDA (.77 at $\beta = 1$, .59 at $\beta = 2.8$ on the recent decoders), so DDA reads a deficit of high-frequency content and not the residual.

Detectors of 2026. Figure 1b places PROBE, DEAR-r and GenSign on the sweep of Section 3.2. PROBE, trained on the SD 2.1 data of B-Free and on detector-guided samples, reaches .96 on the 4-channel decoders at $\beta = 1$ and falls above it (.68 at $\beta = 2.8$), and reads the recent decoders at .67; DEAR-r reaches .89 and .68; GenSign .75 and .73. PROBE and DEAR-r are below chance on a reversed residual (AUROC of $x_{-0.5}$ against x: .38 and .39) and GenSign is at .62, so the three detectors depend on the magnitude and on the sign of the residual of their training data as the detectors of 2025 do.

D.4 Robustness to post-processing

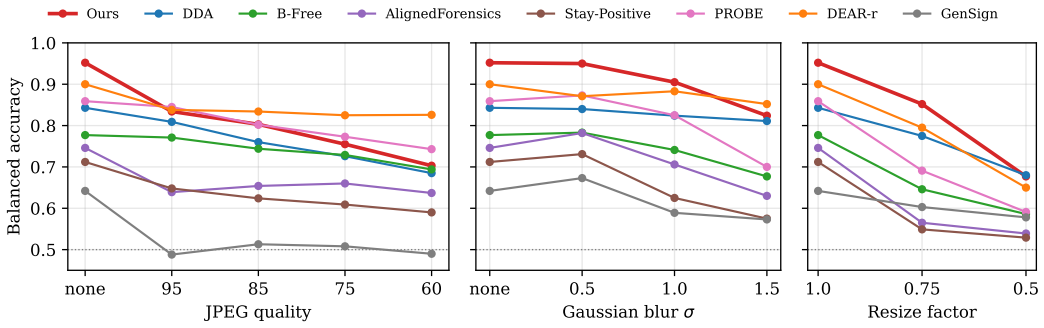

Figure 8: bAcc at 0.5 on the eleven recent sets under post-processing applied to real and generated images alike. 200 pairs per generator.

Table 16: Values of Figure 8, with the true positive and false positive rates of our detector; best per column in bold over the public detectors and our detector. The last two rows are our detector with three passes and our detector with the original patch (224 crop, patch 14), which keeps residual scaling and two passes.

		JPEG quality				Gaussian blur σ			Resize	
Model	None	95	85	75	60	0.5	1.0	1.5	0.75	0.5
Ours	**.952**	.834	.803	.755	.703	**.950**	**.905**	.824	**.852**	.677
DDA	.843	.809	.760	.726	.685	.840	.824	.811	.775	**.680**
B-Free	.777	.771	.744	.729	.694	.783	.741	.677	.646	.586
AlignedForensics	.746	.639	.654	.660	.637	.782	.706	.630	.565	.539
Stay-Positive	.712	.648	.624	.609	.590	.731	.625	.575	.549	.529
PROBE	.859	**.845**	.802	.773	.743	.873	.825	.700	.691	.591
DEAR-r	.900	.838	**.834**	**.825**	**.826**	.871	.883	**.852**	.795	.650
GenSign	.642	.488	.513	.508	.490	.673	.589	.573	.603	.578
Ours, TPR	.933	.713	.653	.575	.491	.954	.848	.709	.710	.360
Ours, FPR	.030	.045	.046	.065	.084	.055	.037	.061	.005	.007
Ours, three passes	.958	.837	.784	.744	.700	.948	.904	.811	.857	.684
Ours, original patch	.845	.795	.754	.701	.639	.866	.829	.750	.710	.620

We apply JPEG compression, Gaussian blur and downscaling to 200 real and 200 generated images of each recent set, the same operation to both classes, and score the detectors at the fixed threshold

(Figure 8, Table 16). Our detector has the highest accuracy in four of the ten conditions (clean, blur at $\sigma = 0.5$ and 1.0, and downscaling by 0.75), DEAR-r in four (JPEG at qualities 85, 75 and 60, and blur at $\sigma = 1.5$), PROBE at JPEG quality 95 and DDA at downscaling by 0.5. Averaged over the ten conditions our detector and DEAR-r are tied (.826 and .828); DEAR-r, trained on aligned real and fake pairs with post-processing, holds under JPEG (.825 at quality 75 against .755) at close to three times our false positive rate on clean images (.082 against .030), and PROBE loses under blur and downscaling (.700 at $\sigma = 1.5$, .691 at downscaling by 0.75). Our detector also loses the most: JPEG at quality 95 removes .12, and every condition but blur at $\sigma = 0.5$ removes more from our detector than from DDA. The loss comes from missed generated images: the true positive rate falls from .933 to .713 at quality 95 and to .491 at quality 60, while the false positive rate stays below .09 in every condition. A post-processed generated image is therefore rated real, and a post-processed photograph is not rated fake. DDA has a false positive rate of .155 to .172 before and after blur, and AlignedForensics and Stay-Positive have a false positive rate near zero with a true positive rate that falls to .08 and .06 after downscaling by 0.5.

D.5 Feature space of the ablation models

For each model of Table 2 we measure where the recent generated images fall in the feature space of that model, the 768-dimensional vector that the linear layer reads. The reference pool consists of 600 photographs of the X-AIGD test split, none in training, and 600 training-like fakes made from 300 of these photographs with the SD 2.1 autoencoder: 200 single reconstructions, 200 second passes, 100 scaled residuals of the first pass and 100 scaled residuals of the second pass, with the seven training values of β in equal numbers. For each query image we find the cosine nearest neighbour in the pool. The *coverage* of a query set is the fraction of query images whose nearest neighbour is a fake, and the miss rate is the fraction of query images below the threshold. The queries are 300 images for each generator.

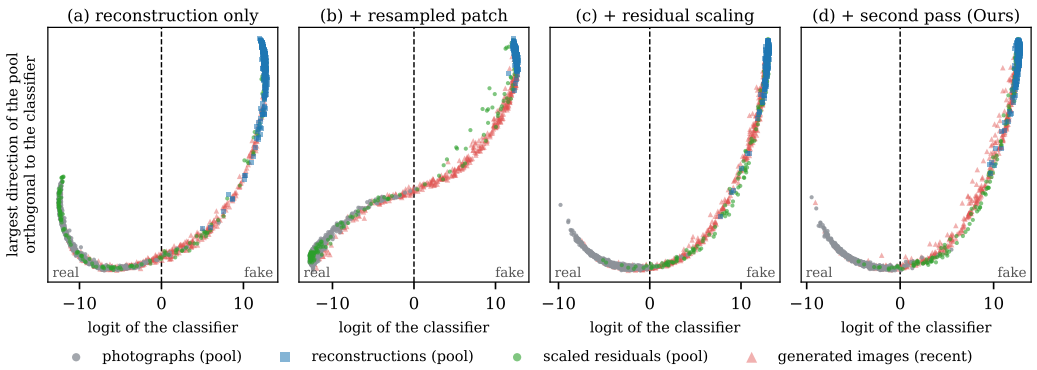

Figure 9: Feature space of the four models of Table 2. Horizontal axis: logit of each model, with the threshold dashed. Vertical axis: largest direction of the pool of photographs and training-like fakes orthogonal to the classifier. Recent generated images are projected onto the same axes.

The resampled patch raises the coverage of the recent generators from .67 to .90 (Table 17): the recent generated images move next to the training fakes. Residual scaling leaves the coverage unchanged and nearly halves the miss rate, from .16 to .09: the same images move further from the photographs. The pool share of each kind of fake is one third for one pass, one third for two passes, and one sixth for the scaled residuals of each pass, of which five sevenths have $\beta > 0$; the nearest fake of a recent generated image is a positive scaling of either pass more often than its share (.22 and .22 against .12 each) and a negative scaling almost never (.03 in total), and the scaled second passes are chosen as often as the scaled first passes. Krea and FLUX.2, the two weakest recent sets, have the lowest coverage (.60 and .81). The GAN sets have a coverage of .17 with the first model and .40 with ours, and the held-out photographs have a fake nearest neighbour in .12 and .03 of the cases. A principal component projection of the pool (Figure 9) shows the same change: the features of every model lie close to one curve, the photographs at one end and the training fakes at the other, and as

Table 17: Coverage and miss rate of the recent generators in the feature space of each model of Table 2 (columns 1 to 4: reconstruction only, + resampled patch, + residual scaling, + second pass), and the kind of the nearest training-like fake for our model: a reconstruction after one or two passes, or a scaled residual of either pass.

	Coverage				Miss rate				Nearest fake (Ours)					
									1 pass	2 passes	scaled, 1 pass		scaled, 2 passes	
Generator	1	2	3	4	1	2	3	4			$\beta > 0$	$\beta < 0$	$\beta > 0$	$\beta < 0$
FLUX.1	.66	.96	.97	.96	.32	.05	.03	.03	.25	.26	.29	.00	.20	.00
FLUX.1 (iPh.)	.57	.98	.96	.96	.42	.05	.03	.03	.24	.21	.23	.00	.31	.00
Krea	.27	.62	.48	.60	.89	.66	.45	.33	.18	.29	.10	.09	.16	.18
Z-Image	.32	.74	.79	.83	.76	.46	.14	.10	.20	.35	.22	.02	.21	.00
SD 3	.94	1.00	1.00	1.00	.03	.00	.00	.00	.48	.39	.03	.00	.10	.00
SD 3.5L	.96	1.00	1.00	.99	.01	.00	.00	.00	.31	.58	.04	.00	.08	.00
SD 3.5L (iPh.)	.94	1.00	1.00	1.00	.03	.00	.00	.00	.37	.39	.06	.00	.18	.00
CogView4	.47	.82	.92	.91	.49	.21	.05	.05	.14	.09	.46	.00	.32	.00
Qwen	.61	.92	.96	.96	.33	.10	.03	.02	.06	.17	.41	.00	.36	.00
FLUX.2	.70	.93	.83	.81	.25	.13	.15	.14	.07	.14	.41	.00	.38	.00
Sana	.89	.90	.84	.90	.09	.14	.15	.06	.29	.34	.22	.01	.13	.00
Mean	.67	.90	.89	.90	.33	.16	.09	.07	.24	.29	.22	.01	.22	.02

the components are added the recent generated images move from the photograph end toward the fake end of the curve.

E LIMITATIONS

Our method targets generators with a learned latent decoder; GANs, pixel-space models, and closed systems fall outside its scope, where broader detectors such as B-Free remain preferable. Residual scaling brings a modest rise in false positives, and one recent generator, Krea, is detected less reliably than others on the same decoder; we leave its cause to future work.