MiLD: Mining Informative Low-Bit Cues for Training-Free AI-Generated Image Detection

Anonymous authors
Paper under double-blind review

Abstract

Low-bit planes have been shown to contain discriminative information for AI-generated image detection. We further observe that the Hamming states induced by low-bit morphological residuals provide useful real/fake cues. However, the Hamming-state occupancy is a static image-level statistic and is sensitive to JPEG compression and Gaussian noise. This motivates us to use Hamming states as spatial cues rather than as the final detection score, and to construct a stable, robust, and efficient training-free detector. We propose MiLD, which selects channel-disagreement locations from low-bit morphological residuals and builds a masked low-bit reversal. MiLD then compares the frozen encoder features of the original and reversed images, using their feature discrepancy as the detection score. Across five benchmarks, MiLD improves the mean AUROC over RIGID by **2.78%** while achieving the highest efficiency score. Extensive ablations further demonstrate the stability, transferability, and robustness of MiLD[1].

1 Introduction

The rapid progress of image generation has made AI-generated image detection an increasingly important forensic task. Modern generators produce highly realistic images, making synthetic content difficult to identify by visual inspection alone. Existing detectors can be broadly divided into training-based and training-free methods. Training-based detectors learn discriminative signals from labeled real and generated images, including CNN-based and frequency-based detectors (Wang et al., 2020; Frank et al., 2020), CLIP-based and hybrid-feature generalization methods (Ojha et al., 2023; Yan et al., 2025), reconstruction-error learning (Wang et al., 2023; Luo et al., 2024), and recent low-bit learning pipelines (Wang et al., 2025; 2026). Training-free detectors avoid fitting a detector on target real/fake collections, and they can be further divided by how they expose forensic cues. Reconstruction-based methods compare how real and generated images behave under autoencoder or diffusion-model reconstruction, as in AEROBLADE (Ricker et al., 2024), Manifold Bias (Brokman et al., 2025), and RDD (Choi et al., 2026). In contrast, non-reconstruction methods probe fixed representations, perturbation sensitivity, uncertainty, or distributional consistency without reconstructing the input, including RIGID (He et al., 2024), MINDER (Tsai et al., 2024), WePe (Nie et al., 2025), ConV (Zhang et al., 2025), and WaRPAD (Choi et al., 2025). These methods show that useful forensic information can be exposed without detector training, but they typically rely on generic reconstructions or generic perturbations.

Existing low-bit studies such as LOTA and LIDA (Wang et al., 2025; 2026) show that low-bit planes contain discriminative information for AI-generated image detection. However, these methods typically rely on learned pipelines to extract and exploit low-bit cues. We further observe that Hamming states induced by low-bit morphological residuals reveal systematic differences between real and generated images. In particular, the intermediate Hamming states reflect channel-disagreement patterns in the low-bit morphology and provide a useful forensic cue without training a detector.

Directly using Hamming-state occupancy as a detection score is not sufficiently stable. It is a static image-level statistic tied to each image and is sensitive to JPEG compression and Gaussian noise. This motivates us to use Hamming states as spatial cues rather than as the final score. The key idea is to locate informative channel-disagreement regions in the low-bit morphology and then measure how a targeted low-bit perturbation changes the frozen image representation.

[1]Code is available: `https://anonymous.4open.science/r/Anonymous_code-B4B6`

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.

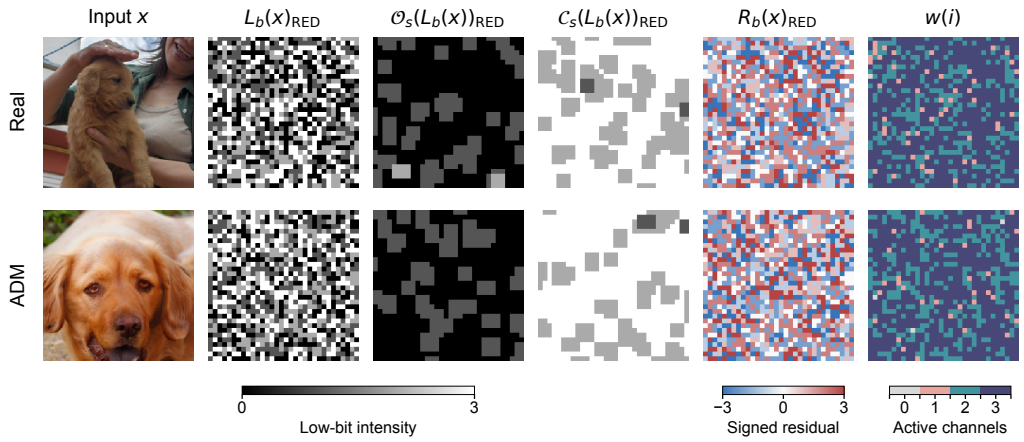


Figure 1: Visualizing low-bit bidirectional morphology. Rows show a real image and an ADM-generated image from the same ImageNet class. After the full input, the columns show $L_b(x)_{\mathrm{RED}}$, $\mathcal{O}_s(L_b(x))_{\mathrm{RED}}$, $\mathcal{C}_s(L_b(x))_{\mathrm{RED}}$, the signed residual $R_b(x)_{\mathrm{RED}}$, and the Hamming weight $w(i)$. Grayscale maps share a common intensity scale, the signed residual uses a zero-centered diverging scale, and Hamming weights use four discrete colors.

Based on this idea, we propose **MiLD** (**M**ining **i**nformative **L**ow-bit cues for training-free AI-generated image **D**etection), a training-free detector based on morphology-guided low-bit reversal. Given an RGB image, MiLD extracts its low-bit planes and applies a signed white and black top-hat morphology operator to obtain a residual field. MiLD then uses the resulting Hamming states to select channel-disagreement locations as a mask and constructs a masked low-bit reversal. The feature discrepancy between the original image and its reversed counterpart defines $E(x)$, which is used directly as the detection score. This design does not train a detector or learn score weights and only requires two encoder forward passes.

We evaluate MiLD on ImageNet (Deng et al., 2009), LSUN-Bedroom (Yu et al., 2015), GenImage (Zhu et al., 2023), AIGCDetectBenchmark (Zhong et al., 2023), and DailyBench (Jiang et al., 2026). Across all five benchmarks, MiLD improves the mean AUROC over RIGID by **2.78%**. Moreover, MiLD has nearly the same processing cost as RIGID while achieving the highest efficiency score among the evaluated methods. Extensive ablations further verify the stability, transferability, and robustness of MiLD across benchmarks, backbones, corruptions, and design choices.

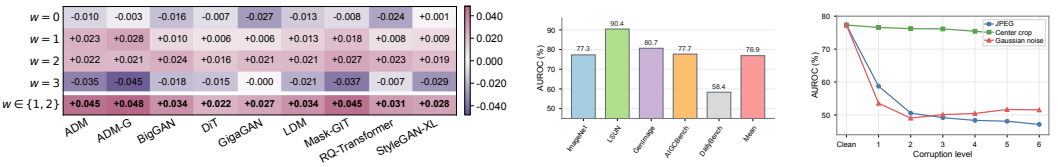
Our contributions are:

- We reveal that low-bit morphological residuals contain useful forensic information for AI-generated image detection.
- We propose MiLD, a training-free detector that uses morphology-guided low-bit reversal to compute a feature-discrepancy score.
- We show through extensive experiments that MiLD achieves a favorable performance-efficiency trade-off and remains robust across benchmarks, backbones, corruptions, and design choices.

2 Preliminary and Motivation

2.1 Preliminary

Low-bit planes can expose image variations that are weak or imperceptible in the original RGB image (Wang et al., 2025). Rather than classify these bit planes with a trained model, MiLD describes their local organization with a fixed morphology operator. Given an RGB image, we first convert it to RGB and bicubically resize it to the encoder input resolution. We denote the resulting image by $x \in [0,1]^{3\times h\times w}$. We write its 8-bit intensities as $I = \mathrm{round}(255x)$ and extract the low-bit image



(a) State shift on ImageNet. (b) Detection. (c) ImageNet robustness.

Figure 2: Analysis of the intermediate Hamming-weight states $w(i) \in \{1, 2\}$. (a) Mean generated-minus-real Hamming-state shifts on ImageNet. The last row combines $w = 1$ and $w = 2$. (b) AUROC of $P(x)$ across five datasets and their mean. (c) AUROC of $P(x)$ on ImageNet under JPEG compression, center cropping, and Gaussian noise, with the horizontal axis indicating increasing corruption levels.

$L_b(x) = I \bmod 2^b$. Here, b is the number of retained least significant bits, so each channel of $L_b(x)$ takes values in $\{0, \ldots, 2^b - 1\}$.

We then separate local bright and dark structures in the low-bit image. For each channel, we use an $s \times s$ kernel for morphological opening and closing. Opening, denoted by $\mathcal{O}$, performs erosion followed by dilation and suppresses bright structures smaller than the structuring element. Closing, denoted by $\mathcal{C}$, reverses this order and fills small dark structures. For any three-channel input field U, define the signed bidirectional morphology operator

$$\mathcal{R}_s(U) = \big(U - \mathcal{O}_s(U)\big) - \big(\mathcal{C}_s(U) - U\big) = 2U - \mathcal{O}_s(U) - \mathcal{C}_s(U). \tag{1}$$

MiLD applies this operator to the low-bit image and obtains $R_b(x) = \mathcal{R}_s(L_b(x)) \in \mathbb{R}^{3\times h\times w}$. The two subtracted terms are the white and black top-hat residuals, respectively. Their difference records the local imbalance between bright and dark low-bit structures. Thus, $R_{b,c}(x, i) = 0$ means that the two residuals balance at pixel i in channel c, including the case where both vanish. The resulting morphology residual $R_b(x)$ supports the low-bit reversal mask defined below. Figure 1 visualizes the intermediate maps.

The signed field $R_b(x)$ contains both residual magnitudes and channel-activity patterns. To describe these patterns, we first binarize whether each channel has a nonzero signed imbalance at each pixel. For each channel, let $a_c(i) = \mathbb{I}\{|R_{b,c}(x, i)| > \epsilon_a\}$, where $c \in \{R, G, B\}$ and ϵ_a is a small activity tolerance. Here, activity denotes a nonzero signed imbalance, rather than the presence of either top-hat residual alone. The binarized vector $a(i) = (a_R(i), a_G(i), a_B(i))$ has eight possible states, and its Hamming weight $w(i) = \sum_c a_c(i)$ counts the active channels. The intermediate states $w(i) \in \{1, 2\}$ mark channel-disagreement locations, where the morphology residual is active in only part of the RGB triplet. They exclude locations where all channels are inactive or active together.

2.2 Motivation

The low-bit morphological residual reveals a difference in channel-activity patterns between real and generated images. Figure 2 summarizes this observation. In Figure 2a, the generated-minus-real shifts are positive for both intermediate Hamming weights, $w = 1$ and $w = 2$, across all nine ImageNet generators. We summarize the frequency of these states in each image by

$$P(x) = \frac{1}{hw} \sum_{i=1}^{hw} \mathbb{I}\{w(i) \in \{1, 2\}\}. \tag{2}$$

When using $P(x)$ alone, larger values are treated as the generated-positive score. We analyze low-bit morphological statistics on a randomly sampled development subset of 10,000 images from the ImageNet-based collection. Despite simplicity, $P(x)$ alone achieves a mean AUROC of 76.89% across the five benchmarks in Figure 2b. This exceeds the 74.41% five-benchmark mean of WePe in Table 4. However, $P(x)$ is a static occupancy statistic rather than a response of the pretrained representation. On ImageNet, its AUROC falls from 77.31% on clean images to 47.14% under the strongest JPEG compression and 51.55% under the strongest Gaussian noise, while retaining 74.53% under the strongest center crop in Figure 2c. Its stand-alone discrimination is therefore not consistently robust across corruptions.

These observations motivate using $w(i) \in \{1, 2\}$ as a spatial cue instead of treating $P(x)$ as the final detection score. We therefore fix the channel-disagreement set $S = \{1, 2\}$ before the main experiments and ask whether a perturbation restricted to these locations can elicit a more robust and transferable response from a frozen representation. MiLD uses the selected states to guide a masked low-bit reversal and measures the resulting feature change. Later ablations test sensitivity to this design rather than tune it after observing benchmark results. The next section defines this construction.

3 METHOD

3.1 MASKED LOW-BIT REVERSAL

Unlike RAID, which reverses the order of all eight bit planes to amplify low-order details for a trained classifier (Cheng et al., 2026), MiLD complements only the low-order bits at morphology-selected channel-disagreement locations, preserves the high-order bits, and measures the resulting feature change with a frozen encoder. Given the channel-disagreement states fixed above, we construct a masked low-bit reversal as the representation probe. Let $I = H_b(x) + L_b(x)$, where $H_b(x) = I - L_b(x)$ stores the high-bit component. Within the selected mask, we reverse the original low-bit value by complementing the retained b bits. Outside the mask, the low-bit value is left unchanged:

$$L'_b(x)_{c,i} = \begin{cases} 2^b - 1 - L_b(x)_{c,i}, & a_c(i) = 1 \text{ and } w(i) \in \{1, 2\}, \\ L_b(x)_{c,i}, & \text{otherwise.} \end{cases} \tag{3}$$

The reversed image is

$$x' = \frac{H_b(x) + L'_b(x)}{255}. \tag{4}$$

This operation keeps the high-bit content fixed and changes only the selected least significant bits.

Let $f(y)$ denote the image-level feature of the frozen encoder after L2 normalization. We define the feature discrepancy score as

$$E(x) = \left\| \frac{f(x) - f(x')}{2} \right\|_2^2 = \frac{1 - \cos(f(x), f(x'))}{2}. \tag{5}$$

Thus, $E(x)$ is large when low-bit reversal at morphology-selected channel-disagreement locations produces a larger image-level representation shift. We use larger $E(x)$ as stronger evidence of generated content. For later ablations, we replace $R_b(x)$ by a three-channel field $U_b(x)$ only when computing the activity indicators and Hamming weights in the reversal mask. We denote the resulting reversed RGB image by x'_{U_b} and its score by $E_{U_b}(x) = \|[f(x) - f(x'_{U_b})]/2\|_2^2$. The default is $U_b(x) = R_b(x)$, so $x'_{R_b} = x'$ and $E_{R_b}(x) = E(x)$. In every setting, the encoder inputs are the resized RGB image x and its masked low-bit-reversed RGB counterpart. The field U_b determines the mask and is not itself fed to the encoder as an image.

3.2 THEORETICAL INTERPRETATION

We interpret the MiLD score through the Jacobian of the frozen pretrained representation. The RGB image is first resized to the encoder input resolution $h \times w$. Low-bit morphology and masked low-bit reversal are then computed at this resolution. The representation maps below include image normalization and feature extraction, with no further spatial resizing.

The feature-discrepancy readout uses low-bit morphology to define a masked low-bit reversal direction, and measures its effect on the image-level representation. Write the resized base image as $\xi = \mathrm{vec}(x) \in \mathbb{R}^n$, where $n = 3hw$. Let $g : \mathbb{R}^n \to \mathbb{R}^q$ denote the unnormalized image-level feature before L2 normalization, and define

$$\phi(\xi) = \frac{g(\xi)}{\|g(\xi)\|_2}. \tag{6}$$

The method uses this normalized image-level feature in Eq. 5. Let $\xi' = \mathrm{vec}(x')$ be the masked low-bit reversal of ξ, and write

$$\mu = \frac{\xi + \xi'}{2}, \qquad \eta = \frac{\xi - \xi'}{2}. \tag{7}$$

Then $\xi = \mu + \eta$ and $\xi' = \mu - \eta$, so

$$E(x) = \left\| \frac{\phi(\mu+\eta) - \phi(\mu-\eta)}{2} \right\|_2^2. \tag{8}$$

If ϕ is three times differentiable in a neighborhood of the segment between $\mu - \eta$ and $\mu + \eta$, a symmetric Taylor expansion gives

$$\frac{\phi(\mu+\eta) - \phi(\mu-\eta)}{2} = J_\phi(\mu)\eta + \mathcal{O}\big(\|\eta\|_2^3\big). \tag{9}$$

Therefore,

$$E(x) \approx \|J_\phi(\mu)\eta\|_2^2 = \eta^\top J_\phi(\mu)^\top J_\phi(\mu)\eta. \tag{10}$$

Thus, $E(x)$ approximates the pullback-metric energy of the normalized image-level representation along the masked low-bit reversal direction η.

This local expression separates the magnitude of the reversal from the representation's sensitivity to its direction. Define the input energy $A(x) = \|\eta\|_2^2$ and, for $A(x) > 0$, the unit direction $u(x) = \eta/\sqrt{A(x)}$. Equation 10 then becomes

$$E(x) \approx A(x)\,\mathcal{R}_\phi(\mu, u(x)), \qquad \mathcal{R}_\phi(\mu, u) = u^\top J_\phi(\mu)^\top J_\phi(\mu) u. \tag{11}$$

Here $A(x)$ measures the squared input change in RGB coordinates, while $\mathcal{R}_\phi$ is the Rayleigh quotient of the local pullback metric along the selected direction. This approximation characterizes the score for small perturbations; it does not by itself establish why the morphology-selected direction distinguishes real from generated images. The L2 normalization also gives $E(x)$ an angular interpretation. When $g(\mu) \neq 0$, differentiating Eq. 6 yields

$$J_\phi(\mu) = \frac{I - \phi(\mu)\phi(\mu)^\top}{\|g(\mu)\|_2} J_g(\mu). \tag{12}$$

Consequently, $E(x)$ removes sensitivity to the feature norm and keeps the component of the image-level response that changes the feature direction. Equivalently, since $\phi(\xi)$ and $\phi(\xi')$ are unit-normalized,

$$E(x) = \left\| \frac{\phi(\xi) - \phi(\xi')}{2} \right\|_2^2 = \frac{1 - \phi(\xi)^\top \phi(\xi')}{2}, \tag{13}$$

which is half of the cosine distance between the original and low-bit-reversed image-level representations. In this view, $E(x)$ captures the angular image-level sensitivity along the masked low-bit reversal direction. Appendix A.2 examines input energy and the finite representation response with energy-matched controls.

4 Experiments

4.1 Setup

Datasets. We evaluate our method on five datasets covering both commonly used benchmarks and more diverse evaluation settings. **ImageNet** (Deng et al., 2009) and **LSUN-Bedroom** (Yu et al., 2015) are adopted as two widely used benchmarks for AI-generated image detection, representing different real-image distributions and generator families. To further assess generalization of our method in more realistic scenarios, we additionally evaluate on two broad benchmarks: **GenImage** (Zhu et al., 2023) and **AIGCDetectBenchmark** (Zhong et al., 2023). Finally, we include the more recent **DailyBench** (Jiang et al., 2026) benchmark to examine the effectiveness of our method on images generated by newly emerging generative models. Detailed dataset information is provided in the Appendix. The 10,000 development images used in the motivation analysis are excluded from the ImageNet test set, and the resulting configuration is kept fixed across all evaluation datasets.

Baselines. We compare our method with representative training-free AI-generated image detectors and group them into two categories according to their detection mechanisms. The first category consists of **Reconstruction-based Methods**, including AEROBLADE (Ricker et al., 2024) and Manifold Bias (Brokman et al., 2025). The second category contains **Non-reconstruction Methods**, including RIGID (He et al., 2024), MINDER (Tsai et al., 2024), WePe (Nie et al., 2025), ConV (Zhang et al., 2025), and WaRPAD (Choi et al., 2025).

Table 1: AI-generated image detection performance (AUROC, %) on ImageNet. **Bold** and underline denote the best and the second best results within each category, respectively.

Method	ADM	ADMG	LDM	DiT	BigGAN	GigaGAN	StyleGAN XL	RQ-Transformer	Mask GIT	Mean
Reconstruction-based Methods										
AEROBLADE	63.84	66.93	59.33	56.24	51.42	44.12	44.40	59.45	60.87	56.29
Manifold Bias	**74.61**	**70.76**	**78.38**	**73.06**	**89.26**	**80.20**	**86.64**	**83.02**	**82.08**	**79.78**
Non-reconstruction Methods										
RIGID	88.28	83.01	80.83	73.22	93.71	88.85	86.00	93.28	92.49	86.63
MINDER	87.55	81.06	83.86	76.30	95.15	91.59	88.77	**94.90**	93.31	88.05
WePe	89.89	83.37	79.37	75.12	92.74	91.90	**93.05**	93.46	87.25	87.35
ConV	87.20	79.46	77.83	71.72	92.37	90.29	91.95	91.11	87.76	85.52
WaRPAD	90.45	85.91	**89.63**	**87.80**	**95.18**	**92.26**	92.84	91.85	**95.36**	**91.26**
MiLD (ours)	**90.52**	**87.10**	84.42	78.53	92.53	89.09	86.67	93.41	92.89	88.35

Table 2: AI-generated image detection performance (AUROC, %) on LSUN-Bedroom. **Bold** and underline denote the best and the second best results within each category, respectively.

Method	ADM	DDPM	iDDPM	Diffusion GAN	Projected GAN	StyleGAN	Unleashing Transformer	Mean
Reconstruction-based Methods								
AEROBLADE	60.19	76.66	70.05	49.56	50.27	40.71	51.97	57.06
Manifold Bias	**75.40**	**88.09**	**86.92**	**96.15**	**97.06**	**94.51**	**92.48**	**90.09**
Non-reconstruction Methods								
RIGID	72.62	87.94	84.53	93.19	93.98	83.11	92.40	86.82
MINDER	72.01	82.98	80.89	97.68	97.90	90.60	96.91	88.43
WePe	72.22	89.18	84.90	94.91	95.05	85.36	94.65	88.04
ConV	71.80	86.62	82.21	93.55	93.87	83.74	92.76	86.36
WaRPAD	**79.58**	93.90	**91.14**	**98.89**	**98.62**	**97.47**	**97.87**	**93.92**
MiLD (ours)	79.48	**94.51**	90.66	96.72	96.65	90.19	96.67	92.13

Table 3: AI-generated image detection performance (AUROC, %) on DailyBench. **Bold** and underline denote the best and the second best results within each category, respectively. G.I.2, N.B.2, Q.I., and Q.I.-E denote GPT-Image2, NanoBanana2, Qwen-Image, and Qwen-Image-E, respectively.

Method	FakeBench							ManipulationBench							Mean
	FLUX	FLUX2	G.I.2	N.B.2	Q.I.	SD35	ZImage	FLUX2-klein	FLUX_O	FLUX_R	G.I.2	N.B.2	Q.I.-E	Step1X-E	
Reconstruction-based Methods															
AEROBLADE	**85.95**	**76.10**	**63.03**	**62.25**	**73.35**	**70.20**	**77.34**	51.26	43.23	43.83	**68.92**	**61.12**	**72.48**	**61.76**	**65.06**
Manifold Bias	56.14	52.29	51.67	45.28	53.60	60.54	56.19	**56.81**	**51.67**	**48.94**	58.52	52.82	60.13	58.31	54.49
Non-reconstruction Methods															
RIGID	55.54	41.89	44.28	28.83	36.36	49.76	46.34	55.79	53.84	47.64	53.94	45.56	60.59	63.66	48.86
MINDER	46.24	38.27	35.06	32.26	34.62	54.40	46.73	53.56	53.28	46.52	54.38	48.81	51.16	51.88	46.23
WePe	26.88	27.63	25.76	30.76	23.46	41.68	34.81	55.58	**57.48**	49.21	50.16	48.21	53.61	53.98	41.37
ConV	22.62	24.87	**64.88**	**44.98**	23.13	36.40	31.30	54.72	56.92	49.97	**81.27**	**56.04**	50.38	45.12	45.90
WaRPAD	31.33	40.29	51.79	35.30	**46.78**	56.37	49.34	**56.32**	56.37	**51.79**	78.32	50.32	**80.53**	**71.16**	**54.00**
MiLD (ours)	**56.49**	**45.55**	53.59	29.59	43.81	**56.98**	**49.57**	54.96	54.84	49.22	61.31	45.21	61.05	62.54	51.77

Metrics and Implementation. We report the area under the receiver operating characteristic curve (AUROC) as the primary evaluation metric for all methods. Following RIGID (He et al., 2024), we adopt DINOv2 ViT-L/14 (Oquab et al., 2023) as the pretrained backbone. For MiLD, we bicubically resize each RGB image to 224×224 before extracting the low-bit morphology. We then construct the masked low-bit reversal for $E(x)$ at this encoder input resolution. We use $b = 2$ low-order bits and a flat square structuring element with side length $s = 3$, unless otherwise specified. These defaults come from the motivation stage design in Section 2 and were fixed before running the main benchmark experiments. The ablations below are post hoc sensitivity analyses and are not used to tune the reported main configuration.

4.2 Main Results

Performance across different benchmarks. Tables 1, 2, 10, 11, and 3 report the results on ImageNet, LSUN-Bedroom, GenImage, AIGCDetectBenchmark, and DailyBench, respectively. MiLD improves over RIGID on all five benchmark groups, with AUROC gains of 1.72, 5.31, 2.77, 1.18, and 2.91 percentage points, respectively. Across the five benchmark groups, MiLD achieves a com-

petitive mean AUROC of 80.14%, outperforming all evaluated baselines except WaRPAD. These results indicate that morphology-guided low-bit reversal captures useful detection cues from low-bit planes.

Average performance and time cost. Table 4 summarizes the five-dataset mean AUROC and time cost of the baselines and MiLD. We separately report the model initialization time and the image processing time required to process 100 images on a single NVIDIA L20 GPU with a batch size of 4. Speed denotes the ratio of each compared method's image processing time to the reported configuration. To quantify the performance-speed trade-off, we compute

$$\text{Eff.} = \frac{\text{AUROC} \times m}{t_{\text{proc}}}, \tag{14}$$

where AUROC is written in $[0, 1]$, $m = 100$ is the number of processed images, and t_{proc} is the processing time in seconds. Across the five benchmark datasets, MiLD trails WaRPAD's mean AUROC by 4.41 percentage points, while outperforming the other evaluated baselines on average. More importantly, because MiLD requires only two encoder forward passes, it uses nearly the same processing time as RIGID but improves its mean AUROC from 77.36% to 80.14%.

Table 4: Mean AI-generated image detection performance (AUROC, %) over five benchmarks, runtime per 100 images (s), and efficiency. **Bold** and underline denote the best and the second best results within each category, respectively. Higher is better for Mean and Eff. Lower is better for Init, Process, and Speed.

Method	Mean	Init	Process	Speed	Eff.
Reconstruction-based Methods					
AEROBLADE	67.05	**4.16**	**25.63**	14.82×	**2.62**
Manifold Bias	**74.00**	7.19	185.34	107.13×	0.40
Non-reconstruction Methods					
RIGID	77.36	4.09	**1.71**	0.99×	45.24
MINDER	76.55	4.34	3.32	1.92×	23.06
WePe	74.41	6.98	1.74	1.01×	42.76
ConV	77.20	3.96	8.36	4.83×	9.23
WaRPAD	**84.55**	**3.31**	23.16	13.39×	3.65
MiLD (ours)	80.14	3.77	1.73	1.00×	**46.32**

4.3 Analysis

Effect of Mask Field. Table 5 evaluates $E_{U_b}(x)$ with different fields U_b used to construct the reversal mask. For the field-based variants, the activity indicators and Hamming weights are computed from the listed field. The selected channels of the original low-bit image $L_b(x)$ are complemented. The frozen encoder always compares the same RGB image x with the resulting RGB image x'_{U_b}. Using x or $L_2(x)$ to construct the mask yields a weaker feature discrepancy than using a low-bit morphology residual. The one-bit residual gives the highest five-dataset mean in this ablation, while the default two-bit residual remains competitive. Global low-bit reversal achieves a mean AUROC of 77.76%, compared with 80.14% for the default morphology-guided mask. This comparison supports the benefit of selectively placing low-bit perturbations over applying them globally.

Effect of Hamming-State Selection. We ablate which morphology Hamming weights enter the reversal mask. For a set $S \subseteq \{0, 1, 2, 3\}$, the selected channel mask is $M_S(c, i) = a_c(i)\,\mathbb{I}\{w(i) \in S\}$. Table 6 reports the AUROC of $E_S(x) = \|[f(x) - f(x'_S)]/2\|_2^2$, where x'_S is constructed with M_S. The default $S = \{1, 2\}$ reaches an 80.14% mean in this controlled run. Among post-resize Hamming-state choices, selecting only $w = 2$ is stronger on all five datasets, reaching 80.51%, whereas adding $w = 3$ lowers the mean to 78.15%. The $w = 0$ mask selects no active channels and gives the expected 50% AUROC. Morphology-selected $w = 1$, $w = 2$, and $w \in \{1, 2\}$ each outperform the corresponding state-count-matched random masks. Extracting the morphology mask at native resolution further improves the mean AUROC from 80.14% to 80.62%. This result indicates that the useful spatial cue is already present before resizing and that extracting it at

Table 5: Mask-field and low-bit ablations on five datasets (AUROC, %).

Mask field U_b	ImageNet	LSUN-Bedroom	GenImage	AIGCDetectBenchmark	DailyBench	Mean
x	57.98	59.27	62.35	55.01	48.13	56.55
$L_2(x)$	87.95	90.34	82.67	83.85	52.06	79.37
$\mathcal{R}_s(L_1(x))$	87.12	**93.86**	83.32	83.56	**54.91**	**80.55**
$\mathcal{R}_s(L_2(x))$	**88.35**	92.13	**83.85**	**84.59**	51.77	80.14
$\mathcal{R}_s(L_3(x))$	87.05	90.78	83.23	84.38	51.61	79.41
$\mathcal{R}_s(L_4(x))$	85.28	90.49	81.54	83.67	51.92	78.58
Global	86.70	89.51	80.80	82.96	48.82	77.76

Table 6: Mask-selection and extraction-scale ablations on five datasets (mean AUROC, %). The state-count-matched random controls retain, for each image, the numbers of pixels assigned each selected Hamming weight while randomizing their locations and RGB channels. The exact input-energy-matched control randomizes locations within each channel and low-bit reversal-amplitude class. It preserves each image's channelwise flip counts and $\|x' - x\|_2^2$ exactly. This input-space energy is distinct from the encoder score $E(x)$. Random-control values average five seeds.

Mask construction	ImageNet	LSUN-Bedroom	GenImage	AIGCDetectBenchmark	DailyBench	Mean
$w = 0$	50.00	50.00	50.00	50.00	50.00	50.00
$w = 1$	84.90	88.31	80.37	77.66	52.37	76.72
$w = 2$	**88.54**	92.36	83.88	**85.22**	**52.55**	80.51
$w = 3$	86.14	89.00	80.19	82.88	49.17	77.48
$w \in \{1, 2\}$	88.35	92.14	83.84	84.61	51.78	80.14
$w \in \{1, 2, 3\}$	86.90	90.02	81.30	83.67	48.83	78.15
Native-derived $w \in \{1, 2\}$	88.41	**92.50**	**84.61**	85.20	52.40	**80.62**
Random, count matched $w = 1$	82.52	84.74	76.46	75.91	50.91	74.11
Random, count matched $w = 2$	86.45	87.85	81.53	81.32	49.89	77.41
Random, count matched $w \in \{1, 2\}$	86.52	87.54	81.61	81.24	49.59	77.30
Random, count and energy matched $w \in \{1, 2\}$	86.90	86.97	81.62	81.33	48.99	77.16

native resolution better preserves its discriminative information. To control perturbation size more tightly, we also randomize the default mask's positions within each RGB channel and each two-bit reversal-amplitude stratum. This preserves every image's channelwise flip counts and exact squared RGB input perturbation energy. Across five random seeds, this control reaches 77.16% mean AUROC, compared with 80.14% for the morphology mask. Thus the placement of low-bit reversals contributes beyond their input energy.

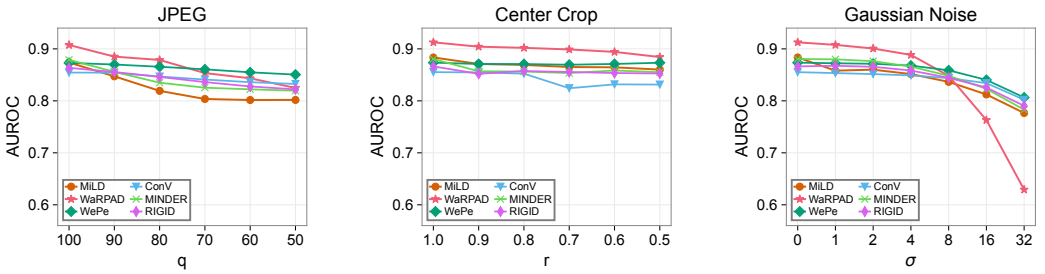
Robustness to Pretrained Backbones. We report the robustness of MiLD to pretrained backbone in Table 7. Specifically, the DINOv2 series is used to evaluate the effect of strong self-supervised visual representations at different model scales. CLIP ViT-L/14 (Radford et al., 2021) tests whether cross-modal pretrained representations are also suitable for MiLD. ResNet50 (He et al., 2016) provides a convolutional counterpart, including both supervised pretraining and SwAV (Caron et al., 2020) self-supervised pretraining. DINO ViT-S/16 (Caron et al., 2021) serves as a reference for earlier self-supervised ViT representations. Across the eight tested backbones, MiLD ranks first or second on seven of them, indicating strong adaptability to different pretrained representations. On the default DINOv2 ViT-L/14 backbone, WaRPAD exceeds MiLD by only 2.91 AUROC points, whereas MiLD surpasses WaRPAD by 14.50 points on CLIP ViT-L/14 and by 10.53 and 15.07 points on the supervised and SwAV ResNet50 backbones.

Table 7: Backbone ablation across non-reconstruction methods on ImageNet (AUROC, %). **Bold** and underline denote the best and second-best results within each backbone row, respectively.

Backbone	RIGID	MINDER	WePe	ConV	WaRPAD	**MiLD**
DINOv2 ViT-S/14	71.24	70.35	66.85	57.53	<u>74.28</u>	**75.37**
DINOv2 ViT-B/14	80.54	<u>81.22</u>	78.22	74.57	80.74	**83.15**
DINOv2 ViT-L/14	86.63	88.05	87.35	85.52	**91.26**	<u>88.35</u>
DINOv2 ViT-g/14	85.34	86.64	86.30	83.12	**88.69**	<u>87.44</u>
CLIP ViT-L/14	<u>77.34</u>	76.92	**77.88**	71.59	59.75	74.25
ResNet50 (supervised, V1)	<u>61.39</u>	60.99	59.77	52.08	55.60	**66.13**
ResNet50 (SwaV)	63.88	63.78	<u>64.59</u>	60.99	57.10	**72.17**
DINO ViT-S/16	<u>62.71</u>	57.82	55.25	55.82	62.64	**66.33**

Robustness to Corruptions. Figure 3 evaluates all non-reconstruction methods on ImageNet under JPEG compression, center cropping, and Gaussian noise, with the three settings shown in Figures 3a, 3b, and 3c, respectively. The Gaussian noise standard deviation is measured on the 0–255 RGB scale. Importantly, each corruption is applied identically to real and generated images, so that both classes undergo the same post-processing transformation. MiLD remains competitive under strong JPEG compression, center cropping and Gaussian noise. In contrast, WaRPAD is sensitive to strong Gaussian noise. At noise $\sigma = 32$, its AUROC drops to 62.92%. Tables 13, 14, and 15 provide the detailed AUROC values and Δ changes for the three corruption settings.

Robustness to Design Choices. Figure 4 summarizes the main design-choice ablations across the five datasets, with numerical values reported in Tables 16, 17, and 18. Figure 4a examines



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

Figure 3: Robustness of non-reconstruction methods under corruptions on ImageNet.



(a) Kernel size (b) Residual operator (c) Resize interpolation

Figure 4: MiLD ablation analysis of MiLD design choices across five datasets.

the effect of enlarging the morphological neighborhood beyond the default 3×3 choice. Larger 5×5 and 7×7 kernels reduce the mean AUROC, indicating that MiLD benefits from fine low-bit morphological structure rather than coarse spatial smoothing. Figure 4b shows that the bidirectional and median residuals are both effective, while the morphological gradient substantially reduces the mean AUROC. These results suggest that MiLD is sensitive to both the residual operator and the local scale of the residual field. Figure 4c shows that MiLD remains stable under common resize interpolation rules.

Detection beyond static occupancy. Table 8 compares static occupancy $P(x)$ with our feature-response score E on DRCT-2M (Chen et al., 2024). Under the same fixed higher-is-fake convention, $P(x)$ yields a mean AUROC of 40.11%, with positive detection directions for only five of the 16 generators. In contrast, E achieves 70.31% mean AUROC and remains above 50% for all 16 generators, including the 11 for which $P(x)$ reverses direction. These results show that E captures useful detection information beyond static state frequency. Fewer channel-disagreement locations in generated images need not imply a weaker representation response to the selected perturbation. Per-generator results are provided in Table 12.

Table 8: Mean AUROC (%) on DRCT-2M.

Score	Mean AUROC (%)	AUROC $> 50\%$
$P(x)$	40.11	5/16
$E(x)$	**70.31**	**16/16**

5 Conclusion

We presented MiLD, a training-free detector that mines useful forensic information from low-bit morphology for AI-generated image detection. Across multiple benchmarks, MiLD shows strong detection performance and offers a compact feature-discrepancy score. It also shows robustness to cropping, noise, and backbone choice.

AI Use Statement

We used generative AI tools to assist with language polishing. The authors reviewed all AI-assisted content and take responsibility for the final paper, including the text, claims, code, and artifacts produced with the aid of generative AI.

Ethics Statement

This paper studies passive detection of AI-generated imagery. Such tools should not be treated as definitive evidence about an individual image. We therefore emphasize that passive detectors should be combined with provenance, watermarking, and human review in high-stakes use cases.

Reproducibility Statement

The method is fully deterministic given an input image, a fixed DINOv2 backbone, and the stated hyperparameters. Section 3 specifies the low-bit decomposition, morphology operator, representation response, and score. Section 4 reports the evaluation protocol, datasets, baselines, metrics, and the fixed hyperparameter setting used for the formal result summaries. We will include anonymized code and result manifests as supplementary material.

References

Jonathan Brokman, Amit Giloni, Omer Hofman, Roman Vainshtein, Hisashi Kojima, and Guy Gilboa. Manifold induced biases for zero-shot and few-shot detection of generated images. In *International Conference on Learning Representations*, volume 2025, pp. 19977–20002, 2025.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *ICML*, volume 2, pp. 6, 2024.

Renxi Cheng, Jie Gui, and Hongsong Wang. Raid: Towards robust ai-generated image detection with bit-reversed images. *arXiv preprint arXiv:2607.28974*, 2026.

Sungik Choi, Hankook Lee, and Moontae Lee. Training-free detection of ai-generated images via cropping robustness. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Sungik Choi, Hankook Lee, Jaehoon Lee, Robin Kim, Stanley Jungkyu Choi, and Moontae Lee. A debiased reconstruction-based framework for training-free detection of ai-generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3254–3263, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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*, pp. 3247–3258. PMLR, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Zhiyuan He, Pin-Yu Chen, and Tsung-Yi Ho. Rigid: A training-free and model-agnostic framework for robust ai-generated image detection. *arXiv preprint arXiv:2405.20112*, 2024.

Xin Jiang, Hao Tang, Junyao Gao, Meiqi Cao, Fei Shen, Dongming Zhang, and Yongdong Zhang. Dailybench: A unified benchmark for ai-generated and manipulated images from modern generative models. *arXiv preprint arXiv:2607.24016*, 2026.

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

Jun Nie, Yonggang Zhang, Tongliang Liu, Yiu-ming Cheung, Bo Han, and Xinmei Tian. Epistemic uncertainty for generated image detection. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489. IEEE, 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. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9130–9140. IEEE, 2024.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

Chung-Ting Tsai, Ching-Yun Ko, I-Hsin Chung, Yu-Chiang Frank Wang, and Pin-Yu Chen. Understanding and improving training-free ai-generated image detections with vision foundation models. *arXiv preprint arXiv:2411.19117*, 2024.

Hongsong Wang, Renxi Cheng, Yang Zhang, Chaolei Han, and Jie Gui. Lota: Bit-planes guided ai-generated image detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17246–17255. IEEE, 2025.

Hongsong Wang, Renxi Cheng, Chaolei Han, and Jie Gui. Attribution as retrieval: model-agnostic ai-generated image attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14062–14072, 2026.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8692–8701. IEEE, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22388–22398. IEEE, 2023.

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*, volume 2025, pp. 70702–70720, 2025.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Yonggang Zhang, Jun Nie, Xinmei Tian, Mingming Gong, Kun Zhang, and Bo Han. Detecting generated images by fitting natural image distributions. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023.

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. *Advances in neural information processing systems*, 36:77771–77782, 2023.

A Additional Details

A.1 Related Work

Training-based detection. Training-based detectors formulate AI-generated image detection as supervised classification over labeled real and generated images. Early CNN-based and frequency-based methods show that synthetic images contain visual, textural, or spectral differences that are often invisible to human observers but can be learned by discriminative models (Wang et al., 2020; Frank et al., 2020). Later methods improve generalization by using stronger pretrained representations, hybrid image features, reconstruction-error supervision, or larger and more diverse training corpora. CLIP-based and hybrid-feature methods aim to transfer across generator families (Ojha et al., 2023; Yan et al., 2025), while reconstruction-error learning exploits discrepancies induced by diffusion-model or latent reconstruction (Wang et al., 2023; Luo et al., 2024). Recent low-bit learning pipelines further indicate that the least significant bit planes contain useful detection and attribution information, but they still rely on learned components such as bit-plane guided inputs, retrieval models, patch selection, or classification heads (Wang et al., 2025; 2026).

Training-free detection. Training-free detectors avoid learning a detector head on target real/fake collections and instead use fixed reconstruction criteria, pretrained-model responses, or hand-designed perturbation rules. Reconstruction-based methods compare real and generated images through autoencoder or diffusion-model reconstruction. AEROBLADE uses autoencoder reconstruction error for latent diffusion images (Ricker et al., 2024), Manifold Bias studies reconstruction-induced biases of generated images (Brokman et al., 2025), and RDD debiases reconstruction-based detection for improved generalization (Choi et al., 2026). Non-reconstruction methods instead probe fixed representations, perturbation sensitivity, uncertainty, or distributional consistency. RIGID measures representation stability under Gaussian perturbations (He et al., 2024), MINDER examines representation-space discrepancy with vision foundation models (Tsai et al., 2024), WePe estimates epistemic uncertainty through weight perturbation (Nie et al., 2025), ConV tests consistency with natural-image distribution fitting (Zhang et al., 2025), and WaRPAD combines cropping robustness with Haar-wavelet high-frequency perturbations (Choi et al., 2025).

A.2 Mechanism Analysis

Magnitude and directional sensitivity. Equation 11 separates the input energy $A(x)$ from local directional sensitivity. To examine whether the observed finite response is explained by input energy alone, we use the diagnostic score

$$\widetilde{E}(x) = \frac{E(x)}{A(x) + \epsilon}, \qquad \epsilon = 10^{-12}. \tag{15}$$

This finite-difference ratio is not an exact measurement of the local Rayleigh quotient $\mathcal{R}_\phi$.

We compare the spatial occupancy P, input energy A, discrepancy E, and normalized response $\widetilde{E}$. Energy is measured in RGB coordinates scaled to $[0, 1]$, before encoder normalization. For two-bit reversal, a selected value l changes by $(3 - 2l)/255$, giving

$$A(x) = \frac{1}{4 \cdot 255^2} \sum_{c,i} M(c, i)\big(3 - 2L_2(x)_{c,i}\big)^2. \tag{16}$$

Thus, equal flip counts need not imply equal energy. The factor $1/4$ follows from the half-difference definition of η. We also normalize the responses of the five existing random controls that preserve each image's channelwise flip counts and reversal-amplitude strata. Their input energies equal the morphology energy exactly for every image.

Information beyond input energy. Table 9 shows that A alone reaches 73.93% mean AUROC, while $\widetilde{E}$ retains 79.22% after normalization. The original E reaches 80.15%, exceeding both A and $\widetilde{E}$ on average and on the first four datasets. Morphology exceeds the energy-matched random control by 2.98 percentage points for E and by 3.29 points for $\widetilde{E}$. The normalized gap is positive on all five datasets, showing that the selected perturbation pattern carries information beyond its scalar input energy.

Table 9: Mechanism diagnostics across five datasets (AUROC, %).

Score	ImageNet	LSUN-Bedroom	GenImage	AIGCDetectBenchmark	DailyBench	Mean
Occupancy P	77.31	90.42	80.66	77.71	58.36	76.89
Input energy A	74.32	85.26	77.35	75.85	56.87	73.93
MiLD E	88.35	92.14	83.84	84.61	51.78	80.14
Normalized $\widetilde{E}$	87.41	90.92	82.50	83.11	52.14	79.22
Energy-matched random E	86.90	86.97	81.62	81.33	48.99	77.16
Energy-matched random $\widetilde{E}$	85.69	85.04	79.99	79.47	49.45	75.93

A.3 Detailed Dataset Descriptions

This section describes the evaluation datasets used in the main paper. Together, the five benchmark groups cover controlled single-domain settings, broad multi-generator collections, and recent open-world generation and manipulation scenarios.

ImageNet. We use the ImageNet-based image collection released by Stein et al. (2023) through the DGM-Eval repository.[2] The collection pairs real ImageNet images with synthetic samples from models trained on the same image domain. Our evaluation covers nine generators: ADM, ADMG, BigGAN, DiT-XL-2, GigaGAN, LDM, StyleGAN-XL, RQ-Transformer, and Mask-GIT.

LSUN-Bedroom. For the bedroom domain, we use real LSUN-Bedroom images and generated samples from the collection of Stein et al. (2023), also distributed through DGM-Eval.[3] Our evaluation includes seven generators: ADM, DDPM, iDDPM, StyleGAN, Diffusion-Projected GAN, Projected GAN, and Unleashing Transformers. This benchmark evaluates detection across generative model families within a shared indoor-scene domain.

GenImage. GenImage (Zhu et al., 2023) is a large-scale benchmark pairing real images from ImageNet with AI-generated images spanning its 1,000 semantic classes. The dataset and download instructions are available from the official repository.[4] We evaluate the eight generator subsets: Midjourney, Stable Diffusion v1.4, Stable Diffusion v1.5, ADM, GLIDE, Wukong, VQDM, and BigGAN.

AIGCDetectBenchmark. We use the evaluation collection released with PatchCraft (Zhong et al., 2023), whose real and generated images can be accessed through the official benchmark repository.[5] Our reported evaluation covers 15 source subsets: ProGAN, StyleGAN, BigGAN, StarGAN, GauGAN, StyleGAN2, WhichFaceIsReal (WFIR), ADM, GLIDE, Midjourney, Stable Diffusion v1.4, Stable Diffusion v1.5, VQDM, Wukong, and DALL-E 2. These subsets combine GAN-based synthesis, image translation, and diffusion-based generation, providing a broader range of image sources than the single-domain collections above.

DailyBench. DailyBench (Jiang et al., 2026) evaluates both fully synthesized images and localized manipulations of real photographs. Data access is provided through its project page.[6] Its real-image pool is drawn from LAION-Aesthetics V2. FakeBench contains images from seven generators: FLUX.1, FLUX.2, GPT-Image 2, Nano Banana 2, Qwen-Image-2512, SD3.5 Large, and Z-Image. ManipulationBench covers FLUX.2-klein, FLUX-Fill, GPT-Image 2, Nano Banana 2, Qwen-Image-Edit-2511, and Step1X-Edit-v1p2. FLUX-Fill is evaluated separately with object masks and random masks, denoted by FLUX_Object and FLUX_Random, yielding seven manipulation subsets. Together, the two subsets test detection of recent generative models and partial synthetic content embedded in otherwise real images.

[2] https://github.com/layer6ai-labs/dgm-eval

[3] https://github.com/layer6ai-labs/dgm-eval

[4] https://github.com/GenImage-Dataset/GenImage

[5] https://github.com/Ekko-zn/AIGCDetectBenchmark

[6] https://dailybench.github.io/

Table 10: AI-generated image detection performance (AUROC, %) on GenImage. **Bold** and underline denote the best and the second best results within each category, respectively.

Method	ADM	BigGAN	GLIDE	Midjourney	SDv1.4	SDv1.5	VQDM	Wukong	Mean
Reconstruction-based Methods									
AEROBLADE	**85.93**	**95.18**	**97.91**	**82.10**	**92.95**	**94.46**	73.92	**93.98**	**89.55**
Manifold Bias	67.21	92.65	81.93	47.72	66.05	65.55	**88.47**	63.77	71.67
Non-reconstruction Methods									
RIGID	87.92	96.30	94.30	76.39	65.80	66.23	92.60	69.13	81.08
MINDER	85.54	87.90	77.63	67.65	68.11	67.89	92.86	71.83	77.43
WePe	82.76	95.55	83.79	52.94	66.27	65.64	95.42	72.03	76.80
ConV	80.41	96.74	79.85	**88.76**	82.00	81.76	94.15	86.15	86.23
WaRPAD	**96.61**	**99.68**	97.47	76.53	**89.71**	**90.58**	**95.85**	**88.45**	**91.86**
MiLD (ours)	89.38	96.64	**98.10**	83.34	69.74	69.99	91.41	72.17	83.85

Table 11: AI-generated image detection performance (AUROC, %) on AIGCDetectBenchmark. **Bold** and underline denote the best and the second best results within each category, respectively.

Method	ProGAN	StyleGAN	BigGAN	StarGAN	GauGAN	StyleGAN2	WFIR	ADM	Glide	Midjourney	SDv1.4	SDv1.5	VQDM	Wukong	DALLE2	Mean
Reconstruction-based Methods																
AEROBLADE	46.27	39.32	37.38	31.83	37.27	38.96	34.18	**89.35**	**99.03**	**86.73**	**96.47**	**96.69**	79.40	**96.85**	**99.43**	67.28
Manifold Bias	**96.80**	**73.85**	**86.22**	**53.86**	**97.87**	**72.72**	**72.63**	68.21	82.13	47.61	66.24	65.89	**88.00**	64.10	73.33	**73.96**
Non-reconstruction Methods																
RIGID	95.41	75.28	88.60	78.21	94.74	83.40	94.85	87.45	94.62	76.24	67.24	67.01	92.16	69.25	86.63	83.41
MINDER	96.93	**82.38**	**92.98**	90.89	96.62	**88.68**	**97.04**	84.53	77.51	66.38	67.68	67.58	92.04	70.47	67.58	82.62
WePe	95.25	75.36	87.02	75.19	93.60	86.23	87.85	82.38	83.56	51.72	65.81	64.91	95.04	72.03	61.49	78.50
ConV	94.85	71.29	88.75	82.06	93.57	80.72	81.20	78.69	78.30	**87.08**	80.31	79.44	93.27	85.34	54.90	81.98
WaRPAD	**97.50**	80.69	92.24	**97.58**	**96.80**	79.53	74.91	**98.59**	**99.15**	80.95	**94.02**	**93.65**	**98.10**	**92.44**	**99.77**	**91.73**
MiLD (ours)	93.80	76.17	84.83	73.44	93.93	84.00	90.28	89.61	98.48	84.07	71.30	71.44	91.40	73.28	92.80	84.59

A.4 Detailed Experimental Results

This section provides the detailed experimental results omitted from the main paper because of space constraints. We first report the complete benchmark tables for GenImage and AIGCDetectBenchmark, then provide the numerical ablations and corruption-robustness results supporting the analysis in Section 4.3. In all MiLD computations, we use $\epsilon_a = 10^{-8}$ for the activity indicator used by the low-bit reversal mask. The implemented detector uses the feature discrepancy $E(x)$ as the image-level detection score.

Additional benchmark results. Tables 10 and 11 report complete per-source AUROC values on GenImage and AIGCDetectBenchmark, respectively. These two benchmark groups complement the ImageNet, LSUN-Bedroom, and DailyBench results in the main paper.

Ablation study protocol. Unless otherwise stated, the following ablations keep the same evaluation datasets, preprocessing, DINOv2 ViT-L/14 backbone, and AUROC metric as the main experiments. They isolate the effects of the low-bit morphology design and common image corruptions. The default MiLD configuration was fixed from the motivation stage design before the main benchmark experiments. These ablations are post hoc sensitivity analyses and are not used for parameter tuning.

Analysis of time cost. The timing comparison in Table 4 uses each method's current full configuration, so the reconstruction-based methods are not normalized to one generative model. AEROBLADE combines the autoencoders from SD v1.1, SD v2.1-base, and Kandinsky 2.1 and includes LPIPS distance computation, giving 25.63s per 100 images. It is still relatively fast because it only uses the autoencoder reconstruction path and does not run diffusion UNet denoising. Manifold Bias uses one SD v1.4 pipeline, but its measured time also includes LLaVA caption generation and CLIP feature computation, giving 185.34s per 100 images. We also implemented and timed RDD under the same 100-image timing protocol. Its current configuration combines SD v1.4, SD v2-base, and MiniSD, and the summed processing time is 1055.18s per 100 images. RDD is much slower because each model performs reconstruction, noise-prediction, decoding, and distance computations. In particular, besides four image reconstructions for the original, blurred, rotated, and blurred-rotated inputs, the decoder-L2 path decodes both predicted and target noise for 16 samples

Table 12: Per-generator AUROC (%) on DRCT-2M. Both scores use the fixed higher-is-fake convention. **Bold** denotes the better result for each generator.

Generator	$P(x)$	$E(x)$
ControlNet-Canny-SDXL	52.91	**82.12**
LCM-LoRA-SDv1.5	41.34	**73.79**
LCM-LoRA-SDXL	40.87	**77.84**
LDM-text2im	50.99	**90.54**
SD-ControlNet-Canny	24.28	**70.46**
SD-Turbo	17.73	**71.39**
SD2.1-ControlNet-Canny	27.85	**77.12**
SDXL-Turbo	17.98	**73.05**
SD2.1	28.54	**53.17**
SD2-Inpainting	**69.08**	64.56
SD-Inpainting	**75.78**	62.09
SD1.4	32.94	**58.19**
SD1.5	32.44	**57.95**
SDXL-Inpainting	**72.60**	63.92
SDXL-Base	35.47	**57.43**
SDXL-Refiner	20.95	**91.41**
Mean	40.11	**70.31**

in two branches, which adds 64 extra decoded samples per model. This pilot timing is about $41.2\times$ the AEROBLADE processing time, so we did not include RDD in the large-scale ablation suite under our computational budget.

For non-reconstruction methods, the cost is instead dominated by the number and size of pretrained-encoder evaluations. RIGID evaluates the original and Gaussian-noised images, and WePe evaluates the original backbone and the weight-perturbed backbone, so both require two encoder forward passes per image and take 1.71s and 1.74s per 100 images, respectively. MINDER compares the original image with a noised image and a blurred image with a sharpened image, resulting in four encoder forward passes and 3.32s per 100 images. ConV evaluates five crop pairs, corresponding to ten crop-level encoder forward passes and 8.36s per 100 images. WaRPAD first resizes each image to 896×896, partitions it into sixteen 224×224 patches, and evaluates both the original and wavelet-perturbed patch sets. This amounts to 32 patch-level encoder evaluations per image and gives 23.16s per 100 images. MiLD processes 100 images in 1.73s under the reported implementation.

Robustness of non-reconstruction baselines. Tables 13, 14, and 15 compare the same external corruption settings across the non-reconstruction methods and report how MiLD behaves under each corruption. In Tables 13, 14, and 15, Δ is the clean AUROC minus the AUROC at the strongest listed corruption, in percentage points. Darker orange indicates a larger drop within each table. The largest drop in each table is bold.

Table 13: Center-crop robustness across non-reconstruction methods on ImageNet (AUROC, %).

Retained area	RIGID	MINDER	WePe	ConV	WaRPAD	MiLD
100%	86.63	88.05	87.35	85.52	91.26	88.35
90%	85.20	85.69	87.09	85.48	90.42	87.10
80%	85.74	85.64	87.08	85.23	90.20	86.91
70%	85.55	85.35	86.95	82.43	89.89	86.51
60%	85.35	85.84	87.10	83.17	89.42	86.45
50%	85.27	85.51	87.33	83.13	88.43	86.01
Δ	1.36	2.54	0.02	2.39	**2.83**	2.34

Morphological kernel size. We choose $s = 3$ a priori because it is the smallest non-trivial odd structuring element for the opening–closing residual. A 1×1 structuring element makes opening and closing reduce to the identity map, yielding a zero morphological residual. We evaluate larger 5×5 and 7×7 kernels only as sensitivity analyses rather than using them for model selection.

Table 14: Gaussian noise robustness across non-reconstruction methods on ImageNet (AUROC, %). Standard deviations use the 0–255 RGB scale.

Noise σ	RIGID	MINDER	WePe	ConV	WaRPAD	MiLD
0	86.63	88.05	87.35	85.52	91.26	88.35
1	86.75	87.99	87.23	85.34	90.77	85.81
2	86.55	87.63	87.13	85.15	90.07	86.00
4	85.84	86.59	86.78	84.89	88.84	85.15
8	84.43	84.82	85.88	84.38	84.83	83.65
16	82.56	82.32	84.03	83.41	76.31	81.23
32	79.04	78.30	80.65	80.21	62.92	77.66
Δ	7.59	9.75	6.70	5.31	**28.34**	10.69

Table 15: JPEG compression robustness across non-reconstruction methods on ImageNet (AUROC, %).

JPEG quality	RIGID	MINDER	WePe	ConV	WaRPAD	MiLD
Clean	86.63	88.05	87.35	85.52	91.26	88.35
100	86.35	87.90	87.30	85.44	90.75	87.37
90	85.55	85.58	86.99	85.42	88.51	84.72
80	84.60	83.49	86.56	84.68	87.87	81.92
70	83.61	82.53	86.06	84.10	85.34	80.36
60	82.77	82.22	85.49	83.59	84.34	80.16
50	82.21	81.95	85.05	83.25	82.48	80.18
Δ	4.42	6.10	2.30	2.27	**8.78**	8.17

Table 16: Morphological-kernel ablations across five datasets (AUROC, %).

Kernel size	ImageNet	LSUN-Bedroom	GenImage	AIGCDetectBenchmark	DailyBench	Mean
3×3	**88.35**	**92.14**	**83.84**	**84.61**	**51.78**	**80.14**
5×5	82.22	88.91	80.61	77.90	54.73	76.88
7×7	73.25	82.59	75.76	70.67	55.29	71.51

Morphological operator. We further replace the signed bidirectional top-hat residual with two local alternatives while keeping the 3×3 structuring element fixed. The morphological gradient uses $\delta_s(L_b) - \varepsilon_s(L_b)$, where δ_s and ε_s denote dilation and erosion with the same $s \times s$ structuring element, and the median residual uses $L_b - \mathrm{median}_s(L_b)$. Table 17 shows that the signed opening–closing residual used by MiLD remains competitive, while the median residual gives a slightly higher mean in this E-only ablation and the morphological gradient is weaker. We retain the signed opening–closing residual as the default because it is the morphology-motivated configuration fixed before the main experiments.

Table 17: Morphological operator ablations across five datasets (AUROC, %).

Residual operator	ImageNet	LSUN-Bedroom	GenImage	AIGCDetectBenchmark	DailyBench	Mean
Bidirectional	**88.35**	**92.13**	**83.85**	**84.59**	**51.77**	**80.14**
Gradient	78.70	83.48	80.57	72.30	53.68	73.75
Median	88.96	92.47	83.67	85.75	50.34	80.24

Resize interpolation. We further test whether the E-only score is sensitive to the interpolation operator used before the low-bit split. In this ablation, each image is first resized to 224×224 with the specified interpolation rule, and the low-bit morphology and feature discrepancy $E(x)$ are then computed exactly as in the main protocol. Table 18 shows that the E-only score is largely stable across common resize operators. Bicubic resizing is the predefined main-protocol preprocessing choice rather than a setting selected from this ablation.

A.5 Score Correlation Heatmaps

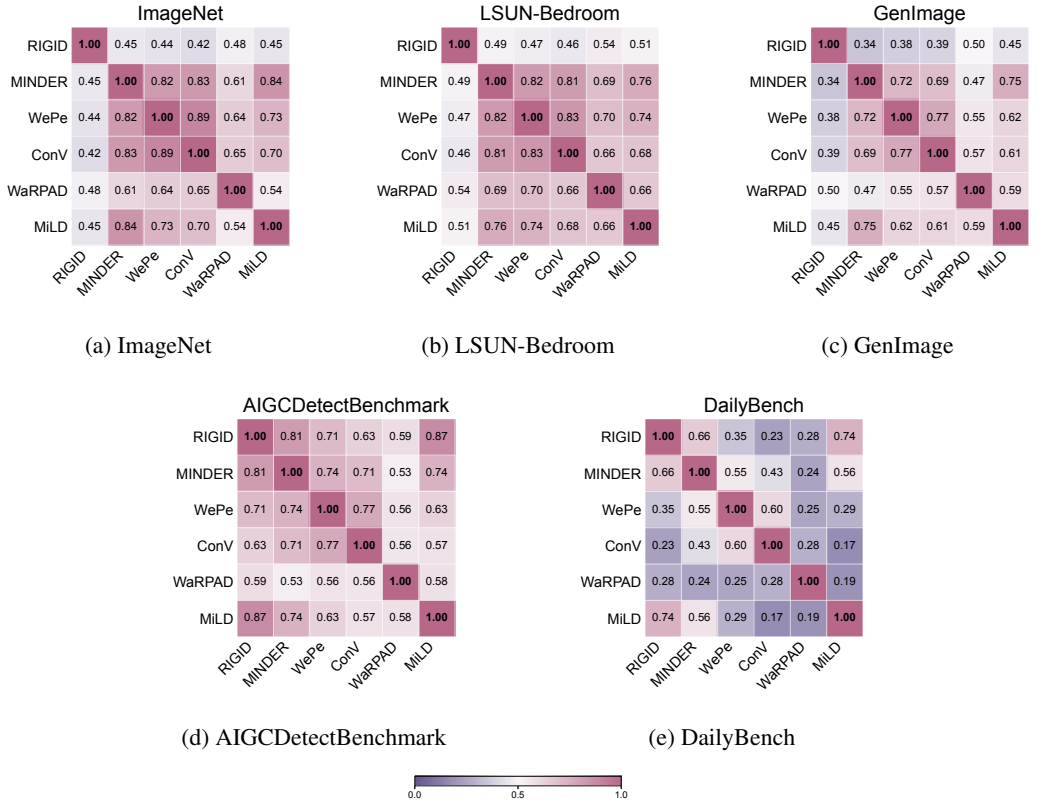
We further examine whether MiLD captures cues that are redundant with existing training-free detectors. For each method, we collect the image-level scores from each benchmark group, convert

Table 18: Resize interpolation ablations across five datasets (AUROC, %).

Resize interpolation	ImageNet	LSUN-Bedroom	GenImage	AIGCDetectBenchmark	DailyBench	Mean
Bicubic	**88.35**	**92.14**	**83.84**	**84.61**	**51.78**	**80.14**
Bilinear	88.30	91.59	83.92	84.63	53.01	80.29
Lanczos	88.21	92.17	83.74	84.33	51.22	79.93

all scores to the real-positive direction, and compute Spearman rank correlations over the pooled score vectors within that group. For MiLD, this uses $-\log E(x)$. This monotonic transformation preserves AUROC order while putting MiLD in the same real-positive convention as the other detectors.

Figure 5 shows two consistent patterns. First, feature-probing baselines such as MINDER, WePe, and ConV form a moderately correlated block on most benchmark groups, suggesting partially overlapping representation cues. Second, MiLD has substantial but imperfect correlation with existing detectors, indicating that the low-bit morphology construction changes the score ordering rather than only reproducing a prior perturbation-based cue.



(a) ImageNet (b) LSUN-Bedroom (c) GenImage

(d) AIGCDetectBenchmark (e) DailyBench

Figure 5: Correlation heatmaps across five benchmark groups. Cell annotations report Spearman correlation values after converting all detectors to the real-positive score direction.

A.6 Supplementary Real-Positive Average Precision

Tables 19–23 report real-positive average precision (AP, %). Following RIGID (He et al., 2024), real images are used as the positive class. For MiLD, we use $-E(x)$ from the mainline CLS discrepancy scores. MiLD also consistently outperforms RIGID in real-positive AP across all five benchmarks, with gains of 2.03, 5.54, 2.39, 1.13, and 4.15 percentage points on ImageNet, LSUN-Bedroom, GenImage, AIGCDetectBenchmark, and DailyBench, respectively.

Table 19: AI-generated image detection performance (real-positive AP, %) on ImageNet. **Bold** and underline denote the best and the second best results within each category, respectively.

Method	ADM	ADMG	LDM	DiT	BigGAN	GigaGAN	StyleGAN XL	RQ-Transformer	Mask GIT	Mean
Reconstruction-based Methods										
AEROBLADE	57.18	60.06	55.05	55.04	50.01	44.88	45.46	57.65	57.51	53.65
Manifold Bias	**74.73**	**71.64**	**78.94**	**73.44**	**88.39**	**79.92**	**86.93**	**82.80**	**81.79**	**79.84**
Non-reconstruction Methods										
RIGID	86.51	80.55	76.79	69.28	93.10	87.09	84.86	92.46	91.56	84.69
MINDER	85.10	78.11	80.07	71.73	**94.32**	89.59	87.05	**93.89**	92.27	85.79
WePe	88.03	80.75	76.06	70.49	92.43	91.38	**92.90**	92.82	86.21	85.67
ConV	84.70	76.27	74.10	66.34	91.56	89.30	91.38	89.58	86.38	83.29
WaRPAD	**89.57**	83.98	**88.06**	**85.65**	94.10	**91.79**	92.52	90.69	**94.64**	**90.11**
MiLD (ours)	89.00	**84.82**	81.81	75.51	91.78	87.44	85.80	92.63	91.65	86.72

Table 20: AI-generated image detection performance (real-positive AP, %) on LSUN-Bedroom. **Bold** and underline denote the best and the second best results within each category, respectively.

Method	ADM	DDPM	iDDPM	Diffusion GAN	Projected GAN	StyleGAN	Unleashing Transformer	Mean
Reconstruction-based Methods								
AEROBLADE	61.82	77.12	71.17	50.46	50.94	44.68	52.13	58.33
Manifold Bias	**75.33**	**88.36**	**87.00**	**96.26**	**97.11**	**94.50**	**92.83**	**90.20**
Non-reconstruction Methods								
RIGID	71.99	88.37	84.96	93.69	94.47	80.91	92.68	86.73
MINDER	68.92	80.88	78.82	97.86	98.09	90.46	97.07	87.44
WePe	69.23	88.68	83.77	95.24	95.37	84.30	94.72	87.33
ConV	68.58	85.49	80.40	93.54	94.03	82.45	92.51	85.29
WaRPAD	78.40	93.66	90.89	**98.97**	**98.73**	**97.62**	**98.00**	**93.75**
MiLD (ours)	**79.07**	**94.88**	**91.27**	96.91	96.70	89.99	97.05	92.27

Table 21: AI-generated image detection performance (real-positive AP, %) on GenImage. **Bold** and underline denote the best and the second best results within each category, respectively.

Method	ADM	BigGAN	GLIDE	Midjourney	SDv1.4	SDv1.5	VQDM	Wukong	Mean
Reconstruction-based Methods									
AEROBLADE	**85.73**	**96.91**	**98.25**	**83.43**	**95.42**	**96.23**	74.15	**96.04**	**90.77**
Manifold Bias	68.37	92.51	82.09	47.17	64.65	63.49	**88.45**	62.55	71.16
Non-reconstruction Methods									
RIGID	86.63	96.12	94.16	77.12	65.18	65.45	91.36	68.16	80.52
MINDER	84.17	87.62	74.70	68.78	66.58	66.20	91.46	69.74	76.16
WePe	81.00	95.34	81.36	53.60	64.06	63.56	95.16	70.74	75.60
ConV	77.81	96.57	77.19	**89.98**	82.58	82.57	93.40	86.87	85.87
WaRPAD	**97.56**	**99.64**	**98.12**	77.98	**91.96**	**92.42**	**96.79**	**90.71**	**93.15**
MiLD (ours)	87.69	96.46	98.01	83.14	68.25	68.58	90.32	70.88	82.92

Table 22: AI-generated image detection performance (real-positive AP, %) on AIGCDetectBenchmark. **Bold** and underline denote the best and the second best results within each category, respectively.

Method	ProGAN	StyleGAN	BigGAN	StarGAN	GauGAN	StyleGAN2	WFIR	ADM	Glide	Midjourney	SDv1.4	SDv1.5	VQDM	Wukong	DALLE2	Mean
Reconstruction-based Methods																
AEROBLADE	47.62	42.51	41.63	46.34	42.16	42.41	41.82	**88.57**	**99.14**	**86.88**	**97.55**	**97.70**	78.79	**97.83**	**99.46**	70.03
Manifold Bias	**95.89**	**66.80**	**85.21**	**51.70**	**97.64**	**68.93**	**71.26**	69.12	81.92	46.92	64.31	64.15	**88.15**	63.10	73.25	**72.56**
Non-reconstruction Methods																
RIGID	95.30	70.77	86.64	78.28	94.57	82.35	95.20	86.00	94.41	76.89	66.39	65.80	90.92	68.23	87.27	82.60
MINDER	**96.73**	**82.26**	**91.66**	92.22	96.14	**88.11**	**97.13**	83.12	73.95	67.29	65.81	65.73	90.63	67.88	68.72	81.83
WePe	94.30	74.64	86.87	78.10	93.33	84.94	87.56	80.28	81.10	52.91	63.56	62.30	94.81	70.37	62.06	77.81
ConV	93.72	70.51	87.77	83.45	93.18	77.83	80.87	75.79	75.25	**88.39**	81.05	79.83	92.47	85.81	54.51	81.36
WaRPAD	96.69	82.00	90.75	**97.93**	**96.67**	78.46	75.17	**98.91**	**99.29**	80.95	**94.80**	**94.29**	**98.32**	**93.37**	**99.80**	**91.83**
MiLD (ours)	93.48	74.14	82.93	74.42	93.73	83.57	90.67	88.01	98.36	83.65	69.38	69.49	90.31	71.19	92.65	83.73

Table 23: AI-generated image detection performance (real-positive AP, %) on DailyBench. **Bold** and underline denote the best and the second best results within each category, respectively. G.I.2, N.B.2, Q.I., and Q.I.-E denote GPT-Image2, NanoBanana2, Qwen-Image, and Qwen-Image-E, respectively.

Method	FakeBench							ManipulationBench							Mean
	FLUX	FLUX2	G.I.2	N.B.2	Q.I.	SD35	ZImage	FLUX2-klein	FLUX_O	FLUX_R	G.I.2	N.B.2	Q.I.-E	Step1X-E	
Reconstruction-based Methods															
AEROBLADE	**87.13**	**75.75**	**63.79**	**68.00**	**75.64**	**71.05**	**78.39**	38.59	33.92	46.79	**62.87**	**53.05**	**67.95**	**51.73**	**62.47**
Manifold Bias	54.43	50.86	51.97	45.95	51.37	58.19	54.61	**42.22**	**39.40**	**49.41**	43.04	40.57	45.75	43.91	47.98
Non-reconstruction Methods															
RIGID	57.84	44.97	45.03	38.07	41.13	50.73	48.18	41.39	41.89	49.28	41.82	35.57	50.75	51.83	45.61
MINDER	49.84	42.98	41.81	39.18	40.56	54.13	48.84	40.10	41.37	48.66	41.87	36.42	42.11	41.80	43.55
WePe	36.88	37.04	36.80	39.12	35.48	45.27	40.45	41.80	44.30	49.74	36.86	37.03	41.28	41.38	40.24
ConV	35.03	35.79	**59.90**	**50.00**	35.22	40.47	38.18	41.11	43.65	50.59	**74.84**	**45.71**	39.32	34.82	44.62
WaRPAD	38.48	43.04	49.15	40.55	47.73	56.13	49.45	**42.99**	45.49	**53.65**	65.86	38.98	**71.50**	**59.84**	**50.20**
MiLD (ours)	**58.86**	**47.94**	56.02	41.41	**49.27**	**58.39**	**52.67**	41.34	**45.60**	53.53	50.50	38.81	50.75	51.48	49.75