ASYMMETRY-GUIDED ITERATIVE ADAPTATION FOR GENERALIZABLE AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The continual emergence of new generative models challenges AI-generated image detectors whose forensic cues may fail beyond known sources. Empirical analysis indicates that target images initially predicted as synthetic provide useful signals for recognizing missed images from the same source, allowing partial cross-generator generalization to seed same-source adaptation. Based on this insight, we propose an iterative framework that repeatedly adapts with synthetic predictions to expand detection coverage. The framework uses asymmetric supervision: predicted-synthetic target images provide source-specific signals, while trusted real-image replay supplies real supervision without reinforcing missed synthetic images as real and suppresses false positives during iterative expansion. To overcome the limited coverage of initial synthetic predictions on difficult source distributions, we further introduce stochastic exploration, provisionally labeling a small subset of the remaining target images as synthetic while replay constrains noisy updates. Exploration accelerates recall expansion and improves final recall on difficult sources. Extensive experiments across unseen generators and diverse synthesis paradigms demonstrate state-of-the-art performance, outperforming recent methods by a substantial margin.

1 INTRODUCTION

Rapid advances in image generation have made realistic visual content increasingly accessible, enabling applications in creative production and digital communication. However, the same capabilities also facilitate visual fabrication and misinformation, undermining trust in digital imagery. Detecting AI-generated images is therefore an increasingly important problem. In practical deployments, a detector must distinguish real images from synthetic images produced by generators that were unavailable during training. As generative architectures, training procedures, and synthesis pipelines continue to evolve, reliable generalization to unseen generative sources remains a central challenge (Ojha et al., 2023; Li et al., 2026b).

Existing research has approached this challenge through transferable forensic cues and increasingly powerful visual representations. Wang et al. (2020) showed that a classifier trained on a single generator can generalize to other CNN-based generators with appropriate data augmentation. DIRE (Wang et al., 2023) instead uses the discrepancy between an image and its diffusion-based reconstruction as a detection representation. Building on large-scale pretraining, UniversalFakeDetect (Ojha et al., 2023) demonstrated the effectiveness of pretrained feature spaces, while FatFormer (Liu et al., 2024), PPM-CLIP (Wang et al., 2026b), and IAPL (Li et al., 2026a) introduce increasingly flexible feature or prompt adaptation. Test-Time Curriculum (TTC) (Zhang et al., 2026) further adapts a detector on an unlabeled test batch through confidence-based curriculum learning. Together, these approaches have advanced cross-generator detection through improved forensic representations and adaptation mechanisms.

Despite this progress, a common paradigm in generalizable detection remains static: a trained detector assigns predictions to target images without using its detection outcomes to improve subsequent decisions. This paradigm relies on cues learned from a finite set of generators remaining informative under potentially different generative distributions, which is difficult to guarantee. A detector may recognize only a subset of images from an unseen source and misclassify the remainder as real. Under static inference, these partial successes are not used to improve detection of the remain-

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.

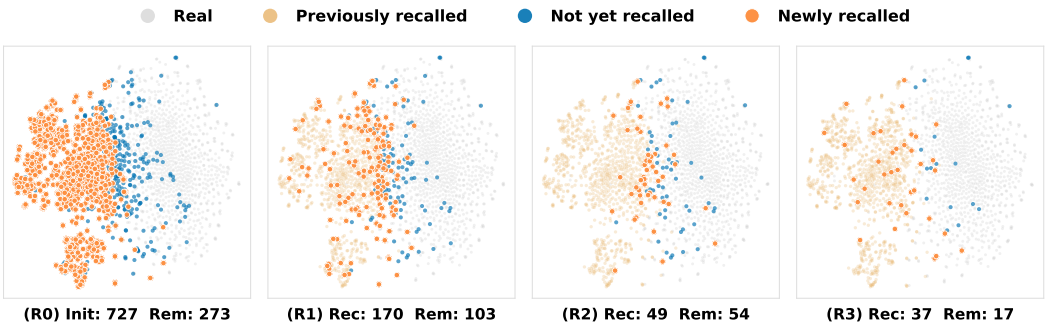


Figure 1: Feature-space evolution of iterative recall on an unseen target source. **Init** denotes the number of images detected at initial detection, **Rec** denotes the number of newly recalled images at each iteration, and **Rem** denotes the number of images that remain undetected. **Leftmost subplot**: initial generalization result. **Right three subplots**: iterative recall results across successive iterations. Undetected images progressively move into the recalled cluster while remaining separate from the real distribution.

ing synthetic images. Recent work has begun to relax this paradigm: IAPL adapts individual test images, while TTC (Zhang et al., 2026) performs curriculum-based adaptation over an unlabeled test batch. We instead focus on using images initially predicted as synthetic as learning signals for identifying additional images from the same source, turning partial detection into iterative recall.

To this end, we propose an iterative adaptation framework based on a central empirical finding: target images predicted as synthetic improve recognition of previously missed images from the same source. Partial cross-generator generalization can therefore seed source-specific learning, as even a small predicted-synthetic subset provides an anchor for adaptation. Updating the detector with these samples yields further detections, which are incorporated as supervision in the next round and progressively expand detection coverage. Figure 1 illustrates this process on a detector trained on the SDv1.4 split of GenImage and evaluated on Midjourney (Zhu et al., 2023), where previously undetected images progressively join the recalled cluster across iterations. This process motivates asymmetric supervision and data sources. Predicted-synthetic targets provide source-specific supervision for the unknown source, although this subset may also contain real images. Trusted real-image replay supplies independent real supervision and counterbalances their influence. Predicted-real targets are excluded because they may contain missed synthetic images whose use as real labels would reinforce existing errors. These asymmetries support iterative recall without assuming either predicted class is error-free.

Our framework implements this process by alternating detection, adaptation, and re-evaluation on an unlabeled target collection. At each iteration, predicted-synthetic images serve as pseudo-labeled samples for learning target-source patterns and recovering missed synthetic images. For difficult sources, however, this subset may cover only a limited portion of the source distribution, limiting subsequent recall expansion. We therefore introduce stochastic exploration, which provisionally labels a small subset of the remaining target images as synthetic to expose the detector to candidates beyond its current coverage. These candidates are learned jointly with predicted-synthetic samples and trusted real-image replay, constraining noise and boundary drift. By broadening the source-specific adaptation signal, exploitation and exploration accelerate recall expansion and progressively recover missed images without target annotations or generator identities. Our main contributions are summarized as follows:

- We show that target images predicted as synthetic during initial cross-generator detection provide learning signals for recognizing previously undetected images from the same source, establishing partial generalization as a starting point for iterative recall improvement.
- We propose an iterative adaptation framework that combines synthetic-prediction exploitation with stochastic exploration to accelerate iterative recall and improve the final recall of difficult sources. Asymmetric supervision from target synthetic predictions and trusted real-image replay supports adaptation while constraining noisy updates and false positives.

- We establish new state-of-the-art results on comprehensive benchmarks, outperforming recent methods by a substantial margin. Extensive experiments across both open-source and closed-source generators, including those with substantially different synthesis mechanisms, demonstrate that our method adapts effectively to evolving generative models and diverse generation paradigms.

2 Related Work

2.1 Evolution of Image Generation

Image generation spans several modeling families. VAEs (Kingma & Welling, 2013) learn probabilistic latent representations, while GANs (Goodfellow et al., 2020), including StyleGAN2 (Karras et al., 2020), synthesize images through adversarial learning. Diffusion models have progressed from iterative denoising with DDPM (Ho et al., 2020) and efficient latent-space generation (Rombach et al., 2022) to scalable transformer backbones such as DiT (Peebles & Xie, 2023) and rectified-flow transformers (Esser et al., 2024). Recent systems such as FLUX.1 Kontext (Labs et al., 2025) further incorporate image context for unified generation and editing. In parallel, VQGAN (Esser et al., 2021) combines discrete representations with autoregressive modeling, while VAR (Tian et al., 2024) and Infinity (Han et al., 2025) develop coarse-to-fine next-scale generation and self-correction. These increasingly diverse architectures, objectives, and generation procedures leave source-dependent forensic traces, making static detectors fragile as generative models evolve.

2.2 Generalizable AI-Generated Image Detection

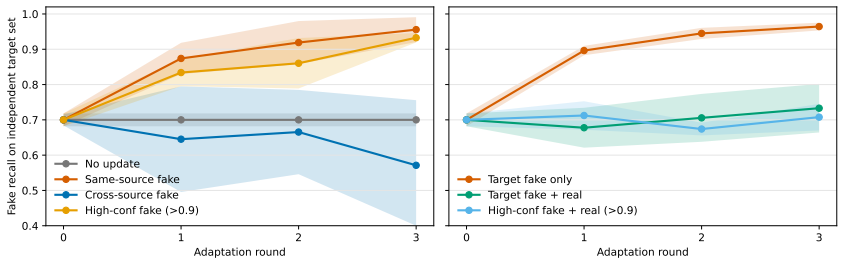
Generalizable detectors seek evidence that transfers across generative sources. CNNDetection (Wang et al., 2020) uses supervised training with augmentation. NPR (Tan et al., 2024b) captures upsampling dependencies. DIRE (Wang et al., 2023) and FakeInversion (Cazenavette et al., 2024) use diffusion reconstruction or inversion. UniversalFakeDetect (Ojha et al., 2023), FatFormer (Liu et al., 2024), AIDE (Yan et al., 2025), PPM-CLIP (Wang et al., 2026b), and IAPL (Li et al., 2026a) exploit pretrained representations, forensic cues, or adaptive prompts, with IAPL adapting each image independently. TTC (Zhang et al., 2026) performs batch-level adaptation with balanced confidence pseudo-labels and curriculum learning. Its symmetric pseudo-labeling may reinforce missed synthetic images as real, while parameter anchoring provides no trusted real-data constraint for stable multi-round adaptation. Its fixed three-round schedule also cannot adjust to target difficulty. We instead exclude predicted-real targets from real supervision, replay trusted real images, and explore beyond current synthetic coverage for stable multi-round recovery.

3 Preliminary Study

Before introducing the full framework, we conduct two controlled studies to examine whether partial detection on an unseen generator can provide useful adaptation signals and how the two predicted classes should be used.

3.1 Partial Generalization Enables Same-Source Learning

Figure 2(a) compares adaptation signals drawn from the same target source and from other synthetic sources. Without updating, recall remains around 0.71. Adapting with the pseudo-labeled Midjourney samples collected by the initial generalization pass steadily raises recall to approximately 0.96 after three rounds, and restricting these seeds to high-confidence predictions produces a similar gain. In contrast, cross-source synthetic images provide no consistent improvement and exhibit substantially larger variation across trials. This behavior indicates that adaptation with a mismatched synthetic source cannot provide stable gains on an unseen target source. The stable same-source improvement therefore cannot be attributed merely to adding more synthetic training images. Instead, the initially predicted synthetic set contains useful source-specific information that turns partial cross-generator generalization into a practical starting point for same-source learning.



(a) Same-source recall benefit (b) Asymmetric target use

Figure 2: Preliminary studies on GenImage (Zhu et al., 2023). A DINOv3 ViT-L detector (Siméoni et al., 2025) trained on SDv1.4 is adapted to unlabeled Midjourney images. Same-source seeds are predictions collected after one inference pass without ground-truth filtering. Target labels are used only to evaluate recall. Curves and shading show the mean and standard deviation over three independently sampled runs. The three cross-source trials in (a) use SDv1.4, ADM, and VQDM, respectively. (a) Same-source predictions improve recall, whereas cross-source adaptation is unstable. (b) Target synthetic predictions alone perform best. Adding target real pseudo-labels impedes recovery, and restricting them to high-confidence predictions provides no consistent improvement.

3.2 Target Pseudo-Labels Have Asymmetric Effects

We next examine whether target predictions should be used symmetrically. As shown in Figure 2(b), target synthetic predictions alone support effective adaptation, whereas additionally assigning real pseudo-labels to target images predicted as real leaves recall near its initial level. Restricting the predicted-real subset to high-confidence predictions does not resolve this problem: it yields no consistent improvement over using the full subset and remains far below synthetic-only adaptation. This limitation is structural. The predicted-real subset is an unknown mixture of genuine real images and undetected synthetic images, and high confidence does not guarantee that these two groups are correctly separated. Consequently, even confidence-filtered real pseudo-labels can reinforce undetected synthetic images as real and make them harder to recover in subsequent rounds. These results motivate asymmetric supervision: target synthetic predictions provide source-specific synthetic supervision, whereas real supervision should be supplied by an independent trusted real-image replay set rather than target real predictions.

4 Method

In this section, we first formulate generalizable AI-generated image detection in an unlabeled, mixed-source target setting. We then present our asymmetry-guided iterative adaptation framework, including synthetic-prediction exploitation, stochastic exploration, trusted real-image replay, and the iterative update and termination procedure.

4.1 Problem Formulation

Let $\mathcal{D}_S = \{(x_i^s, y_i^s)\}$ denote a labeled source dataset, where $y = 1$ indicates a synthetic image and $y = 0$ a real image. A binary detector f_0, trained on $\mathcal{D}_S$, outputs a synthetic probability $p_0(x)$. At deployment, the detector receives an unlabeled target collection $\mathcal{D}_T = \{x_i^t\}_{i=1}^N$ containing real images and synthetic images produced by one or more generators that are unavailable during source training. Neither target labels nor generator identities are given. Generalizable AI-generated image detection aims to correctly classify $\mathcal{D}_T$ despite this source-to-target shift. Adapting in this setting therefore requires constructing supervision for both classes without labeled target data.

The two classes provide asymmetric supervision sources. Verified real images can be collected independently and broadly cover the real-image distribution. In contrast, target synthetic images originate from unknown generators for which no labeled samples are available in advance. Their supervision must be inferred from $\mathcal{D}_T$. We retain the verified real images as a trusted replay set $\mathcal{D}_R$ and use it as an independent real-class reference during target adaptation. This requires no target annotations or generator assumptions, and the framework remains detector-agnostic.

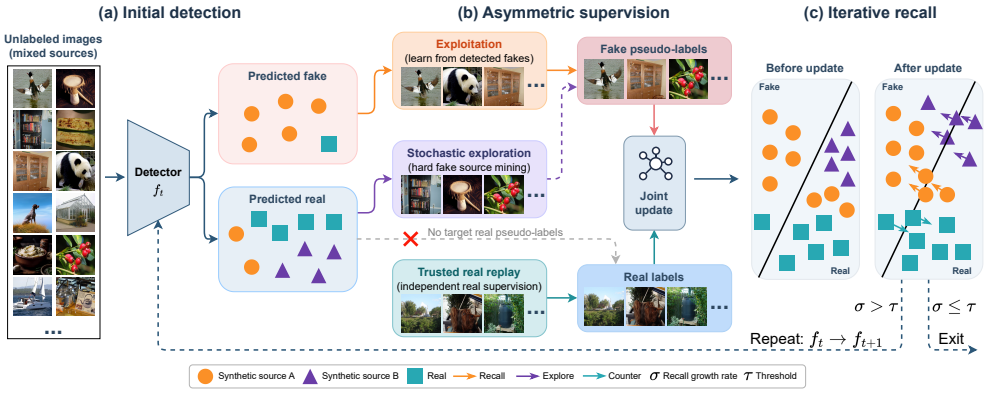


Figure 3: **Overview of the proposed asymmetry-guided iterative adaptation framework.** (a) The initial detector partitions the unlabeled target collection into predicted-synthetic and predicted-real subsets. (b) Exploitation uses predicted-synthetic images, stochastic exploration samples candidates from the remaining pool, and trusted real replay provides real supervision. Predicted-real targets receive no real pseudo-labels. (c) The updated detector recalls additional synthetic images while replay constrains drift. After at least $T_{\min}$ rounds, iteration continues when the growth rate σ of predicted-synthetic images exceeds τ and stops otherwise.

4.2 Asymmetry-Guided Iterative Adaptation

Figure 3 presents the overall procedure, which alternates target detection and model adaptation. Each round obtains synthetic supervision from predicted-synthetic targets through exploitation and from randomly sampled remaining candidates through stochastic exploration, while real supervision comes exclusively from trusted replay. This asymmetry is consistent with the preliminary study, as it avoids reinforcing the errors that iterative recall is intended to correct.

Initial detection and exploitation. Let $p_t(x)$ be the synthetic probability produced by detector f_t, and let η be its decision threshold. At initialization, applying f_0 to the target collection yields

$$\mathcal{F}_0 = \{x \in \mathcal{D}_T \mid p_0(x) \geq \eta\}. \tag{1}$$

Although $\mathcal{F}_0$ may cover only part of the target synthetic distribution, it provides target-specific cues absent from the source training data. At iteration t, applying the current detector f_t to obtain the pseudo-labeled set $\mathcal{F}_t$ which provides the synthetic supervision for the next update. Exploitation uses only predicted-synthetic targets. Although predicted-real targets could provide real supervision, they contain missed synthetic images whose assignment as real would reinforce existing errors and hinder later recall (Figure 2(b)). We instead obtain real supervision from the trusted replay set $\mathcal{D}_R$, which counterbalances false positives in $\mathcal{F}_t$, constrains boundary drift, and preserves real-class accuracy as synthetic coverage expands.

Stochastic exploration beyond current coverage. Exploitation is limited to patterns represented in $\mathcal{F}_t$, so difficult sources may remain undercovered and recall expansion may slow or converge prematurely. Let

$$\mathcal{M}_t = \mathcal{D}_T \setminus \mathcal{F}_t \tag{2}$$

denote the remaining pool, which contains both real and undetected synthetic images. We uniformly sample a small subset $\mathcal{E}_t \subset \mathcal{M}_t$ and provisionally assign synthetic labels. Resampling across iterations exposes patterns beyond the current detection coverage, helping recover difficult missed samples without generator identities or clustering assumptions.

We keep $\mathcal{E}_t$ small so exploration supplements exploitation rather than dominates the update. Since $\mathcal{M}_t$ contains genuine real images, random exploration inevitably assigns synthetic labels to some real samples. Excessive expansion of $\mathcal{E}_t$ may cause sampled real images to outnumber missed synthetics and dominate pseudo-fake supervision. Trusted real replay counteracts this noise and prevents boundary drift, but larger exploration requires more replay and raises update cost. Sampling a small subset balances broader coverage and faster recall against replay cost and noisy adaptation.

Asymmetric joint update. We combine the exploitation and the exploration set as $\mathcal{P}_t = \mathcal{F}_t \cup \mathcal{E}_t$, assigning every $x \in \mathcal{P}_t$ the synthetic label $y = 1$. The trusted real replay set $\mathcal{D}_R$ provides real samples with label $y = 0$. In each round, we draw a trusted real subset $\mathcal{R}_t \subset \mathcal{D}_R$ with the same number of samples as the pseudo-labeled synthetic set, i.e., $|\mathcal{R}_t| = |\mathcal{P}_t|$. The detector is initialized from its current parameters θ_t and updated by minimizing the cross-entropy loss over the union of both sets:

$$\mathcal{L}_t(\theta) = -\frac{1}{|\mathcal{P}_t| + |\mathcal{R}_t|} \left[\sum_{x \in \mathcal{P}_t} \log p_\theta(x) + \sum_{x \in \mathcal{R}_t} \log(1 - p_\theta(x)) \right], \tag{3}$$

where $p_\theta(x)$ denotes the synthetic probability produced by the detector with parameters θ. The resulting parameters define f_{t+1}. Replay counteracts drift caused both by false positives in $\mathcal{F}_t$ and by real images sampled during exploration in $\mathcal{E}_t$, preventing the noisy update from trivially moving all target images toward the synthetic class.

Recall update and termination. After adaptation, the updated detector f_{t+1} re-evaluates $\mathcal{D}_T$ and refreshes $\mathcal{F}_{t+1}$. The refreshed set incorporates newly predicted synthetic images and provides supervision for the next round, enabling the iterative expansion illustrated in Figure 3(c). Since true target recall is unavailable, we monitor the growth rate of the number of images predicted as synthetic,

$$\sigma_t = \frac{|\mathcal{F}_{t+1}| - |\mathcal{F}_t|}{\max(|\mathcal{F}_t|, 1)}. \tag{4}$$

Trusted replay constrains boundary drift, so growth in the predicted-synthetic set primarily comes from recalling missed synthetic images rather than accumulating false positives. Let τ be the growth threshold and $T_{\min}, T_{\max}$ the minimum and maximum rounds. We run at least $T_{\min}$ rounds to resample exploration and give difficult samples multiple opportunities. After this minimum budget, iteration continues only when $\sigma_t > \tau$ and $t + 1 < T_{\max}$. Otherwise, adaptation terminates. At the termination round T, f_T is used to classify the target collection, with $\mathcal{F}_T$ as the predicted synthetic subset and $\mathcal{D}_T \setminus \mathcal{F}_T$ as the predicted real subset.

5 Experiments

5.1 Experimental Setup

Evaluation benchmarks. Following the comprehensive evaluation setting of DDA (Chen et al., 2025), our primary comparison uses its eleven-benchmark protocol, comprising seven manually curated benchmarks and four in-the-wild benchmarks. These benchmarks cover GAN-, diffusion-, and autoregressive-generation paradigms, as well as web images with unknown generators and post-processing pipelines. We additionally evaluate AIGIBench (Li et al., 2026b) and Treasure (Wang et al., 2026a), yielding thirteen benchmarks in total and extending the evaluation to recent open- and closed-source generators, editing pipelines, and real-world image distributions. Details of all compared methods are provided in Appendix B.

Evaluation metrics. For the primary eleven benchmarks, we report balanced accuracy, i.e., the mean real- and synthetic-class accuracy. AIGIBench and Treasure retain their original mean-accuracy protocols. Adaptation is evaluated on a disjoint held-out subset, while Table 1 also reports adaptation-subset performance. Appendix A details detector/source training, trusted-real replay sources, and target adaptation/evaluation splits.

5.2 Main Results

Overall and class-wise results. Tables 1 and 2 show that our method improves all eleven benchmarks, raising average balanced accuracy from 87.2% to 98.5% and outperforming DDA by 7.8%. The class-wise results clarify that these gains arise primarily from recalling missed synthetic images: average AI accuracy increases from 79.7% to 98.4%, whereas Non-AI accuracy changes from 94.7% to 98.5%. Trusted real replay constrains false positives and decision-boundary drift as synthetic coverage expands, while also providing additional real supervision that generally preserves real-image accuracy. Detailed results for each benchmark are provided in Appendix F.

Table 1: Balanced accuracy (%) on eleven benchmarks. Results for the comparison methods are directly taken from DDA (Chen et al., 2025). Ours (adapt) and Ours (held-out) report final-round performance on the 80% adaptation subset and the disjoint 20% held-out subset, respectively. The baseline is the initial detector at round zero evaluated on held-out data. **Avg** and **Min** denote the average and minimum performance across benchmarks. The best result in each column is bolded.

Method	Manually Curated Benchmarks							In-the-Wild Benchmarks				Avg	Min
	GenImage	DRCT-2M	DDA-COCO	EvalGEN	Synthbuster	ForenSynths	AIGCDet.	Chameleon	SynthWildX	WildRF	BFree		
NPR (CVPR 2024b)	51.5	37.3	42.2	2.9	50.0	47.9	53.1	59.9	49.8	63.5	49.5	46.1	2.9
UnivFD (CVPR 2023)	64.1	61.8	52.4	15.4	67.8	77.7	72.5	50.7	52.3	55.3	49.0	56.3	15.4
FatFormer (CVPR 2024)	62.8	52.2	51.7	45.6	56.1	90.0	85.0	51.2	52.1	58.9	50.0	59.6	45.6
SAFE (KDD 2025)	50.3	59.3	49.9	1.1	46.5	49.7	50.3	59.2	49.1	57.2	50.5	47.6	1.1
C2P-CLIP (AAAI 2025)	74.4	59.2	51.3	38.9	68.5	92.0	81.4	51.1	57.1	59.6	50.0	62.1	38.9
AIDE (ICLR 2025)	61.2	64.6	50.0	19.1	53.9	59.4	63.6	63.1	48.8	58.4	53.1	54.1	19.1
DRCT (ICML 2024)	84.7	90.5	60.2	77.8	84.8	73.9	81.4	56.6	55.1	50.6	55.7	70.1	50.6
AlignedForensics (ICLR 2025)	79.0	95.5	86.5	68.0	77.4	53.9	66.6	71.0	78.8	80.1	68.5	75.0	53.9
DDA (NeurIPS 2025)	91.7	98.1	92.2	97.2	90.1	81.4	87.8	82.4	90.9	90.3	95.1	90.7	81.4
Baseline ($t = 0$)	93.3	93.2	83.3	80.8	93.8	85.2	91.3	79.3	87.6	92.8	78.6	87.2	78.6
Ours (adapt)	**99.2**	**99.7**	**94.7**	**99.9**	**99.9**	**99.3**	**99.4**	**98.5**	**97.5**	**99.5**	**100.0**	**98.9**	**94.7**
Ours (held-out)	99.1	**99.7**	94.6	**99.9**	**99.9**	99.1	99.2	97.8	94.6	99.4	**100.0**	98.5	94.6

Table 2: AI and Non-AI class accuracies (%) on the eleven benchmarks in Table 1.

Method	GenImage		DRCT-2M		DDA-COCO		EvalGEN		Synthbuster		ForenSynths		AIGCDet.		Chameleon		SynthWildX		WildRF		BFree		**Avg**	
	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI
Baseline ($t = 0$)	88.2	98.4	87.9	98.5	68.2	**98.4**	62.7	98.9	89.3	98.2	79.3	91.1	86.0	96.5	70.0	88.5	82.4	92.8	90.7	94.8	71.7	85.5	79.7	94.7
Ours (adapt)	**98.5**	**99.9**	**99.9**	**99.5**	92.6	96.8	**99.8**	**100.0**	**99.9**	99.9	**99.2**	**99.5**	**99.6**	**99.2**	**97.1**	99.9	**99.5**	**95.5**	**99.2**	99.9	**100.0**	**100.0**	**98.7**	**99.1**
Ours (held-out)	98.4	99.8	**99.9**	**99.5**	**93.4**	95.8	**99.8**	**100.0**	**99.9**	**100.0**	98.9	99.3	99.4	99.1	95.5	**100.0**	98.9	90.4	98.7	**100.0**	**100.0**	**100.0**	98.4	98.5

Table 3: Mean accuracy (%) on AIGIBench and Treasure.

Method	AIGIBench	Treasure
FreqNet (AAAI 2024a)	66.2	73.4
CLIPDetection (CVPR 2023)	72.5	74.2
AIDE (ICLR 2025)	77.6	84.8
SAFE (KDD 2025)	79.9	86.7
PLM (AAAI 2026)	76.6	88.1
Ours	**99.4**	**98.9**

Table 4: Component ablation on three representative benchmarks.

Setting	AIGIBench	Chameleon	GenImage	Average
Initial detector	81.4	86.3	87.5	85.1
w/o exploitation	96.5	92.3	97.7	95.5
w/o exploration	99.1	94.0	98.0	97.0
replay $\rightarrow$ target-real	80.4	85.8	84.1	83.4
Ours	**99.4**	**98.0**	**99.3**	**98.9**

Comparison on AIGIBench and Treasure. Table 3 compares our method with five detectors reported by AIGIBench (Li et al., 2026b) and Fleet (Wang et al., 2026a). Without target labels, our method reaches 99.4% on AIGIBench and 98.9% on Treasure, exceeding the strongest comparison results by 19.5% and 10.8%, respectively. These results demonstrate strong generalization to recent open- and closed-source generators. Detailed source-wise results are provided in Appendix F.

5.3 Robustness to Common Post-processing

Following DDA (Chen et al., 2025), we evaluate robustness on GenImage under JPEG compression, image resizing, and Gaussian blurring. As shown in Figure 4, our method maintains approximately 99% balanced accuracy across all tested settings, while the comparison methods degrade under stronger perturbations. Images under the same degradation retain a shared within-class distribution, allowing initially detected degraded synthetic images to provide degradation-specific adaptation signals. The detector can therefore adapt to the degraded target distribution and iteratively recall the remaining synthetic images, following the same paradigm as in the original setting.

5.4 Ablation Studies

Component ablation. Table 4 confirms the importance of exploitation, exploration, and trusted real replay. Exploitation maximally leverages existing synthetic priors to establish within-class recall patterns, whereas exploration mines hard samples beyond current coverage to increase the attainable recall on difficult target distributions. In the replay ablation, all test images not used as fake supervision are labeled as real and replace trusted replay. Removing real supervision would degenerate training toward synthetic. Its drop from 98.9% to 83.4% shows that missed synthetic images are reinforced as real, consistent with Table 5. The absence of exploitation and exploration removes target-specific synthetic supervision and reduces the procedure to real-data diversification examined separately in Appendix E. Appendix C further presents an ablation across detector architectures, demonstrating the broad architectural applicability of our method.

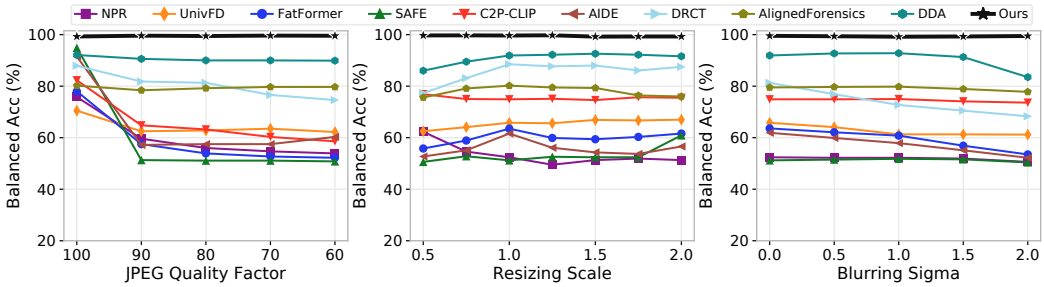


Figure 4: Robustness analysis on GenImage under JPEG compression, image resizing, and Gaussian blurring. Performance is measured by balanced accuracy (%). Our method remains consistently robust across all settings.

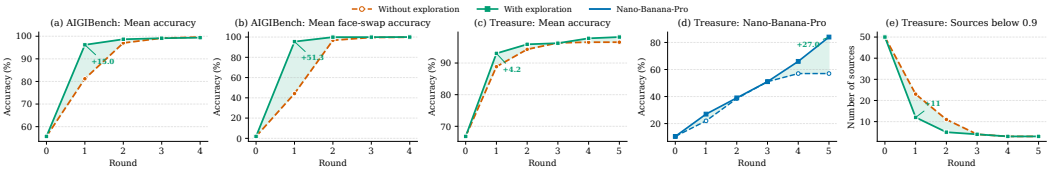


Figure 5: Effect of stochastic exploration across adaptation rounds. We report fake-class accuracy to illustrate the progression of synthetic-image recall. (a) Mean accuracy on AIGIBench, (b) mean accuracy over its five face-swap categories, which pose substantial generalization challenges, (c) mean accuracy on Treasure, (d) the test-time accuracy trajectory of Nano-Banana-Pro, a particularly difficult Treasure category, and (e) the number of Treasure categories with accuracy below 0.9.

Convergence and stopping bounds. Figure 6 shows rapid convergence across benchmarks. The first round provides the largest gain, while later rounds mainly recover the remaining difficult cases. By $T_{\min} = 3$, most benchmarks are within 1% of their final accuracy, and all satisfy the stopping criterion by round seven. The diminishing round-wise gains also support using predicted-synthetic growth as a practical convergence signal when target labels are unavailable. This adaptive rule allows easier benchmarks to terminate once expansion saturates, while retaining additional rounds for sources whose missed images are recovered more gradually. Thus, $T_{\min}$ captures the dominant recall expansion without premature termination, while $T_{\max} = 10$ provides sufficient margin for slower targets and bounds adaptation cost.

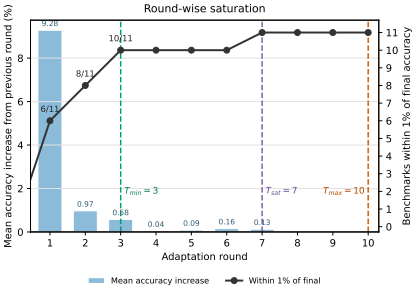


Figure 6: Round-wise convergence across eleven benchmarks.

Effect of stochastic exploration. We ablate exploration on AIGIBench (Li et al., 2026b) and Treasure (Wang et al., 2026a) while keeping exploitation and real replay unchanged. Figure 5 shows that exploration provides its largest benefit during early adaptation, when the predicted-synthetic subset covers only part of the target distribution. Sampling candidates beyond this subset exposes the detector to difficult regions that exploitation alone may reach slowly or fail to recover, thereby reducing the number of poorly detected sources. Although exploitation can narrow the gap on easier sources over successive rounds, exploration still improves the final recovery of the most difficult cases. These results confirm that exploration complements exploitation by accelerating convergence and improving final recall on difficult sources.

Hard target source count and proportion. Using Midjourney as a hard target, we study how count and proportion affect recall. Figure 7 shows distinct effects. When the Midjourney count is small, limited same-source evidence cannot establish a stable within-source pattern, so noise

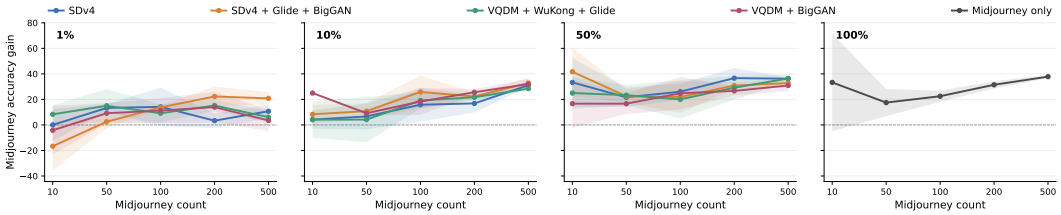


Figure 7: Midjourney accuracy gain after adaptation across proportions (panels) and counts (x-axis). Tests use 50,000 real images, counts of $10, 50, 100, 200$, and 500, and proportions of $1\%, 10\%, 50\%$, and 100%. Colors denote remaining-fake source combinations with random source counts. The 100% panel contains only Midjourney. Shading denotes variance over three random-seed runs.

Table 5: Symmetric and asymmetric target supervision on three benchmarks. F, R, and B denote fake, real, and balanced accuracy (%). Symmetric variants use predicted-real targets as real supervision, with confidence filtering and trusted replay added as indicated. High confidence means prediction confidence > 0.9. Our asymmetric method excludes predicted-real targets.

Setting	AIGIBench			Chameleon			GenImage			Avg
	F	R	B	F	R	B	F	R	B	
Initial detector ($t = 0$)	68.6	94.1	81.4	75.0	97.5	86.3	75.4	99.7	87.5	85.1
All predicted-real	74.8	89.2	82.0	83.0	93.0	88.0	79.8	99.5	89.6	86.5
High-confidence symmetric	75.9	92.9	84.4	86.5	92.0	89.3	75.4	99.7	87.5	87.1
High-confidence symmetric + replay	79.2	98.9	89.1	87.4	99.5	93.5	79.6	**99.8**	89.7	90.7
Ours (asymmetric)	**99.7**	**99.1**	**99.4**	**96.0**	**100.0**	**98.0**	**98.8**	**99.8**	**99.3**	**98.9**

prevents consistent gains. At low proportions, prevalent fake sources obscure this pattern, while limited global-growth contribution may trigger premature stopping. Gains become larger and more stable as both factors increase. Above 10%, they approach the single-source setting, indicating that competing sources no longer substantially dilute Midjourney cues. Above 100 Midjourney samples, adaptation yields gains even at 1%, showing that sufficient absolute evidence supports rare-source adaptation. Appendix D extends the analysis across sources and compositions.

Symmetric target supervision. Table 5 extends the controlled single-source experiment in Figure 2(b) to three full benchmarks. All settings other than the real-supervision strategy are held fixed across variants. Across benchmarks, symmetric supervision provides only limited improvement because missed fake images in the predicted-real set receive real labels, reinforcing the original false negatives and obstructing subsequent recall. Confidence filtering does not reliably remove these missed fakes. Adding trusted real replay preserves real accuracy and counteracts false-positive drift from noisy pseudo-fake supervision, but it cannot undo the contradictory real labels assigned to missed fakes, so fake recall remains limited. In contrast, our asymmetric supervision excludes predicted-real targets and consistently recovers the missed fake distribution. These results show that asymmetric supervision is decisive for iterative fake recall, while trusted replay plays a complementary role in controlling false positives and preserving the real boundary.

6 Conclusion

We presented an asymmetry-guided iterative adaptation framework for generalizable AI-generated image detection. Target images predicted as synthetic contain source-specific signals for recovering missed images from the same source, converting partial cross-generator generalization into improved same-source detection. The framework exploits these predictions, uses trusted real-image replay to stabilize adaptation, and incorporates stochastic exploration to accelerate recall and improve difficult-source coverage without target annotations or generator identities. Across thirteen benchmarks, the method substantially outperforms the initial detector and existing methods on held-out data, recent generators, and common degradations. These results show that reusing detection outcomes as adaptation signals provides an effective direction for robust detection under continuously evolving generative distributions.

AI USE STATEMENT

Generative AI tools were used solely to aid and polish the writing of this manuscript. Specifically, they assisted with English translation and language editing to improve clarity and readability. The authors reviewed and approved all AI-assisted revisions and take full responsibility for the final text.

ETHICS STATEMENT

This work studies the detection of AI-generated images and involves no human-subject experiments. Experiments use established research benchmarks and publicly available images. The technology is dual-use: it may support content provenance and moderation, but false positives may flag authentic images and false negatives may create unwarranted confidence in synthetic content. Performance may vary across generators, domains, post-processing, and underrepresented groups. Detector outputs should therefore be treated as probabilistic evidence, not the sole basis for high-stakes decisions. Deployment should include domain-specific validation, privacy and data-governance safeguards, human oversight, and compliance with licenses, platform policies, and applicable laws.

REPRODUCIBILITY STATEMENT

We provide the problem formulation and complete adaptation procedure, including exploitation, exploration, trusted real replay, the optimization objective, and the stopping rule, in Section 4. Section 5 describes the evaluation benchmarks and metrics, while detailed replay-data sources, detector architecture, source-training settings, target-data split, and adaptation hyperparameters are provided in Appendix A. The main experiments report comparisons, component ablations, the effect of stochastic exploration, round-wise convergence, and robustness tests. The appendix additionally provides descriptions of every comparison method (Appendix B), architecture dependence (Appendix C), the applicability boundary of iterative recall under different target-set sizes and source compositions (Appendix D), the effect of replay real supervision (Appendix E), and detailed source-wise results on both recent benchmarks and the eleven primary benchmarks (Appendix F). All external datasets, pretrained backbones, and comparison methods are cited, and held-out evaluation is used throughout to separate adaptation from assessment. Together, these materials specify the experimental protocol and provide detailed results for independently checking the reported findings.

REFERENCES

George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion. In *CVPR*, pp. 10759–10769. IEEE, 2024.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pp. 3557–3567. 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.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, et al. Dual data alignment makes ai-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*, 2025.

Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. Raise: A raw images dataset for digital image forensics. In *MMSys*, pp. 219–224, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. Ieee, 2009.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pp. 12873–12883, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity∞: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *CVPR*, pp. 15733–15744. IEEE, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851, 2020.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pp. 4396–4405. IEEE, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pp. 8107–8116. IEEE, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv: 1312.6114*, 2013.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 128(7):1956–1981, 2020.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *KDD*, pp. 2405–2414, 2025.

Yiheng Li, Zichang Tan, Guoqing Xu, Zhen Lei, Xu Zhou, and Yang Yang. Towards generalizable ai-generated image detection via image-adaptive prompt learning. In *CVPR*, pp. 21262–21272, 2026a.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? *NeurIPS*, 38, 2026b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755. Springer, 2014.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *CVPR*, pp. 10770–10780. IEEE, 2024.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, pp. 24480–24489. IEEE, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4172–4182. IEEE, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10674–10685. ieee, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. In *ICLR*, volume 2025, pp. 63304–63322, 2025.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *AAAI*, volume 38, pp. 5052–5060, 2024a.

Chuangchuang Tan, 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 *CVPR*, pp. 28130–28139, 2024b.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *AAAI*, volume 39, pp. 7184–7192, 2025.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *NeurIPS*, 37:84839–84865, 2024.

Jiaan Wang, Sirui Liu, Yu Li, Kaiyuan Yang, Juan Cao, and Sheng Tang. Fleet: Few shots lead effective ai-generated image detection. *arXiv preprint arXiv:2606.31082*, 2026a.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, pp. 8692–8701. IEEE, 2020.

Xinyuan Wang, Yingxin Lai, Zhiming Luo, and Zhihui Liu. Ppm-clip: Probabilistic prompt modeling for generalizable ai-generated image detection. In *CVPR*, pp. 21316–21325, 2026b.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *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 *ICLR*, volume 2025, pp. 70702–70720, 2025.

Yiqian Zhang, Zheyuan Gu, Xiangzhao Hao, Zefeng Zhang, Jingjia Mao, Jiahao Hu, Jiaxu Miao, Jun Yu, Zhenyu Zhang, Shuohuan Wang, et al. Test-time curriculum for open-set aigc detection. *arXiv preprint arXiv:2608.00559*, 2026.

Chenming Zhou, Jiaan Wang, Yu Li, Lei Li, Juan Cao, and Sheng Tang. Beyond semantic features: Pixel-level mapping for generalized ai-generated image detection. In *AAAI*, volume 40, pp. 36101–36109, 2026.

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. *NeurIPS*, 36:77771–77782, 2023.

APPENDIX

A IMPLEMENTATION DETAILS

Detector and source training. We use a DINOv3 ViT-L image encoder (Siméoni et al., 2025) with a binary classification head and consider two source-training settings. For the eleven-benchmark comparison in Table 1, the initial detector is trained using the source data adopted by DDA (Chen et al., 2025), enabling a direct comparison under its evaluation protocol. These data pair MS COCO photographs with aligned synthetic counterparts reconstructed by the Stable Diffusion 2.1 VAE. Before reconstruction, each image is center-cropped to the largest spatial size divisible by eight, and JPEG compression is applied to the reconstructed counterpart with probability 0.5 so that the source set contains both JPEG and PNG synthetic images. For experiments outside this eleven-benchmark protocol, including the AIGIBench and Treasure comparison, component analyses, and auxiliary ablations, the initial detector is trained on the SDv1.4 split of GenImage (Zhu et al., 2023), using balanced real and synthetic samples. In both settings, inputs are resized to 224×224 for evaluation, while training uses a 224×224 random resized crop with scale $[0.8, 1.0]$ and random horizontal flipping.

Trusted real replay sources. We construct the trusted real replay set $\mathcal{D}_R$ from ImageNet (Deng et al., 2009), MS COCO (Lin et al., 2014), LAION-5B (Schuhmann et al., 2022), RAISE (Dang-Nguyen et al., 2015), Open Images (Kuznetsova et al., 2020), Conceptual 12M (Changpinyo et al., 2021), FFHQ (Karras et al., 2019), CelebA-HQ (Karras et al., 2017), and additional photographs collected from public image-hosting websites. The collection is curated to cover diverse semantic content, including natural scenes, everyday objects, indoor and outdoor environments, and human and facial imagery. We additionally preserve diversity in image resolution, compression, acquisition source, and hosting platform. The replay data are deduplicated against the evaluation benchmarks. The resulting replay pool is fixed across all target benchmarks and contains no target annotations.

Target adaptation and evaluation split. After source training, the detector is adapted to each target benchmark independently. For every benchmark, we randomly divide the target collection into an 80% unlabeled adaptation subset and a disjoint 20% held-out evaluation subset. Only the adaptation subset is used to construct the exploitation and exploration sets and update the detector. The held-out images never participate in optimization. Each adaptation round performs one training epoch with AdamW, a batch size of 32, a learning rate of 2×10^{-4}, weight decay of 10^{-4}, and mixed-precision training. Synthetic probabilities are thresholded at 0.5. Exploration uniformly samples from the current remaining pool with a budget matched to the predicted-synthetic set, $|\mathcal{E}_t| = \min(|\mathcal{F}_t|, |\mathcal{M}_t|)$. This prevents real images sampled as exploratory pseudo-fakes from dominating the synthetic supervision. Each update uses an equal number of randomly sampled trusted real replay images, $|\mathcal{R}_t| = |\mathcal{P}_t|$. We use $T_{\min} = 3$, $T_{\max} = 10$, and a predicted-synthetic growth threshold of $\tau = 1\%$ for every benchmark. The same adaptation configuration and trusted real replay pool are used across all benchmarks. Following DDA (Chen et al., 2025), detailed-table Avg values are sample-weighted benchmark aggregates, not unweighted means of source results.

B COMPARED METHODS

We compare with all detectors reported in Tables 1 and 3. For the eleven primary benchmarks, comparison results are reproduced from DDA (Chen et al., 2025). Results on AIGIBench and Treasure follow their corresponding benchmark reports and released checkpoints (Li et al., 2026b; Wang et al., 2026a). The methods cover spatial- and frequency-domain forensic cues, pretrained vision-language representations, reconstruction-based data alignment, and pixel-level mappings.

NPR. NPR (Tan et al., 2024b) models neighboring-pixel relationships introduced by generator upsampling operations. It trains a ResNet-50 detector on low-level local residual patterns rather than relying primarily on image semantics.

FreqNet. FreqNet (Tan et al., 2024a) learns frequency-aware representations designed to expose spectral differences between real and generated images and improve transfer across unseen generators.

UniversalFakeDetect. UniversalFakeDetect (Ojha et al., 2023) uses pretrained CLIP visual features with a lightweight classifier and is denoted CLIPDetection in the AIGIBench and Treasure comparisons. Its large-scale pretrained representation reduces dependence on artifacts from a single training generator.

FatFormer. FatFormer (Liu et al., 2024) augments a vision-transformer representation with forgery-aware adapters operating in the image and frequency domains and uses language-guided alignment to encourage transferable forensic features.

SAFE. SAFE (Li et al., 2025) studies synthetic-image detection through image transformations and frequency-sensitive cues. Its training strategy uses transformation-based augmentation to reduce overfitting to generator-specific traces.

C2P-CLIP. C2P-CLIP (Tan et al., 2025) introduces category-common prompts into CLIP-based detection. Caption-derived textual information and contrastive training are used to align image features with shared real and synthetic concepts, while inference uses the adapted image encoder and classifier.

AIDE. AIDE (Yan et al., 2025) combines low-level and semantic evidence. It selects frequency-extreme patches using DCT scores, extracts noise-sensitive features with SRM filters, and fuses them with CLIP semantic embeddings for classification.

DRCT. DRCT (Chen et al., 2024) reconstructs real images with diffusion models to create content-aligned synthetic training samples. Contrastive training encourages the detector to focus on reconstruction fingerprints that generalize across diffusion generators.

AlignedForensics. AlignedForensics (Sundara Rajan et al., 2025) constructs aligned real–synthetic pairs by passing real images through the autoencoder of a latent diffusion model. The alignment suppresses semantic shortcuts and emphasizes artifacts introduced by the VAE decoder.

DDA. DDA (Chen et al., 2025) extends reconstruction-based alignment to both pixel and frequency domains. It trains on MS COCO images and SD2.1-VAE reconstructions with dual-domain alignment, reducing distribution bias between real and synthetic training data.

PLM. PLM (Zhou et al., 2026) learns a pixel-level mapping that exposes low-level discrepancies between real and generated images beyond semantic pretrained features.

C Detector Architecture Ablation

We evaluate whether the proposed iterative adaptation depends on a particular detector architecture. The comparison covers convolutional networks, CLIP, DINOv2, and DINOv3 at different model scales. ResNet detectors are fully fine-tuned, while the transformer-based detectors are adapted using low-rank adaptation (LoRA). For each architecture, we independently source-train a checkpoint on the same SDv1.4 data and report held-out balanced accuracy before adaptation and at the final round selected by the stopping rule.

Table 6 shows that the iterative framework transfers across substantially different detector families. Despite weak initial generalization, the three ResNet variants improve their average balanced accuracy by 27.2–29.0% and reach 90.5–92.9%, demonstrating that the recall mechanism is not restricted to pretrained vision transformers. The three CLIP backbones achieve comparable final performance, while the DINOv2 variants reach 94.4–97.7% after adaptation. The framework also consistently improves all DINOv3 variants: DINOv3-S and DINOv3-B gain 32.1% and 26.7% on average, reaching 95.4% and 97.3%, respectively. Larger DINOv3 models provide the strongest final results, with DINOv3-L and DINOv3-H achieving 98.9% and 99.3%. The remaining differences in final performance may reflect not only model capacity but also the backbone architecture and pretraining regime. Compared with conventionally supervised CNNs, ViT backbones pretrained on larger and more diverse datasets with vision-language or self-supervised objectives can provide

Table 6: Architecture ablation on three representative benchmarks using the same SDv1.4 source-training data and disjoint held-out evaluation splits. Each entry reports balanced accuracy (%) as the initial detector at $t = 0 \rightarrow$ the detector at the final round selected by the stopping rule. Average is the unweighted mean across the three benchmarks. FT denotes full fine-tuning. Transformer backbones use LoRA.

Architecture	Adaptation	AIGIBench	Chameleon	GenImage	Average
ResNet-50	FT	58.5 → 93.6	63.8 → 91.5	71.3 → 93.5	64.5 → 92.9
ResNet-101	FT	57.8 → 94.5	65.5 → 88.5	71.7 → 93.5	65.0 → 92.2
ResNet-152	FT	57.5 → 89.4	58.0 → 89.8	69.0 → 92.3	61.5 → 90.5
CLIP ViT-B/16	LoRA	64.8 → 98.9	53.8 → 96.5	69.8 → 98.6	62.8 → 98.0
CLIP ViT-B/32	LoRA	66.8 → 98.2	69.5 → 95.5	71.8 → 97.3	69.3 → 97.0
CLIP ViT-L/14	LoRA	71.4 → 98.3	61.8 → 97.0	68.3 → 99.4	67.1 → 98.2
DINOv2-S	LoRA	56.6 → 97.4	55.0 → 89.8	67.4 → 96.0	59.7 → 94.4
DINOv2-B	LoRA	61.9 → 98.9	61.8 → 92.3	71.7 → 97.2	65.1 → 96.1
DINOv2-L	LoRA	67.0 → 99.3	58.8 → 95.3	76.9 → 98.6	67.6 → 97.7
DINOv3-S	LoRA	57.1 → 97.1	61.5 → 93.3	71.3 → 95.8	63.3 → 95.4
DINOv3-B	LoRA	67.9 → 99.1	70.0 → 95.0	73.7 → 97.7	70.5 → 97.3
DINOv3-L	LoRA	81.4 → 99.4	86.3 → 98.0	87.5 → 99.3	85.1 → 98.9
DINOv3-H	LoRA	83.1 → **99.7**	89.8 → **98.8**	97.9 → **99.6**	90.2 → **99.3**

more transferable representations of subtle synthetic-image patterns. Within the same DINO family, the stronger results of larger variants further suggest that feature-extraction capacity raises the attainable adaptation ceiling. Because architecture, pretraining objective, data scale, and model size are not independently controlled in this comparison, these results should be interpreted as evidence of their combined influence rather than as an isolated causal attribution. Overall, the consistent gains across CNN, CLIP, DINOv2, and DINOv3 backbones support the architecture-agnostic nature of the proposed framework, while the absolute final performance remains higher for models with stronger pretrained representations. We use DINOv3-L in the main experiments because it approaches the performance of DINOv3-H with a smaller backbone.

D Applicability Boundary of Iterative Recall

Experimental setting. In every setting, the real portion of the unlabeled target collection is fixed at 10,000 images. We vary the synthetic portion by per-source image count and source combination, using the same candidate counts for every included source. The initial detector, real-image pool, adaptation hyperparameters, and stopping rule remain unchanged, isolating the availability and diversity of target synthetic evidence.

Figures 8 and 9 analyze complementary aspects of the applicability boundary. The first measures how the amount of synthetic evidence affects aggregate performance, while the second examines how accompanying sources affect individual-source recall. Comparisons with the round-zero detector also reveal when adaptation is preferable to retaining the initial one-pass prediction.

Iterative recall fundamentally learns a source-specific class pattern from target images currently predicted as synthetic and then uses this pattern to recover missed images from the same source. Its effectiveness therefore depends on the amount of usable within-class information contained in the predicted-synthetic set. When a source contributes only a few fake images, the initially detected subset contains too few consistent examples to reveal a reliable shared pattern. In this regime, adaptation may fail to form an effective recall pattern and can move the decision boundary away from the round-zero solution. Figure 8 exhibits this behavior for Midjourney and ADM with only 10 images, where final recall can fall below round zero.

As the number of fake images increases, more consistent within-source evidence becomes available, the effective signal-to-noise ratio of adaptation improves, and recall rises sharply, reaching high lev-

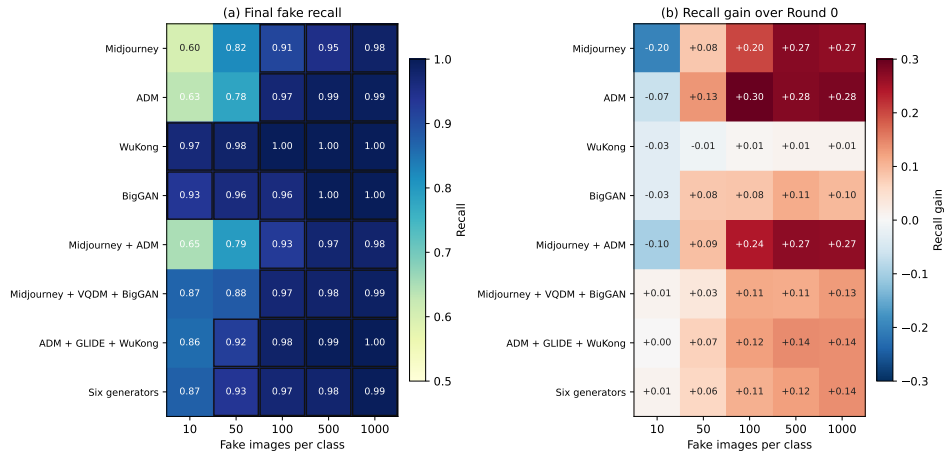


Figure 8: Effect of target synthetic-image count on iterative recall under different source-composition settings, with the real portion fixed at 10,000 images. Rows denote target-source compositions and columns denote the number of synthetic images per source. Cells report (a) aggregate final fake recall and (b) recall change relative to the round-zero detector. Black borders in (a) indicate final recall of at least 0.9, while negative values in (b) identify settings in which iteration performs worse than one-pass generalization. Across compositions, target collections containing more synthetic images exhibit more stable gains, with performance gradually approaching saturation.

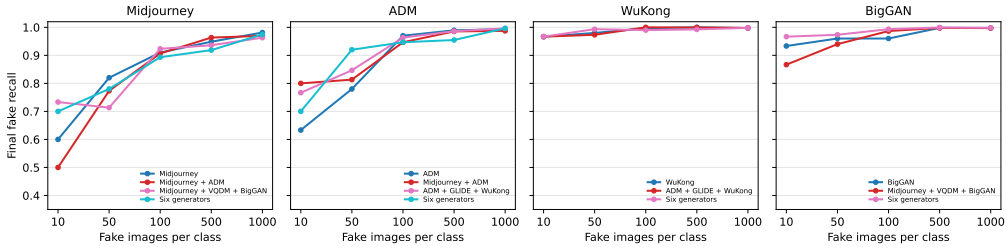


Figure 9: Effect of target source composition on the recall of individual synthetic sources. Each panel reports the final fake recall for one of four evaluated sources (Midjourney, ADM, Wukong, and BigGAN) as the number of images per source increases. Each curve denotes a target composition containing the source named above the panel. Source-wise recall varies only modestly across compositions at the same sample count, whereas it changes consistently with the number of images per source, showing that target fake count is the dominant factor.

els by 100 images per source. Across source compositions, performance then approaches saturation in the moderate- and high-count regimes. In contrast, Wukong and BigGAN remain strong even at the smallest count because their source patterns are already well covered by the initial detector and require little additional target evidence.

Figure 9 further shows that the recovery of an individual source is comparatively insensitive to which other synthetic sources are present. At a fixed sample count, the curves for different compositions remain close, with only modest fluctuations for the more difficult Midjourney and ADM sources in the lowest-count regime. Wukong and BigGAN exhibit even smaller composition-dependent differences because their round-zero coverage is already high. In contrast, all four sources improve consistently as their own sample count increases, and the curves approach the same high-recall region once 100 or more images per source are available. These results indicate that source composition has a relatively small effect, while the amount of target fake evidence largely governs the reliability and final recall of iterative adaptation.

Together, the two analyses show that fake count imposes a fundamental boundary on iterative recall. The method requires enough same-source evidence for the detector to distinguish a repeatable

Table 7: AI and Non-AI accuracy (%) of static detectors under original and replay real supervision. Within each detector pair, the synthetic training images and optimization protocol are fixed, and only the real-image distribution is changed. **Avg** is the unweighted class-wise mean over the three benchmarks.

Detector	Real Supervision	GenImage		AIGIBench		Chameleon		**Avg**	
		AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI
NPR	SDv1.4 real	38.8	90.1	25.7	83.4	9.9	85.7	24.8	86.4
NPR	Replay real	30.2	90.8	19.6	78.2	11.1	79.4	20.3	82.8
UnivFD	SDv1.4 real	52.2	88.3	59.3	73.1	85.7	17.7	65.8	59.7
UnivFD	Replay real	68.3	55.8	49.7	76.1	44.5	32.4	54.2	54.8
PLM	SDv1.4 real	96.6	100.0	63.3	90.2	4.4	99.8	54.8	96.6
PLM	Replay real	96.9	100.0	58.3	94.2	4.3	100.0	53.2	98.1
Ours baseline (DINOv3 ViT-L)	SDv1.4 real	75.4	99.7	68.6	94.1	75.0	97.5	73.0	97.1
Ours baseline (DINOv3 ViT-L)	Replay real	71.7	99.8	47.6	99.4	40.8	99.1	53.4	99.5

synthetic pattern from pseudo-label and exploration noise. Changing the accompanying source composition has a relatively small effect and cannot reliably replace missing within-source information. When a difficult source is represented by only a few images, fine-tuning is therefore more likely to amplify noise than to learn a useful recall pattern. We recommend retaining the round-zero detector and using one-pass generalization in this regime, or postponing adaptation until more target images are accumulated. Iteration is also unnecessary when the initial detector already provides near-saturated coverage, as observed for Wukong and BigGAN, for which the additional gain is small relative to the adaptation cost.

E Effect of Replay Real Supervision

This experiment isolates whether replay real supervision affects AI and Non-AI recognition without adding synthetic sources or target-specific supervision. For each detector, the synthetic training data, optimization, and evaluation remain fixed. Only the real data change from the original SDv1.4 split to the replay collection. Table 7 shows that replay real supervision alone does not consistently improve either class. Average AI accuracy decreases by 4.5% for NPR, 11.6% for UnivFD, 1.6% for PLM, and 19.6% for our DINOv3 ViT-L baseline. For DINOv3 ViT-L, Non-AI accuracy improves from 97.1% to 99.5%, but AI accuracy falls from 73.0% to 53.4%. PLM exhibits the same trade-off on a smaller scale, while both class averages decrease for NPR and UnivFD.

The results reflect the distribution alignment of the original training pair. The SDv1.4 split uses ImageNet real images and synthetic images generated from corresponding class conditions (Zhu et al., 2023), allowing the detector to learn a boundary from generation-related differences between semantically aligned real and fake data. Replacing the original real supervision with replay data without a corresponding expansion of the fake side can contract the region assigned to fake images. The mean predicted fake probability consequently decreases for all four detectors. This can improve Non-AI accuracy, as observed for PLM and DINOv3 ViT-L, but reduces generalization to unseen fake images and may also introduce distribution mismatch that harms both classes. In our framework, replay is instead paired with target-specific synthetic expansion. It prevents this expansion from absorbing real images, thereby reducing false positives and improving real accuracy while iterative recall preserves and expands fake coverage. This is a benefit of the coupled update rather than a general gain from replay real supervision alone.

F Detailed Results on Individual Benchmarks

This section reports two groups of detailed results under their respective source-training settings. For the eleven benchmarks in Table 1, we follow the DDA training setting and use the same source-training data adopted by DDA (Chen et al., 2025). Both our round-zero baseline and final detector are evaluated on the disjoint held-out split. For AIGIBench and Treasure in Table 3, our detector is instead trained on the SDv1.4 split of GenImage. Tables 8 and 9 report AI and Non-AI accuracy separately for each source, with benchmark-level class accuracies in the Avg columns. The subse-

Table 8: Detailed AI and Non-AI accuracy (%) on AIGIBench.

Method	ProGAN		R3GAN		StyleGAN3		StyleGAN-XL		StyleSwim		WFIR		BlendFace	
	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI
FreqNet	99.4	99.3	59.9	64.7	98.2	68.0	95.5	64.3	97.1	64.7	89.3	30.2	0.3	46.2
CLIPDetection	98.9	97.9	94.1	72.9	82.6	76.5	96.7	72.5	98.1	74.7	91.9	48.0	5.5	64.6
AIDE	95.3	99.1	99.0	86.7	91.1	85.1	91.7	85.7	82.0	85.4	42.9	99.9	23.2	79.7
SAFE	**99.9**	**100.0**	91.2	96.6	92.9	96.6	89.7	96.5	99.3	96.3	20.7	**100.0**	0.8	93.8
PLM	98.5	96.9	**100.0**	87.7	73.0	88.7	92.5	89.1	76.2	89.3	0.0	**100.0**	0.0	81.3
Ours	99.5	99.5	**100.0**	**99.5**	**100.0**	**100.0**	**99.5**	**100.0**	**100.0**	**99.0**	**100.0**	**100.0**	**100.0**	**99.5**

Method	E4S		FaceSwap		InSwap		SimSwap		FLUX1-dev		Midjourney-V6	
	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI
FreqNet	1.1	50.3	6.2	73.4	0.9	72.6	0.6	72.0	92.4	64.7	83.6	25.3
CLIPDetection	46.9	67.0	27.3	78.4	8.2	78.4	8.6	78.8	86.6	73.3	80.6	50.0
AIDE	6.6	82.1	14.3	89.0	11.4	89.6	21.5	88.2	90.0	86.0	79.8	73.0
SAFE	28.9	94.8	3.3	97.1	56.8	97.7	1.1	97.0	**99.8**	96.3	97.2	91.0
PLM	0.0	80.9	3.3	93.0	0.0	93.3	0.2	92.4	99.6	89.7	98.6	67.3
Ours	**100.0**	**99.5**	**100.0**	**99.5**	**100.0**	**100.0**	**100.0**	**99.5**	98.5	**100.0**	**100.0**	**100.0**

Method	GLIDE		DALLE-3		Imagen3		SD3		SDXL		BLIP	
	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI
FreqNet	79.7	71.8	68.2	64.2	81.5	65.8	88.1	66.6	98.9	66.6	**100.0**	87.7
CLIPDetection	75.2	78.2	75.2	74.9	84.2	73.6	90.6	78.4	91.0	78.4	92.1	85.1
AIDE	98.4	88.4	24.5	85.7	93.9	85.7	99.3	89.3	97.6	89.3	**100.0**	92.8
SAFE	87.8	97.2	1.8	96.1	97.0	96.4	91.7	96.6	99.9	96.6	**100.0**	99.4
PLM	**99.9**	93.8	0.0	88.2	**99.5**	88.4	99.6	92.9	**100.0**	90.5	**100.0**	**100.0**
Ours	97.0	**99.0**	**100.0**	**100.0**	97.0	**100.0**	**100.0**	**99.5**	**100.0**	**100.0**	99.5	98.0

Method	Infinite-ID		InstantID		IP-Adapter		PhotoMaker		SocialRF		CommunityAI		Avg	
	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI
FreqNet	92.7	65.4	93.9	65.8	93.0	65.8	88.6	65.5	39.3	68.5	12.2	98.9	66.4	65.9
CLIPDetection	93.8	75.2	96.9	74.0	92.0	73.3	65.2	73.3	55.5	53.3	51.2	82.8	71.5	73.3
AIDE	97.5	87.1	97.0	86.6	93.5	86.6	97.5	85.9	18.4	97.2	9.3	99.0	67.0	88.1
SAFE	99.8	96.3	**99.9**	96.5	89.8	95.9	98.0	96.0	16.4	99.6	8.5	**100.0**	63.0	96.8
PLM	**100.0**	91.2	**99.9**	88.2	**99.9**	89.7	**100.0**	86.9	19.8	**99.7**	8.0	99.9	62.7	90.4
Ours	**100.0**	**99.0**	99.5	**100.0**	99.0	**100.0**	**100.0**	**99.5**	**98.5**	94.0	**96.5**	**100.0**	**99.4**	**99.4**

quent tables provide source-level results corresponding to Table 1. Comparison results, including their reported Avg entries, are reproduced from the detailed tables of DDA, with deviations omitted for compactness. Unless a caption states otherwise, source columns report balanced accuracy and Avg is the benchmark-level aggregate score under the original evaluation protocol, which may differ from an unweighted mean when source sizes or reporting conventions differ. The four in-the-wild benchmarks are grouped in Table 17 and discussed jointly: SynthWildX and WildRF provide source or platform breakdowns, whereas Chameleon and BFree-Online provide only aggregate results because generator identities are unavailable.

AIGIBench. Our method reaches 99.4% mean accuracy, with 99.4% AI and 99.4% Non-AI accuracy. Its source-wise balanced accuracy is at least 96.3% on all 25 sources. The improvement is especially consistent across the five face-swap sources and the SocialRF and CommunityAI subsets, where several zero-shot detectors remain close to chance on AI images. Performance is also near saturation across GAN, diffusion, personalized generation, and image-conditioned generation sources, showing that the aggregate gain is not dominated by a single family or obtained by favoring one class.

Treasure. Treasure provides a broader test over recent open- and closed-source generators. Our method reaches 98.9% mean accuracy, with 98.2% AI and 99.6% Non-AI accuracy, while using no target annotations. Crucially, the gain is driven by substantially improved recall of AI-generated images rather than by increasing real-image accuracy: our AI accuracy exceeds the strongest comparison result by 21.6%, while Non-AI accuracy remains essentially saturated at 99.6%. Thus, the high mean accuracy reflects expanded synthetic-image coverage instead of a bias toward classifying more images as real. AI accuracy is at least 95.0% on 61 of the 64 sources. The remaining difficult sources are Midjourney V7, Nano Banana Pro, and PixArt-α, with AI accuracies of 89.5%, 84.0%, and 60.0%, respectively. Their diversity leaves challenging modes underrepresented by the initially recalled subset, motivating additional exploration.

Table 9: Detailed AI and Non-AI accuracy (%) on Treasure.

Method	ADM		BAGEL		BRIA v3.2		BigGAN		CogView2		CogView4	
	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI
FreqNet	32.7	75.0	82.9	75.0	84.4	75.0	97.9	75.0	65.5	75.0	21.7	75.0
CLIPDetection	14.3	75.8	73.3	75.8	83.3	75.8	77.0	75.8	90.5	75.8	46.7	75.8
AIDE	92.9	98.6	78.9	98.6	80.0	98.6	73.8	98.6	11.8	98.6	0.7	98.6
SAFE	62.4	**99.8**	96.8	**99.8**	98.2	99.8	0.0	**99.8**	55.7	99.8	0.1	99.8
PLM	98.2	**99.8**	98.2	**99.8**	98.7	**99.9**	97.3	99.7	1.4	**99.9**	0.0	**99.9**
Ours	**99.5**	99.6	**99.5**	99.6	**99.5**	99.6	**100.0**	99.6	**100.0**	99.6	**98.5**	99.6

Method	CogView3+		DALL-E 2		DALL-E 3		DF-GAN		DeepFloyd IF		FLUX.1-dev		FLUX.2	
	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI
FreqNet	61.0	75.0	84.0	75.0	60.9	75.0	75.0	75.0	27.1	75.0	41.0	75.0	84.7	75.0
CLIPDetection	71.6	75.8	45.2	75.8	55.8	75.8	99.8	75.8	57.2	75.8	51.1	75.8	74.1	75.8
AIDE	82.5	98.6	84.2	98.6	6.0	98.6	90.9	98.6	94.0	98.6	1.5	98.6	79.3	98.6
SAFE	94.6	99.8	85.6	99.8	0.1	**99.8**	94.9	**99.8**	52.6	**99.8**	0.0	99.8	86.5	**99.8**
PLM	96.9	**100.0**	86.3	**99.9**	0.1	**99.8**	97.8	**99.8**	96.3	**99.8**	0.1	**99.9**	**99.0**	**99.8**
Ours	**100.0**	99.6	**100.0**	99.6	**99.5**	99.6	**100.0**	99.6	**99.0**	99.6	**100.0**	99.6	97.0	99.6

Method	GLIDE		GPT-4o		GigaGAN		HiDream-I1-Dev		HunyuanDiT		HunyuanImage 3.0		Imagen	
	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI
FreqNet	80.8	75.0	83.8	75.0	98.5	75.0	98.6	75.0	67.9	75.0	88.2	75.0	90.3	75.0
CLIPDetection	81.5	75.8	70.5	75.8	79.5	75.8	84.1	75.8	80.8	75.8	86.8	75.8	89.9	75.8
AIDE	98.6	98.6	75.2	98.6	97.6	98.6	99.9	98.6	75.2	98.6	99.5	98.6	96.1	98.6
SAFE	95.0	**99.8**	88.1	99.8	92.5	**99.8**	99.5	**99.8**	93.5	99.8	96.7	**99.8**	54.9	99.8
PLM	**99.7**	**99.8**	90.8	**99.9**	99.3	**99.8**	99.9	**99.8**	96.9	**99.9**	98.6	99.7	86.5	**99.9**
Ours	99.5	99.6	**100.0**	99.6	**100.0**	99.6	**100.0**	99.6	**98.0**	99.6	**100.0**	99.6	**99.5**	99.6

Method	Imagen 4		Infinity		Janus-Pro-7B		Kolors		LlamaGen		LongCat-Image		LUMINA-Image 2.0	
	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI
FreqNet	67.7	75.0	94.2	75.0	98.9	75.0	91.8	75.0	86.7	75.0	48.7	75.0	79.1	75.0
CLIPDetection	73.8	75.8	80.5	75.8	85.8	75.8	81.2	75.8	89.9	75.8	86.9	75.8	85.7	75.8
AIDE	16.2	98.6	97.2	98.6	97.3	98.6	92.3	98.6	94.5	98.6	87.3	98.6	84.2	98.6
SAFE	0.6	**99.8**	**99.9**	**99.8**	99.8	**99.8**	98.4	**99.8**	99.0	**99.8**	95.4	99.8	90.4	**99.8**
PLM	0.1	**99.8**	98.7	**99.8**	99.2	**99.8**	99.2	**99.8**	99.2	**99.8**	96.1	**100.0**	96.1	**99.8**
Ours	**96.0**	99.6	99.5	99.6	**100.0**	99.6	**100.0**	99.6	**100.0**	99.6	**98.0**	99.6	**100.0**	99.6

Method	MAE		Midjourney V6.1		Midjourney V7		Midjourney V4		Midjourney V5		Midjourney V6		Nano Banana	
	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI
FreqNet	84.2	75.0	35.1	75.0	26.1	75.0	81.8	75.0	81.0	75.0	67.0	75.0	81.0	75.0
CLIPDetection	70.0	75.8	51.2	75.8	41.7	75.8	67.3	75.8	60.6	75.8	65.5	75.8	73.6	75.8
AIDE	97.9	98.6	1.0	98.6	1.0	98.6	71.7	98.6	63.9	98.6	69.5	98.6	90.1	98.6
SAFE	**100.0**	**99.8**	0.2	**99.8**	0.3	**99.8**	85.3	**99.8**	90.7	**99.8**	89.1	**99.8**	97.3	**99.8**
PLM	96.9	**99.8**	0.1	**99.8**	0.3	**99.8**	96.0	**99.8**	**97.5**	**99.8**	97.2	**99.8**	98.6	**99.8**
Ours	**100.0**	99.6	**97.0**	99.6	**89.5**	99.6	**100.0**	99.6	95.5	99.6	**97.5**	99.6	**99.0**	99.6

Method	Nano Banana Pro		NextStep		OmniGen		OmniGen2		Playground V2		Playground V2.5		ProGAN	
	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI
FreqNet	57.6	75.0	99.3	75.0	84.2	75.0	61.7	75.0	92.5	75.0	87.0	75.0	**100.0**	75.0
CLIPDetection	56.6	75.8	87.1	75.8	84.8	75.8	72.7	75.8	81.4	75.8	75.1	75.8	98.8	75.8
AIDE	4.5	98.6	97.4	98.6	79.5	98.6	83.1	98.6	95.0	98.6	89.5	98.6	99.5	98.6
SAFE	0.1	**99.8**	**100.0**	**99.8**	97.7	99.8	94.8	**99.8**	97.3	**99.8**	97.4	99.8	**100.0**	**99.8**
PLM	0.0	99.7	98.3	**99.8**	98.5	**99.9**	96.5	**99.8**	99.2	**99.8**	99.4	**100.0**	99.0	**99.8**
Ours	**84.0**	99.6	**100.0**	99.6	**100.0**	99.6	**99.5**	99.6	**100.0**	99.6	**100.0**	99.6	**100.0**	99.6

Method	Qwen-Image		SD3-Medium		SDXL		SD v1.4		SD v1.5		SD v2.1		SANA v1.5	
	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI
FreqNet	63.3	75.0	84.3	75.0	47.3	75.0	99.9	75.0	**99.8**	75.0	89.8	75.0	61.4	75.0
CLIPDetection	75.8	75.8	75.6	75.8	74.2	75.8	94.2	75.8	93.8	75.8	68.3	75.8	73.6	75.8
AIDE	95.9	98.6	80.5	98.6	93.8	98.6	99.9	98.6	99.7	98.6	97.6	98.6	95.8	98.6
SAFE	98.6	**99.8**	83.3	**99.8**	98.6	**99.8**	99.9	**99.8**	**99.8**	**99.8**	96.2	**99.8**	99.3	99.8
PLM	99.5	**99.8**	95.7	99.7	**99.8**	**99.8**	98.7	**99.8**	98.1	**99.8**	96.8	**99.8**	97.5	**99.9**
Ours	**100.0**	99.6	**99.0**	99.6	99.5	99.6	**100.0**	99.6	99.5	99.6	**99.0**	99.6	**99.5**	99.6

Method	Show-o		Show-o2		StarGAN		StyleGAN3		VQDM		Wukong		Z-Image-Turbo	
	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI
FreqNet	93.2	75.0	98.3	75.0	42.7	75.0	16.2	75.0	34.6	75.0	98.1	75.0	82.4	75.0
CLIPDetection	83.7	75.8	90.1	75.8	62.4	75.8	31.5	75.8	46.6	75.8	78.6	75.8	64.3	75.8
AIDE	89.4	98.6	98.0	98.6	35.5	98.6	3.7	98.6	89.8	98.6	98.8	98.6	94.5	98.6
SAFE	99.6	**99.8**	99.8	**99.8**	35.2	**99.8**	0.0	**99.8**	84.9	**99.8**	**99.6**	**99.8**	90.3	**99.8**
PLM	99.2	**99.8**	99.5	**99.8**	34.5	**99.8**	0.0	**99.8**	98.0	**99.8**	98.1	**99.8**	97.8	**99.8**
Ours	**100.0**	99.6	**100.0**	99.6	**100.0**	99.6	**100.0**	99.6	**100.0**	99.6	99.5	99.6	**99.0**	99.6

Method	Doubao Seedream 3.0		Doubao Seedream 4.0		GPT-image-1.5		Ideogram		Ovis-U1		PixArt-α		Sora-image	
	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI	AI	Non-AI
FreqNet	33.3	75.0	45.8	75.0	72.3	75.0	48.9	75.0	74.6	75.0	60.6	75.0	73.9	75.0
CLIPDetection	54.3	75.8	74.7	75.8	65.8	75.8	60.1	75.8	82.7	75.8	**72.8**	75.8	66.8	75.8
AIDE	0.4	98.6	0.0	98.6	75.7	98.6	2.1	98.6	78.6	98.6	41.3	98.6	75.2	98.6
SAFE	0.0	**99.8**	0.3	**99.8**	**98.4**	**99.8**	1.6	**99.8**	96.9	**99.8**	45.8	**99.8**	**98.7**	**99.8**
PLM	0.1	99.7	0.0	**99.8**	97.8	**99.8**	1.6	99.7	96.1	**99.8**	52.2	99.7	98.3	**99.8**
Ours	**99.0**	99.6	**97.5**	99.6	95.0	99.6	**99.0**	99.6	**99.5**	99.6	60.0	99.6	97.0	99.6

Method	Wan2.2-T2I-Flash		Wan2.5-T2I-Preview		Avg	
	AI	Non-AI	AI	Non-AI	AI	Non-AI
FreqNet	76.6	75.0	66.7	75.0	71.8	75.0
CLIPDetection	86.9	75.8	83.9	75.8	72.6	75.8
AIDE	70.6	98.6	92.1	98.6	70.9	98.6
SAFE	98.2	**99.8**	99.5	**99.8**	73.5	**99.8**
PLM	76.5	99.7	56.4	**99.8**	76.6	**99.8**
Ours	**98.5**	99.6	**100.0**	99.6	**98.2**	99.6

Table 10: Detailed balanced accuracy (%) on the eight GenImage sources.

Method	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg
NPR	53.4	55.1	55.0	43.8	41.2	57.4	48.4	57.7	51.5
UnivFD	55.1	55.6	55.7	62.5	61.3	61.1	76.9	84.4	64.1
FatFormer	52.1	53.6	53.8	61.4	65.5	60.9	72.5	82.2	62.8
SAFE	49.0	49.7	49.8	49.5	53.0	50.3	50.2	50.9	50.3
C2P-CLIP	56.6	77.5	76.9	71.6	73.5	79.4	73.7	85.9	74.4
AIDE	58.2	77.2	77.4	50.4	54.6	70.5	50.8	50.6	61.2
DRCT	82.4	88.3	88.2	76.9	86.1	87.9	85.4	87.0	84.7
AlignedForensics	**97.5**	99.7	99.6	52.4	57.6	99.6	75.0	50.6	79.0
DDA	95.6	98.7	98.6	89.5	89.6	98.7	76.5	86.5	91.7
Baseline ($t = 0$)	78.5	97.3	98.5	86.5	93.0	98.3	96.3	98.0	93.3
Ours	94.5	**99.8**	**99.8**	**99.5**	**99.8**	**99.8**	**100.0**	**99.8**	**99.1**

GenImage. Our method reaches at least 94.5% on every GenImage source and raises the aggregate balanced accuracy from 93.3% to 99.1%. The largest gains over the baseline occur on Midjourney and ADM, improving by 16.0% and 13.0%, respectively. These are also the two sources with the weakest initial coverage, supporting the view that adaptation primarily recovers distributions not already captured by the source-trained detector. Stable Diffusion, Wukong, VQDM, and BigGAN begin from stronger baselines and are preserved or brought close to saturation, indicating that the recall expansion does not require sacrificing sources that already generalize well.

DRCT-2M. Performance after adaptation is at least 99.0% on every DRCT-2M subset, with the aggregate score increasing from 93.2% to 99.7%. The most substantial recovery is observed on SDXL, whose balanced accuracy increases from 72.8% to 100.0%. SD2 and SDXL-Turbo also show clear gains from weaker initial coverage. At the same time, the method maintains near-saturated performance on inpainting and diffusion-reconstruction subsets. The consistent result across standard diffusion, refiner, turbo, latent-consistency, ControlNet, and reconstruction variants shows that the adaptation signal transfers across substantially different diffusion pipelines.

Table 11: Detailed balanced accuracy (%) on the 16 DRCT-2M subsets.

Method	LDM	SD1.4	SD1.5	SD2	SDXL	Ref.	SD-T	XL-T	LCM1.5	LCM-XL	SD1-C	SD2-C	XL-C	SD1-DR	SD2-DR	XL-DR	Avg
NPR	33.0	29.1	29.0	35.1	33.2	28.4	27.9	27.9	29.4	30.2	28.4	28.3	34.7	67.9	67.4	66.1	37.3
UnivFD	85.4	56.8	56.4	58.2	63.2	55.0	56.5	53.0	54.5	65.9	68.0	65.4	75.9	64.6	56.2	53.9	61.8
FatFormer	55.9	48.2	48.2	48.2	48.2	48.3	48.2	48.2	48.3	50.6	49.7	49.9	59.8	66.3	60.6	56.0	52.2
SAFE	50.3	50.1	50.0	50.0	49.9	50.1	50.0	50.0	50.1	50.0	49.9	50.0	54.7	98.2	98.5	97.3	59.3
C2P-CLIP	83.0	51.7	51.7	52.9	51.9	64.6	51.7	50.6	52.0	66.1	56.9	54.7	77.8	67.2	57.1	56.7	59.2
AIDE	64.4	74.9	75.1	58.5	53.5	66.3	52.8	52.8	70.0	54.3	65.9	53.6	53.9	95.3	73.3	69.0	64.6
DRCT	96.7	96.3	96.3	94.9	96.2	93.5	93.4	92.9	91.2	95.0	95.6	92.7	92.0	94.1	69.6	57.4	90.5
AlignedForensics	99.9	99.9	**99.9**	99.6	90.2	81.3	**99.7**	89.4	**99.7**	90.0	**99.9**	99.2	87.6	**99.9**	**99.8**	92.6	95.5
DDA	99.2	98.9	99.0	98.3	98.0	96.8	97.9	94.8	95.9	98.2	98.7	99.0	99.4	99.0	99.5	96.3	98.1
Baseline ($t = 0$)	98.5	95.0	95.0	82.0	72.8	93.5	94.3	86.0	95.5	89.8	96.0	96.8	98.5	99.5	99.3	98.5	93.2
Ours	**100.0**	**100.0**	99.8	**99.8**	**100.0**	**99.8**	99.5	**100.0**	99.0	**99.8**	99.5	**99.8**	**99.8**	99.8	99.3	**99.8**	**99.7**

Table 12: Detailed balanced accuracy (%) on the 17 AIGCDetectionBenchmark sources.

Method	ADM	DALL-E2	GLIDE	MJ	VQDM	BigGAN	CycleGAN	GauGAN	ProGAN	SDXL	SD1.4	SD1.5	StarGAN	StyleGAN	StyleGAN2	WFR	Wukong	Avg
NPR	43.8	20.0	41.2	53.4	48.4	53.1	76.6	42.2	58.7	59.6	55.1	55.0	67.4	57.9	54.6	58.8	57.4	53.1
UnivFD	62.5	50.0	61.3	55.1	76.9	87.5	96.9	98.8	99.4	58.2	55.6	55.7	95.1	80.0	69.4	69.2	61.1	72.5
FatFormer	80.2	68.5	91.1	54.4	88.0	99.2	**99.5**	99.1	98.5	71.7	67.5	67.2	99.4	98.0	98.8	88.3	75.6	85.0
SAFE	49.5	49.5	53.0	49.0	50.2	52.2	51.9	50.0	50.0	49.8	49.7	49.8	50.1	50.0	50.0	49.8	50.3	50.3
C2P-CLIP	71.6	52.3	73.5	56.6	73.7	98.4	96.8	98.8	99.3	62.3	77.5	76.9	99.6	93.1	79.4	94.8	79.4	81.4
AIDE	52.9	51.1	60.2	49.8	69.3	70.1	93.6	60.6	89.0	49.6	51.6	51.0	72.1	66.5	59.0	80.6	54.5	63.6
DRCT	79.9	89.2	89.2	85.5	88.6	81.4	91.0	93.8	71.1	88.3	91.4	91.0	53.0	62.7	63.8	73.9	90.8	81.4
AlignedForensics	51.6	52.0	55.6	96.2	72.1	51.2	49.5	50.8	50.7	95.1	**99.7**	**99.6**	53.8	52.7	51.6	50.0	99.6	66.6
DDA	89.5	94.6	89.6	95.6	76.6	91.0	72.5	92.7	92.8	99.4	98.7	98.6	72.7	87.8	90.2	52.1	98.8	87.8
Baseline ($t = 0$)	83.5	87.3	91.5	78.0	97.3	98.0	96.3	98.5	91.3	94.5	99.0	98.0	99.8	83.8	76.5	80.3	98.0	91.3
Ours	**98.8**	**99.5**	**99.8**	**96.3**	**100.0**	**100.0**	98.8	**100.0**	**100.0**	**100.0**	99.3	99.3	**100.0**	**100.0**	**100.0**	**95.3**	**100.0**	**99.2**

Table 13: Detailed balanced accuracy (%) on ForenSynths.

Method	BigGAN	CRN	CycleGAN	DeepFake	GauGAN	IMLE	ProGAN	SAN	SeeingDark	StarGAN	StyleGAN	StyleGAN2	WFR	Avg
NPR	53.1	0.4	76.6	35.7	42.2	5.3	58.7	48.4	63.6	67.4	57.9	54.6	58.8	47.9
UnivFD	87.5	55.7	96.9	69.4	98.8	68.1	99.4	58.2	62.2	95.1	80.0	69.4	69.2	77.7
FatFormer	99.3	72.1	99.5	93.0	99.3	72.1	98.4	70.8	81.9	99.4	98.1	98.9	88.3	90.0
SAFE	52.2	50.0	51.9	50.1	50.0	50.0	50.0	50.9	41.1	50.1	50.0	50.0	49.8	49.7
C2P-CLIP	98.4	93.3	96.8	92.6	98.8	93.2	99.3	63.2	94.7	99.6	93.1	79.4	94.8	92.0
AIDE	70.1	12.2	93.6	53.2	60.6	15.9	89.0	55.3	44.2	72.1	66.5	59.0	80.6	59.4
DRCT	81.4	78.4	91.0	51.5	93.8	82.6	71.1	84.9	72.2	53.0	62.7	63.8	73.9	73.9
AlignedForensics	51.2	50.4	49.5	71.7	50.8	49.7	50.7	67.6	51.4	53.8	52.7	51.6	50.0	53.9
DDA	91.0	87.0	72.5	76.5	92.7	89.7	92.8	**94.7**	58.6	72.7	87.8	90.2	52.1	81.4
Baseline ($t = 0$)	98.0	78.0	98.3	55.3	98.8	80.0	90.5	65.9	52.8	99.8	86.8	79.5	82.8	85.2
Ours	**100.0**	**100.0**	**100.0**	**99.3**	**100.0**	**100.0**	**100.0**	67.0	**95.8**	**100.0**	**99.8**	**100.0**	**98.5**	**99.1**

Table 14: Detailed balanced accuracy (%) on the nine Synthbuster generators.

Method	DALL-E2	DALL-E3	Firefly	GLIDE	Midjourney	SD1.3	SD1.4	SD2	SDXL	Avg
NPR	51.1	49.3	46.5	48.5	52.8	51.4	51.8	46.0	52.8	50.0
UnivFD	83.5	47.4	89.9	53.3	52.5	70.4	69.9	75.7	68.0	67.8
FatFormer	59.4	39.5	60.3	72.7	44.4	53.7	54.0	52.3	69.1	56.1
SAFE	58.0	9.9	10.3	52.2	56.7	59.4	59.1	53.0	59.5	46.5
C2P-CLIP	55.6	63.2	59.5	86.7	52.9	75.2	76.7	69.2	77.7	68.5
AIDE	34.9	33.7	24.8	65.0	57.5	74.1	73.7	53.2	68.4	53.9
DRCT	77.2	86.6	84.1	82.6	73.7	86.6	86.6	83.2	71.3	84.8
AlignedForensics	50.2	48.9	51.7	53.5	98.7	98.8	98.8	98.6	97.3	77.4
DDA	86.3	90.0	91.9	76.5	93.5	92.9	92.7	93.3	93.5	90.1
Baseline ($t = 0$)	87.3	97.5	78.0	96.0	95.0	99.3	98.8	94.5	97.5	93.8
Ours	**100.0**	**100.0**	**99.5**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**99.9**

AIGCDetectionBenchmark. The final detector exceeds 95% on all 17 AIGCDetectionBenchmark sources and reaches 100% on nine of them, increasing the aggregate score from 91.3% to 99.2%. The largest improvements occur for StyleGAN2 and Midjourney, showing that iterative recall benefits both GAN- and diffusion-based sources with weak initial coverage. WhichFaceIsReal and Midjourney remain the lowest final entries, but both improve substantially, while the initially strong Stable Diffusion, StarGAN, and Wukong subsets remain near saturation. The breadth of these gains indicates that the final average is not driven only by sources closely related to the source-training distribution.

ForenSynths. Our method raises aggregate balanced accuracy from 85.2% to 99.1% and reaches at least 95.8% on every source except SAN. DeepFake and SeeingDark exhibit the largest recoveries, increasing from 55.3% to 99.3% and from 52.8% to 95.8%, respectively. SAN improves only from 65.9% to 67.0% and remains the clear failure case, suggesting that its artifacts are insufficiently represented by both the initial predictions and the explored target samples.

Table 15: Detailed results (%) on DDA-COCO.

Method	Real	XL	EMA	MSE	SD2.1	SD3.5	FLUX.1	Avg
NPR	55.4	16.1	31.1	41.3	41.2	24.9	19.2	42.2
UnivFD	99.2	3.9	9.6	7.3	7.4	3.4	1.8	52.4
FatFormer	96.4	5.4	6.9	10.4	10.3	6.6	2.8	51.7
SAFE	98.8	0.6	0.9	0.9	1.0	0.3	1.8	49.9
C2P-CLIP	99.5	2.0	2.7	4.3	4.2	4.0	1.3	51.3
AIDE	98.8	0.6	2.2	1.7	1.8	0.3	0.6	50.0
DRCT	94.2	16.9	34.8	33.5	33.6	21.7	17.2	60.2
AlignedForensics	**99.8**	82.5	99.2	99.0	99.1	55.4	3.6	86.5
DDA	99.0	95.0	99.3	99.7	99.7	68.1	50.2	92.2
Baseline ($t = 0$)	98.4	88.5	95.5	97.5	99.0	13.0	15.5	83.3
Ours	95.8	**99.0**	**100.0**	**100.0**	**100.0**	**75.5**	**86.0**	**94.6**

Table 16: Detailed accuracy (%) on the five EvalGEN sources.

Method	Flux	GoT	Infinity	NOVA	OmniGen	Avg
NPR	0.7	0.2	6.5	4.7	2.2	2.9
UnivFD	4.0	9.2	15.7	8.3	39.6	15.4
FatFormer	9.9	47.9	44.7	98.3	27.3	45.6
SAFE	1.0	0.5	1.9	0.6	1.6	1.1
C2P-CLIP	8.7	49.6	35.3	86.4	14.5	38.9
AIDE	17.9	24.7	3.4	16.3	33.4	19.1
DRCT	72.5	81.4	77.9	84.6	72.5	77.8
AlignedForensics	32.0	72.3	74.0	84.8	77.0	68.0
DDA	89.9	99.5	97.8	99.5	99.5	97.2
Baseline ($t = 0$)	39.6	48.6	56.8	84.7	83.8	62.7
Ours	**100.0**	**100.0**	**99.1**	**100.0**	**100.0**	**99.8**

Table 17: Detailed balanced accuracy (%) on four in-the-wild benchmarks.

(a) SynthWildX

Method	DALL-E3	Firefly	Midjourney	Avg
NPR	43.6	61.3	44.5	49.8
UnivFD	45.4	65.3	46.2	52.3
FatFormer	46.5	61.6	48.3	52.1
SAFE	49.4	48.2	49.6	49.1
C2P-CLIP	56.9	61.4	53.0	57.1
AIDE	63.4	48.8	51.9	48.8
DRCT	58.3	56.4	50.5	55.1
AlignedForensics	85.5	58.5	92.2	78.8
DDA	92.3	87.3	93.1	90.9
Baseline ($t = 0$)	**94.9**	83.2	86.2	87.6
Ours	**94.9**	**98.4**	**100.0**	**94.6**

(b) WildRF

Method	Facebook	Reddit	Twitter	Avg
NPR	78.1	61.0	51.3	63.5
UnivFD	49.1	60.2	56.5	55.3
FatFormer	54.1	68.1	54.4	58.9
SAFE	50.9	74.1	37.5	57.2
C2P-CLIP	54.4	68.4	55.9	59.6
AIDE	57.8	71.5	45.8	58.4
DRCT	46.6	53.1	55.2	50.6
AlignedForensics	89.4	69.1	81.8	80.1
DDA	93.1	86.4	91.5	90.3
Baseline ($t = 0$)	92.2	94.7	86.6	92.8
Ours	**100.0**	**99.7**	**98.2**	**99.4**

(c) Aggregate-only sets

Method	Chameleon	BFree
NPR	59.9	49.5
UnivFD	50.7	49.0
FatFormer	51.2	50.0
SAFE	59.2	50.5
C2P-CLIP	51.1	50.0
AIDE	63.1	53.1
DRCT	56.6	55.7
AlignedForensics	71.0	68.5
DDA	82.4	95.1
Baseline ($t = 0$)	79.3	78.6
Ours	**97.8**	**100.0**

Synthbuster. The detector becomes nearly saturated across Synthbuster, increasing the aggregate score from 93.8% to 99.9%, reaching 99.5% on Firefly, and reaching 100% on the other eight generators. Firefly exhibits the largest gain of 21.5%, while DALL-E 2 also benefits from a weaker initial result. Performance on DALL-E 3, GLIDE, Midjourney, and the Stable Diffusion variants is preserved or improved to saturation, demonstrating consistent adaptation across both closed-source systems and multiple diffusion generations.

DDA-COCO. The main gains arise on SD3.5 and FLUX.1, whose fake-image accuracies increase by 62.5% and 70.5%. The four older reconstruction subsets reach at least 99.0%, while the recent generators remain harder despite substantial recovery. Aggregate balanced accuracy rises from 83.3% to 94.6%, showing that replay limits boundary drift while allowing synthetic coverage to expand.

EvalGEN. Iterative adaptation raises accuracy on every source to at least 99.1% and increases mean accuracy from 62.7% to 99.8%. The recovery of Infinity and NOVA demonstrates transfer to autoregressive generation, while gains on Flux, GoT, and OmniGen cover recent diffusion and multimodal pipelines, indicating that the iterative signal is not tied to one generation architecture.

In-the-Wild Benchmarks. Across the four benchmarks, SynthWildX rises from 87.6% to 94.6%, with the main gains on Firefly and Midjourney, while WildRF rises from 92.8% to 99.4% and exceeds 98% on every platform. Chameleon improves from 79.3% to 97.8%, and BFree-Online from 78.6% to 100.0%. These results demonstrate effective adaptation under unknown generators, platform processing, and web-distribution shifts without source identities.