TEST-TIME CURRICULUM FOR OPEN-SET AIGC DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated image detectors deployed in open-world environments inevitably face distribution shifts as new and stronger generative models continue to emerge. Although existing methods improve cross-generator generalization through better representations or training data construction, they typically follow a static train-once-and-deploy paradigm and cannot adapt after deployment. In this work, we study open-set AIGC image detection from a test-time adaptation perspective, where pretrained detectors adapt to unlabeled target data from unseen generators. We propose Test-Time Curriculum (TTC), a simple and model-agnostic framework that performs curriculum-based self-training for open-set AIGC detection. TTC separates pseudo-label reliability from adaptation difficulty by starting from reliable balanced samples and progressively incorporating harder informative cases under generator shift. We further introduce Cross-Scale Pseudo-Label Refinement to improve pseudo-label reliability by aggregating complementary evidence across multiple resolutions. In addition, we construct AIGCGuard, a benchmark containing 3,100 representative real images and 124,000 generated images from 40 of the most advanced open-source and proprietary text-to-image models. Extensive experiments on five benchmarks demonstrate that TTC substantially improves detection performance under diverse unseen-generator shifts.

1 INTRODUCTION

The rapid progress of image generation models has made AI-generated images increasingly difficult to distinguish from real photographs (Wang et al., 2026; Seedream et al., 2025). This creates a practical challenge for web-scale visual data ecosystems: AI-generated images may enter training corpora, distort real-data statistics, and introduce generation artifacts into future training cycles (Team et al., 2025; Cao et al., 2025). Reliable AIGC image detection is therefore becoming an important component of trustworthy visual learning systems (Wang, 2024).

The core difficulty is open-set generator shift. During training, a detector only observes images from a limited set of known generators. After deployment, it must detect images produced by newer and often stronger generators whose artifacts may differ substantially from those seen during training. As generators evolve, previously learned artifacts may disappear, become weaker, or be replaced by new patterns, causing the test distribution to shift over time.

Existing approaches mainly improve detector generalization through designing stronger representations or constructing better training data. Recent approaches exploit pretrained visual features (Tan et al., 2025), frequency cues (Yan et al., 2024), patch-level artifacts (Zhong et al., 2023), reconstruction discrepancies (Luo et al., 2024), or aligned real-generated pairs (Cazenavette et al., 2024). Although robust to certain distribution shifts, they typically follow a static train-once-and-deploy paradigm and cannot adapt to emerging generator distributions without labeled target data or retraining.

We investigate test-time adaptation (TTA) (Liang et al., 2025; Yuan et al., 2023) as a potential solution for maintaining AIGC detectors after deployment under unseen-generator shifts. Unlike the conventional train-once-and-deploy paradigm, TTA allows a pretrained detector to exploit unlabeled target data collected after deployment without additional annotation or retraining from scratch. This setting aligns naturally with industrial AIGC detection pipelines, which periodically accumulate mixed-domain image batches reflecting recent generator distributions and thus provide a practical data source for offline adaptation. Since generator shifts do not necessarily affect all target samples

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.

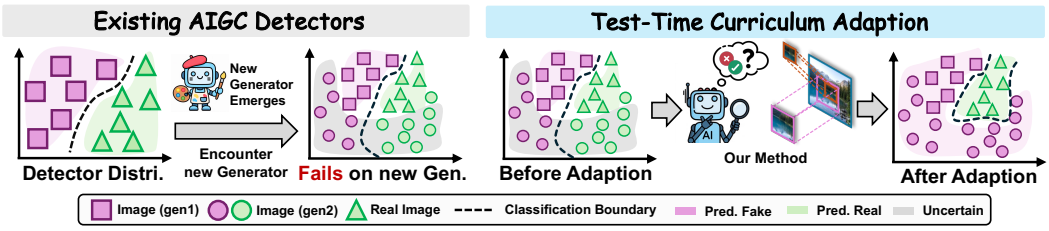

Figure 1: Illustration of open-set AIGC image detection under evolving generator distributions. Existing detectors are trained on a limited set of known generators and often fail when new generators emerge, causing a mismatch between the training and test distributions. In contrast, the proposed Test-Time Curriculum (TTC) adapts the detector using unlabeled test samples, progressively shifting the decision boundary to better separate real images from images generated by unseen models.

equally, a pretrained detector may still retain reliable predictions on a subset of the target batch, providing a natural starting point for pseudo-label-based adaptation.

Building on this observation, we instantiate TTA for AIGC detection through pseudo-label-based self-training (Amini et al., 2025; Rosenberg et al., 2005). However, pseudo-label reliability and adaptation difficulty are not equivalent: highly confident predictions provide safer supervision, whereas more difficult samples may be more informative for adaptation but are also more vulnerable to pseudo-label errors. This motivates Test-Time Curriculum (TTC), a simple and model-agnostic framework that explicitly separates these two factors. TTC first restricts adaptation to class-balanced reliable candidates, then prioritizes informative hard-within-reliable samples, while progressively relaxing the reliability criterion across curriculum rounds to cover a broader portion of the target distribution. To further improve pseudo-label reliability, we integrate TTC with Cross-Scale Pseudo-Label Refinement, which aggregates complementary predictions across resolutions to provide more robust supervision for curriculum adaptation.

Finally, we observe that many existing benchmarks do not fully cover the latest high-fidelity text-to-image generators. To support evaluation under more realistic generator shifts, we construct **AIGCGuard**, a new benchmark containing 3,100 representative real images and generated counterparts from 40 recent open-source and proprietary text-to-image models, resulting in 127,100 images in total, covering diverse semantic categories, resolutions, and generator families.

In all, our contribution can be summarized as follows:

- We investigate test-time adaptation as a practical paradigm for maintaining open-set AIGC detectors under unseen-generator shifts. Our setting allows pretrained detectors to leverage periodically collected unlabeled target batches without additional annotation or retraining from scratch.
- We propose Test-Time Curriculum (TTC), a model-agnostic adaptation framework tailored to generator shifts in AIGC detection. TTC introduces hard-within-reliable selection to separate pseudo-label reliability from adaptation difficulty, and incorporates Cross-Scale Pseudo-Label Refinement to improve supervision robustness across resolutions.
- Comprehensive experiments demonstrate that TTC achieves consistent and significant improvements across multiple benchmarks. We construct AIGCGuard, a modern benchmark containing 127.1K images from 40 recent open-source and proprietary text-to-image models.

2 Related Work

2.1 AI-Generated Image Detection

Existing AIGC image detection methods mainly follow several routes. Early forensic methods exploit low-level spatial or frequency artifacts, such as color statistics, co-occurrence patterns, upsampling traces, and high-frequency fingerprints (McCloskey & Albright, 2018; Nataraj et al., 2019; Frank et al., 2020; Tan et al., 2024). Later works improve generalization through stronger pretrained representations (Wang et al., 2020b; Ojha et al., 2023; Liu et al., 2024; Tan et al., 2025), patch-level analysis (Zhong et al., 2023; Chen et al., 2024b), reconstruction discrepancy (Wang et al., 2023;

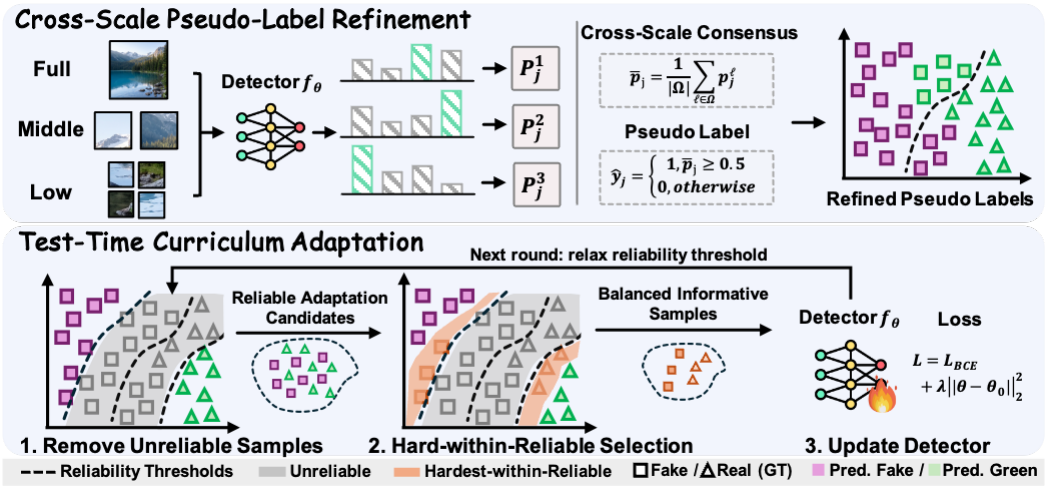

Figure 2: Overview of the proposed TTC framework. Given unlabeled test images, Cross-Scale Pseudo-Label Refinement aggregates multi-scale predictions into a cross-scale consensus for more reliable pseudo labels. TTC then removes unreliable samples and applies hard-within-reliable selection to form a balanced and informative adaptation subset. The detector is then updated, and the procedure repeats across curriculum rounds with progressively relaxed reliability criteria.

Ricker et al., 2024), and frequency modeling (Li et al., 2025; Yan et al., 2024). Recent dataset alignment methods further reduce training or evaluation bias by matching real and generated images in content, style, resolution, or format (Grommelt et al., 2024; Yu et al., 2024; Guillaro et al., 2025). These methods have substantially improved cross-generator generalization, but they are typically deployed with fixed parameters and therefore cannot exploit newly collected unlabeled target data after deployment. Motivated by this limitation, we investigate test-time adaptation as a practical mechanism for maintaining AIGC detectors under unseen-generator shifts.

2.2 Test-Time Adaptation

Test-time adaptation (TTA) adapts a pretrained model using unlabeled target data available at test time. Existing methods adapt normalization statistics (Nado et al., 2020; Schneider et al., 2020), minimize unsupervised objectives such as prediction entropy (Wang et al., 2020a; Niu et al., 2023a), select reliable target samples (Niu et al., 2022; 2023a), or use prototype-based adaptation (Iwasawa & Matsuo, 2021). Related self-training work has also explored class-balanced pseudo-labeling (Zou et al., 2018) and curriculum-based pseudo-labeling (Cascante-Bonilla et al., 2021). TTA has also been studied for DeepFake detection. OST (Chen et al., 2022) performs one-shot test-time training for individual test samples, while T^2A (Nguyen-Le et al., 2025) considers online adaptation over sequentially arriving mini-batches. In contrast, we study offline mixed-domain adaptation for general AIGC image detection, where samples from multiple unseen generators are jointly adapted within a single unlabeled target batch, matching the industrial deployment scenario considered in this work.

Methodologically, TTC separates pseudo-label reliability from adaptation difficulty. It first forms a class-balanced reliable candidate set and then applies hard-within-reliable selection to prioritize informative samples within this constraint. Cross-Scale Pseudo-Label Refinement further improves pseudo-label supervision by aggregating complementary predictions across resolutions.

3 Method

We first present **(1) Test-Time Curriculum Adaptation Framework (TTC)** as the overall framework for adapting detectors at test time. Then, we introduce a **(2) Cross-Scale Pseudo-Label Refinement** to enhance TTC. Together, these two components form a practical pipeline for robust adaptation under evolving AIGC test distributions. The overall framework is shown in Fig. 2.

3.1 PRELIMINARY

The task is to distinguish AI-generated images from real images in an open-set deployment scenario. During training, a labeled set $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{N}$ is drawn from the training distribution $p_{\text{train}}(x)$, where N denotes the number of training samples and $y_i \in \{0, 1\}$ denotes the real/fake label. Fake training images are generated by a limited set of known generators $\{G_1, \ldots, G_K\}$. At test time, the detector is given an unlabeled test set $\mathcal{D}_{test} = \{x_j\}_{j=1}^{M}$, where M denotes the number of test samples and $x_j \sim p_{\text{test}}(x)$. The test set may contain images from previously unseen generators $\{G_{K+1}, \ldots\}$, leading to a distribution shift $p_{\text{test}}(x) \neq p_{\text{train}}(x)$. As new generators emerge, $p_{\text{test}}(x)$ may further evolve during deployment.

The objective is to learn a detector $f_\theta : x \mapsto y \in \{0, 1\}$ that generalizes to unseen generators under the shifted test distribution $p_{\text{test}}(x)$. In a test-time adaptation (TTA) setting (Niu et al., 2022; Liu et al., 2021), the initial pretrained detector f_θ is allowed to update its parameters using the unlabeled test set $\mathcal{D}_{test}$, without access to ground-truth labels. In this work, TTA is instantiated as pseudo-label-based self-training. At the beginning of adaptation, the pretrained detector f_θ generates pseudo labels $\hat{y}_j$ for unlabeled test samples x_j. A generic self-training update at the adaptation stage t can be written as

$$\theta' = \theta - \eta \nabla_\theta \mathcal{L}(f_\theta(x_j), \hat{y}_j), \quad x_j \in \mathcal{D}_{test}, \tag{1}$$

where $\mathcal{L}(\cdot)$ is a loss function, and η is the learning rate. Due to distribution shifts, the detector f_θ may produce incorrect predictions, which manifest as noise in the pseudo labels and can degrade model performance. To mitigate this issue, we propose Test-Time Curriculum (TTC), which explicitly controls error propagation during adaptation.

3.2 TEST-TIME CURRICULUM ADAPTATION FRAMEWORK

Pseudo-Labeling Selection. We first focus on selecting reliable pseudo-labeled samples at curriculum round r, which are critical for safe adaptation under distribution shift. A reliable sample is one for which the detector's prediction is likely correct, thereby minimizing the risk of propagating early errors under distribution shift. To quantify reliability, we introduce the binary entropy $H(\cdot)$ (Lewis, 1995) of the predicted probability p_j for each test sample:

$$H(p_j) = -p_j \log p_j - (1 - p_j) \log(1 - p_j), \tag{2}$$

where $p_j = f_\theta(x_j)$ denotes the predicted probability of sample x_j being generated. Lower entropy indicates a prediction farther from the decision boundary and thus more trustworthy. Based on this reliability measure, TTC selects samples whose entropy falls below an uncertainty threshold τ_r:

$$\mathcal{R}_r = \{x_j \in \mathcal{D}_{test} \mid H(p_j) \leq \tau_r\}. \tag{3}$$

This set contains both reliably predicted real and fake samples. Within $\mathcal{R}_r$, samples are separated into two classes based on the predicted label: the reliable fake candidates $\mathcal{R}_r^{\text{fake}}$ and the reliable real candidates $\mathcal{R}_r^{\text{real}}$. Formally,

$$\begin{aligned} \mathcal{R}_r^{\text{fake}} &= \{x_j \in \mathcal{R}_r \mid \hat{y}_j = 1\}, \\ \mathcal{R}_r^{\text{real}} &= \{x_j \in \mathcal{R}_r \mid \hat{y}_j = 0\}, \\ \hat{y}_j &= \mathbf{1}[p_j > 0.5]. \end{aligned} \tag{4}$$

In practice, real images are often easier to classify confidently, while images from newly emerging generators tend to produce higher uncertainty. Because entropy-based selection relies on prediction uncertainty rather than the number of samples, the numbers of reliably predicted real and fake samples may therefore differ, *i.e.*, $|\mathcal{R}_r^{\text{real}}| \neq |\mathcal{R}_r^{\text{fake}}|$, especially in early curriculum rounds with strict thresholds. Such imbalance can induce class-wise gradient bias, causing the adaptation to be dominated by the over-represented class and potentially degrading performance. To enforce class balance, TTC selects an equal number of samples from each class. Specifically, the number of samples per class is given by $N_r = \min(|\mathcal{R}_r^{\text{fake}}|, |\mathcal{R}_r^{\text{real}}|)$, ensuring that both classes contribute equally to the adaptation. Consequently, all reliable samples from the smaller predicted class are retained, while only N_r samples are selected from the larger class. Class balancing is used as an optimization constraint rather than as an estimate of the target class prior. The original target distribution is not reweighted at evaluation time; balancing only prevents the adaptation gradient from being dominated by one predicted class. If either reliable class is empty, the current round is skipped and the detector retains the last safe parameters, preventing degenerate updates on effectively single-class target batches.

Once class balance is established, a new challenge arises: which samples within each equally sized class subset should be used for adaptation to maximize learning? To this end, TTC performs a second-stage selection within each class, focusing on the informative samples. Although low-uncertainty samples in $\mathcal{R}_r$ are reliable, they are often already well-classified and provide limited new information for adaptation. To focus adaptation on informative samples, TTC selects a hard-within-reliable subset $\mathcal{S}_r$ from the reliable set $\mathcal{R}_r$, prioritizing samples with higher uncertainty within each class:

$$\mathcal{S}_r = \underbrace{\arg\max_{\substack{S\subseteq\mathcal{R}_r^{\text{fake}} \\ |S|=N_r}} \sum_{x_j\in S} H(p_j)}_{\text{hard fake samples}} \cup \underbrace{\arg\max_{\substack{S\subseteq\mathcal{R}_r^{\text{real}} \\ |S|=N_r}} \sum_{x_j\in S} H(p_j)}_{\text{hard real samples}}, \tag{5}$$

where samples closest to the decision boundary are prioritized within each class. This ensures adaptation focuses on challenging yet trustworthy examples, improving the detector's ability to capture subtle cues in generated images.

The above selective sampling rule separates label reliability from learning difficulty: the two confidence regions provide reliable pseudo labels, hard-within-reliable sampling focuses adaptation on informative samples. Balanced selection controls class-wise gradient bias. High-entropy samples with predictions near 0.5 are excluded from the reliable set, reducing the risk of propagating early errors. This set $\mathcal{S}_r$ then forms the input for the curriculum update in round r, allowing the detector to progressively incorporate harder samples as it adapts to the evolving test distribution p_{test}.

Curriculum Adaptation. In the presence of distribution shifts, directly adapting a detector to a new, unseen test distribution p_{test} can be challenging due to the limited availability of reliable samples in the early stages of adaptation. To address this, TTC adopts a curriculum approach, where the detector is gradually adapted in multiple rounds using progressively relaxed uncertainty thresholds. The first round focuses on the most reliable samples with a strict threshold τ_0, as the model is still distant from the test distribution and the available trustworthy data $\mathcal{S}_0$ is limited. In subsequent rounds, the threshold is relaxed, allowing the detector to incorporate harder, more uncertain samples that are closer to the decision boundary. As each round brings the model closer to p_{test}, the set of reliable samples expands, enabling more comprehensive adaptation.

For the selected hard-within-reliable set $\mathcal{S}_r$ in round r, TTC updates the detector using pseudo-label supervision (Cox, 1958) while constraining deviation from the initial pretrained parameters θ_0. The loss function for round r is

$$\mathcal{L}^{(r)} = \text{BCE}(f_\theta(x_j), \hat{y}_j^{(r)}) + \lambda\|\theta - \theta_0\|_2^2, \quad x_j \in \mathcal{S}_r, \tag{6}$$

where λ controls the strength of L2 regularization (Krogh & Hertz, 1991), preventing catastrophic drift from noisy pseudo labels while enabling stable adaptation.

After each round, the detector recomputes the predicted probabilities for all test samples, updating the selected subset for the next round. This iterative process forms a curriculum: reliable samples guide the initial adaptation, and the adapted detector progressively incorporates harder, moderately uncertain samples. This progressive design is intended to reduce the influence of uncertain pseudo labels in early rounds while gradually expanding the adaptation set as the detector becomes better aligned with the target distribution, while a label-free safety check rejects unstable rounds and restores the last safe model when necessary.

3.3 Cross-Scale Pseudo-Label Refinement

Effective pseudo-labeling for Test-Time Curriculum (TTC) requires capturing artifacts present at different visual granularities, from fine-grained local traces (Yang et al., 2025) to coarse structural patterns (Tan et al., 2025). Single-scale predictions often fail to detect all relevant cues, leading to noisy supervision. We therefore introduce a cross-scale refinement strategy that combines multi-resolution predictions, improving pseudo-label reliability for curriculum-based adaptation.

Multi-Scale Prediction. Artifacts introduced by AI-generated images often appear at different visual granularities, ranging from subtle local traces (e.g., high-frequency noise, edges) to coarse structural or semantic inconsistencies. To capture such multi-scale evidence, each test image x_j is processed at a predefined set of scales Ω, which may include the full image, intermediate downsampled resolutions, or localized patches extracted from regions likely to contain artifacts.

Table 1: Comparison on the AIGCDetect Benchmark. We report balanced accuracy (%) for each generator subset and the average across subsets. "Base" denotes the detector initialized with the same backbone and training data as DDA for fair comparison, but without using DDA's alignment strategy. "Ours" applies the proposed TTC adaptation on top of this Base detector. The best result and the second-best result are marked in **bold** and underline, respectively.

Method	ADM	DALLE2	GLIDE	Midjourney	VQDM	BigGAN	CycleGAN	GauGAN	ProGAN	SDXL	SD14	SD15	StarGAN	StyleGAN	StyleGAN2	WFR	Wukong	Avg.
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
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
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
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
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
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
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
Base	70.1	88.5	73.2	66.5	71.5	90.7	60.0	99.0	87.4	97.2	99.1	99.0	49.4	80.2	75.9	81.0	99.1	81.6
Ours	**89.7**	**99.2**	**95.7**	**98.7**	**98.4**	93.7	69.4	97.4	97.2	99.3	98.7	98.8	76.6	90.9	87.1	91.9	98.8	**93.0**

For patch-based scales, informative patches $\{x_{j,k}^{\ell}\}_{k=1}^{N_\ell}$ are selected based on simple heuristics such as local gradient magnitude, high-frequency energy, and texture complexity, while also including a small number of random patches to preserve spatial diversity.

At each scale $\ell \in \Omega$, the detector produces a patch-level probability $p_{j,k}^{\ell}$ for each patch. These patch-level probabilities are aggregated into a single scale-level probability p_j^{ℓ} using logit-space log-sum-exp pooling (Ilse et al., 2018; Fatima et al., 2023):

$$p_j^{\ell} = \sigma(T_\ell \log \frac{1}{N_\ell} \sum_{k=1}^{N_\ell} \exp\left(\frac{\sigma^{-1}(p_{j,k}^{\ell})}{T_\ell}\right)). \tag{7}$$

Here, $\sigma^{-1}(\cdot)$ denotes the inverse sigmoid function (logit), $\sigma(\cdot)$ is the sigmoid function mapping logit back to probability. T_ℓ is a hyperparameter controlling the selectivity of aggregation. This aggregation preserves strong localized evidence while allowing multiple moderately suspicious patches to contribute, improving the reliability of individual scale predictions.

Importantly, this multi-scale prediction framework is model-agnostic: it can be applied to any pre-trained detector without modifying its architecture or training procedure. By extracting complementary evidence across multiple resolutions (Lin et al., 2017; Liu et al., 2018), it provides a robust foundation for the subsequent Cross-Scale Pseudo-Label Refinement, improving pseudo-label reliability for Test-Time Curriculum (TTC) adaptation.

Cross-Scale Pseudo-Label Refinement. Single-scale predictions, even when aggregated at the patch level, may still be unreliable due to calibration inconsistencies across scales or variations in generator artifacts. To improve the reliability of pseudo labels for Test-Time Curriculum adaptation, we aggregate scale-level predictions into a cross-scale mean probability $\bar{p}_j = \frac{1}{|\Omega|} \sum_{\ell \in \Omega} p_j^{\ell}$, where p_j^{ℓ} are the scale-level probabilities computed from multi-scale predictions. The cross-scale mean provides a consensus prediction that is less sensitive to unreliable responses from an individual scale.

At curriculum round r, the cross-scale mean $\bar{p}_j$ is used to compute prediction entropy and to assign the hard pseudo label, $\hat{y}_j = \mathbf{1}[\bar{p}_j > 0.5]$. Thus, reliable candidate sets and their predicted class assignments are determined by cross-scale consensus. Within each reliable class, hard-within-reliable selection instead ranks samples using the corresponding scale-specific probability p_j^{ℓ}, allowing each scale to focus on samples close to its own decision boundary. In the full cross-scale setting, Eqs. (2)–(4) use $\bar{p}_j$ for reliability estimation and pseudo-label assignment, whereas Eq. (5) is instantiated per scale using $H(p_j^{\ell})$ for hardness ranking. The selected samples are optimized using the shared hard pseudo label $\hat{y}_j$ in the BCE loss.

For final inference, we combine all scales using noisy-or fusion (Pearl, 2014):

$$s_j = 1 - \prod_{\ell \in \Omega} (1 - p_j^{\ell}). \tag{8}$$

Table 2: Comparison on the AIGI-Holmes Benchmark. The best result and the second-best result are marked in **bold** and underline, respectively.

Method	Janus	J-Pro-1B	J-Pro-7B	Show-o	LlamaGen	Infinity	VAR	PixArt-XL	SD3.5-L	FLUX	Avg.
UnivFD	70.6	68.3	59.4	57.9	75.2	56.7	66.7	53.9	74.0	51.6	63.4
NPR	30.1	29.8	29.8	79.6	79.5	79.8	79.3	78.7	79.4	79.1	64.5
FatFormer	48.8	48.6	48.8	73.2	84.3	53.9	90.1	67.1	71.0	54.6	64.0
SAFE	49.9	50.0	50.0	**99.7**	99.4	**99.9**	95.3	**99.7**	95.7	**99.4**	83.9
DRCT	71.1	71.7	70.9	70.0	70.8	71.3	61.2	70.8	69.6	67.2	69.5
AIDE	50.6	52.6	51.5	98.7	**99.7**	99.7	98.6	98.8	99.0	95.4	84.5
C2P-CLIP	72.4	86.9	86.0	99.0	98.8	99.1	94.0	98.9	98.6	96.1	93.0
AlignedForensics	93.9	92.3	92.4	81.3	94.0	94.9	89.3	95.1	84.9	92.0	91.0
DDA	98.5	**99.3**	98.7	94.8	99.3	98.9	80.4	99.4	97.0	97.2	96.3
Base	95.5	98.8	97.3	95.4	98.0	64.4	84.1	94.6	50.5	50.6	82.9
Ours	**99.2**	**99.3**	**99.1**	99.1	99.2	99.2	**99.0**	99.0	**99.2**	99.1	**99.1**

Because predictions from different scales are correlated, s_j is not interpreted as a calibrated posterior probability. We use it only as a fixed heuristic aggregation score that emphasizes images receiving a strong fake response at any scale. The noisy-or score s_j is used only for final inference and does not participate in pseudo-label generation, sample selection, or parameter updates.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate on five evaluation suites: three established benchmark collections (AIGCDetectionBenchmark (Wang et al., 2020b), AIGI-Holmes (Zhou et al., 2025) and GenImage (Zhu et al., 2023)), one in-the-wild collection, Chameleon (Yan et al., 2024), and our controlled modern-generator benchmark, AIGCGuard. These datasets comprise real images from multiple sources and fake images generated by a variety of generative models, including diffusion models, GANs, auto-regressive models, and other unseen generators. The datasets differ in format, content, resolution and curation procedures, providing a comprehensive evaluation. Detailed curation procedures for AIGCGuard are provided in the Appendix.

Unless otherwise stated, we evaluate mixed-domain offline batch TTA. For each benchmark, TTC receives the full unlabeled target split as a single mixed-generator batch and performs one adaptation process without access to generator identities or target labels. The adapted detector is then evaluated on the same target collection, making the primary protocol transductive. To examine whether improvements are caused mainly by fitting the adaptation images, Table 12 additionally adapts on a sampled subset and evaluates on the non-overlapping remainder after removing all adaptation samples. The difference between full-set and non-overlapping evaluation is at most 0.10 BAcc in this analysis. Generator-specific adaptation is reported only as a diagnostic protocol in the Appendix.

Implementation Details. TTC is model-agnostic and can be applied to different AIGC detectors. For fair comparison, we use the same pretrained backbone and training data as DDA, but do not apply its data-alignment strategy; this detector is used as the Base model for TTC adaptation. Unless otherwise stated, TTC is adapted once on the full unlabeled test split of each benchmark without generator identities. We use 3 curriculum rounds with confidence thresholds (0.8, 0.7, 0.6), converted to corresponding binary-entropy thresholds τ for reliable pseudo-label selection. The detector is updated using SGD with a learning rate of 5e-4 and momentum of 0.9, and 3 adaptation epochs. All benchmarks use the same hyperparameters without dataset-specific tuning. Following reliable and shift-aware TTA practices (Niu et al., 2023b; Lim et al., 2023; Park et al., 2023), we apply a label-free safety check to skip or roll back adaptation when prediction statistics indicate severe drift or confidence collapse. Multi-scale prediction uses three input scales: 70×70, 224×224, and 336×336 with 24, 6, and 1 patch per image, respectively. The LSE pooling temperature is set to $T_\ell = 1.5$. More details are provided in the Appendix.

Evaluation Metrics. Following common practice in AIGC detection, we report balanced accuracy—the average of real and fake image accuracies—as the evaluation metric to ensure fair comparison across potentially skewed datasets. This metric computes the unweighted arithmetic mean of the individual detection accuracies for authentic and generated images, providing a comprehensive

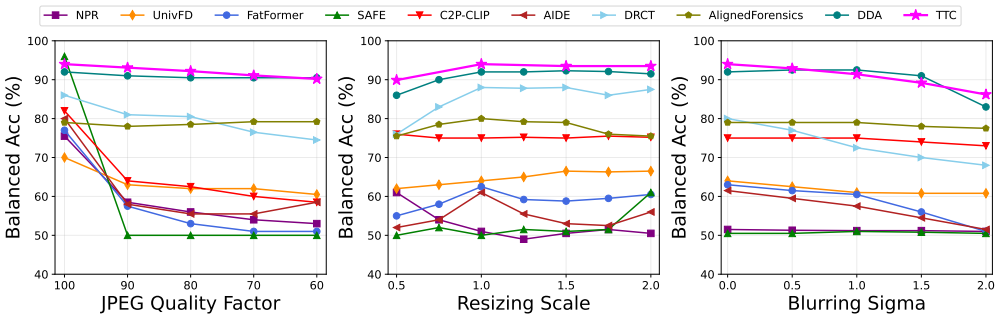

Figure 3: Robustness under varying JPEG compression, resizing, and Gaussian blur on GenImage. The experiment evaluates adaptation to changes in perturbation severity and mixture.

and unbiased measure of each model's fundamental discriminative capability. Results on the main benchmark evaluations (AIGCDetection, AIGI-Holmes, and Chameleon) are averaged over three independent adaptation runs to reduce randomness from sample selection and optimization.

Comparative Methods. We evaluate 9 advanced AIGC detection methods covering a range of representative approaches. The comparative baselines include data-alignment-based methods such as DDA (Chen et al., 2025b), DRCT (Chen et al., 2024a), and AlignedForensics (Rajan et al., 2024); advanced frequency-based detectors including NPR (Tan et al., 2024), SAFE (Li et al., 2025), and AIDE (Yan et al., 2024); and vision-language model-based approaches such as UnivFD (Ojha et al., 2023), Fatformer (Liu et al., 2024), and C2P-CLIP (Tan et al., 2025). Following DDA (Chen et al., 2025b), we evaluate all methods under an open-set evaluation setting. To isolate the contribution of the proposed adaptation strategy, we further compare TTC with representative general-purpose TTA methods. Additional analyses of detector generalization, adaptation efficiency, disjoint adaptation–evaluation splits, hyperparameter sensitivity, and run-to-run stability are provided in the Appendix.

4.2 Main Results

Comparison on AIGCDetection.

As shown in Table 1, TTC achieves the best average balanced accuracy of 93.0%, outperforming the Base model by 11.4 points and DDA by 5.2 points. TTC delivers especially large gains on challenging generators such as Midjourney (+32.2), StarGAN (+27.2), VQDM (+26.9), GLIDE (+22.5), and ADM (+19.6) over the Base model. It also surpasses DDA substantially on several difficult cases, including WFR (+39.8) and VQDM (+21.8). These results indicate that TTC effectively adapts the detector to generator shifts beyond the fixed training distribution.

Comparison on AIGI-Holmes.

Table 2 shows that TTC achieves the best average balanced accuracy of 99.1%, improving over the Base model by 16.2 points and over DDA by 2.8 points. The gains are particularly notable on challenging subsets where the Base detector degrades severely, including Infinity (64.4%→99.2%), SD3.5-L (50.5%→99.2%), and FLUX (50.6%→99.1%). TTC also achieves near-saturated performance across diverse generator families, including autoregressive, diffusion, and multimodal generators, with all subset accuracies reaching at least 99.0%. These results show that TTC remains effective under mixed-generator adaptation without relying on generator identities.

Comparison on Chameleon and AIGCGuard.

We further evaluate TTC on Chameleon, an in-the-wild collection, and AIGCGuard, a benchmark covering recent high-fidelity generators. As shown in Table 4, TTC improves over the strongest previous baseline DDA from 82.4% to 91.1% on Chameleon. On the more challenging AIGCGuard benchmark, where many existing detectors suffer substantial performance drops, TTC achieves the best result of 78.4%, outperforming both the Base model and all compared baselines. These results highlight the advantage of adapting to evolving real-world test distributions at test time.

Table 3: Ablation study on the AIGI-Holmes Benchmark. BRS, HWR, and CSPLR denote Balanced Reliable Selection, Hard-within-Reliable Selection, and Cross-Scale Pseudo-Label Refinement, respectively. Rows with "+" indicate cumulative additions, and "Full" additionally applies noisy-or fusion at inference.

	BAcc (%)	Δ
Base	82.9	–
+ BRS	92.9	+10.0
+ HWR	83.8	+0.9
+ CSPLR	97.0	+14.1
Full	99.1	+16.2
Full w/o HWR	95.0	+12.1

Table 4: Comparison on the Chameleon Benchmark and our AIGCGuard Benchmark.

Method	Chameleon	AIGCGuard
UnivFD	50.7	50.4
NPR	59.9	56.6
FatFormer	51.2	50.6
SAFE	59.2	49.7
DRCT	56.6	47.5
AIDE	63.1	54.4
C2P-CLIP	51.1	55.5
AlignedForensics	71.0	60.8
DDA	<u>82.4</u>	**77.2**
Base	62.4	60.6
Ours	**91.1**	**78.4**

4.3 Ablation Studies

We conduct ablation studies on the AIGI-Holmes benchmark to evaluate the contribution of each component in TTC. Table 3 reveals that effective test-time adaptation requires jointly controlling pseudo-label reliability and sample informativeness. Balanced reliable selection (BRS) first establishes a stable candidate set for adaptation, improving BAcc from 82.9% to 92.9% by preventing biased updates from over-represented pseudo-label classes. However, introducing HWR before cross-scale refinement decreases performance to 83.8%, indicating that HWR selection without reliable pseudo-label refinement is vulnerable to noisy supervision. After introducing CSPLR, the same hard-within-reliable strategy improves performance to 97.0%, demonstrating that reliable pseudo-label supervision is essential for effectively exploiting informative hard samples. Furthermore, removing HWR from the full TTC framework reduces performance from 99.1% to 95.0%, suggesting that HWR provides complementary gains when combined with CSPLR-refined pseudo-label supervision. Noisy-or fusion further improves the final inference performance to 99.1%.

4.4 Robustness to Perturbations

We evaluate robustness on GenImage under common image perturbations , including JPEG compression, resizing, and Gaussian blur. We first construct an unlabeled perturbed target split by randomly applying these perturbations to the GenImage test set, and perform TTC adaptation on this mixed split. We then evaluate the adapted detector under different perturbation types and severity levels. This experiment examines whether TTC can adapt to deployment batches containing diverse image-quality degradation, rather than evaluating general corruption robustness. As shown in Fig. 3, TTC maintains stronger performance than static detectors under different perturbation types and severity levels. It degrades gracefully under JPEG compression, from 94.0% BAcc at $Q = 100$ to 90.2% at $Q = 60$, and remains the strongest method under resizing. Gaussian blur is more challenging, reducing TTC to 86.2% at $\sigma = 2.0$, while DDA achieves 83.0%.

5 Conclusion

We present TTC, a model-agnostic test-time adaptation framework that investigates how AIGC detectors adapt to unseen-generator shifts using unlabeled target data. TTC improves adaptation stability by separating pseudo-label reliability from adaptation difficulty through balanced reliable selection and hard-within-reliable curriculum updates, while Cross-Scale Pseudo-Label Refinement improves the reliability of adaptation supervision. We also introduce AIGCGuard, a benchmark built from recent high-fidelity text-to-image models for evaluating adaptation under modern generator shifts. Extensive experiments show that TTC substantially improves detection performance across diverse benchmarks and unseen-generator settings. These results demonstrate the potential of test-time adaptation for improving AIGC detector robustness under unseen-generator shifts.

AI USE STATEMENT

In this work, we used generative AI tools for literature assistance and manuscript polishing. Specifically, GPT-5.5 was used to assist with literature search and summarization, and to improve the clarity, grammar, and readability of the manuscript. Generative AI tools were not used to formulate the research ideas, design the method, design or conduct experiments, or interpret experimental results. We also used generative AI tools to assist with implementing and reproducing baseline methods from prior work. The authors remain responsible for validating the experimental results and the final manuscript. AI-assisted content was reviewed and revised as needed, and cited references were checked against their original publications. The generative models used to construct the AIGCGuard benchmark, including 40 recent open-source and proprietary text-to-image models, were used solely as data generators for benchmark construction and evaluation. They were not used as research assistants or tools for developing the method, designing experiments, or analyzing results. The benchmark curation, evaluation protocol, and analyses were performed by the authors. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the full formulation of TTC and Cross-Scale Pseudo-Label Refinement in Sec. 3, including reliable pseudo-label selection, class balancing, hard-within-reliable sampling, parameter anchoring, multi-scale aggregation, and inference fusion. Sec. 4.1 specifies the evaluation protocol and the default implementation settings, including the mixed-domain offline batch adaptation protocol, curriculum thresholds, optimizer configuration, adaptation epochs, input scales, patch budgets, and LSE pooling temperature. The Appendix further documents the construction and preprocessing of AIGCGuard, mixed-domain and generator-specific adaptation protocols, comparisons with general-purpose test-time adaptation methods, evaluations with additional detector backbones and metrics, adaptation data efficiency and computational cost, non-overlapping adaptation–evaluation analysis, run-to-run stability, and hyperparameter sensitivity. Unless otherwise stated, the same TTC hyperparameters are used across benchmarks without dataset-specific tuning.

Upon acceptance, we will release the source code, model checkpoints for the reported experiments, and the AIGCGuard benchmark to support independent reproduction and further research.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Lies Hadjadj, Emilie Devijver, and Yury Maximov. Self-training: A survey. *Neurocomputing*, 616:128904, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, et al. Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer. *arXiv preprint arXiv:2605.11061*, 2026.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 6912–6920, 2021.

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

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024a.

Jiaxuan Chen, Jieteng Yao, and Li Niu. A single simple patch is all you need for ai-generated image detection. *arXiv preprint arXiv:2402.01123*, 2024b.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024c.

Liang Chen, Yong Zhang, Yibing Song, Jue Wang, and Lingqiao Liu. Ost: Improving generalization of deepfake detection via one-shot test-time training. *Advances in Neural Information Processing Systems*, 35:24597–24610, 2022.

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*, 2025b.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025c.

David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232, 1958.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

Samman Fatima, Sikandar Ali, and Hee-Cheol Kim. A comprehensive review on multiple instance learning. *Electronics*, 12(20):4323, 2023.

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.

Fengyi Fu, Mengqi Huang, Shaojin Wu, Yunsheng Jiang, Yufei Huo, Hao Li, Yinghang Song, Fei Ding, Jianzhu Guo, Qian He, Zheren Fu, Zhendong Mao, and Yongdong Zhang. Lance: Unified multimodal modeling by multi-task synergy, 2026. URL https://arxiv.org/abs/2605.18678.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or jpeg? revealing common biases in generated image detection datasets. In *European Conference on Computer Vision*, pp. 80–95. Springer, 2024.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18685–18694, 2025.

Eyal Gutflaish, Eliran Kachlon, Hezi Zisman, Tal Hacham, Nimrod Sarid, Alexander Visheratin, Saar Huberman, Gal Davidi, Guy Bukchin, Kfir Goldberg, et al. Generating an image from 1,000 words: Enhancing text-to-image with structured captions. *arXiv preprint arXiv:2511.06876*, 2025.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15733–15744, 2025.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34:2427–2440, 2021.

Jiachun Jin, Zetong Zhou, Xiao Yang, Hao Zhang, Pengfei Liu, Jun Zhu, and Zhijie Deng. Latentum: Unleashing the potential of interleaved cross-modal reasoning via a latent-space unified model. *arXiv preprint arXiv:2604.02097*, 2026.

Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm sigir forum*, volume 29, pp. 13–19. ACM New York, NY, USA, 1995.

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

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. *arXiv preprint arXiv:2302.05155*, 2023.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. *Advances in Neural Information Processing Systems*, 34:22968–22981, 2021.

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

Jiaxiang Liu, Zhida Feng, Pengyu Zou, Zhenyu Qian, Tianrui Zhu, Jun Xia, Yuehu Dong, Yanzheng Lin, Honglin Xiong, Anqi Chen, et al. Ernie-image technical report. *arXiv preprint arXiv:2605.25347*, 2026.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8759–8768, 2018.

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

Scott McCloskey and Michael Albright. Detecting gan-generated imagery using color cues. *arXiv preprint arXiv:1812.08247*, 2018.

Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.

Lakshmanan Nataraj, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, Amit K Roy-Chowdhury, and BS Manjunath. Detecting gan generated fake images using co-occurrence matrices. *arXiv preprint arXiv:1903.06836*, 2019.

Hong-Hanh Nguyen-Le, Van-Tuan Tran, Dinh-Thuc Nguyen, and Nhien-An Le-Khac. Think twice before adaptation: Improving adaptability of deepfake detection via online test-time adaptation. *arXiv preprint arXiv:2505.18787*, 2025.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pp. 16888–16905. PMLR, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023a.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023b.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24480–24489, 2023.

Sunghyun Park, Seunghan Yang, Jaegul Choo, and Sungrack Yun. Label shift adapter for test-time adaptation under covariate and label shifts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16421–16431, 2023.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

Jie Qin, Jiancheng Huang, Limeng Qiao, and Lin Ma. Star: Stacked autoregressive scheme for unified multimodal learning. *arXiv preprint arXiv:2512.13752*, 2025a.

Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and vision. *arXiv preprint arXiv:2508.05606*, 2025b.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. *arXiv preprint arXiv:2410.11835*, 2024.

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

Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. 2005.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33:11539–11551, 2020.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Xiangwei Shen, Zhimin Li, Zhantao Yang, Shiyi Zhang, Yingfang Zhang, Donghao Li, Chunyu Wang, Qinglin Lu, and Yansong Tang. Directly aligning the full diffusion trajectory with fine-grained human preference, 2025. URL https://arxiv.org/abs/2509.06942.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28130–28139, 2024.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025.

Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, et al. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing. *arXiv preprint arXiv:2603.09877*, 2026.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020a.

Haifeng Wang, Hua Wu, Tian Wu, Yu Sun, Jing Liu, Dianhai Yu, Yanjun Ma, Jingzhou He, Zhongjun He, Dou Hong, et al. Ernie 5.0 technical report. *arXiv preprint arXiv:2602.04705*, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695–8704, 2020b.

Yu Wang. Survey for detecting ai-generated content. *Advances in Engineering Technology Research*, 11(1):643–643, 2024.

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

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International conference on machine learning*, pp. 1954–1963. PMLR, 2015.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*, 2025c.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *Advances in Neural Information Processing Systems*, 38:138867–138907, 2026.

Zheng Yang, Ruoxin Chen, Zhiyuan Yan, Ke-Yue Zhang, Xinghe Fu, Shuang Wu, Xiujun Shu, Taiping Yao, Shouhong Ding, Zequn Qin, et al. All patches matter, more patches better: Enhance ai-generated image detection via panoptic patch learning. *arXiv preprint arXiv:2504.01396*, 2025.

Xiao Yu, Kejiang Chen, Kai Zeng, Han Fang, Zijin Yang, Xiuwei Shang, Yuang Qi, Weiming Zhang, and Nenghai Yu. Semgir: Semantic-guided image regeneration based method for ai-generated image detection and attribution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8480–8488, 2024.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 2023.

Yichen Zhang, Da Peng, Zonghao Guo, Zijian Zhang, Xuesong Yang, Tong Sun, Shichu Sun, Yidan Zhang, Yanghao Li, Haiyan Zhao, et al. Cheers: Decoupling patch details from semantic representations enables unified multimodal comprehension and generation. *arXiv preprint arXiv:2603.12793*, 2026.

Dian Zheng, Manyuan Zhang, Hongyu Li, Hongbo Liu, Kai Zou, Kaituo Feng, and Hongsheng Li. Uni-edit: Intelligent editing is a general task for unified model tuning. *arXiv preprint arXiv:2605.21487*, 2026.

Nan Zhong, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang. Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 3(6):1, 2023.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18746–18758, 2025.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in neural information processing systems*, 36:77771–77782, 2023.

Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 289–305, 2018.

A Appendix

A.1 Results on GenImage

We further evaluate TTC on the GenImage benchmark under the same mixed-domain adaptation protocol. As shown in Table 5, TTC achieves the best average balanced accuracy of 92.5%, improving over the Base detector by 12.4 points and outperforming all baselines. TTC brings particularly large gains on generators where the Base detector generalizes poorly, such as Midjourney, GLIDE, and VQDM, improving balanced accuracy from 66.5% to 98.6%, from 73.2% to 94.6%, and from 71.5% to 98.2%, respectively. These results show that TTC can effectively adapt to mixed unseen generator distributions and improve overall performance on GenImage using only unlabeled test data.

Table 5: Comparison on the GenImage Benchmark. We report balanced accuracy (%) for each generator subset and the average across subsets. The best result and the second-best result are marked in **bold** and underline, respectively.

Method	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg.
UnivFD	55.1	55.6	55.7	62.5	61.3	61.1	76.9	84.4	64.1
NPR	53.4	55.1	55.0	43.8	41.2	57.4	48.4	57.7	51.5
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
DRCT	82.4	88.3	88.2	76.9	86.1	87.9	85.4	**87.0**	84.7
AIDE	58.2	77.2	77.4	50.4	54.6	70.5	50.8	50.6	61.2
C2P-CLIP	56.6	77.5	76.9	71.6	73.5	79.4	73.7	85.9	74.4
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
Base	66.5	99.1	99.0	70.1	73.2	99.1	71.5	62.0	80.1
Ours	**98.6**	98.9	98.8	87.6	**94.6**	98.8	**98.2**	64.8	**92.5**

Table 6: Comparison on the AIGCDetect Benchmark under different adaptation protocols. Ours (MD-TTC) denotes the main mixed-domain protocol, and Ours (GS-TTC) denotes generator-specific adaptation. We report balanced accuracy (%) for each generator subset and the average across subsets.

Method	ADM	DALLE2	GLIDE	Midjourney	VQDM	BigGAN	CycleGAN	GauGAN	ProGAN	SDXL	SD14	SD15	StarGAN	StyleGAN	StyleGAN2	WFR	Wukong	Avg.
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
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
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
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
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
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
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
Base	70.1	88.5	73.2	66.5	71.5	90.7	60.0	99.0	87.4	97.2	99.1	99.0	49.4	80.2	75.9	81.0	99.1	81.6
Ours (MD-TTC)	**89.7**	99.2	**95.7**	**98.7**	**98.4**	93.7	69.4	97.4	97.2	99.3	98.7	98.8	76.6	90.9	87.1	91.9	98.8	**93.0**
Ours (GS-TTC)	84.0	**99.4**	92.6	96.7	98.1	98.9	72.4	**99.4**	**99.4**	**99.6**	99.3	99.5	62.0	86.4	85.1	81.9	99.5	91.4

A.2 Comparison with Generator-Specific TTC

The main experiments use mixed-domain TTC (MD-TTC), where adaptation is performed once on the full unlabeled test split without generator identities. Here we compare it with generator-specific TTC (GS-TTC), where TTC is adapted independently on each generator subset, to study whether separating target domains by generator is beneficial.

Tables 6 and 7 show that MD-TTC achieves higher average performance than GS-TTC on both benchmarks: 93.0% vs. 91.4% on AIGCDetect, and 99.1% vs. 98.9% on AIGI-Holmes. This suggests that mixed-domain adaptation can exploit artifacts shared across generators through reliable pseudo labels from the full target set. However, this shared supervision also introduces a trade-off: it may dilute generator-specific cues and cause negative transfer on some subsets. For example, on AIGCDetect, MD-TTC slightly degrades from the Base detector on GauGAN, SD14, SD15, and Wukong, whereas GS-TTC avoids these drops and improves all subsets in these tables over the Base detector. This contrast shows that MD-TTC favors shared cross-generator cues, while GS-TTC better preserves generator-dependent evidence.

AIGI-Holmes further illustrates this trade-off under a more diverse generator composition. Rather than uniformly benefiting all subsets, the mixed target set makes the shared-versus-specific cue trade-off more apparent: MD-TTC is slightly worse than GS-TTC on several subsets, but substantially better on difficult ones such as SD3.5-L and FLUX. This suggests that cross-generator supervision helps challenging subsets borrow useful cues from related generators, while generator-specific adaptation better preserves isolated domain-specific evidence.

Overall, these results support our main protocol: TTC does not require generator identities and can benefit from shared artifacts in mixed-generator test distributions, while generator-specific adaptation provides a useful diagnostic view of the trade-off between shared and generator-specific cues.

Table 7: Comparison on the AIGI-Holmes Benchmark under different adaptation protocols. Ours (MD-TTC) denotes the main mixed-domain protocol, and Ours (GS-TTC) denotes generator-specific adaptation. We report balanced accuracy (%) for each generator subset and the average across subsets.

Method	Janus	J-Pro-1B	J-Pro-7B	Show-o	LlamaGen	Infinity	VAR	PixArt-XL	SD3.5-L	FLUX	Avg.
UnivFD	70.6	68.3	59.4	57.9	75.2	56.7	66.7	53.9	74.0	51.6	63.4
NPR	30.1	29.8	29.8	79.6	79.5	79.8	79.3	78.7	79.4	79.1	64.5
FatFormer	48.8	48.6	48.8	73.2	84.3	53.9	90.1	67.1	71.0	54.6	64.0
SAFE	49.9	50.0	50.0	**99.7**	99.4	**99.9**	95.3	**99.7**	95.7	**99.4**	83.9
DRCT	71.1	71.7	70.9	70.0	70.8	71.3	61.2	70.8	69.6	67.2	69.5
AIDE	50.6	52.6	51.5	98.7	**99.7**	99.7	98.6	98.8	99.0	95.4	84.5
C2P-CLIP	72.4	86.9	86.0	99.0	98.8	99.1	94.0	98.9	98.6	96.1	93.0
AlignedForensics	93.9	92.3	92.4	81.3	94.0	94.9	89.3	95.1	84.9	92.0	91.0
DDA	98.5	99.3	98.7	94.8	99.3	98.9	80.4	99.4	97.0	97.2	96.3
Base	95.5	98.8	97.3	95.4	98.0	64.4	84.1	94.6	50.5	50.6	82.9
Ours (MD-TTC)	99.2	99.3	99.1	99.1	99.2	99.2	99.0	99.0	**99.2**	99.1	**99.1**
Ours (GS-TTC)	**99.4**	**99.5**	**99.4**	**99.7**	**99.9**	99.1	**99.6**	99.5	98.3	94.6	98.9

Table 8: Comparison with representative general test-time adaptation methods on AIGCDetect Benchmark. All methods start from the same pretrained Base detector using the DDA architecture and use the same unlabeled mixed-domain target data.

Method	BAcc (%)
Base	81.6
TENT	50.0
EATA	65.4
SAR	50.0
CoTTA	83.8
T3A	84.4
TTC (Ours)	**93.0**

Table 9: Stability of TTC across multiple (3) adaptation runs. We repeat TTC three times under the same mixed-domain protocol and report the mean and standard deviation of balanced accuracy (%) across runs.

Benchmark	BAcc (%)
AIGCDetect	93.0 ± 0.42
AIGI-Holmes	99.1 ± 0.15
Chameleon	91.1 ± 0.31

A.3 Comparison with General TTA Methods.

We further evaluate representative general-purpose test-time adaptation methods on our Base detector built with the DDA architecture. Recent AIGC detectors increasingly rely on large pretrained vision or vision-language backbones with lightweight task-specific adaptation modules, making the parameter-update strategies assumed by conventional classification-oriented TTA methods difficult to transfer directly across detector architectures. We therefore preserve the principal adaptation objective of each method while mapping its optimization scope to the trainable modules available in our detector. Specifically, TENT, SAR, and CoTTA optimize the DDA classification head, EATA optimizes the DDA LoRA modules together with the classification head, and T3A retains its gradient-free prototype adaptation. All methods start from the same pretrained Base detector and receive the same unlabeled mixed-domain target data.

As shown in Table 8, TENT and SAR obtain 50.0% BAcc, while EATA also degrades the Base detector. Further inspection shows that TENT and SAR converge to an all-real prediction regime, with 100% real recall and 0% fake recall, yielding exactly 50.0% BAcc. CoTTA and T3A provide moderate improvements, reaching 83.8% and 84.4% BAcc, respectively. In comparison, TTC achieves 93.0% BAcc, outperforming the strongest general TTA baseline by 8.6 percentage points. Overall, conventional classification-oriented TTA methods either rely on architecture-specific adaptation mechanisms that are difficult to transfer across AIGC detectors or provide limited gains when transferred. In contrast, TTC is model-agnostic in design and achieves substantially stronger performance under the same mixed-domain adaptation protocol.

A.4 Generalization to Different Detector Backbones

To demonstrate the model-agnostic nature of TTC, we train additional Base detectors initialized from C2P-CLIP and AIDE backbones using the same training data as DDA but without DDA's alignment

Table 10: Generalization of TTC to additional detector backbones, including C2P-CLIP and AIDE. For each backbone, we train a corresponding Base detector using the same training data as DDA but without DDA's alignment strategy, and then apply TTC adaptation. DDA is included as a strong reference baseline.

Method	AIGCDetect	AIGI-Holmes
DDA (reference)	87.8	96.3
AIDE Base	72.0	73.3
TTC w/ AIDE	89.0	99.5
C2P-CLIP Base	74.0	87.5
TTC w/ C2P-CLIP	**93.2**	**99.6**

Table 11: Complementary evaluation using AUC (%) on the AIGCDetect and AIGI-Holmes benchmarks. We report all generator subsets together with the average over subsets. Base denotes the pretrained detector before test-time adaptation, and DDA is included as the primary strong reference detector. The best result in each column is marked in bold.

(a) AIGCDetect

Method	ProGAN	StyleGAN	StyleGAN2	BigGAN	CycleGAN	StarGAN	GauGAN	SD1.4	SD1.5	Wukong	ADM	GLIDE	MJ	SDXL	VQDM	DALL-E2	WhichFace	Avg.
DDA	98.9171	94.0797	94.7913	93.9068	83.2125	72.3112	96.5938	99.9308	99.9069	99.9397	97.4290	97.6913	99.5282	99.9957	92.6721	99.2111	87.4526	94.5629
Base	96.3477	89.8806	87.5490	98.3533	87.9854	55.7227	99.9456	99.9571	99.9481	99.9610	91.8152	95.9726	94.4117	99.9200	95.1208	99.5294	**99.2002**	93.6247
TTC (Ours)	**99.8471**	**97.9316**	**95.6687**	**99.7163**	**89.3735**	**90.5816**	**99.9809**	**99.9997**	**99.9851**	**99.9990**	**97.9490**	**99.2079**	**99.9831**	**100.0000**	**99.8601**	**99.9553**	98.2876	**98.1369**

(b) AIGI-Holmes

Method	Janus	J-Pro-1B	J-Pro-7B	Show-o	LlamaGen	Infinity	VAR	PixArt-XL	SD3.5-L	FLUX	Avg.
DDA	99.8925	99.9708	99.8976	99.2781	99.9875	99.9486	95.2220	99.9903	99.7453	99.7035	99.3636
Base	99.7476	99.9626	99.9299	99.8424	99.9287	97.0698	97.6090	99.7889	86.5336	86.1627	96.6575
TTC (Ours)	**99.9990**	**99.9999**	**100.0000**	**99.9999**	**100.0000**	**100.0000**	**99.9542**	**99.9999**	**99.9993**	**99.9869**	**99.9939**

strategy. We then apply TTC adaptation to these detectors. We evaluate these variants on two representative generator-shift benchmarks, which cover diverse generator families including GAN-, diffusion-, and autoregressive-based models. As shown in Table 10, TTC consistently improves both corresponding Base detectors. With AIDE, TTC improves balanced accuracy from 72.0% to 89.0% on AIGCDetect and from 73.3% to 99.5% on AIGI-Holmes. With C2P-CLIP, TTC further increases performance from 74.0% to 93.2% and from 87.5% to 99.6%, respectively. Notably, both TTC variants outperform DDA, the strongest baseline in Table 1 and Table 2, on both benchmarks. These results support the model-agnostic design of TTC, showing that the same adaptation protocol can effectively improve detectors built on different backbones across diverse generator distributions.

A.5 Complementary Evaluation Metrics

To complement balanced accuracy, we additionally report AUC, which evaluates real/fake discrimination independently of a fixed decision threshold. Table 11 presents the results on all generator subsets of AIGCDetect and AIGI-Holmes. We include the pretrained Base detector to quantify the effect of test-time adaptation and DDA as the primary strong reference detector.

TTC improves the average AUC from 93.6247% to 98.1369% on AIGCDetect and from 96.6575% to 99.9939% on AIGI-Holmes. The improvements are highly consistent across generators: TTC improves over the Base detector on 16 of 17 AIGCDetect subsets and all 10 AIGI-Holmes subsets, while outperforming DDA on every evaluated subset. The gains are particularly pronounced on challenging generators, including StarGAN (55.7227% to 90.5816%), SD3.5-L (86.5336% to 99.9993%), and FLUX (86.1627% to 99.9869%). These results provide complementary evidence that TTC improves the discriminative capability of the detector under unseen-generator shifts.

Table 12: Adaptation data efficiency on AIGCDetect Benchmark. TTC is adapted using K unlabeled target images and evaluated on the full test set. "Excl. adapt." reports R3 performance after removing all adaptation samples from evaluation. Results are reported from a single adaptation run for each adaptation budget to compare the trade-off between adaptation data size and computational cost.

Adaptation size	Coverage	Base	R1	R2	R3	Excl. adapt.	Time (GPU h)
1K	0.7%	81.60	86.91	86.00	88.93	88.93	0.18
10K	6.6%	81.60	91.33	93.03	94.46	94.42	0.79
30K	19.7%	81.60	92.13	93.54	**94.62**	**94.52**	2.15
76.25K	50.0%	81.60	91.69	92.58	93.00	92.91	5.27
152.598K (Full)	100.0%	81.60	91.52	92.49	93.59	–	11.16

A.6 Adaptation Data Efficiency.

To examine whether TTC requires access to the full test set, we randomly sample K unlabeled images from the mixed AIGCDetect Benchmark, using a fixed seed and perform TTC adaptation only on this sampled subset. Each setting is independently initialized from the same Base detector, while performance is evaluated on both the full test set and the non-overlapping remainder excluding all adaptation samples. Runtime is measured as GPU hours on a single NVIDIA H800 card.

As shown in Table 12, performance is not monotonic in the number of adaptation images. Only 10K adaptation images improve BAcc from 81.60% to 94.46% in 0.79 GPU hours, while 30K images achieve the best result of 94.62% in 2.15 GPU hours, compared with 11.16 GPU hours for full-set adaptation. This suggests that a larger heterogeneous target batch may introduce redundant or noisy pseudo labels. We therefore interpret this experiment as evidence that TTC does not require access to the full test set, rather than as evidence that adaptation performance necessarily scales with target-set size. Evaluating only on samples not used for adaptation produces nearly identical results, with differences of at most 0.10 percentage points. This shows that the gains transfer beyond the adaptation samples and are not mainly caused by same-sample fitting. The 10K and 30K settings substantially reduce adaptation cost compared with full-set TTC.

A.7 Stability Across Multiple Runs

To evaluate the run-to-run stability of TTC, we repeat the test-time adaptation process three times under the same mixed-domain protocol. Table 9 reports the mean and standard deviation of balanced accuracy across runs. TTC exhibits small variance on all evaluated benchmarks, achieving 93.0 ± 0.42 on AIGCDetect, 99.1 ± 0.15 on AIGI-Holmes, and 91.1 ± 0.31 on Chameleon. These results indicate that TTC is stable across independent adaptation runs and is not sensitive to a particular random seed or stochastic sampling order.

A.8 Hyperparameter Sensitivity Analysis

Sensitivity to curriculum schedules. We conduct an extended five-round schedule study to analyze the effect of threshold relaxation. Here, "Ours" denotes the extended five-round schedule used only for this sensitivity analysis, rather than the default three-round configuration adopted in the main experiments.

As shown in Fig. 4, all schedules start from the Base detector and produce substantial gains after the first round. Among the evaluated schedules, Ours achieves the best final BAcc of 94.87%. Starting from a higher threshold yields comparable early performance but a lower final result, while slower decay leads to more conservative adaptation. The fixed and lower-start schedules perform substantially worse, showing that overly permissive early selection or the absence of progressive relaxation limits adaptation.

Specifically, the extended Ours uses $(0.8, 0.7, 0.6, 0.55, 0.5)$; Higher Start uses $(0.9, 0.8, 0.7, 0.6, 0.5)$; Slower Decay uses $(0.8, 0.75, 0.7, 0.65, 0.6)$; Fixed uses 0.7 throughout; and Lower Start uses $(0.7, 0.65, 0.6, 0.55, 0.5)$.

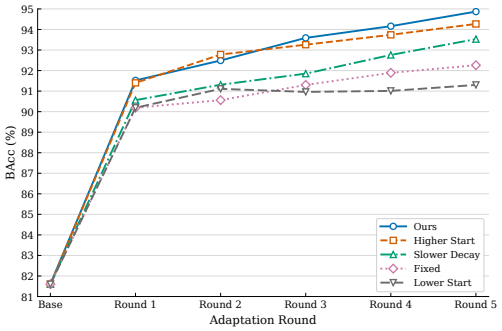

Figure 4: Sensitivity to five-round curriculum schedules on the AIGCDetect Benchmark. All schedules start from the same Base detector. Here, "Ours" denotes the extended five-round schedule used only for this sensitivity analysis, whereas the main experiments use the default three-round schedule (0.8, 0.7, 0.6). Among the evaluated five-round schedules, Ours achieves the highest final BAcc.

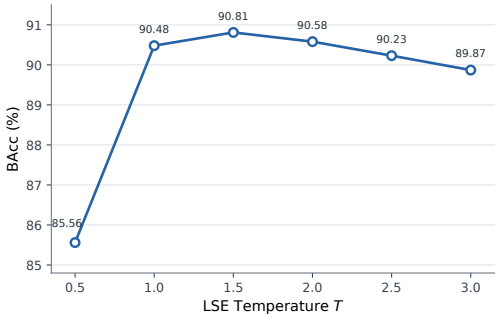

Figure 5: Sensitivity to the LSE temperature T on AIGCDetect. We vary the T while keeping all other settings fixed. BAcc is reported after the first adaptation round.

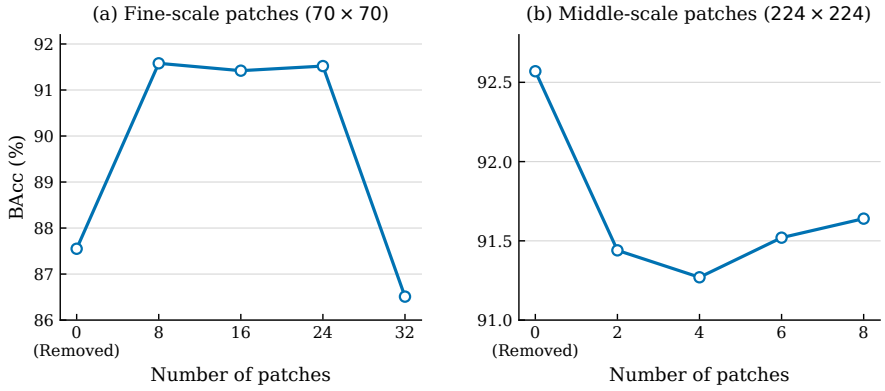

Figure 6: Sensitivity to patch numbers on AIGCDetect. We vary the number of fine-scale (70×70) and middle-scale (224×224) patches while keeping the remaining scales fixed. BAcc is reported after the first adaptation round.

Sensitivity to patch numbers. We vary the number of patches at one scale while fixing the other scales to study the sensitivity of TTC to the multi-scale patch budget.

As shown in Fig. 6(a), removing the fine-scale (70×70) patches substantially reduces the first-round performance, whereas using 8–24 patches yields stable results. Increasing the number to 32 leads to a clear degradation, suggesting that excessive local patches may introduce redundant or noisy supervision. In contrast, Fig. 6(b) shows only modest variation across different numbers of middle-scale (224×224) patches, indicating that TTC is relatively insensitive to the middle-scale patch budget. Overall, the method remains stable over a broad range of practical patch configurations without requiring precise tuning.

Sensitivity to LSE temperature. We further study the effect of the LSE temperature T used in multi-scale prediction aggregation. As shown in Fig. 5, performance remains stable for $T \in [1.0, 2.5]$, where BAcc varies by only 0.58 percentage points. The setting $T = 2.0$ achieves 90.58% BAcc, only 0.23 points below the best result in this range. In contrast, an overly low temperature of $T = 0.5$ causes a clear degradation, while larger temperatures lead to only a mild decline. These results

Table 13: Sensitivity to patch selection strategies on AIGCDetect. All variants use the same patch size and patch budget. BAcc is reported after the first adaptation round.

Patch selection strategy	BAcc (%)
Random	**91.16**
Score-only	90.07
Score + diversity	90.43
Full	90.55

Table 14: Effect of pseudo-label quality on AIGCDetect. Panel (a) reports TTC's three adaptation rounds; panel (b) fixes the adaptation samples and controls pseudo-label accuracy using ground-truth labels.

Setting	Pseudo-label accuracy (%)	BAcc (%)
(a) Our		
Round 1	92.6	91.0
Round 2	91.7	91.9
Round 3	92.6	93.0
(b) Oracle-controlled pseudo labels		
100% accuracy	100	97.1
95% accuracy	95	95.1
90% accuracy	90	95.1
85% accuracy	85	89.9
80% accuracy	80	79.3

indicate that the method is robust to the choice of T over a broad practical range and does not rely on precise temperature tuning.

Sensitivity to patch selection strategies. We compare four patch selection strategies under the same patch budget.

Random uniformly samples crops from the image. *Score-only* ranks sliding-window candidates using a weighted combination of high-frequency energy, gradient magnitude, and local texture variation, and selects the top-ranked patches. *Score + diversity* additionally applies greedy spatial suppression to reduce redundant overlapping crops. *Full* selects most patches using the score-and-diversity strategy and fills the remaining slots with randomly sampled crops to maintain broader image coverage.

As shown in Table 13, all four strategies achieve comparable Round-1 performance, with BAcc ranging from 90.07% to 91.16%. Random selection performs slightly better in this setting, while the full strategy remains competitive with the guided variants. This result indicates that TTC is not strongly dependent on a particular patch selection strategy.

A.9 Effect of Pseudo-label Quality.

We analyze both naturally selected and synthetically corrupted pseudo labels to quantify when TTC remains stable and when erroneous supervision causes degradation. Confidence alone cannot guarantee correctness under severe generator shift. TTC uses cross-scale consensus to improve pseudo-label reliability, while class-balanced selection, parameter anchoring, and a label-free safety check mitigate the impact of biased or erroneous supervision. We run the default three-round TTC and use ground-truth labels only for post-hoc evaluation of the selected pseudo labels. As shown in Table 14, the selected pseudo labels maintain accuracies of 92.63%, 91.69%, and 92.55% across the three rounds, and consistently improve performance across rounds.

We then conduct an oracle-controlled analysis to isolate the effect of pseudo-label accuracy. We fix the adaptation samples, synthetically construct pseudo labels at varying accuracy levels using ground-truth labels, and perform one adaptation round. TTC achieves 97.11%, 95.10%, 95.10%, and 79.34% BAcc with pseudo-label accuracies of 100%, 95%, 90%, and 80%, respectively. These results reveal a clear reliability regime: once pseudo-label accuracy reaches 90%, further improvements in label accuracy have only a limited effect on the adaptation outcome. The selected pseudo labels of our TTC remain within this effective range across all three rounds. Performance degrades sharply below this regime, motivating the safety check and rollback mechanism.

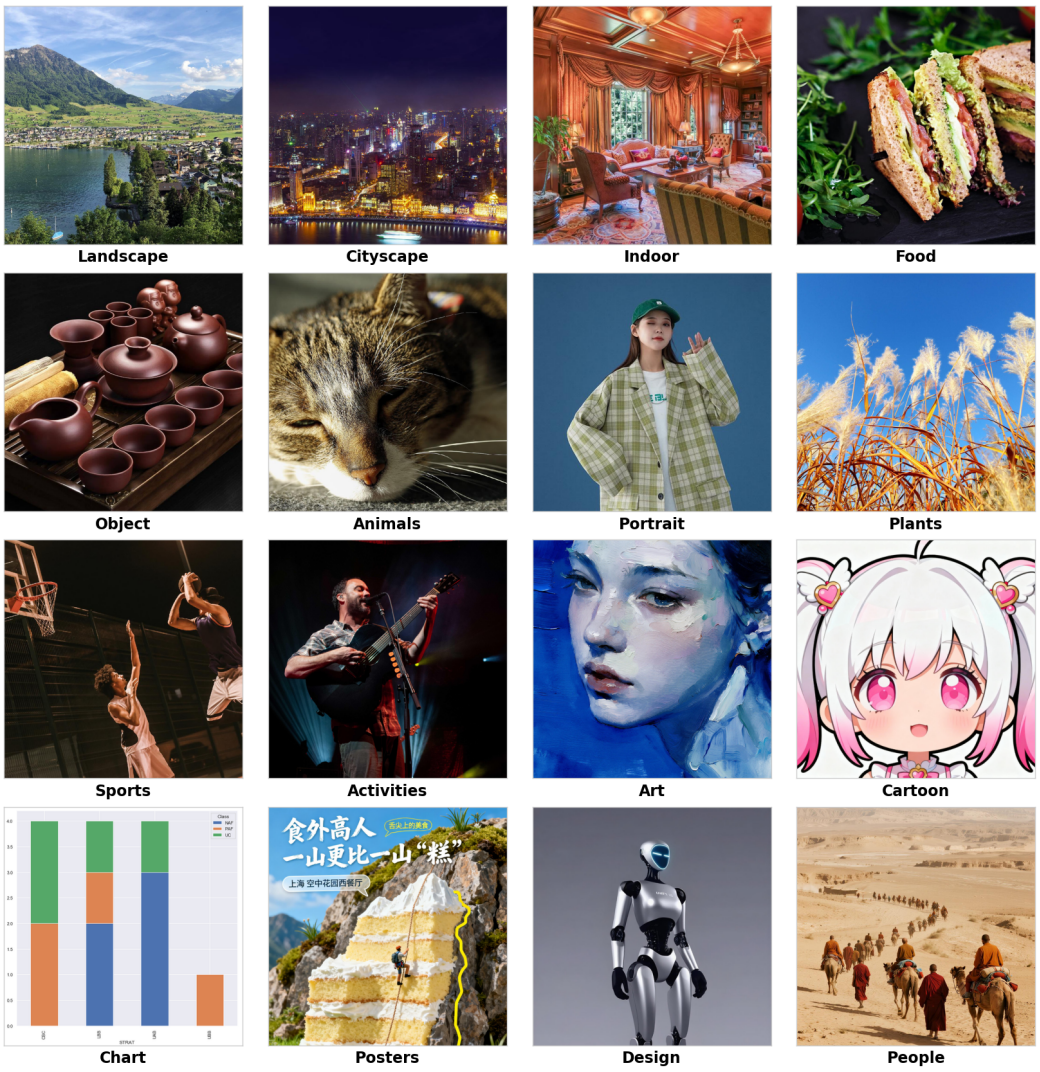

Figure 7: Representative real images from 16 semantic categories in AIGCGuard.

A.10 Additional Implementation Details

Base Detector Training. Although TTC itself is applied only at test time, the Base detector is trained with lightweight image-level perturbations to improve robustness before adaptation. Specifically, JPEG compression, Gaussian blur, and resizing are randomly applied during source training. These perturbations are not used during TTC adaptation on clean benchmarks, except when explicitly evaluating robustness to perturbations. This separation keeps the reported clean benchmark results faithful to the original test distribution, while providing a more robust initialization for adaptation under natural distribution shifts.

Patch Selection Details. For each local scale, candidate crops are extracted using a sliding window with a stride equal to half the patch size. Each candidate is scored using high-frequency energy, average gradient magnitude, and local intensity standard deviation, with weights 0.4, 0.3, and 0.3, respectively. Candidates are traversed in descending score order, and a crop is retained only if its Chebyshev distance from all previously selected crops is at least half the patch size. We set the random sampling ratio to 0.2, and randomly sampled crops also fill any remaining slots when spatial suppression returns fewer candidates than required.

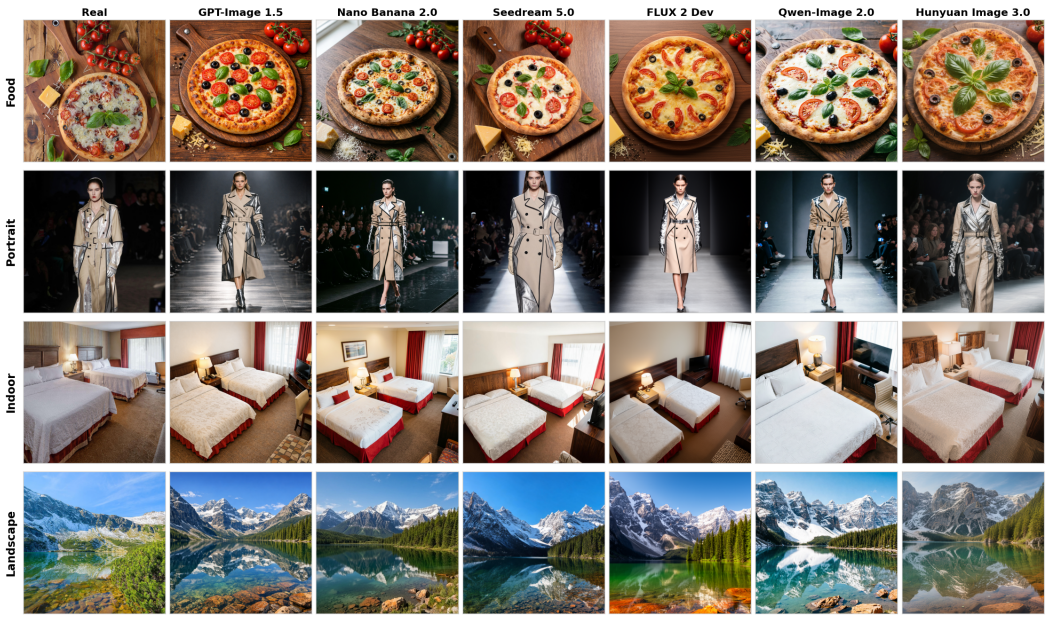

Figure 8: Real images and samples generated by recent text-to-image models from the same prompts.

Label-Free Safety Check Since ground-truth labels are unavailable during test-time adaptation, unreliable pseudo labels may cause the detector to drift or collapse. We therefore use a label-free safety check that compares each adapted curriculum round against the frozen Base detector using two prediction-only statistics: the predicted fake prior and the average prediction certainty. A round is considered unsafe in two cases. First, a large prior shift together with reduced certainty suggests that adaptation may be losing discriminability. Second, when the Base detector is already highly confident, a prior shift together with increased certainty may indicate overconfident pseudo-label confirmation. Because curriculum rounds are sequentially dependent, we use early stopping rather than treating rounds independently: once the first unsafe round is detected, TTC selects the previous round for evaluation, with the frozen Base detector serving as the fallback when no adapted round is safe. This prevents over-updating when pseudo labels are unreliable, while preserving adaptation on shifted target distributions. The safety check is applied uniformly across benchmarks using fixed thresholds and prediction-only statistics. The parameter-anchoring coefficient in Eq. (6) is fixed to $\lambda = 0.001$ across all experiments.

A.11 Details of Benchmark Curation

Principles. Our benchmark is designed to provide a high-quality, semantically diverse, and visually realistic dataset for evaluating AIGC detection methods. The construction emphasizes semantic coverage, intra-category diversity, and inclusion of advanced generative models, reflecting the current state of the most powerful text-to-image generative models.

Real Image Sampling. Real images are sampled from tens of billions of industrial Pre-Training and Mid-Training text-to-image datasets. To ensure semantic diversity, we assign images to 22 categories using the Qwen3-VL (Bai et al., 2025) vision-language model, covering a broad range of real-world content such as objects, landscapes, animals, portraits, and charts. Within each category, we perform contrastive-learning-based K-means clustering (Chen et al., 2020; Caron et al., 2018) to capture fine-grained semantic distinctions. From these clusters, a representative subset is selected via Facility Location selection (Wei et al., 2015; Sener & Savarese, 2017), resulting in a representative subset of high-quality real images.

Selected images undergo quality control including image quality assessment, manual AIGC detection and NSFW filtering, and secondary caption re-annotation to ensure accurate textual descriptions. Ad-

ditional manual inspection is performed to remove any remaining visually ambiguous or semantically inconsistent samples. After filtering and validation, the dataset contains 3,100 real images.

Image Generation. Each real image is paired with counterparts generated by 40 state-of-the-art text-to-image models with generated image resolutions randomly varying from 512 pixels up to 4K, covering both open-source and proprietary systems: GPT-Image-2, GPT-Image-1.5, Nano banana, Seedream-4.0 (Seedream et al., 2025), Seedream-5.0, Flux2-Dev, Flux2-Dev-Turbo, Qwen-Image-2.0, Qwen-Image-max (Wu et al., 2025a), Z-Image (Cai et al., 2025), Z-Image-Turbo, Ernie-Image (Liu et al., 2026), Ernie-Image-Turbo, Longcat-Image (Team et al., 2025), Hunyuan-Image-3.0 (Cao et al., 2025), Cosmos3 (Agarwal et al., 2026), GLM-Image, Omnigen2 (Wu et al., 2025b), HiDream-o1 (Cai et al., 2026), Ideogram-4-fp8, openUni (Wu et al., 2025c), SRPO (Shen et al., 2025), FIBO (Gutflaish et al., 2025), Infinity (Han et al., 2025), Kolors (Team, 2024), StableDiffusion3.5-Large, PixArt-Sigma (Chen et al., 2024c), Janus-Pro-7B (Chen et al., 2025c), BLIP3o (Chen et al., 2025a), MMaDA (Yang et al., 2026), BAGEL (Deng et al., 2025), UniCoT (Qin et al., 2025b), Uni-Edit (Zheng et al., 2026), Emu3.5 (Cui et al., 2025), Internvl-U (Tian et al., 2026), Cheers (Zhang et al., 2026), Lance (Fu et al., 2026), STAR (Qin et al., 2025a), LatentUM (Jin et al., 2026), and SenseNova-U1-8B-MoT (Diao et al., 2026). Each model receives the same prompt independently, producing 40 generated images per real image, resulting in a total of 127,100 images. The dataset is balanced across generators, equalizing each generator's contribution to the aggregate metric and providing a robust and high-quality benchmark for AIGC detection evaluation.

To reduce low-level shortcut cues arising from differences in image encoding and resolution, mild JPEG compression and resizing are incorporated into the construction of the AIGCGuard evaluation set.

A.12 Visualization of Our Benchmark

To qualitatively illustrate the diversity and difficulty of AIGCGuard, we provide two visualizations. Figure 7 shows representative real images from 16 semantic categories, covering natural scenes, city scenes, indoor environments, food, objects, animals, portraits, plants, sports, activities, people, artwork, cartoon-style images, posters, design content, and charts. These examples demonstrate the broad semantic coverage of the real-image split.

Figure 8 further visualizes real images and generated samples from representative recent text-to-image models. Each row corresponds to the same prompt, with the first column showing the real image and the remaining columns showing generated counterparts. The generated samples preserve the high-level semantics of the real images while exhibiting model-specific variations in style, composition, and details. Their high visual fidelity makes AIGCGuard a challenging benchmark for open-set AIGC image detection.

A.13 Limitations and Future Work

TTC has several limitations. First, our primary protocol is offline and transductive: an unlabeled target batch is available during adaptation, and the current experiments do not establish online or streaming adaptation to future generator distributions. Second, TTC relies on the Base detector producing sufficiently reliable predictions for both classes. Systematic high-confidence errors, extreme target imbalance, or one-class prediction collapse may prevent useful adaptation, in which case TTC skips the corresponding round or falls back to the last safe checkpoint. Third, multi-scale prediction and iterative optimization introduce additional computation. Moreover, the noisy-or output is a heuristic decision score rather than a calibrated posterior probability. Finally, AIGCGuard complements in-the-wild evaluation with controlled, prompt-matched coverage of recent generators. Its generator-balanced construction is designed to isolate cross-generator detection performance rather than reproduce the naturally occurring real/fake prior or generator frequencies of a deployment stream.

Future work will focus on more efficient and reliable test-time adaptation for AIGC detection, including reducing adaptation overhead, improving label-free reliability estimation, and designing safer update rules. Another important direction is streaming adaptation, where the detector must adapt to continuously changing distributions under limited memory and computation.