DECUE: LEARNING GENERATOR-INVARIANT FORENSIC REPRESENTATIONS VIA FEATURE DECOUPLING FOR AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid growth of image generative models, from GANs to diffusion models, has enabled the widespread creation of photorealistic synthetic images, posing serious threats to digital media authenticity and information integrity. A fundamental bottleneck of AI-generated image detection is the poor cross-generator generalization of existing detectors. We attribute this generalization failure to the inherent entanglement of deep representations. Detectors tend to overfit to superficial semantic content features rather than learning intrinsic generation-invariant forensic traces that transfer across architectures. To address this issue, we present DeCue, a novel feature decoupling framework that explicitly disentangles entangled representations into forensic cue features encoding generation-specific discriminative artifacts, and content features that capture image semantic information. The proposed DeCue achieves robust feature decoupling through hard fake sample construction and cross-reconstruction constraint. The hard fake samples are strategically synthesized to eliminate trivial content-level disparities between real and fake images, compelling the model to learn generalizable forensic cues instead of relying on content shortcuts. The cross-reconstruction constraint further provides explicit supervisory signals to guaranty effective feature decoupling. Extensive evaluations across 8 benchmarks, including 5 standard academic benchmarks and 3 challenging in-the-wild benchmarks, show that our DeCue achieves a mean accuracy of **99.23%**, substantially outperforming the state-of-the-art detector by **11.29%**, demonstrating its strong robustness and superior cross-generator generalization capability in practical forensic scenarios.

1 INTRODUCTION

Recent advances in image generative models, ranging from GANs to diffusion models, have enabled the widespread creation of photorealistic synthetic content (Dhariwal & Nichol, 2021; Nichol et al., 2022). While these models have greatly facilitated creative production, they have also raised serious concerns about digital media authenticity and the spread of misinformation, posing tangible threats to public trust and information security (Ojha et al., 2023).

AI-Generated Image(AIGI) detection has grown increasingly challenging as generative artifacts become subtler and more diverse across architectures (Bie et al., 2025). Existing detectors perform well on standard benchmarks such as GenImage (Zhu et al., 2023), but suffer drastic performance degradation in real-world scenarios (Li et al., 2026a). A representative case is the Chameleon benchmark (Yan et al., 2025), whose high-fidelity samples can pass human visual Turing tests, yet no existing detector exceeds 85% accuracy on it (Chen et al., 2025). This generalization gap has become the core bottleneck hindering practical deployment (Li et al., 2025a). With novel generative architectures emerging rapidly, training dedicated detectors for each generator is neither practical nor scalable, making cross-generator generalization a critical and urgent research problem (Zhang et al., 2025a).

Early forensic detectors rely on handcrafted generative artifacts such as frequency-domain statistics (Frank et al., 2020), local pixel patterns (Tan et al., 2024), and bit-plane splicing (Wang et al., 2025). While achieving satisfactory accuracy on specific generators, they degrade sharply on unseen

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.

architectures, as handcrafted cues are highly sensitive to non-causal factors like resolution and compression. To break this limitation, pre-trained Vision Foundation Models (VFMs) have become the mainstream, forming two dominant paradigms as depicted in Figure 1: methods leveraging high-level feature inconsistencies from models like CLIP (Ojha et al., 2023), and reconstruction-based methods exploiting low-level residuals from generative VFMs like diffusion models (Chen et al., 2024). Recent works such as SynerDetect (Li et al., 2026b) further fuse the two paradigms for complementary enhancement.

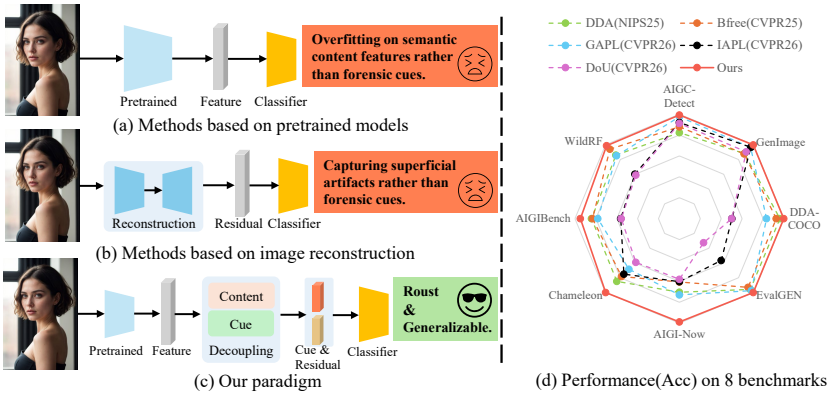


Figure 1: Comparison of different AIGI detection paradigms and performance. (a) Methods based on pre-trained VFMs are prone to capturing content features rather than intrinsic forensic cues. (b) Reconstruction-based detection methods suffer from poor generalization across unseen generative architectures. (c) Our proposed feature decoupling framework achieves strong robustness and cross-generator generalization capability. (d) Quantitative results on 8 benchmarks verify the superior performance of our method over state-of-the-art approaches.

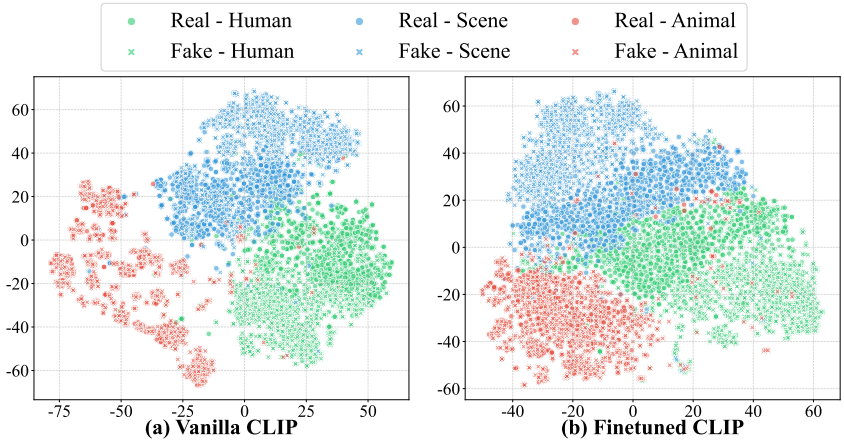


Figure 2: t-SNE visualization of CLIP features on the Chameleon. Colors indicate semantic categories, and shapes denote image authenticity. (a) Features extracted by the vanilla CLIP vision encoder. (b) Features extracted by the fine-tuned CLIP vision encoder.

Despite these advances, a fundamental limitation persists across all methods: the exploited features remain inherently entangled, with semantic content and forensic cues mixed together, which fundamentally caps the generalization upper bound. As visualized in Figure 2, CLIP features (Radford et al., 2021) cluster predominantly by semantic content rather than authenticity, regardless of fine-tuning. Consequently, even with powerful generalizable backbones, detectors operating on entangled features cannot deliver satisfactory cross-generator performance. This motivates our core research question: **How can we effectively disentangle authentic forensic cue features from entangled deep representations for robust cross-generator detection?**

To address this challenge, we propose DeCue, a feature decoupling framework that explicitly disentangles deep representations into two distinct components: forensic cue features capturing generation-specific discriminative artifacts and content features encoding image semantic information. Our core design philosophy is formulating AIGI detection as a single-class classification task. Since fake image distributions are highly diverse and unstable with evolving generators, real image distributions governed by physical imaging laws remain relatively stable (Healey & Kondepudy, 1994; Lukas et al., 2006), making a real-centric decision boundary more generalizable (Liu et al., 2025a). Under this paradigm, the proposed DeCue achieves reliable feature decoupling via two synergistic strategies: (1) a hard fake sample construction scheme that eliminates trivial content-level discrepancies, forcing the model to learn generalizable forensic cues instead of content shortcuts; (2) a cross-reconstruction constraint that provides explicit supervision for thorough feature decoupling.

As illustrated in Figure 3, our DeCue comprises a training data construction pipeline and a feature decoupling learning module. For data construction, we reconstruct real images via a pre-trained diffusion model with latent perturbations, then blend real and fake samples following the DDA paradigm (Chen et al., 2025) to generate hard fake samples that preserve diverse generative artifacts while sharing highly similar semantic content with real images. For the feature decoupling pipeline, a pre-trained VFM first extracts image features, which are then mapped into separate forensic cue and content subspaces via a feature decoupling module. The final forensic representation combines forensic cue features with reconstruction residuals for authenticity judgment, with cross-reconstruction constraint supervising the entire training process.

Our main contributions are summarized as follows:

- We propose DeCue, a novel feature decoupling framework for AIGI detection that explicitly disentangles forensic cue features from semantic content features in deep representations. It tackles the cross-generator generalization bottleneck from the feature decoupling perspective, offering a new paradigm for mitigating content overfitting in detection models.
- We design an integrated training pipeline combining hard fake sample construction and cross-reconstruction constraint. The two components work synergistically to enable effective feature decoupling under the single-class classification paradigm.
- Extensive experiments on 8 benchmarks (5 standard academic and 3 real-world in-the-wild) demonstrate that the proposed DeCue achieves state-of-the-art performance with **99.23%** mean accuracy, outperforming existing methods by **11.29%** and verifying its superior cross-generator generalization and practical forensic value.

2 Related Work

2.1 AI-Generated Image Detection

AI-Generated Image (AIGI) detection has evolved from artifact-driven approaches to vision foundation model (VFM)-based methods, while cross-generator generalization remains a central challenge (Yan et al., 2025). Early detectors exploit handcrafted traces, including GAN fingerprints (Yu et al., 2019), frequency-domain statistics (Frank et al., 2020), upsampling artifacts (Tan et al., 2024), frequency inconsistencies (Li et al., 2025c), and low-bit plane patterns (Wang et al., 2025). Although effective on known generators, these cues are sensitive to compression, resizing, and architectural changes, resulting in poor generalization. VFM-based methods mainly follow three paradigms. Semantic contrastive methods, including UnivFD (Ojha et al., 2023), FatFormer (Liu et al., 2024), C2P-CLIP (Tan et al., 2025), and GAPL (Qin et al., 2026), leverage high-level representations from contrastive vision models but tend to capture semantic content rather than intrinsic forensic cues. Reconstruction-based methods, such as FakeInversion (Cazenavette et al., 2024), DRCT (Chen et al., 2024), SemGIR (Yu et al., 2024), and DDA (Chen et al., 2025), detect generation traces through reconstruction residuals, yet their effectiveness often depends on the reconstruction architecture. Hybrid methods, including CO-SPY (Cheng et al., 2025) and SynerDetect (Li et al., 2026b), fuse semantic and residual features for complementary enhancement. However, these methods still operate on entangled representations in which semantic content and forensic cues are mixed, leaving detectors vulnerable to content shortcuts and limiting cross-generator generalization.

2.2 Feature Decoupling

Feature decoupling separates independent variation factors in representation space and has been applied to forensic tasks such as face forgery detection (Li et al., 2025b; Zhang et al., 2025b; Cui et al., 2025). Several recent works have introduced feature decoupling ideas into AIGI detection. DoU (He et al., 2026) separates forgery and superfluous cues through a cue information bottleneck; GSD (Fu et al., 2025) and ODP-Net (Wang et al., 2026) employ QR and orthogonal decomposition, respectively; and OmniAID (Guo et al., 2025) constructs semantic and artifact experts through data partitioning. Nevertheless, these methods mainly rely on algebraic decomposition or loosely constrained network designs, without explicit supervision to ensure effective disentanglement. Moreover, they do not jointly perform sample-level content alignment and feature-level decoupling. In contrast, the proposed DeCue combines hard fake sample construction with cross-reconstruction constraint under a real-centric single-class classification paradigm, explicitly separating forensic cues from semantic content to improve cross-generator generalization.

3 Methodology

We present DeCue, a feature decoupling framework for cross-generator generalizable AI-generated image detection. Its core goal is to explicitly disentangle entangled deep representations into two complementary components: discriminative forensic cue features encoding generation-specific artifacts and forensic-agnostic content features capturing authenticity-invariant semantic information. This is achieved via two synergistic designs: sample-level hard fake sample construction for content alignment, and feature-level cross-reconstruction constraint for feature decoupling supervision.

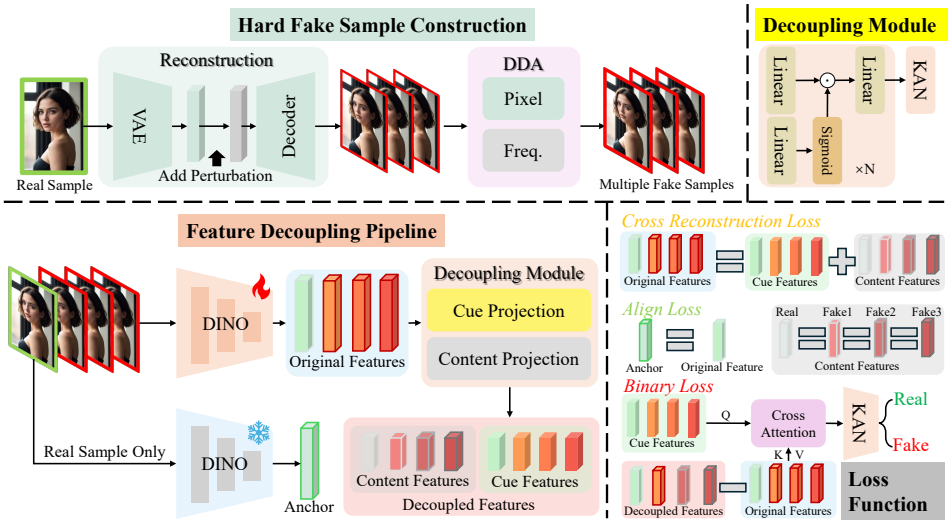


Figure 3: Overall architecture of the proposed DeCue framework, consisting of a hard fake sample construction pipeline for content-aligned training data and a feature decoupling pipeline with cross-reconstruction constraint.

As shown in Figure 3, the proposed DeCue first extracts high-dimensional representations with a pre-trained vision foundation model. A feature decoupling module then projects entangled features into separate forensic cue and content subspaces. Finally, disentangled forensic cues are fused with feature-level reconstruction residuals and fed into a Kolmogorov–Arnold Network (KAN) (Liu et al., 2025b) based classifier for authenticity judgment. We elaborate each component below.

3.1 Hard Fake Sample Construction

The ever-shifting distribution of fake images caused by emerging generative architectures leads to severe out-of-distribution degradation for detectors trained on limited generators. In contrast, real image distributions governed by physical imaging principles remain relatively stable and generalizable. Motivated by this asymmetry, we formulate AIGI detection as a real-centric single-class

classification task, where the model characterizes the intrinsic distribution of real images and treats deviations as fake.

To prevent the detector from relying on content-level shortcuts, we construct hard negative samples that share nearly identical semantic content with real images but carry distinct generative artifacts via a two-stage pipeline. First, we reconstruct real images through a pre-trained variational autoencoder with controlled latent perturbations. Given a real image $\boldsymbol{x}_r$, we encode it into latent space $\boldsymbol{x} = E(\boldsymbol{x}_r)$ and inject zero-mean Gaussian noise:

$$\boldsymbol{z}' = \boldsymbol{z} + \boldsymbol{\delta}, \quad \boldsymbol{\delta} \sim \mathcal{N}(0, \epsilon^2 \boldsymbol{I}), \tag{1}$$

where ϵ controls the magnitude of perturbation to preserve semantic fidelity while introducing generation artifacts. The perturbed latent code is decoded to yield an initial reconstructed fake sample $\boldsymbol{x}_f = D(\boldsymbol{z}')$.

Second, following the dual data alignment paradigm (Chen et al., 2025), we blend real and reconstructed fake samples in both pixel and frequency domains, further eliminating trivial content discrepancies while retaining diverse generative artifacts to form final hard fake samples.

3.2 Feature Decoupling Pipeline

With content-aligned training samples, we introduce the feature decoupling pipeline to separate forensic cues and semantic content from entangled deep representations.

Backbone and Feature Extraction. We adopt DINO (Siméoni et al., 2025) as the feature backbone for its strong representation capacity and robustness to common distortions (compression, resizing), which suits real-world detection scenarios.

To avoid semantic drift during fine-tuning, we employ a dual-branch strategy: a learnable DINO branch $f(\cdot)$ and a frozen anchor branch $g(\cdot)$. For a real sample $\boldsymbol{x}_r$ and its i-th hard fake sample $\boldsymbol{x}_f^i$, we extract deep features $\boldsymbol{h}_r = f(\boldsymbol{x}_r)$, $\boldsymbol{h}_f^i = f(\boldsymbol{x}_f^i)$, and anchor feature $\boldsymbol{h}_a = g(\boldsymbol{x}_r)$. An L1 anchor alignment loss constrains the learner to preserve the generalizable semantic structure of the frozen backbone:

$$\mathcal{L}_{\text{anchor}} = \left\| \boldsymbol{h}_r - \boldsymbol{h}_a \right\|_1, \tag{2}$$

where $\| \cdot \|_1$ denotes the L^1 norm.

Decoupling Module. The entangled feature $\boldsymbol{h}$ is fed into a decoupling module with a gating mechanism to adaptively select the dimensions of the features for different variation factors. A shared linear layer first produces a transformed feature $\boldsymbol{h}'$ and gating weight vector $\boldsymbol{w}$:

$$\boldsymbol{h}', \boldsymbol{w} = \text{Linear}(\boldsymbol{h}). \tag{3}$$

The Sigmoid-normalized gating weights are element-wise multiplied with $\boldsymbol{h}'$ to activate artifact-relevant dimensions, then projected via a linear layer and a KAN projection layer to yield forensic cue features:

$$\boldsymbol{f}^{\text{cue}} = \text{KAN}\left(\text{Linear}\left(\boldsymbol{h}' \odot \sigma(\boldsymbol{w})\right)\right), \tag{4}$$

where $\sigma(\cdot)$ is the Sigmoid function and $\odot$ denotes element-wise multiplication. The non-linear KAN mapping enhances the model's ability to learn complex disentangled representations. The semantic content features $\boldsymbol{f}^{\text{content}}$ are obtained via an independent projection branch with the same gating structure but separate parameters, optimized to capture only authenticity-invariant semantic information.

3.3 Cross-Reconstruction Constraint

Dual-branch projection alone does not guarantee meaningful feature decoupling: without explicit supervision, the two branches may still learn entangled representations, reducing the framework to a trivial feature transformation. We thus propose a cross-reconstruction constraint paired with a content alignment constraint to supervise the feature decoupling process.

Cross-Reconstruction Loss. For a group of samples sharing identical semantic content (one real + multiple hard fakes), content features should be interchangeable across samples, while forensic cue features are sample-specific. Hence, combining the cue feature of any sample i with the content feature of any sample j should reconstruct the original feature of sample i. For a group with n samples, the cross-reconstruction loss is defined as:

$$\mathcal{L}_{\text{cross}} = \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| \boldsymbol{h}_i - \left(\boldsymbol{f}_i^{\text{cue}} + \boldsymbol{f}_j^{\text{content}} \right) \right\|_1, \tag{5}$$

where $\boldsymbol{h}_i$, $\boldsymbol{f}_i^{\text{cue}}$, $\boldsymbol{f}_i^{\text{content}}$ denote the original feature, forensic cue feature, and content feature of the i-th sample. This loss enforces content invariance across semantically identical samples and disentangles artifacts from semantics in the feature space.

Content Alignment Loss. To further strengthen feature decoupling, we introduce a content alignment loss that directly constrains the content features of all hard fake samples to match that of the corresponding real sample:

$$\mathcal{L}_{\text{content}} = \frac{1}{n-1} \sum_{i=1}^{n-1} \left\| \boldsymbol{f}_r^{\text{content}} - \boldsymbol{f}_{f_i}^{\text{content}} \right\|_1, \tag{6}$$

where $\boldsymbol{f}_r^{\text{content}}$ and $\boldsymbol{f}_{f_i}^{\text{content}}$ denote content features of the real sample and the i-th fake sample. Combined with the anchor loss, the overall alignment loss is:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{anchor}} + \lambda \mathcal{L}_{\text{content}}, \tag{7}$$

where λ is a balancing hyperparameter. Together, the two constraints provide multi-level supervision for principled feature decoupling.

3.4 Decision Boundary Construction

Using forensic cue features alone may lack robustness under complex real-world post-processing. We thus introduce feature reconstruction residuals as a complementary forensic signal, and design a cross-attention fusion module with a KAN-based classifier for robust decision boundaries.

Reconstruction Residual and Feature Fusion. The feature reconstruction residual is computed as the difference between the reconstructed feature (sum of cue and content features) and the original entangled feature:

$$\boldsymbol{f}^{\text{residual}} = \left(\boldsymbol{f}^{\text{cue}} + \boldsymbol{f}^{\text{content}} \right) - \boldsymbol{h}. \tag{8}$$

This residual captures discriminative reconstruction error that complements explicit forensic cues. We employ a Cross-Attention (CA) mechanism to fuse the two signals, using cue features as queries and residuals as keys and values, so that cues guide attention to the most discriminative residual regions:

$$\boldsymbol{f}^{\text{fused}} = \text{CA}\left(\boldsymbol{f}^{\text{cue}}, \boldsymbol{f}^{\text{residual}}, \boldsymbol{f}^{\text{residual}} \right). \tag{9}$$

KAN Classifier and Total Loss. The fused representation is fed into a KAN classifier to predict authenticity logits $p = \text{KAN}(\boldsymbol{f}^{\text{fused}})$. Compared with traditional linear classifiers, KAN provides stronger non-linear fitting capacity for a more sophisticated real-centric decision boundary. We use standard Binary Cross-Entropy (BCE) loss for supervision:

$$\mathcal{L}_{\text{bce}} = -\left[y \log \hat{y} + (1-y) \log(1-\hat{y}) \right], \tag{10}$$

where $y \in \{0, 1\}$ is the ground-truth label and $\hat{y} = \sigma(p)$ is the predicted probability. The total loss is the weighted sum of three components:

$$\mathcal{L} = \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{align}} + \alpha \mathcal{L}_{\text{cross}}, \tag{11}$$

where α is a balancing hyperparameter. Joint optimization of these losses enables effective feature decoupling for superior detection performance and cross-generator generalization.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. To thoroughly evaluate the performance and cross-generator generalization of the proposed DeCue, we conduct experiments on 8 benchmarks, covering 5 standard academic benchmarks and 3 challenging in-the-wild benchmarks: (1) **AIGCDetection** (Zhong et al., 2024): synthetic images from 17 GAN and diffusion models; (2) **GenImage** (Zhu et al., 2023): large-scale benchmark with 8 semantic categories; (3) **DDA-COCO** (Chen et al., 2025): reconstructed images from 5 diffusion models; (4) **EvalGEN** (Chen et al., 2025): covers 5 advanced autoregressive and diffusion generators; (5) **AIGI-Now** (Chen et al., 2026): 9 latest commercial generative models; (6) **Chameleon** (Yan et al., 2025): high-resolution (720P–4K) real-world images curated via human perceptual Turing tests; (7) **AIGIBench** (Li et al., 2025d): real-world samples collected from AI creation communities, we adopt two subsets, Community AI and SocialRF, for our in-the-wild evaluation; (8) **WildRF** (Cavia et al., 2024): images crawled from mainstream social media platforms. To rigorously test generalization in real-world scenarios, we do not use any training split of the above benchmarks. Instead, we use images from MSCOCO (Lin et al., 2015) for training and evaluate exclusively on the test sets of all 8 benchmarks.

Baseline Methods. We compare DeCue with a wide range of state-of-the-art detectors: UnivFD (Ojha et al., 2023), NPR (Tan et al., 2024), FatFormer (Liu et al., 2024), DRCT (Chen et al., 2024), SAFE (Li et al., 2025c), C2P-CLIP (Tan et al., 2025), AIDE (Yan et al., 2025), AlignedForensics (Sundara Rajan et al., 2025), CO-SPY (Cheng et al., 2025), Bfree (Guillaro et al., 2025), DDA (Chen et al., 2025), GAPL (Qin et al., 2026), IAPL (Li et al., 2026c), LTD (Yang et al., 2026), DoU (He et al., 2026) and PM (Zhou et al., 2026). For fair comparison, we use official publicly available code and pre-trained weights whenever possible; otherwise, we re-implement the methods strictly following their original training protocols.

Implementation Details. We adopt DINOv3 (Siméoni et al., 2025) as the feature backbone, fine-tuned via LoRA (Hu et al., 2021) to balance representation capacity and training efficiency. We use Stable Diffusion v2.1 (Rombach et al., 2022) to construct fake samples. More details of the experimental configuration of DeCue are provided in the Appendix A.

4.2 QUANTITATIVE ANALYSIS

Table 1 summarizes the accuracy on all 8 benchmarks. The proposed DeCue achieves a mean accuracy of 99.23%, outperforming the second-best method by a large margin (99.23% vs. 87.94%). Artifact-driven methods (e.g., NPR, SAFE) suffer severe performance drops across benchmarks due to their fragile handcrafted cues. VFM-based semantic methods (e.g., C2P-CLIP, CO-SPY, GAPL) exhibit better generalization but still degrade sharply on challenging benchmarks such as Chameleon. DDA benefits from dual data alignment to partially suppress content shortcuts, yet it still fails to exceed 85% accuracy on Chameleon. Notably, the performance lead of DeCue is more pronounced on in-the-wild benchmarks than on standard academic datasets, reflecting its stronger practical applicability in open scenarios. Even on benchmarks dominated by unseen generative paradigms such as autoregressive models, DeCue maintains consistently high accuracy, verifying that the learned forensic cues are architecture-agnostic. These results collectively demonstrate that without explicit feature decoupling, detectors inevitably overfit to superficial semantic content, which limits their cross-generator generalization. Detailed results are provided in Appendix B.

4.3 QUALITATIVE ANALYSIS

To visually verify the effect of feature decoupling, we visualize the distributions of the features at different stages of the pipeline in Figure 4, including original entangled features, forensic cue features, reconstruction residuals, content features, and final fused forensic features. Original features are heavily entangled and cluster primarily by semantic content rather than authenticity. The content features are grouped by semantic category under the constraints of cross-reconstruction and content alignment, consistent with our design goal. Remarkably, the forensic cue features form two clearly separated clusters for real and fake samples without explicit classification supervision, directly val-

idating the effectiveness of our feature decoupling framework. The clear clustering pattern further confirms that our framework effectively eliminates semantic interference and purifies forensic representations at the feature level. Although reconstruction residuals remain partially entangled, the final fused forensic features exhibit a distinct decision boundary, providing a solid discriminative basis for authenticity judgment. This also intuitively explains the performance gain of the fusion strategy: cue features and residuals capture complementary discriminative information from different perspectives. We provide additional qualitative results in Appendix B.

Table 1: Performance comparison with state-of-the-art methods on 8 benchmarks (5 standard academic and 3 challenging in-the-wild) in terms of classification accuracy (%).

Method	Ref	AIGC-Detection	Gen-Image	DDA-COCO	EvalGEN	AIGI-Now	Cham-eleon	AIGI-Bench	WildRF	Mean
UnivFD	CVPR23	77.40	69.96	52.74	41.48	52.26	57.40	52.99	55.14	58.11
NPR	ECCV24	82.65	77.22	51.27	49.56	55.00	59.80	55.01	65.05	62.75
FatFormer	CVPR24	84.91	76.90	52.09	48.95	51.61	58.03	54.43	59.53	62.01
DRCT	ICML24	83.21	89.49	63.01	88.66	73.88	53.86	52.66	51.98	75.13
SAFE	KDD25	93.24	95.35	50.34	32.00	64.86	59.11	56.39	58.91	65.65
C2P-CLIP	AAAI25	86.22	81.62	52.44	57.86	50.24	58.46	53.64	59.33	64.28
AIDE	ICLR25	93.86	99.65	50.64	44.30	67.24	65.75	64.30	65.83	70.17
AlignedForensics	ICLR25	67.82	79.91	92.65	72.27	57.42	71.99	72.56	81.56	74.51
CO-SPY	CVPR25	73.37	78.00	50.66	94.65	75.32	69.06	66.27	75.75	73.91
Bfree	CVPR25	87.94	87.68	92.67	93.00	60.63	78.31	83.19	94.04	84.96
DDA	NeurIPS25	82.32	88.06	96.16	95.81	70.44	84.81	84.10	86.05	86.33
GAPL	CVPR26	97.50	96.37	83.38	97.42	72.96	68.45	78.13	85.17	87.94
IAPL	CVPR26	92.02	96.68	50.14	56.64	60.36	75.09	56.20	60.33	69.54
LTD	CVPR26	91.27	91.61	50.04	37.27	53.85	57.63	55.14	53.83	63.16
DoU	CVPR26	91.19	90.35	50.05	32.63	58.04	59.04	55.64	58.95	63.35
PM	AAAI26	92.23	97.55	49.98	31.13	65.54	59.49	57.58	60.64	65.97
Ours	-	**99.23**	**99.89**	**99.83**	**99.93**	**98.90**	**99.94**	**94.80**	**98.67**	**99.23**

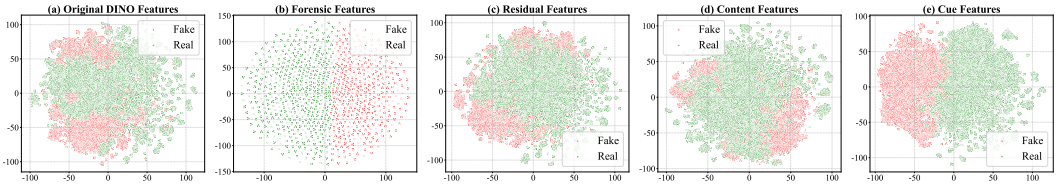


Figure 4: t-SNE visualization of features on Chameleon. Colors indicate image authenticity. (a) Original DINOv3 features. (b) Final discriminative features. (c) Reconstruction residual features. (d) Disentangled cue features. (e) Disentangled content features.

4.4 Ablation Study

To verify the contribution of each component, we conduct ablation studies on AIGI-Now and AIGIBench, with mean accuracy (mAcc) reported in Table 2 and Table 3.

4.4.1 Effectiveness of the Feature Decoupling Framework

As shown in Table 2, removing the KAN projection leads to a notable performance drop, confirming that non-linear transformation facilitates both feature decoupling and discriminative boundary learning. Specifically, the KAN head with learnable activation functions enables the model to capture intricate nonlinear dependencies across feature dimensions, which is crucial for disentangling deeply entangled representations. The vanilla DINOv3 baseline yields 95.86% mAcc, indicating that while the backbone already captures partial forensic traces, our feature decoupling framework effectively extracts more discriminative cues from entangled representations. Removing cross-reconstruction loss also causes evident performance degradation, verifying its critical role in enforcing effective feature decoupling.

Table 2: Ablation study of core modules on AIGI-Now.

Module Variant	mAcc (%)
Vanilla DINOv3	95.86
w/o cross-reconstruction loss	97.63
w/o KAN projection	95.49
Ours (full)	**98.20**

Table 3: Ablation study of fusion strategies on AIGIBench.

Fusion Strategy	mAcc (%)
Element-wise addition	89.72
Concatenation	88.10
w/o residual features	87.85
Ours (cross-attention)	**94.10**

4.4.2 Analysis of Fusion Strategies

Table 3 shows that the fusion strategy strongly affects the final performance. Simple operations (addition, concatenation) cause an approximately 5% accuracy drop, compared to using cue features alone. We attribute this to noise in decoupled features: even with successful feature decoupling, forensic cues inevitably carry noise from feature extraction. The cross-attention mechanism models the correlation between the two complementary signals and suppresses noise interference, thus achieving superior detection performance.

4.5 Robustness Evaluation

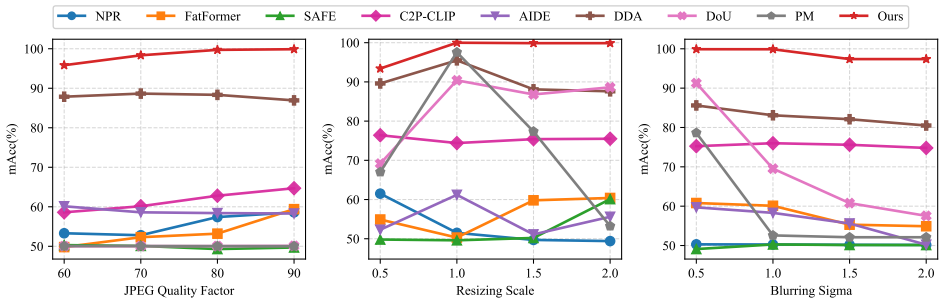


Figure 5: Robustness evaluation under common image degradations on GenImage.

We further evaluate the robustness of the proposed DeCue against common image degradations following the protocol of (Chen et al., 2025), including JPEG compression, resizing, and Gaussian blur, all tested on GenImage. Most baseline methods, especially artifact-driven methods, experience dramatic accuracy drops under severe distortions, particularly high-rate JPEG compression and aggressive resizing. As shown in Figure 5, DeCue consistently achieves the best performance under all degradation settings, verifying that it maintains stable discriminative representations under real-world perturbations. This strong robustness is attributed to the disentangled forensic cues, which focus on generator-invariant generative artifacts rather than fragile superficial features vulnerable to post-processing.

5 Conclusion

We present DeCue, a feature decoupling framework that explicitly disentangles forensic cue features from entangled deep representations for robust, cross-generator AI-generated image detection. Through hard fake sample construction and cross-reconstruction constraint, the proposed DeCue separates generation-invariant forensic artifacts from semantic content features, effectively mitigating content overfitting. Extensive experiments on 8 diverse benchmarks demonstrate state-of-the-art performance across both standard academic and challenging real-world scenarios. We hope this work provides a new perspective for building generalizable visual forensic systems.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with paper writing, including language polishing, grammar correction, and improving the clarity and readability of the manuscript. We did not use generative AI tools for developing research ideas, designing the methodology, conducting experiments, implementing code, analyzing data, generating results, creating figures, or performing the literature review and citation process. All AI-assisted text was carefully reviewed, revised and verified by the authors to ensure its accuracy, originality, and consistency with research findings. We take full responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on AI-generated image detection and is intended to support digital media authenticity, misinformation mitigation and responsible use of generative techniques. The study does not involve human subjects, personally identifiable information, or private data. All experiments are conducted using publicly available datasets in accordance with their intended research purposes and applicable licenses. The proposed method is defensive in nature and does not generate, manipulate, or facilitate the creation of harmful content. We are not aware of any concerns related to privacy, discrimination, fairness, legal compliance, conflicts of interest, or sponsorship. The experiments and results are reported faithfully and transparently. Therefore, we believe that this work does not raise specific issues under the Code of Ethics.

REPRODUCIBILITY STATEMENT

We have taken several measures to ensure the reproducibility of this work. All training and evaluation data are obtained from publicly available datasets, which are described and cited in the main paper. Detailed model architectures, data-processing procedures, hyperparameter settings, training configurations, and evaluation protocols are provided in the Appendix A. In addition, source code together with the trained model weights is publicly available in the anonymous repository at `https://anonymous.4open.science/r/DeCue1012-D060`. The repository also provides complete environment specifications, configuration files, data-preparation instructions, and step-by-step scripts for reproducing the reported results.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant U25A20440, the Zhejiang Provincial Natural Science Foundation of China under Grant LY24F020004, the Key Research and Development Program of Zhejiang Province under Grant 2023C01039, the National College Students Innovation and Entrepreneurship Training Program under Grant 202510353027, and the Zhejiang University Students Science and Technology Innovation Activity Plan and New Talent Plan under Grant 2026R408B073.

REFERENCES

Fengxiang Bie, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Golnari, David A. Clifton, Yuxiong He, Dacheng Tao, and Shuaiwen Leon Song. Renaissance: A survey into ai text-to-image generation in the era of large model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2212–2231, 2025.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world, 2024.

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

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 7621–7639, 2024.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes AI-generated image detector easier generalizable. In *Proceedings of the thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Ruoxin Chen, Jiahui Gao, Kaiqing Lin, Keyue Zhang, Yandan Zhao, Isabel Guan, Taiping Yao, and Shouhong Ding. Alignggemini: Generalizable ai-generated image detection through task-model alignment, 2026.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13455–13465, 2025.

Jialei Cui, Jianwei Du, Yanzhe Li, Lei Gao, Hui Jiang, and Chenfu Bao. Hamlet-ffd: Hierarchical adaptive multi-modal learning embeddings transformation for face forgery detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1327–1336, 2025.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, pp. 8780–8794, 2021.

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

Xinghe Fu, Zhiyuan Yan, Zheng Yang, Taiping Yao, Yandan Zhao, Shouhong Ding, and Xi Li. PiD: Generalized AI-generated images detection with pixelwise decomposition residuals. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 17894–17908, 2025.

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

Yuncheng Guo, Junyan Ye, Chenjue Zhang, Hengrui Kang, Haohuan Fu, Conghui He, and Weijia Li. Omniaid: Decoupling semantic and artifacts for universal ai-generated image detection in the wild, 2025.

Qinghui He, Haifeng Zhang, Qiao Qin, Bo Liu, Xiuli Bi, and Bin Xiao. Diversity over uniformity: Rethinking representation in generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40407–40417, 2026.

G.E. Healey and R. Kondepudy. Radiometric ccd camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(3):267–276, 1994.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Chunxiao Li, Xiaoxiao Wang, Meiling Li, Boming Miao, Peng Sun, Yunjian Zhang, Xiangyang Ji, and Yao Zhu. Bridging the gap between ideal and real-world evaluation: Benchmarking ai-generated image detection in challenging scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20379–20389, 2025a.

Kai Li, Wenqi Ren, Jianshu Li, Wei Wang, and Xiaochun Cao. Critical forgetting-based multi-scale disentanglement for deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 424–432, 2025b.

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, 2025c.

Shuaibo Li, Zhaohu Xing, Hongqiu Wang, Pengfei Hao, Xingyu Li, Zekai Liu, and Lei Zhu. Toward medical deepfake detection: A comprehensive dataset and novel method. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention*, pp. 626–637, 2026a.

Shuaibo Li, Yijun Yang, Zhaohu Xing, Hongqiu Wang, Pengfei Hao, Xingyu Li, Zekai Liu, Qing Zhang, and Lei Zhu. Synerdetect: Hierarchical synergistic learning for generalizable ai-generated image detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026b.

Yiheng Li, Zichang Tan, Guoqing Xu, Zhen Lei, Xu Zhou, and Yang Yang. Towards generalizable ai-generated image detection via image-adaptive prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21262–21272, 2026c.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? *Advances in Neural Information Processing Systems*, 38, 2025d.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

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.

Ruiqi Liu, Yi Han, Zhengbo Zhang, Liwei Yao, Zhiyuan Yan, Jialiang Shen, ZhiJin Chen, Boyi Sun, Lubin Weng, Jing Dong, Yan Wang, and Shu Wu. Beyond artifacts: Real-centric envelope modeling for reliable ai-generated image detection, 2025a.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Hou, and Max Tegmark. Kan: Kolmogorov–arnold networks. In *Proceedings of the International Conference on Learning Representations*, volume 2025, pp. 70367–70413, 2025b.

J. Lukas, J. Fridrich, and M. Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214, 2006.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 16784–16804, 2022.

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.

Ziheng Qin, Yuheng Ji, Renshuai Tao, Yuxuan Tian, Yuyang Liu, Yipu Wang, and Xiaolong Zheng. Scaling up ai-generated image detection with generator-aware prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43008–43017, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

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

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. In *Proceedings of the International Conference on Learning Representations*, volume 2025, pp. 63304–63322, 2025.

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, June 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.

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

Zhiyuan Wang, Yanxiang Chen, Pengcheng Zhao, Yunfeng Diao, and Xin Liao. Decoupling semantics and fingerprints: A universal representation for ai-generated image detection, 2026.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *Proceedings of the International Conference on Learning Representations*, volume 2025, pp. 70702–70720, 2025.

Yawen Yang, Feng Li, Shuqi Kong, Yunfeng Diao, Xinjian Gao, Zenglin Shi, and Meng Wang. Layer consistency matters: Elegant latent transition discrepancy for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38111–38121, 2026.

Ning Yu, Larry Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7555–7565, 2019.

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.

Laixin Zhang, Shuaibo Li, Wei Ma, and Hongbin Zha. Truemoe: Dual-routing mixture of discriminative experts for synthetic image detection, 2025a.

Tianshuo Zhang, Li Gao, Siran Peng, Xiangyu Zhu, and Zhen Lei. Devfd : Developmental face forgery detection by learning shared and orthogonal lora subspaces. *Advances in Neural Information Processing Systems*, 38:4103–4128, 2025b.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection, 2024.

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

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

A Implementation Details

Training Details. We adopt DINOv3-ViT-L (Siméoni et al., 2025) as the feature backbone, fine-tuned via LoRA (Hu et al., 2021) to balance representational capacity and training efficiency. For hard fake sample construction, we employ Stable Diffusion v2.1 (Rombach et al., 2022) with perturbation magnitudes $\epsilon \in \{0.01, 0.05, 0.1\}$, generating 3 hard fake samples per real image. The entire framework is trained end-to-end using the Adam optimizer with an initial learning rate of 5×10^{-5}, decayed to 2×10^{-5} via a cosine schedule. Loss weights are set to $\lambda = 0.5$ and $\alpha = 0.5$. Training proceeds for 1 epoch with a batch size of 64 on a single NVIDIA RTX 3090 GPU. Common data augmentations—JPEG compression, random horizontal flip, Gaussian noise, and Gaussian blur—are applied to improve robustness.

Evaluation Protocol. Following standard practice (Ojha et al., 2023; Chen et al., 2025; Zhou et al., 2026), we adopt classification accuracy (Acc) and average precision (AP) as the evaluation metrics. Owing to space constraints, the main paper reports benchmark-level results and overall average performance primarily in terms of Acc. Detailed results for individual subsets of each benchmark, evaluated using both Acc and AP, are provided in the Appendix B.

B Additional Experimental Results

Test Benchmark Details. We provide detailed descriptions of the 8 test benchmarks used in the main paper. All evaluations of our method strictly follow the zero-shot cross-benchmark protocol: models are trained exclusively on real images from MSCOCO (Lin et al., 2015) training set and tested directly on the held-out test splits of each benchmark without dataset-specific fine-tuning.

- **AIGCDetection (Zhong et al., 2024)**. AIGCDetection covers both GAN and diffusion paradigms, including synthetic images from 8 GAN models (BigGAN, CycleGAN, GauGAN, ProGAN, StarGAN, StyleGAN, StyleGAN2, WFIR) and 9 diffusion models (ADM, DALLE2, GLIDE, Midjourney, VQDM, SD XL, SD v1.4, SD v1.5, Wukong). We use all 17 generative sources as our test set.
- **GenImage (Zhu et al., 2023)**. GenImage is a large-scale benchmark comprising synthetic images from 8 advanced generative architectures, including 7 diffusion models (Midjourney, SD v1.4, SD v1.5, ADM, GLIDE, Wukong, VQDM) and BigGAN. We evaluate on all 8 model subsets.
- **DDA-COCO (Chen et al., 2025)**. Built upon the MSCOCO (Lin et al., 2015) validation set, DDA-COCO mainly contains reconstructed images from 5 advanced diffusion models: SD XL, SD EMA, SD MSE, SD v2.1, and SD v3.5.
- **EvalGEN (Chen et al., 2025)**. EvalGEN consists of 5 state-of-the-art autoregressive and diffusion models (Flux, GoT, Infinity, OmiGen, NOVA), designed to evaluate cross-architecture generalization with high visual fidelity and detection complexity.
- **AIGI-Now (Chen et al., 2026)**. AIGI-Now covers 9 latest commercial generative models (NanoBanana, GPT-4o, Jimeng, Keling, Minimax, Flux Pro, Flux Krea, Flux Dev, Flux Kontext), presenting prominent challenges to existing forensic methods due to their high-fidelity output and diverse generation mechanisms.
- **Chameleon (Yan et al., 2025)**. Chameleon is a highly challenging real-world benchmark with high-resolution images (720P–4K) curated via human perceptual Turing tests, serving as a touchstone for the practical generalization of detectors.
- **AIGIBench (Li et al., 2025d)**. AIGIBench includes 11 GAN models, 12 diffusion models, and 2 in-the-wild subsets (Community AI, SocialRF) collected from AI creation communities and social platforms. We adopt the two real-world subsets for our in-the-wild evaluation.
- **WildRF (Cavia et al., 2024)**. WildRF collects synthetic images from three mainstream social media platforms (Facebook, Reddit, Twitter) with real-world post-processing such as compression and resizing, simulating practical detection scenarios. We use all three platform subsets as our test set.

Full Benchmark Results. We report the complete per-benchmark comparison results. All experimental protocols strictly follow the settings described in the main paper. Tables 4 to 18 detail the accuracy and AP across all standard academic and in-the-wild benchmarks, covering every competing method and test subset. Furthermore, we provide additional visualization results in Figure 6.

B.1 RESULTS ON AIGCDTECTION

Table 4: Comparison on 9 diffusion models from AIGCDetection in terms of accuracy(%).

Method	Ref	ADM	DALLE2	GLIDE	Midjourney	VQDM	SD-XL	SD-v1.4	SD-v1.5	Wukong	Mean
UnivFD	CVPR23	66.92	50.75	62.52	56.26	85.42	60.17	63.73	63.52	71.07	64.48
NPR	ECCV24	69.70	64.90	78.35	77.84	78.13	78.53	78.63	78.89	76.11	75.68
FatFormer	CVPR24	78.44	69.70	88.03	56.09	86.88	73.50	67.83	68.06	73.06	73.51
DRCT	ICML24	79.42	92.55	89.18	91.50	90.03	89.90	95.01	94.41	94.67	90.74
SAFE	KDD25	82.00	94.45	95.82	94.56	96.21	98.85	99.23	99.09	98.05	95.36
C2P-CLIP	AAAI25	76.83	65.00	88.13	58.72	86.82	63.32	82.46	82.47	79.83	75.95
AIDE	ICLR25	99.55	97.60	99.82	98.79	99.93	98.88	99.79	99.83	99.58	99.31
Aligned-Forensics	ICLR25	54.04	56.15	63.25	96.25	74.80	99.65	99.78	99.66	99.67	82.58
CO-SPY	CVPR25	58.77	84.65	81.41	69.88	72.14	82.83	93.04	92.98	91.21	80.77
Bfree	CVPR25	75.90	82.00	81.87	94.22	86.79	99.52	99.05	99.05	99.00	90.82
DDA	NeurIPS25	88.40	94.25	86.54	96.27	63.08	99.33	98.01	98.14	97.74	91.31
GAPL	CVPR26	95.75	95.75	98.00	90.60	97.95	99.75	97.85	97.90	98.15	96.86
IAPL	CVPR26	75.15	98.50	93.15	87.10	97.30	95.85	**100.00**	99.85	**100.00**	94.10
LTD	CVPR26	89.38	83.90	98.27	61.18	92.29	67.07	95.09	94.61	95.16	86.33
DoU	CVPR26	82.93	93.00	92.37	72.91	97.94	91.20	99.99	99.86	99.83	92.23
PM	AAAI26	98.72	**100.00**	99.84	99.71	99.48	**99.95**	99.92	99.90	99.88	99.71
Ours	-	**99.87**	99.60	**99.87**	**99.87**	**99.87**	99.60	**99.87**	**100.00**	99.87	**99.82**

Table 5: Comparison on 9 diffusion models from AIGCDetection in terms of average precision(%).

Method	Ref	ADM	DALLE2	GLIDE	Midjourney	VQDM	SD-XL	SD-v1.4	SD-v1.5	Wukong	Mean
UnivFD	CVPR23	87.13	63.04	84.27	74.62	96.65	88.05	86.57	86.19	91.34	84.21
NPR	ECCV24	74.64	76.69	85.72	85.36	81.19	87.73	84.03	84.59	80.47	82.27
FatFormer	CVPR24	91.73	81.84	95.99	62.76	96.99	86.95	81.12	81.09	85.86	84.93
DRCT	ICML24	88.96	97.67	94.64	97.03	96.54	95.45	99.65	99.49	99.37	96.53
SAFE	KDD25	96.59	99.38	99.14	99.40	99.49	99.92	99.95	99.89	99.74	99.28
C2P-CLIP	AAAI25	94.47	90.90	98.21	79.44	97.79	91.17	96.31	96.25	94.25	93.20
AIDE	ICLR25	**100.00**	99.75	**100.00**	99.98	**100.00**	99.96	**100.00**	99.97	99.96	99.96
Aligned-Forensics	ICLR25	78.37	96.99	84.94	99.46	95.46	**100.00**	99.99	99.93	99.98	95.01
CO-SPY	CVPR25	84.56	97.90	96.49	90.85	93.81	97.67	99.18	99.04	98.85	95.37
Bfree	CVPR25	92.42	98.22	96.07	98.99	97.07	**100.00**	99.97	99.97	99.99	98.08
DDA	NeurIPS25	97.08	98.80	96.83	99.55	81.29	**100.00**	99.90	99.90	99.89	97.03
GAPL	CVPR26	99.29	99.33	99.94	97.76	99.95	**100.00**	99.98	99.89	99.98	99.57
IAPL	CVPR26	98.51	**100.00**	99.78	99.71	99.97	99.98	**100.00**	99.98	**100.00**	99.77
LTD	CVPR26	98.73	98.77	99.84	80.78	99.28	94.38	99.45	99.32	99.48	96.67
DoU	CVPR26	99.35	99.96	99.89	99.18	99.98	99.97	**100.00**	99.99	**100.00**	99.81
PM	AAAI26	99.99	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	99.96	**100.00**	99.99
Ours	-	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

Diffusion Models. As shown in Table 4, the proposed DeCue achieves a mean accuracy of 99.82% across the 9 diffusion generators. Its accuracy remains between 99.60% and 100.00% on all subsets, including ADM, DALLE2, GLIDE, Midjourney, VQDM, SD XL, two Stable Diffusion versions, and Wukong. This consistency is important because these generators differ in architecture, training data, and sampling procedure. The corresponding AP is 100.00% on every subset after rounding to two decimal places, as reported in Table 5, indicating nearly perfect separation of real and generated images across decision thresholds.

Table 6: Comparison on 8 GAN models from AIGCDetection in terms of accuracy(%).

Method	Ref	Big-GAN	Cycle-GAN	Gau-GAN	Pro-GAN	Star-GAN	Style-GAN	Style-GAN2	WFR	Mean
UnivFD	CVPR23	95.08	98.33	99.47	99.81	95.72	84.96	74.95	87.05	91.92
NPR	ECCV24	84.35	96.10	82.50	99.79	99.35	97.70	98.38	65.80	90.50
FatFormer	CVPR24	99.50	99.36	99.43	99.89	99.75	97.13	98.80	88.10	97.75
DRCT	ICML24	82.67	94.51	78.39	76.83	63.13	71.58	67.75	63.10	74.75
SAFE	KDD25	89.22	98.49	91.35	99.81	99.75	97.71	98.32	52.10	90.84
C2P-CLIP	AAAI25	99.12	97.31	99.17	**99.98**	99.60	96.43	95.69	94.80	97.76
AIDE	ICLR25	84.90	91.98	87.67	92.75	88.07	88.62	93.63	74.20	87.73
AlignedForensics	ICLR25	51.08	49.43	50.85	51.81	51.53	53.86	51.19	50.00	51.22
CO-SPY	CVPR25	71.62	58.48	69.62	74.67	62.01	63.68	59.63	60.70	65.05
Bfree	CVPR25	90.15	70.78	95.92	95.28	82.89	79.85	75.93	86.80	84.70
DDA	NeurIPS25	76.25	61.36	88.08	84.67	61.61	72.69	81.35	51.65	72.21
GAPL	CVPR26	98.55	98.30	99.70	99.90	96.75	98.05	**99.70**	94.90	98.23
IAPL	CVPR26	98.45	**99.55**	95.00	97.10	**100.00**	86.35	68.50	72.55	89.69
LTD	CVPR26	96.83	98.75	96.79	**99.98**	99.70	97.43	94.99	90.15	96.83
DoU	CVPR26	95.45	98.90	86.98	96.83	**100.00**	83.77	68.82	89.40	90.02
PM	AAAI26	75.00	99.05	68.02	92.03	98.87	94.57	92.97	50.00	83.81
Ours	-	**99.60**	96.40	**99.92**	94.00	99.62	**99.67**	99.61	**99.60**	**98.55**

GAN Models. On the 8 GAN subsets, the proposed DeCue obtains a mean accuracy of 98.55% and an average AP of 99.85%, as shown in Tables 5 and 7. The detector achieves at least 99.60% accuracy on 6 of the 8 subsets, while the lower results on CycleGAN and ProGAN are 96.40% and 94.00%, respectively. Despite being trained with diffusion-reconstructed hard negatives, DeCue generalizes strongly to GAN-generated images, supporting the conclusion that its representations are not tied exclusively to artifacts produced by the training generator. The high AP on ProGAN further indicates that its lower fixed-threshold accuracy is primarily associated with score calibration rather than a failure to rank real and fake samples.

Table 7: Comparison on 8 GAN models from AIGCDetection in terms of average precision(%).

Method	Ref	Big-GAN	Cycle-GAN	Gau-GAN	Pro-GAN	Star-GAN	Style-GAN	Style-GAN2	WFR	Mean
UnivFD	CVPR23	99.27	99.80	99.98	**100.00**	99.37	97.57	97.90	97.27	98.90
NPR	ECCV24	87.80	98.48	85.49	99.99	99.94	99.78	99.84	65.32	92.08
FatFormer	CVPR24	99.98	**100.00**	**100.00**	**100.00**	**100.00**	99.75	99.92	98.48	99.77
DRCT	ICML24	93.51	98.68	86.62	91.03	96.29	79.44	73.80	91.04	88.80
SAFE	KDD25	95.60	99.71	97.07	**100.00**	**100.00**	99.91	99.97	52.45	93.09
C2P-CLIP	AAAI25	99.96	**100.00**	**100.00**	**100.00**	**100.00**	99.50	99.86	99.54	**99.86**
AIDE	ICLR25	97.70	98.86	98.94	98.75	98.96	95.25	98.60	92.55	97.45
AlignedForensics	ICLR25	46.60	40.78	52.42	81.99	69.30	65.31	64.26	38.24	57.36
CO-SPY	CVPR25	83.88	55.83	83.44	78.09	94.33	69.80	62.92	65.59	74.24
Bfree	CVPR25	97.72	86.88	99.19	98.64	88.80	92.94	90.73	99.88	94.35
DDA	NeurIPS25	87.97	72.34	95.24	97.66	66.70	84.18	91.76	68.78	83.08
GAPL	CVPR26	**100.00**	99.94	**100.00**	**100.00**	99.87	**99.97**	**99.99**	99.14	**99.86**
IAPL	CVPR26	99.46	99.97	99.29	99.59	100.00	95.90	85.45	99.92	97.45
LTD	CVPR26	99.98	**100.00**	99.97	**100.00**	**100.00**	99.90	99.46	98.17	99.69
DoU	CVPR26	99.69	99.98	95.02	99.98	**100.00**	98.07	96.24	99.90	98.61
PM	AAAI26	84.91	99.98	83.35	99.16	**100.00**	99.90	99.93	51.74	89.87
Ours	-	**100.00**	99.37	**100.00**	99.53	**100.00**	99.96	99.97	**100.00**	99.85

Overall Analysis. Averaging the diffusion and GAN groups yields the benchmark-level results reported in the main paper and Table 18. The proposed DeCue reaches 99.23% accuracy and 99.93% AP on AIGCDetection, outperforming the strongest baseline in benchmark-level accuracy. The small difference between its diffusion and GAN results suggests that the learned forensic representation transfers effectively across the two major generative paradigms.

B.2 RESULTS ON GENIMAGE

As reported in Table 8, the proposed DeCue achieves 99.89% mean accuracy on GenImage, with all generator-specific accuracies ranging from 99.87% to 100.00%. It obtains the best average result and remains consistently accurate on both the seven diffusion subsets and the BigGAN subset. This result contrasts with several baselines whose performance varies considerably across generators, indicating sensitivity to generator-specific artifacts.

Table 8: Comparison on GenImage in terms of accuracy(%).

Method	Ref	Mid-journey	SD-v1.4	SD-v1.5	ADM	GLI-DE	Wu-kong	VQ-DM	Big-GAN	Mean
UnivFD	CVPR23	56.26	63.73	63.52	66.92	62.52	71.07	85.42	90.21	69.96
NPR	ECCV24	77.84	78.63	78.89	69.70	78.35	76.11	78.13	80.08	77.22
FatFormer	CVPR24	56.09	67.83	68.06	78.44	88.03	73.06	86.88	96.79	76.90
DRCT	ICML24	91.50	95.03	94.41	79.42	89.18	94.67	90.03	81.67	89.49
SAFE	KDD25	94.56	99.23	99.09	82.00	95.82	98.05	96.21	97.82	95.35
C2P-CLIP	AAAI25	58.72	82.46	82.47	76.83	88.13	79.83	86.82	97.68	81.62
AIDE	ICLR25	98.79	99.79	99.83	99.55	99.82	99.58	**99.93**	**99.94**	99.65
AlignedForensics	ICLR25	96.25	99.78	99.66	54.04	63.25	99.67	74.80	51.80	79.91
CO-SPY	CVPR25	69.84	93.08	92.96	58.77	81.36	91.17	72.16	64.68	78.00
Bfree	CVPR25	94.22	99.05	99.08	75.90	81.87	99.00	86.79	65.55	87.68
DDA	NeurIPS25	96.27	98.01	98.14	88.40	86.54	97.74	63.08	76.33	88.06
GAPL	CVPR26	89.55	97.80	97.90	95.05	97.85	97.40	98.05	97.35	96.37
IAPL	CVPR26	95.10	99.90	99.80	85.40	95.90	99.80	98.80	98.70	96.68
LTD	CVPR26	62.97	97.23	97.23	90.38	97.63	96.33	92.53	98.61	91.61
DoU	CVPR26	72.97	**99.99**	99.85	83.14	92.40	99.83	97.88	76.70	90.35
PM	AAAI26	99.71	99.92	99.90	98.72	99.84	**99.88**	99.48	82.91	97.55
Ours	-	**99.87**	99.87	**100.00**	**99.87**	**99.87**	99.87	99.87	99.91	**99.89**

The AP results in Table 9 show 100.00% for the proposed DeCue on every subset after rounding. Together with the uniformly high accuracy, this result indicates both strong score separability and a stable fixed decision boundary. Since no GenImage samples are used for training or dataset-specific adaptation, the results provide evidence of zero-shot cross-generator transfer rather than within-benchmark fitting.

Table 9: Comparison on GenImage in terms of average precision(%).

Method	Ref	Mid-journey	SD-v1.4	SD-v1.5	ADM	GLI-DE	Wu-kong	VQ-DM	Big-GAN	Mean
UnivFD	CVPR23	74.62	86.57	86.19	87.13	84.27	91.34	96.65	98.21	88.12
NPR	ECCV24	85.36	84.03	84.59	74.64	85.72	80.47	81.19	88.18	83.02
FatFormer	CVPR24	62.76	81.12	81.09	91.73	95.99	85.86	96.99	99.55	86.89
DRCT	ICML24	97.03	99.65	99.49	88.96	94.64	99.37	96.54	90.44	95.77
SAFE	KDD25	99.40	99.95	99.89	96.59	99.14	99.74	99.49	99.81	99.25
C2P-CLIP	AAAI25	79.44	96.31	96.25	94.47	98.21	94.25	97.79	99.84	94.57
AIDE	ICLR25	99.98	**100.00**	99.97	**100.00**	**100.00**	99.96	**100.00**	**100.00**	99.99
AlignedForensics	ICLR25	99.46	99.99	99.93	78.37	84.94	99.98	95.46	90.22	93.54
CO-SPY	CVPR25	90.88	99.18	99.04	84.56	96.53	98.85	93.77	90.66	94.18
Bfree	CVPR25	98.99	99.97	99.97	92.42	96.07	99.99	97.07	92.56	97.13
DDA	NeurIPS25	99.55	99.90	99.90	97.08	96.83	99.89	81.29	92.32	95.85
GAPL	CVPR26	97.32	99.96	99.95	99.14	99.96	99.92	99.95	99.89	99.51
IAPL	CVPR26	99.82	**100.00**	99.99	98.60	99.66	99.98	99.96	99.94	99.74
LTD	CVPR26	80.86	99.46	99.32	98.70	99.84	99.50	99.24	99.91	97.10
DoU	CVPR26	99.26	**100.00**	99.98	99.40	99.91	**100.00**	99.98	99.76	99.79
PM	AAAI26	**100.00**	**100.00**	99.96	99.99	**100.00**	**100.00**	**100.00**	99.71	99.96
Ours	-	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

B.3 RESULTS ON DDA-COCO

Table 10 shows that the proposed DeCue achieves 99.83% mean accuracy on DDA-COCO, substantially exceeding the 96.16% result of the strongest baseline. Its accuracy ranges from 99.76% to 99.94% across SD XL, SD EMA, SD MSE, SD v2.1, and SD v3.5. In particular, the 99.94% accuracy on SD v3.5 represents a large improvement over the second-best result of 86.30% on that subset.

Table 10: Comparison on DDA-COCO in terms of accuracy(%).

Method	Ref	XL	EMA	MSE	SDv2.1	SDv3.5	Mean
UnivFD	CVPR23	51.52	54.42	53.21	53.27	51.29	52.74
NPR	ECCV24	43.80	52.15	52.39	52.49	55.54	51.27
FatFormer	CVPR24	50.80	51.50	53.20	53.30	51.65	52.09
DRCT	ICML24	56.31	66.22	65.70	65.63	61.17	63.01
SAFE	KDD25	50.27	50.48	50.45	50.42	50.08	50.34
C2P-CLIP	AAAI25	51.63	52.13	52.89	52.84	52.69	52.44
AIDE	ICLR25	50.30	51.12	50.81	50.86	50.12	50.64
AlignedForensics	ICLR25	89.69	99.50	99.40	99.42	75.25	92.65
CO-SPY	CVPR25	50.90	50.80	50.74	50.76	50.09	50.66
Bfree	CVPR25	96.65	99.12	99.24	99.23	69.13	92.67
DDA	NeurIPS25	96.81	99.17	99.26	99.27	86.30	96.16
GAPL	CVPR26	78.50	90.00	88.75	88.90	70.75	83.38
IAPL	CVPR26	50.00	50.60	50.05	49.95	50.10	50.14
LTD	CVPR26	49.84	50.21	50.08	50.04	50.04	50.04
DoU	CVPR26	50.00	50.07	50.01	50.08	50.07	50.05
PM	AAAI26	49.97	49.97	49.98	49.98	49.98	49.98
Ours	-	**99.92**	**99.76**	**99.76**	**99.76**	**99.94**	**99.83**

DDA-COCO is challenging because its reconstructed images retain semantic content closely aligned with the corresponding real images. The strong performance of DeCue therefore supports the intended role of hard fake sample construction and feature decoupling in suppressing content-based shortcuts. As shown in Table 11, DeCue also obtains 100.00% AP on all five subsets after rounding, demonstrating consistently strong ranking performance across different diffusion reconstruction pipelines.

Table 11: Comparison on DDA-COCO in terms of average precision(%).

Method	Ref	XL	EMA	MSE	SDv2.1	SDv3.5	Mean
UnivFD	CVPR23	65.14	76.13	72.88	72.86	64.51	70.30
NPR	ECCV24	42.80	50.96	51.61	51.57	55.21	50.43
FatFormer	CVPR24	48.37	53.50	58.19	58.19	54.20	54.49
DRCT	ICML24	64.07	77.04	76.25	76.22	71.35	72.99
SAFE	KDD25	53.09	56.59	56.12	56.44	50.50	54.55
C2P-CLIP	AAAI25	56.48	58.99	62.16	62.15	60.41	60.04
AIDE	ICLR25	53.83	60.08	58.43	58.46	52.78	56.72
AlignedForensics	ICLR25	99.18	99.98	99.96	99.96	98.13	99.44
CO-SPY	CVPR25	60.16	59.69	59.53	59.53	49.64	57.71
Bfree	CVPR25	99.65	99.94	99.96	99.96	90.52	98.01
DDA	NeurIPS25	99.62	99.98	99.98	99.98	96.75	99.26
GAPL	CVPR26	96.45	98.99	98.89	98.72	94.36	97.48
IAPL	CVPR26	55.14	67.16	64.78	64.15	62.03	62.65
LTD	CVPR26	41.72	61.43	58.71	58.79	54.68	55.07
DoU	CVPR26	75.53	84.59	83.61	83.64	81.22	81.72
PM	AAAI26	46.19	46.30	45.41	45.37	47.12	46.08
Ours	-	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

B.4 RESULTS ON EVALGEN

As shown in Table 12, the proposed DeCue achieves 99.93% mean accuracy across Flux, GoT, Infinity, OmiGen, and NOVA. The per-generator results vary by only 0.01%, from 99.92% to 99.93%, whereas the strongest competing method reaches 97.42% on average. DeCue also obtains 100.00% AP on all 5 subsets after rounding, as reported in Table 13.

Table 12: Comparison on EvalGEN in terms of accuracy(%).

Method	Ref	Flux	GoT	Infinity	OmiGen	NOVA	Mean
UnivFD	CVPR23	33.60	37.91	41.39	36.21	58.27	41.48
NPR	ECCV24	44.25	39.07	60.42	52.83	51.22	49.56
FatFormer	CVPR24	31.31	41.39	43.32	35.25	93.49	48.95
DRCT	ICML24	86.90	89.74	90.30	87.17	89.19	88.66
SAFE	KDD25	31.87	31.53	32.65	32.37	31.59	32.00
C2P-CLIP	AAAI25	37.15	64.25	55.35	41.47	91.10	57.86
AIDE	ICLR25	43.46	48.13	33.44	54.09	42.36	44.30
AlignedForensics	ICLR25	47.29	74.02	78.53	78.84	82.69	72.27
CO-SPY	CVPR25	89.36	97.47	91.76	98.95	95.69	94.65
Bfree	CVPR25	67.23	99.77	98.59	99.69	99.74	93.00
DDA	NeurIPS25	92.42	95.73	96.09	96.54	98.29	95.81
GAPL	CVPR26	94.65	98.65	98.60	96.80	98.40	97.42
IAPL	CVPR26	52.10	53.75	61.70	58.25	57.40	56.64
LTD	CVPR26	33.30	36.61	33.23	32.99	50.24	37.27
DoU	CVPR26	31.62	33.37	31.77	31.96	34.44	32.63
PM	AAAI26	31.11	31.11	31.13	31.17	31.12	31.13
Ours	-	**99.93**	**99.93**	**99.93**	**99.92**	**99.92**	**99.93**

EvalGEN includes recent generators with heterogeneous autoregressive and diffusion architectures that are not used to construct the training samples. The uniformly high results therefore provide evidence that DeCue does not merely identify artifacts specific to Stable Diffusion v2.1. Instead, the learned representation transfers to unseen generation mechanisms with substantially different architectural characteristics.

Table 13: Comparison on EvalGEN in terms of average precision(%).

Method	Ref	Flux	GoT	Infinity	OmiGen	NOVA	Mean
UnivFD	CVPR23	81.33	88.72	90.48	87.42	96.15	88.82
NPR	ECCV24	68.93	65.63	79.71	74.87	73.01	72.43
FatFormer	CVPR24	60.80	81.29	80.67	71.83	99.26	78.77
DRCT	ICML24	97.27	97.57	98.18	97.20	97.66	97.58
SAFE	KDD25	73.60	76.24	73.20	77.27	74.75	75.01
C2P-CLIP	AAAI25	87.17	97.34	95.21	89.25	99.57	93.71
AIDE	ICLR25	91.65	95.40	78.67	95.02	91.21	90.39
AlignedForensics	ICLR25	98.69	99.69	99.66	99.56	99.83	99.49
CO-SPY	CVPR25	99.64	99.93	99.75	99.98	99.87	99.83
Bfree	CVPR25	97.65	**100.00**	99.97	99.99	**100.00**	99.52
DDA	NeurIPS25	99.63	99.84	99.84	99.87	99.95	99.83
GAPL	CVPR26	99.75	99.94	99.95	99.82	99.92	99.88
IAPL	CVPR26	86.84	89.39	93.00	92.40	93.85	91.10
LTD	CVPR26	82.55	89.20	85.11	52.69	96.31	81.17
DoU	CVPR26	90.53	94.67	95.42	92.67	96.15	93.89
PM	AAAI26	67.17	65.85	67.72	70.26	66.13	67.43
Ours	-	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

B.5 RESULTS ON AIGI-NOW

Table 14 reports the results on nine recent commercial generators. The proposed DeCue achieves 98.90% mean accuracy, improving upon the strongest baseline by 23.58%. It obtains at least 98.38% accuracy on 8 subsets; the most challenging subset is Flux Krea, on which DeCue still reaches 97.60%. In contrast, the baseline results vary substantially across generators, with the best baseline average limited to 75.32%.

Table 14: Comparison on AIGI-Now in terms of accuracy(%).

Method	Ref	Nano-Banana	GPT-4o	Ji-meng	Ke-ling	Mini-max	Flux Pro	Flux Krea	Flux Dev	Flux Kontext	Mean
UnivFD	CVPR23	51.20	53.55	51.42	52.65	51.25	51.92	51.58	50.05	56.75	52.26
NPR	ECCV24	61.52	62.48	38.12	62.90	57.00	40.40	47.33	62.50	62.75	55.00
FatFormer	CVPR24	53.02	50.68	49.33	53.40	49.50	49.33	50.12	51.75	57.38	51.61
DRCT	ICML24	75.48	75.48	71.78	72.35	72.25	71.30	76.90	77.92	71.43	73.88
SAFE	KDD25	74.67	75.08	50.22	74.83	49.98	50.22	50.68	75.45	82.65	64.86
C2P-CLIP	AAAI25	49.12	49.23	49.25	49.80	49.58	50.35	49.40	50.68	54.77	50.24
AIDE	ICLR25	75.30	64.05	57.63	76.72	52.73	56.93	53.65	78.97	89.20	67.24
AlignedForensics	ICLR25	67.62	50.48	63.82	56.57	51.73	56.17	53.12	61.65	55.65	57.42
CO-SPY	CVPR25	67.05	71.70	79.70	83.83	72.10	75.08	81.92	81.70	64.83	75.32
Bfree	CVPR25	57.90	54.43	65.77	72.78	51.23	61.52	53.37	63.10	65.55	60.63
DDA	NeurIPS25	71.55	79.30	74.98	80.65	67.03	68.05	52.38	73.32	66.72	70.44
GAPL	CVPR26	74.30	73.70	66.50	77.45	66.75	68.20	70.05	81.45	78.20	72.96
IAPL	CVPR26	67.65	56.15	50.55	66.85	51.35	52.85	50.85	71.40	75.55	60.36
LTD	CVPR26	55.67	56.43	50.05	57.93	49.95	49.85	49.93	57.10	57.77	53.85
DoU	CVPR26	70.35	60.40	49.95	61.50	50.02	50.00	50.10	65.72	64.35	58.04
PM	AAAI26	74.98	74.98	49.98	74.90	49.98	49.98	50.00	75.00	90.08	65.54
Ours	-	**99.28**	**99.20**	**99.20**	**98.80**	**99.20**	**99.20**	**97.60**	**99.20**	**98.38**	**98.90**

The corresponding AP results in Table 15 range from 99.87% to 99.99%, with an average of 99.96%. The small gap between AP and accuracy indicates that the model separates real and generated samples reliably, while the remaining errors are concentrated near the fixed classification threshold. These results demonstrate strong transfer to recent commercial systems whose architectures and training procedures may differ substantially from publicly available academic generators.

Table 15: Comparison on AIGI-Now in terms of average precision(%).

Method	Ref	Nano-Banana	GPT-4o	Ji-meng	Ke-ling	Mini-max	Flux Pro	Flux Krea	Flux Dev	Flux Kontext	Mean
UnivFD	CVPR23	58.35	68.90	60.85	65.98	59.84	61.58	59.63	52.94	71.82	62.21
NPR	ECCV24	67.22	67.38	34.44	69.45	58.48	37.18	42.43	66.46	66.61	56.63
FatFormer	CVPR24	58.85	55.84	38.60	61.57	43.49	40.32	46.56	53.95	65.61	51.64
DRCT	ICML24	85.17	83.05	75.81	78.02	77.52	78.80	84.81	85.26	76.09	80.50
SAFE	KDD25	85.02	84.84	58.65	84.91	57.53	52.20	56.73	86.28	96.54	73.63
C2P-CLIP	AAAI25	56.79	48.32	42.66	61.73	49.19	47.95	49.72	56.89	62.51	52.86
AIDE	ICLR25	85.33	85.44	80.00	89.61	68.37	81.31	69.00	94.04	98.27	83.49
AlignedForensics	ICLR25	83.45	70.54	86.97	82.93	79.09	80.36	80.04	82.09	75.56	80.11
CO-SPY	CVPR25	91.56	92.81	96.12	97.32	92.67	94.44	97.10	97.26	87.87	94.13
Bfree	CVPR25	81.29	74.45	88.45	92.36	61.08	83.86	68.51	82.59	84.85	79.72
DDA	NeurIPS25	87.22	92.39	89.90	92.95	79.87	84.30	59.42	84.72	78.26	83.23
GAPL	CVPR26	92.17	94.43	89.50	94.90	90.19	92.29	93.54	96.10	95.39	93.17
IAPL	CVPR26	89.71	84.77	56.67	89.23	60.23	64.76	63.87	88.52	94.38	76.90
LTD	CVPR26	71.59	73.47	46.23	70.84	40.35	43.79	38.83	67.26	76.94	58.81
DoU	CVPR26	85.11	81.64	46.49	84.07	55.38	51.92	61.13	83.37	88.67	70.86
PM	AAAI26	81.42	81.96	44.91	81.21	48.31	46.90	47.74	82.35	98.94	68.19
Ours	-	**99.99**	**99.99**	**99.94**	**99.99**	**99.93**	**99.96**	**99.87**	**99.98**	**99.96**	**99.96**

B.6 RESULTS ON CHAMELEON, AIGIBENCH AND WILDRF

Chameleon. As reported in Table 16, the proposed DeCue achieves 99.94% accuracy on Chameleon, exceeding the second-best result of 84.81% by 15.13%. Chameleon contains high-resolution, perceptually convincing images selected through human evaluation, and thus exposes the limitations of detectors that rely on conspicuous low-level artifacts or dataset-specific semantic correlations. DeCue further obtains 100.00% AP after rounding, compared with 91.23% for the strongest baseline, as shown in Table 17. The simultaneous improvement in accuracy and AP indicates that the gain cannot be explained solely by a more favorable classification threshold. Instead, DeCue produces a substantially more separable score distribution on this challenging real-world benchmark.

Table 16: Comparison on Chameleon, AIGIBench and WildRF in terms of accuracy(%).

Method	Ref	Chameleon	AIGIBench			WildRF				Mean
			CommunityAI	SocialRF	Avg.	Facebook	Reddit	Twitter	Avg.	
UnivFD	CVPR23	57.40	51.39	54.58	52.99	49.06	60.27	56.08	55.14	54.80
NPR	ECCV24	59.80	55.19	54.83	55.01	77.81	65.07	52.27	65.05	60.83
FatFormer	CVPR24	58.03	51.92	56.93	54.43	54.37	69.47	54.76	59.53	57.58
DRCT	ICML24	53.86	54.85	50.47	52.66	49.06	55.20	51.68	51.98	52.52
SAFE	KDD25	59.11	54.43	58.35	56.39	50.62	74.13	51.98	58.91	58.10
C2P-CLIP	AAAI25	58.46	51.88	55.40	53.64	53.12	66.73	58.13	59.33	57.29
AIDE	ICLR25	65.75	66.17	62.42	64.30	61.56	74.00	61.93	65.83	65.31
AlignedForensics	ICLR25	71.99	71.85	73.27	72.56	91.88	70.80	81.99	81.56	76.96
CO-SPY	CVPR25	69.06	64.86	67.67	66.27	72.81	77.73	76.72	75.75	71.48
Bfree	CVPR25	78.31	80.37	86.00	83.19	95.00	89.33	**97.80**	94.04	87.80
DDA	NeurIPS25	84.81	88.54	79.65	84.10	86.88	82.40	88.87	86.05	85.19
GAPL	CVPR26	68.45	76.25	80.00	78.13	87.00	79.50	89.00	85.17	80.03
IAPL	CVPR26	75.09	54.60	57.80	56.20	53.50	75.50	52.00	60.33	61.42
LTD	CVPR26	57.63	51.75	58.52	55.14	50.94	71.13	39.43	53.83	54.90
DoU	CVPR26	59.04	53.71	57.57	55.64	53.12	72.20	51.54	58.95	57.86
PM	AAAI26	59.49	54.83	60.33	57.58	50.00	79.93	51.98	60.64	59.43
Ours	-	**99.94**	**99.87**	**89.73**	**94.80**	**100.00**	**99.07**	96.93	**98.67**	**97.59**

Table 17: Comparison on Chameleon, AIGIBench and WildRF in terms of average precision(%).

Method	Ref	Chameleon	AIGIBench			WildRF				Mean
			CommunityAI	SocialRF	Avg.	Facebook	Reddit	Twitter	Avg.	
UnivFD	CVPR23	47.68	57.93	57.80	57.87	44.42	66.98	60.75	57.38	55.93
NPR	ECCV24	49.60	56.66	56.35	56.51	86.33	69.33	52.54	69.40	61.80
FatFormer	CVPR24	57.87	71.26	65.11	68.19	75.68	80.68	65.33	73.90	69.32
DRCT	ICML24	60.43	67.42	54.82	61.12	47.89	66.67	56.57	57.04	58.97
SAFE	KDD25	50.83	62.92	65.72	64.32	55.31	83.79	61.01	66.70	63.26
C2P-CLIP	AAAI25	50.70	58.49	59.67	59.08	61.45	75.40	63.39	66.75	61.52
AIDE	ICLR25	70.09	80.75	69.24	75.00	67.25	83.51	69.18	73.31	73.34
AlignedForensics	ICLR25	82.24	92.77	83.25	88.01	97.09	79.50	91.85	89.48	87.78
CO-SPY	CVPR25	68.87	70.83	73.55	72.19	84.42	86.79	91.21	87.47	79.28
Bfree	CVPR25	83.41	91.24	93.28	92.26	99.07	96.81	99.76	98.55	93.93
DDA	NeurIPS25	91.23	95.66	88.92	92.29	96.84	91.70	96.28	94.94	93.44
GAPL	CVPR26	79.91	84.06	91.75	87.91	94.34	95.91	95.88	95.38	90.31
IAPL	CVPR26	53.29	56.24	64.80	60.52	55.90	91.48	56.59	67.99	63.05
LTD	CVPR26	35.31	51.17	67.79	59.48	70.05	83.94	74.90	76.30	63.86
DoU	CVPR26	46.58	51.13	65.82	58.48	65.02	89.51	66.49	73.67	64.09
PM	AAAI26	43.52	55.42	66.26	60.84	49.43	85.50	51.96	62.30	58.68
Ours	-	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**99.99**	**100.00**	**100.00**

AIGIBench. On the two in-the-wild AIGIBench subsets, the proposed DeCue achieves 99.87% accuracy on CommunityAI and 89.73% on SocialRF, resulting in a mean accuracy of 94.80% (Table 16). It outperforms the strongest baseline average of 84.10% by 10.70%. The lower accuracy on SocialRF relative to CommunityAI suggests that social-media data introduce a more pronounced distribution shift, potentially due to heterogeneous acquisition sources and post-processing operations. Despite this fixed-threshold accuracy gap, DeCue obtains 100.00% AP on both subsets after rounding (Table 17). This shows that real and generated samples remain almost perfectly rank-separable, while the SocialRF errors are mainly associated with score calibration under domain shift. Consequently, the results support both the discriminative strength of the learned representation and the importance of calibration when deploying detectors across heterogeneous online platforms.

WildRF. As shown in Table 16, the proposed DeCue obtains accuracies of 100.00%, 99.07%, and 96.93% on Facebook, Reddit, and Twitter, respectively, yielding an average of 98.67%. This exceeds the strongest baseline average of 94.04%. The relatively lower accuracy on Twitter suggests that platform-specific processing or content distributions produce a larger shift at the fixed classification threshold, although DeCue remains highly accurate. The AP results are 100.00% on Facebook and Reddit and 99.99% on Twitter after rounding, with an average of 100.00% (Table 17). These results indicate that the forensic representation remains highly discriminative after real-world compression, resizing, and platform-dependent processing. They are also consistent with the robustness evaluation in the main paper, where DeCue maintains strong performance under controlled image degradations.

B.7 Overall Comparison Across Benchmarks

Table 18 summarizes AP across all 8 benchmarks. The proposed DeCue achieves an overall average AP of 99.98%, exceeding the second-best result of 96.68% by 3.30%. More importantly, its performance remains stable across standard academic datasets, recent commercial generators, and in-the-wild social-media data. The corresponding mean accuracy of 99.23% reported in Table 1 exceeds the second-best result of 87.94% by 11.29%. The smaller margin in AP than in accuracy indicates that some baseline methods can rank samples reasonably well but lack a decision threshold that transfers reliably across datasets. DeCue improves both ranking quality and fixed-threshold classification, suggesting that its real-centric decision boundary is more stable under generator and domain shifts.

Table 18: Comparison on all 8 benchmarks in terms of average precision(%).

Method	Ref	AIGC-Detection	Gen-Image	DDA-COCO	Eval-GEN	AIGI-Now	Chame-leon	AIGI-Bench	WildRF	Mean
UnivFD	CVPR23	91.12	88.12	70.30	88.82	62.21	47.68	57.87	57.38	76.08
NPR	ECCV24	86.89	83.02	50.43	72.43	56.63	49.60	56.51	69.40	68.53
FatFormer	CVPR24	91.91	86.89	54.49	78.77	51.64	57.87	68.19	73.90	72.17
DRCT	ICML24	92.89	95.77	72.99	97.58	80.50	60.43	61.12	57.04	83.12
SAFE	KDD25	96.37	99.25	54.55	75.01	73.63	50.83	64.32	66.70	77.01
C2P-CLIP	AAAI25	96.33	94.57	60.04	93.71	52.86	50.70	59.08	66.75	76.50
AIDE	ICLR25	98.78	99.99	56.72	90.39	83.49	70.09	75.00	73.31	83.78
AlignedForensics	ICLR25	77.30	93.54	99.44	99.49	80.11	82.24	88.01	89.48	89.61
CO-SPY	CVPR25	85.43	94.18	57.71	99.83	94.13	68.87	72.19	87.47	85.09
Bfree	CVPR25	96.32	97.13	98.01	99.52	79.72	83.41	92.26	98.55	94.10
DDA	NeurIPS25	90.46	95.85	99.26	99.83	83.23	91.23	92.29	94.94	93.68
GAPL	CVPR26	99.71	99.51	97.48	99.88	93.17	79.97	87.91	95.38	96.68
IAPL	CVPR26	98.68	99.74	62.65	91.10	76.90	53.29	60.52	67.99	82.02
LTD	CVPR26	98.09	97.10	55.07	81.17	58.81	35.31	59.48	76.30	75.68
DoU	CVPR26	99.25	99.79	81.72	93.89	70.86	46.58	58.48	73.67	84.93
PM	AAAI26	95.23	99.96	46.08	67.43	68.19	43.52	60.84	62.30	72.59
Ours	-	**99.93**	**100.00**	**100.00**	**100.00**	**99.96**	**100.00**	**100.00**	**100.00**	**99.98**

B.8 Additional Qualitative Results

Figure 6 presents Grad-CAM visualizations on representative human, animal, and scene images from Chameleon. The discriminative responses are concentrated in localized image regions rather

than uniformly covering the dominant semantic objects. This activation pattern is consistent with the intended behavior of DeCue: the detector bases its predictions on localized forensic evidence while reducing its dependence on global semantic content. Although Grad-CAM visualizations do not by themselves establish causal feature disentanglement, they provide complementary qualitative evidence for the feature-distribution analysis presented in Figure 4.

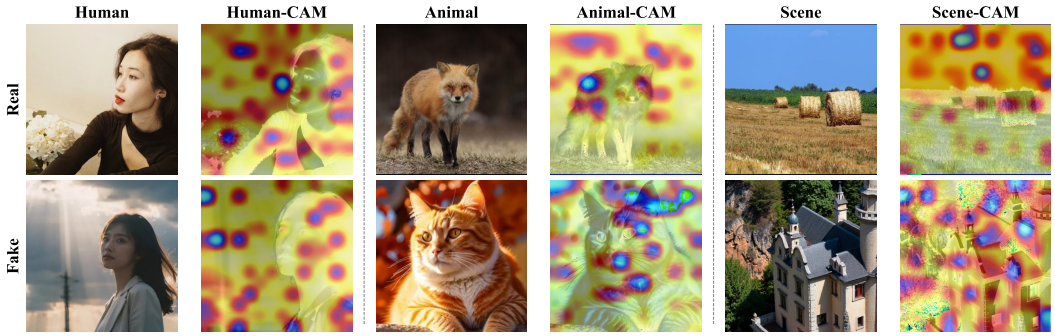


Figure 6: Grad-CAM activation visualizations of the proposed DeCue model on the Chameleon benchmark. We display representative samples across three common semantic categories: human, animal, and scene. The results show that discriminative activations concentrate on local regions carrying generative artifacts rather than the global semantic content of images, which validates that our feature decoupling framework effectively drives the model to learn intrinsic forensic cues instead of exploiting semantic content shortcuts.