PROBING AI-GENERATED IMAGE DETECTORS VIA BLACK-BOX ANTI-FORENSICS ON SEMANTIC AND ARTIFACT CUES

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in generative models have made AI-generated images increasingly difficult to distinguish from real ones, motivating detectors that exploit semantic representations from vision foundation models (VFMs) and synthesis-specific artifacts left by generative models. However, it remains unclear whether these cues capture intrinsic evidence of forgery or merely reflect incidental characteristics of the generation process. In this work, we investigate this question through systematic anti-forensic probing, finding that detectors can be substantially compromised by independently manipulating these two types of evidence: the semantic representations used for discrimination can be steered toward authentic semantics, while synthesis-specific artifacts can be suppressed without noticeably altering image content. Based on these findings, we propose CRAS, a black-box anti-forensic framework that jointly targets semantic and artifact evidence. Specifically, a cross-modal reconstruction module steers forged images embeddings toward authentic text anchors in the VFM feature space, followed by an artifact suppression module that reduces the synthesis-specific artifacts left by the reconstruction process. Extensive experiments show that CRAS achieves the best anti-forensic performance among existing methods while preserving high visual fidelity. Our results suggest that current AIGI detectors rely heavily on both types of evidence that are not invariant to carefully controlled transformations, highlighting the importance of learning manipulation-invariant forensic evidence.

1 INTRODUCTION

With the rapid advancement of generative models, particularly Generative Adversarial Networks (GANs) and diffusion models, the barrier to generating realistic, diverse, and high-quality images has been substantially lowered (He et al., 2025; Zhou et al., 2025a; Ding et al., 2026a). However, the potential misuse of these models has also raised growing concerns regarding issues such as the spread of misinformation and financial fraud (Ding et al., 2026b; Fan et al., 2023).

To distinguish AI-generated images from natural ones, numerous detection models have been proposed. Conventional supervised paradigms (Wang et al., 2020; Liu et al., 2020; He et al., 2016; Li et al., 2025; Tan et al., 2024) train deep neural networks from scratch to capture low-level statistical artifacts, achieving strong performance when training and test distributions are similar but degrading substantially on unseen generation methods. More recently, pre-trained VFMs such as CLIP (Radford et al., 2021) and MetaCLIP2 (Chuang et al., 2026) have been widely adopted as detection backbones (Ojha et al., 2023; Yan et al., 2026; 2025) owing to their robust semantic priors. Other works (Zhou et al., 2024) have probed the nature of detector from an anti-forensic perspective, crafting adversarial images that deceive both detectors and human perception to drive the development of more robust detectors. However, it remains unclear whether the two forms of forensic evidence exploited by these detectors, namely semantic representations from VFMs and synthesis-specific artifacts left by generative models, capture intrinsic evidence of image forgery or merely reflect incidental characteristics of the generation process.

To investigate this question, we conduct systematic anti-forensic probing in this work. We summarize the underlying working mechanisms of existing SOTA detectors into two broad categories of cues: *1)* with the significant enhancement of visual representation capability in modern vision

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.

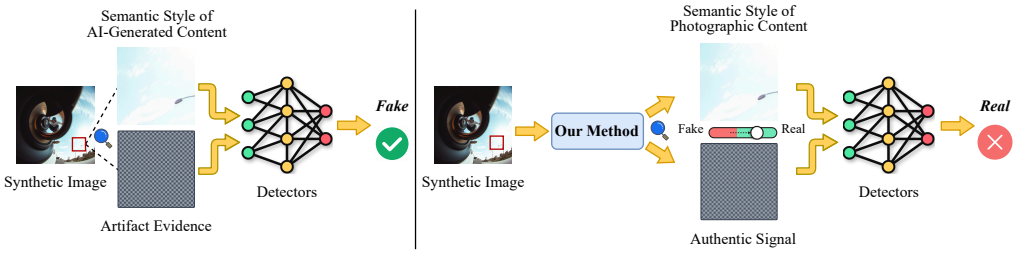


Figure 1: Overview of our method attacking the semantic and artifact evidence. **Left:** the detector correctly classifies the synthetic image as fake by exploiting its AI-generated style and diffusion-specific artifacts. **Right:** our method causes the image to be misclassified as real by steering it toward authentic semantics and suppressing synthesis-specific artifacts.

foundation models, detectors are often able to correctly classify most content exhibiting pronounced AI-generated styles by relying solely on global visual semantics; *2)* many existing works heavily rely on simulating the imperceptible, fine-grained artifacts incidental to the generation process of diffusion models (DMs) to promote generalization across different generators.

As shown in Fig. 1, by independently manipulating these two types of evidence, we find that the performance of modern detectors can be substantially compromised. Based on these observations, we propose CRAS, a detector-agnostic, black-box anti-forensic framework that jointly targets semantic and artifact evidence. CRAS consists of two cascaded modules: a Cross-Modal Reconstruction module and an Artifact Suppression module. The Cross-Modal Reconstruction module targets semantic evidence, using a multimodal guidance loss to drive forged image embeddings within the VFM feature space toward authentic text anchors. The Artifact Suppression module targets artifact evidence, training a network specifically designed to simulate pristine cues in order to suppress the artifacts present in the DM-reconstructed images. Extensive experiments show that CRAS achieves the best anti-forensic performance among existing methods while preserving high visual fidelity. Our main contributions are summarized as follows:

- We summarize detector decision-making into two broad categories of cues, namely semantic representations from VFMs and synthesis-specific artifacts, and find that independently manipulating each type of evidence leads to substantial performance degradation.
- We propose CRAS, a detector-agnostic, black-box anti-forensic framework in which a Cross-Modal Reconstruction Module and an Artifact Suppression Module are cascaded to jointly attack semantic and artifact evidence. Extensive experiments demonstrate that, compared with existing anti-forensic methods, our method achieves the best anti-forensic performance while preserving high visual fidelity.
- Our findings suggest that current detectors rely heavily on semantic and synthesis-specific cues that are not invariant to carefully controlled transformations, highlighting the need for manipulation-invariant forensic evidence in future research.

2 Related Works

2.1 AI-Generated Image Forensics

Research on AI-generated image detection has evolved from conventional supervised detectors to approaches leveraging pretrained vision foundation models. Conventional methods train classifiers to capture statistical anomalies and generation artifacts. CNNSpot (Wang et al., 2020) demonstrated the effectiveness of CNN classifiers in identifying low-level artifacts, GramNet (Liu et al., 2020) exploited texture statistics, NPR (Tan et al., 2024) leveraged upsampling artifacts in CNN-based generators, and SAFE (Li et al., 2025) encouraged transformation-invariant forensic representations. However, reliance on training-time artifacts limits generalization to unseen architectures. Recent work addresses this by leveraging pretrained vision foundation models. UniFD (Ojha et al., 2023) showed that a simple classifier on frozen CLIP features can generalize across generators. DGS-Net (Yan et al., 2026) applied gradient surgery to preserve transferable knowledge during CLIP fine-tuning, and Effort (Yan et al., 2025) introduced orthogonal subspaces to retain pretrained knowledge

while adapting forgery-related representations. Together, these detectors draw on both semantic features from VFMs and synthesis artifacts to distinguish real from generated images, yet it remains unclear whether these cues are inherently indicative of forgery or arise from incidental properties of the generation process.

2.2 Anti-Forensics Attacks

Image forensics is crucial for curbing misuse, motivating extensive research into the vulnerabilities of forensic approaches and anti-forensic techniques. Existing work (Fang et al., 2025) has explored VFM vulnerability to universal perturbations via multimodal contrastive losses. DiffAttack (Chen et al., 2024b) leveraged diffusion models to generate highly transferable adversarial perturbations. StealthDiffusion (Zhou et al., 2024) jointly optimized the latent space and frequency domain to make generated images visually indistinguishable from real ones. UDTF (Lai et al., 2026) proposed a multiplicative attack that fundamentally eliminates generative model fingerprints to achieve untraceability. FTM (Liang et al., 2025) enhanced targeted attack transferability by combining random and optimized feature-space perturbations. FPR (Ren et al., 2025) improved adversarial transferability on Vision Transformers by refining attention maps and token embeddings during forward propagation. In this paper, we propose a detector-agnostic black-box attack that reveals the heavy reliance of current detectors on semantic and artifact cues that lack invariance under carefully controlled transformations.

3 Methodology

Overview. We hypothesize that modern detectors rely on two broad categories of cues: semantic representations derived from VFMs and synthesis-specific artifacts left by generative models. To investigate the vulnerability of these forms of evidence, as illustrated in Fig. 2, we design a cross-modal reconstruction module, which steers forged images toward authentic text anchors in the VFM feature space to attack semantic evidence, and an artifact suppression module, which reduces generation-specific artifacts in the reconstructed images to attack artifact evidence, disentangling and manipulating them independently.

Objective. The fundamental task of image forensics is to distinguish real images from AI-generated ones. Given an input AI-generated image x with label $y = 1$ (fake), the goal of a detector D is to correctly determine the authenticity of the image, i.e., $D(x) = 1$. In this work, we focus exclusively on anti-forensic attacks against clean fake images and do not consider real images. Specifically, we generate adversarial images x' that mislead the detector into classifying AI-generated images as real, i.e., $D(x') = 0$ (real).

3.1 Cross-Modal Reconstruction Module

To attack the strategy by which modern detectors exploit VFM semantic representations for discrimination, we design a cross-modal reconstruction module (**CRM**) that leverages the reconstruction capability of Stable Diffusion to steer forged images toward authentic text anchors in the VFM feature space. Specifically, we first map the clean input image x_{ori} into the latent space of Stable Diffusion using the encoder E of a Variational Autoencoder (VAE), which compresses the input image into an initial latent representation, denoted as z_0. Subsequently, we apply DDIM inversion to progressively map z_0 into a highly noisy latent state z_T. DDIM inversion is adopted instead of the standard forward diffusion process, which injects random Gaussian noise, because it exploits the noise prediction capability of the pretrained diffusion model to deterministically invert the latent variable along the diffusion trajectory, thereby gradually increasing the noise level while largely preserving the semantic and structural information of the original image. At the t-th timestep, the pretrained denoising network ϵ_θ first predicts the noise from the current latent variable z_t. Based on the predicted noise, the corresponding clean latent estimate is obtained as

$$\hat{z}_0^{(t)} = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(z_t, t, c)}{\sqrt{\bar{\alpha}_t}}. \tag{1}$$

Here, $\bar{\alpha}_t$ denotes the cumulative coefficient determined by the predefined noise schedule, and c represents the textual conditioning embedding used by Stable Diffusion. In our setting, we adopt an empty text prompt, whose embedding is used as c. Then, according to the deterministic DDIM inversion update rule, the current latent variable z_t is mapped to a noisier latent state z_{t+1}:

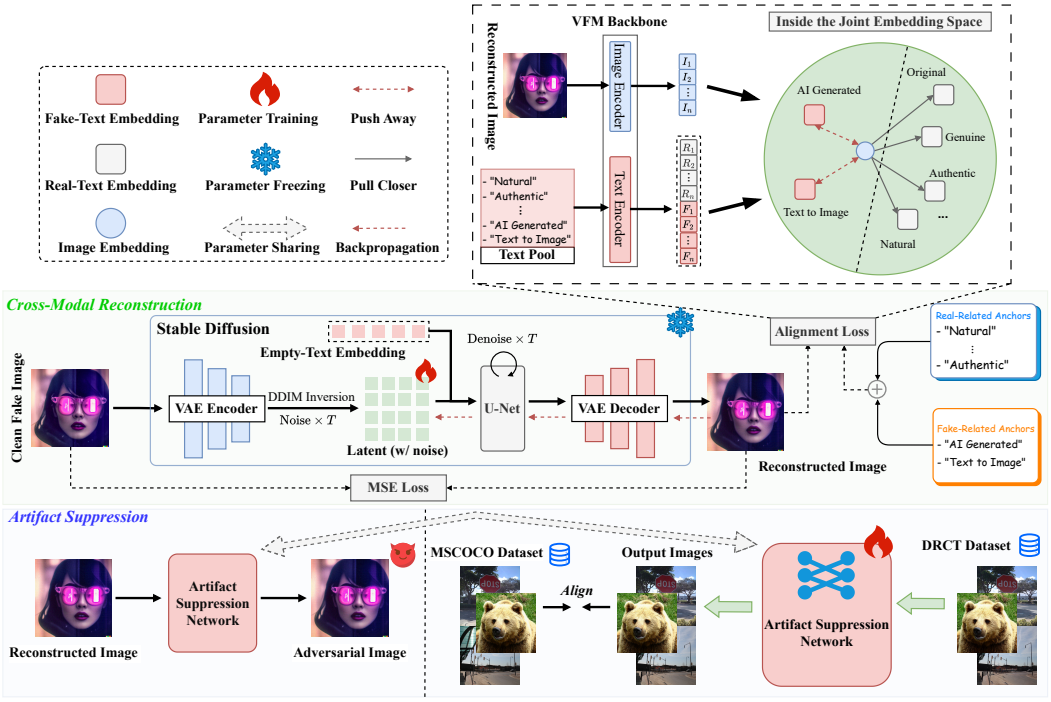


Figure 2: **Overview of the proposed framework.** Cross-Modal Reconstruction targets semantic evidence by optimizing DM latent variables with an MSE loss to preserve visual fidelity and an alignment loss to drive reconstructed image embeddings toward authentic text anchors in the VFM feature space. Artifact Suppression targets artifact evidence by training a network to simulate pristine cues, aligning the output with real images to suppress DM-induced artifacts.

$$z_{t+1} = \sqrt{\bar{\alpha}_{t+1}}\,\hat{z}_0^{(t)} + \sqrt{1-\bar{\alpha}_{t+1}}\,\epsilon_\theta(z_t, t, c). \tag{2}$$

By iteratively applying this inversion step along the predefined diffusion timesteps, the inverted latent representation z_T corresponding to the input image can be obtained. The inverted latent representation z_T is subsequently used as the optimization variable for adversarial optimization. To map the optimized latent back to the image space, we perform deterministic DDIM reverse denoising from z_T toward the low-noise latent space. At the t-th reverse step, the clean latent estimate $\hat{z}_0^{(t)}$ is first computed using Eq. 1, and the latent state at the next denoising timestep is then obtained as

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,\hat{z}_0^{(t)} + \sqrt{1-\bar{\alpha}_{t-1}}\,\epsilon_\theta(z_t, t, c). \tag{3}$$

This reverse update is iteratively applied along the predefined DDIM timesteps until the final low-noise latent representation z_0' is obtained. The resulting latent is then decoded into the image space through the VAE decoder D, yielding the reconstructed image x_{rec}.

To enhance the adversarial effectiveness of the reconstructed image while preserving its original visual content, we jointly constrain the optimization of the latent variable z_T from two complementary perspectives: pixel-level fidelity and vision-text feature alignment. First, a pixel-level mean squared error constraint is introduced to keep the optimized image x_{rec} close to the clean input image x_{ori}:

$$\mathcal{L}_{\text{mse}} = \frac{1}{N_p}\left\|x_{rec} - x_{ori}\right\|_2^2. \tag{4}$$

where N_p denotes the total number of image pixels. Subsequently, to attack the strategy by which modern detectors exploit VFM semantic representations for discrimination, we design a multimodal guidance loss that steers forged images toward authentic text anchors in the VFM feature space.

Specifically, the normalized image feature of the reconstructed image x_{rec} is extracted by the VFM's image encoder E_{img} and denoted as $\mathbf{v}$. Meanwhile, the real-related textual descriptions are encoded by the VFM's text encoder E_{text} into a real semantic pool $\mathcal{R} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{N_r}\}$, while textual descriptions associated with AI-generated image characteristics are encoded by the same text encoder into a fake semantic pool $\mathcal{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{N_f}\}$. Please refer to Appendix A.4 for the specific textual descriptions.

For the reconstructed image feature $\mathbf{v}$, its similarities to the real and fake textual features are computed as $s_i^r = \mathbf{v}^\top \mathbf{r}_i$ and $s_j^f = \mathbf{v}^\top \mathbf{f}_j$, respectively. When all features are L_2-normalized, the inner product is equivalent to cosine similarity. We then regard the real and fake text pools as a unified candidate set and normalize their similarity scores using a temperature-scaled Softmax. The overall probability that the reconstructed image matches the real semantic pool is defined as

$$P_{\text{real}}(x_{rec}) = \frac{\sum_{i=1}^{N_r} \exp\left(s_i^r/\tau\right)}{\sum_{i=1}^{N_r} \exp\left(s_i^r/\tau\right) + \sum_{j=1}^{N_f} \exp\left(s_j^f/\tau\right)}. \tag{5}$$

Based on this probability, we define the real-semantic alignment loss as $\mathcal{L}_{\text{align}} = -\log P_{\text{real}}(x_{rec})$. The proposed loss uses multiple real-related descriptions as positive references and fake-related descriptions as competing negatives, encouraging the reconstructed image to move toward real-oriented semantics in the VFM feature space.

Finally, we jointly optimize the high-noise latent variable z_T obtained from DDIM inversion using the pixel-level fidelity constraint and the semantic alignment constraint:

$$z_T^* = \arg\min_{z_T} \left[\lambda_{\text{mse}} \mathcal{L}_{\text{mse}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}\right], \tag{6}$$

where λ_{mse} and λ_{align} control the trade-off between visual fidelity and real-semantic alignment, respectively. During optimization, Stable Diffusion remains frozen, and only z_T is updated through back-propagation over the DDIM denoising process and VAE decoder. The final reconstructed image is obtained as $x_{\text{rec}} = D(\mathcal{F}_{\text{DDIM}}(z_T^*))$.

3.2 Artifact Suppression Module

To further attack the strategy by which detectors exploit synthesis-specific artifacts left by generative models for discrimination, we design an artifact suppression module (**ASM**) to isolate and suppress synthesis-specific forensic evidence. Moreover, the artifact suppression module can be integrated with the aforementioned cross-modal reconstruction module to reduce artifacts in the DM-reconstructed image, further improving anti-forensic performance.

The artifact suppression module Φ is trained end-to-end on pairs of real and synthetic images (x_r, x_s) to reduce artifacts in synthetic images while preserving perceptual fidelity. The module Φ adopts an encoder-decoder architecture. The encoder consists of 8 convolutional downsampling blocks, and the decoder employs 8 symmetric transposed-convolution upsampling blocks. Please refer to Appendix A.3 for more details. In our implementation, the encoder is frozen and only the decoder is updated during training, restricting the learnable transformation to minimal, artifact-targeted edits rather than holistic regeneration of image content. The training pair (x_r, x_s) consists of a fake image x_s reconstructed by the Stable Diffusion in DRCT (Chen et al., 2024a) and its content-matched real MS-COCO (Lin et al., 2014) image x_r. This construction of the training pair determines the targeted nature of the module's intervention: since x_s and x_r are matched in semantic content, the only variation Φ can learn is the discrepancy that distinguishes x_s from x_r in the eyes of the detector, namely the synthesis-specific artifacts introduced by the diffusion model itself. Given a synthetic image x_s, ASM outputs an artifact-suppressed image $\hat{y} = \Phi(x_s)$. ASM is optimized using a combination of adversarial loss and paired L_1 reconstruction loss:

$$\mathcal{L}_\Phi = -\mathbb{E}_{x_s \sim p_{\text{data}}} \left[\log D(x_s, \Phi(x_s))\right] + \lambda_{L1} \mathbb{E}_{(x_s, x_r) \sim p_{\text{data}}} \left[\|\Phi(x_s) - x_r\|_1\right]. \tag{7}$$

Table 1: Anti-forensic performance of the proposed method against CLIP-based and conventional detectors across datasets. $\mathrm{Acc}_{\mathrm{clean}}$ and $\mathrm{Acc}_{\mathrm{adv}}$ denote the classification accuracy on clean fake images and adversarial images, respectively. Negative Δ_{acc} indicates reduced detection accuracy.

Dataset	Metrics	CLIP-based Detectors					Conventional Detectors				
		UniFD	DGS-Net	Effort	CLIP+Linear	CLIP+LoRA	CNNSpot	GramNet	ResNet50	SAFE	NPR
GenImage	$\mathrm{Acc}_{\mathrm{clean}}$	0.70	0.61	0.87	0.64	0.69	0.43	0.54	0.33	0.70	0.69
	$\mathrm{Acc}_{\mathrm{adv}}\downarrow$	0.33	0.33	0.22	0.05	0.03	0.08	0.22	0.00	0.20	0.37
	$\boldsymbol{\Delta}_{\mathrm{acc}}\downarrow$	**-0.37**	**-0.28**	**-0.65**	**-0.59**	**-0.66**	**-0.35**	**-0.32**	**-0.33**	**-0.50**	**-0.32**
AIGCDetectBench	$\mathrm{Acc}_{\mathrm{clean}}$	0.68	0.73	0.89	0.64	0.71	0.39	0.52	0.29	0.71	0.60
	$\mathrm{Acc}_{\mathrm{adv}}\downarrow$	0.37	0.36	0.26	0.06	0.03	0.05	0.11	0.00	0.14	0.25
	$\boldsymbol{\Delta}_{\mathrm{acc}}\downarrow$	**-0.31**	**-0.37**	**-0.63**	**-0.58**	**-0.68**	**-0.34**	**-0.41**	**-0.29**	**-0.57**	**-0.35**
AIGIBench	$\mathrm{Acc}_{\mathrm{clean}}$	0.69	0.42	0.88	0.87	0.56	0.24	0.27	0.11	0.56	0.43
	$\mathrm{Acc}_{\mathrm{adv}}\downarrow$	0.49	0.40	0.47	0.17	0.05	0.01	0.02	0.00	0.10	0.26
	$\boldsymbol{\Delta}_{\mathrm{acc}}\downarrow$	**-0.20**	**-0.02**	**-0.41**	**-0.70**	**-0.51**	**-0.23**	**-0.25**	**-0.11**	**-0.46**	**-0.17**
AIGI-Holmes	$\mathrm{Acc}_{\mathrm{clean}}$	0.62	0.36	0.83	0.77	0.44	0.14	0.23	0.17	0.63	0.59
	$\mathrm{Acc}_{\mathrm{adv}}\downarrow$	0.26	0.31	0.31	0.13	0.04	0.01	0.05	0.00	0.16	0.34
	$\boldsymbol{\Delta}_{\mathrm{acc}}\downarrow$	**-0.36**	**-0.05**	**-0.52**	**-0.64**	**-0.40**	**-0.13**	**-0.18**	**-0.17**	**-0.47**	**-0.25**
Avg.	$\mathrm{Acc}_{\mathrm{clean}}$	0.67	0.53	0.86	0.73	0.60	0.30	0.39	0.22	0.65	0.57
	$\mathrm{Acc}_{\mathrm{adv}}\downarrow$	0.36	0.35	0.31	0.10	0.03	0.03	0.10	0.00	0.15	0.30
	$\boldsymbol{\Delta}_{\mathrm{acc}}\downarrow$	**-0.31**	**-0.18**	**-0.55**	**-0.63**	**-0.57**	**-0.27**	**-0.29**	**-0.22**	**-0.50**	**-0.27**

where D adopts a conditional PatchGAN architecture, taking as input the channel-wise concatenation of x_s with either the real image x_r or $\Phi(x_s)$, and producing a patch-level probability map.

Through training the artifact suppression model, we inject an inductive bias for artifact reduction into Φ, enabling it to remove synthesis-specific fingerprints while preserving visual fidelity, thereby isolating and suppressing synthesis-specific forensic evidence.

4 Experiments

4.1 Experimental Setups

We evaluate the proposed framework against CLIP (Radford et al., 2021) as the target VFM, attacking both CLIP-based detectors (UniFD (Ojha et al., 2023), DGS-Net (Yan et al., 2026), Effort (Yan et al., 2025), CLIP+Linear, CLIP+LoRA) and conventional detectors (CNNSpot (Wang et al., 2020), GramNet (Liu et al., 2020), ResNet50 (He et al., 2016), SAFE (Li et al., 2025), NPR (Tan et al., 2024)), and compare against five SOTA anti-forensic baselines including DiffAttack (Chen et al., 2024b), UDTF (Lai et al., 2026), FTM (Liang et al., 2025), FPR (Ren et al., 2025) and StealthDiffusion (Zhou et al., 2024). Experiments are conducted on four representative datasets, including GenImage (Zhu et al., 2023), AIGCDetectBench (Zhong et al., 2023), AIGI-Holmes (Zhou et al., 2025b), and AIGIBench (Li et al., 2026). We report classification accuracy on clean fake (Acc_{clean}) and adversarial (Acc_{adv}) images, together with the relative accuracy change $\Delta_{\mathrm{acc}} = Acc_{\mathrm{adv}} - Acc_{\mathrm{clean}}$, where a more negative value indicates a stronger attack. Please refer to Appendix A.2 for the complete experimental setup, including implementation details.

4.2 Attack on Forensic Detectors

Main Results. Tab. 1 reports the anti-forensic performance of the proposed method against various detectors across multiple datasets. For CLIP-based detectors, the performance of CLIP+Linear and CLIP+LoRA drops to nearly zero. UniFD, DGS-Net, and Effort likewise suffer substantial degradation despite adapting CLIP in different ways, indicating that the VFM representations these detectors rely on remain vulnerable to our attack. Furthermore, conventional detectors are similarly compromised, showing that the suppression of forensic artifacts renders detectors relying on low-level statistical anomalies ineffective as well. These results demonstrate that both the semantic cues and synthesis-specific forensic artifacts exploited by existing detectors can be effectively disrupted by our carefully crafted transformations. To further evaluate the effectiveness of our method across different VFMs, please refer to Appendix A.1.2 for experiments using CLIP2 as the attack objective.

Cross-Generator Generalization. While ASM is trained on SD-reconstructed images, this does not limit its generalization across generators, as it suppresses artifacts introduced by the fixed SD1.4 reconstruction process shared by all inputs, regardless of their source generators. The cross-generator experiments in Tab. 2, covering both GAN and diffusion architectures, empirically confirm this generalization. Both CLIP-based and conventional detectors suffer substantial performance degradation. Please refer to Appendix A.1.1 for the complete experimental results across detectors.

Table 2: Cross-Generator Generalization. Detection accuracy of two types of detectors on specific subsets of the AIGCDetectBench dataset, covering both Diffusion and GAN architectures.

Diffusion Models						GAN Models					
Generator	Metrics	CLIP-based Detectors		Conventional Detectors		Generator	Metrics	CLIP-based Detectors		Conventional Detectors	
		Effort	CLIP+Linear	GramNet	SAFE			Effort	CLIP+Linear	GramNet	SAFE
ADM	$\text{Acc}_{\text{clean}}$	0.66	0.19	0.16	0.22	BigGAN	$\text{Acc}_{\text{clean}}$	0.98	0.41	0.09	0.55
	$\text{Acc}_{\text{adv}}\downarrow$	**0.16**	**0.04**	**0.00**	**0.08**		$\text{Acc}_{\text{adv}}\downarrow$	**0.10**	**0.04**	**0.00**	**0.04**
DALL·E 2	$\text{Acc}_{\text{clean}}$	0.95	0.84	0.35	0.93	CycleGAN	$\text{Acc}_{\text{clean}}$	0.92	0.51	0.96	0.96
	$\text{Acc}_{\text{adv}}\downarrow$	**0.40**	**0.02**	**0.00**	**0.20**		$\text{Acc}_{\text{adv}}\downarrow$	**0.16**	**0.04**	**0.04**	**0.08**
GLIDE	$\text{Acc}_{\text{clean}}$	0.87	0.73	0.51	0.69	GauGAN	$\text{Acc}_{\text{clean}}$	0.87	0.21	0.01	0.44
	$\text{Acc}_{\text{adv}}\downarrow$	**0.16**	**0.02**	**0.00**	**0.06**		$\text{Acc}_{\text{adv}}\downarrow$	**0.28**	**0.06**	0.02	**0.08**
Midjourney	$\text{Acc}_{\text{clean}}$	0.76	0.70	0.34	0.68	ProGAN	$\text{Acc}_{\text{clean}}$	0.75	0.87	0.93	0.82
	$\text{Acc}_{\text{adv}}\downarrow$	**0.24**	**0.08**	**0.04**	**0.26**		$\text{Acc}_{\text{adv}}\downarrow$	**0.29**	**0.10**	**0.02**	**0.19**
VQ-Diffusion	$\text{Acc}_{\text{clean}}$	0.88	0.37	0.07	0.43	StarGAN	$\text{Acc}_{\text{clean}}$	0.99	0.71	0.89	0.26
	$\text{Acc}_{\text{adv}}\downarrow$	**0.14**	**0.02**	**0.02**	**0.14**		$\text{Acc}_{\text{adv}}\downarrow$	**0.42**	**0.12**	**0.00**	**0.00**
Stable Diffusion v1.4	$\text{Acc}_{\text{clean}}$	0.99	0.98	0.99	0.95	StyleGAN	$\text{Acc}_{\text{clean}}$	0.78	0.46	1.00	0.92
	$\text{Acc}_{\text{adv}}\downarrow$	**0.28**	**0.12**	**0.60**	**0.28**		$\text{Acc}_{\text{adv}}\downarrow$	**0.14**	**0.00**	**0.00**	**0.08**
Stable Diffusion v1.5	$\text{Acc}_{\text{clean}}$	0.99	0.98	0.99	0.95	StyleGAN2	$\text{Acc}_{\text{clean}}$	0.75	0.51	1.00	0.83
	$\text{Acc}_{\text{adv}}\downarrow$	**0.48**	**0.10**	**0.66**	**0.32**		$\text{Acc}_{\text{adv}}\downarrow$	**0.14**	**0.04**	**0.02**	**0.10**
Avg.	$\text{Acc}_{\text{clean}}$	0.87	0.68	0.49	0.69	**Avg.**	$\text{Acc}_{\text{clean}}$	0.86	0.52	0.69	0.68
	$\text{Acc}_{\text{adv}}\downarrow$	**0.26**	**0.05**	**0.18**	**0.19**		$\text{Acc}_{\text{adv}}\downarrow$	**0.21**	**0.05**	**0.01**	**0.08**

Table 3: Comparison between our method and SOTA anti-forensic methods across datasets. Acc_{adv} is reported, with lower values indicating better anti-forensic performance.

Dataset	Anti-forensic Methods	CLIP-based Detectors					Conventional Detectors				
		UniFD	DGS-Net	Effort	CLIP+Linear	CLIP+LoRA	CNNSpot	GramNet	ResNet50	SAFE	NPR
GenImage	DiffAttack	0.71	0.90	0.92	0.68	0.48	0.23	0.24	0.00	0.79	0.19
	StealthDiffusion	0.48	0.35	0.64	0.34	0.57	0.13	0.27	0.00	0.46	0.27
	UDTF	0.93	0.98	0.39	0.29	0.84	0.09	0.26	0.00	0.54	0.04
	FTM	0.71	0.58	0.85	0.64	0.77	0.43	0.56	0.35	0.73	0.70
	FPR	0.35	0.37	0.61	0.36	0.21	0.15	0.24	0.00	0.25	**0.03**
	Ours	**0.33**	**0.33**	**0.22**	**0.05**	**0.03**	**0.08**	**0.22**	**0.00**	**0.20**	0.37
AIGCDetectBench	DiffAttack	0.69	0.90	0.92	0.65	0.36	0.13	0.13	0.00	0.83	0.32
	StealthDiffusion	0.51	0.38	0.68	0.35	0.57	0.08	0.16	0.00	0.41	0.28
	UDTF	0.92	0.98	0.30	0.22	0.88	0.07	0.04	0.00	0.60	0.32
	FTM	0.74	0.72	0.88	0.67	0.83	0.40	0.55	0.31	0.72	0.63
	FPR	0.44	0.49	0.74	0.34	0.16	0.08	**0.03**	0.00	0.17	0.27
	Ours	**0.37**	**0.36**	**0.26**	**0.06**	**0.03**	**0.05**	0.11	**0.00**	**0.14**	**0.25**
AIGIBench	DiffAttack	0.71	0.87	0.97	0.85	0.31	0.07	0.05	0.00	0.87	0.36
	StealthDiffusion	0.55	0.44	0.80	0.66	0.57	0.01	0.03	0.00	0.44	0.26
	UDTF	0.96	0.96	0.49	0.36	0.82	0.01	0.03	0.00	0.56	0.32
	FTM	0.73	0.42	0.88	0.87	0.65	0.29	0.33	0.11	0.55	0.41
	FPR	0.58	0.49	0.83	0.68	0.14	0.03	0.05	0.00	0.15	0.31
	Ours	**0.49**	**0.40**	**0.47**	**0.17**	**0.05**	**0.01**	**0.02**	**0.00**	**0.10**	**0.26**
AIGI-Holmes	DiffAttack	0.70	0.91	0.96	0.85	0.36	0.03	0.05	0.00	0.81	0.38
	StealthDiffusion	0.46	0.32	0.77	0.46	0.57	0.04	0.04	0.00	0.52	0.35
	UDTF	0.93	0.95	0.38	0.23	0.83	0.01	0.02	0.00	0.47	0.37
	FTM	0.68	0.33	0.82	0.75	0.53	0.19	0.31	0.18	0.61	0.58
	FPR	0.35	0.31	0.66	0.46	0.05	0.02	**0.01**	0.00	0.25	0.43
	Ours	**0.26**	**0.31**	**0.31**	**0.13**	**0.04**	**0.01**	0.05	**0.00**	**0.16**	**0.34**

4.3 Comparisons with other Methods

Anti-Forensic Performance Comparison. Tab. 3 presents a performance comparison with SOTA anti-forensic attacks, with lower detection accuracy indicating greater attack effectiveness. UDTF uses real images from GenImage as the training set, while StealthDiffusion uses the ResNet-50 weights released by the authors. To ensure fair comparison, the methods that require surrogate models, namely DiffAttack, FTM, and FPR, all use ResNet-50 as the surrogate model, with all other settings following the original source code. The results show that our method achieves the best anti-forensic performance across both detector families without requiring any detector knowledge.

Image Quality Comparison. To further demonstrate the image quality of our adversarial generation method, we conducted both qualitative and quantitative analyses of the adversarial images produced by the baseline attack method and our proposed method. Fig. 3 shows the generated adversarial images. It can be clearly observed that DiffAttack, UDTF, FTM, and FPR introduce visible noise and local semantic distortion, whereas the images generated by our method and StealthDiffusion exhibit no noticeable noise patterns or distortions. Tab. 4 reports the quantitative results of various methods, including the average PSNR, SSIM, and LPIPS. Although our method performs slightly worse than StealthDiffusion, its image quality still outperforms the other methods while achieving the best anti-forensic performance.

4.4 Analysis of Individual Modules

Further Analysis of CRM. *1) t-SNE Visualization:* Fig. 4(a) shows the t-SNE projections of embeddings obtained by encoding clean real, clean fake, and CRM-processed images (using only the CRM Module) from the GenImage and AIGCDetectBench datasets through the pretrained CLIP encoder.

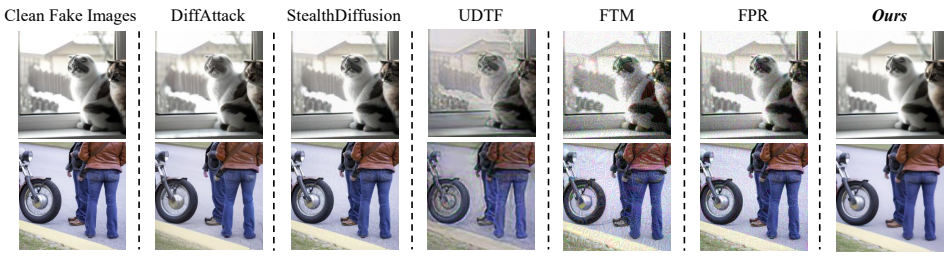


Figure 3: Qualitative assessment of adversarial images generated by DiffAttack, StealthDiffusion, UDTF, FTM, FPR, and our method on the AIGCDetectBench dataset.

Table 4: Quantitative evaluation of image quality. We use AlexNet-based LPIPS following the standard implementation.

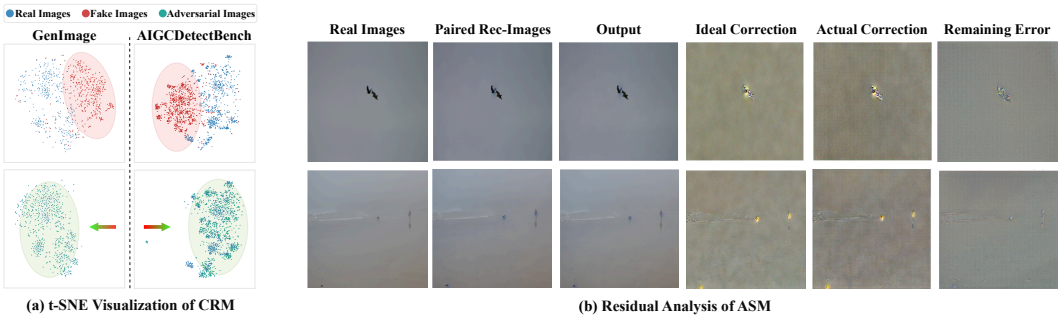
Dataset		DiffAttack	StealthDiffusion	UDTF	FTM	FPR	Ours
GenImage	PSNR ↑	25.11	34.20	19.21	24.53	26.22	32.25
	SSIM ↑	0.74	0.90	0.77	0.52	0.64	0.86
	LPIPS ↓	0.09	0.10	0.16	0.44	0.15	0.09
AIGCDetectBench	PSNR ↑	24.91	34.32	18.98	24.47	26.21	31.91
	SSIM ↑	0.73	0.90	0.76	0.53	0.64	0.85
	LPIPS ↓	0.09	0.09	0.17	0.44	0.13	0.11
AIGIBench	PSNR ↑	25.89	34.44	18.69	24.44	26.25	32.13
	SSIM ↑	0.78	0.90	0.76	0.52	0.64	0.85
	LPIPS ↓	0.08	0.10	0.17	0.46	0.14	0.09
AIGI-Holmes	PSNR ↑	24.33	34.11	18.43	24.46	26.17	30.81
	SSIM ↑	0.77	0.90	0.77	0.52	0.64	0.82
	LPIPS ↓	0.08	0.09	0.17	0.42	0.13	0.10

The pretrained CLIP backbone inherently separates real and fake images into distinct clusters, but the adversarial embeddings (green) directionally migrate toward the real cluster, demonstrating that the CRM can effectively steers fake images to shift semantically closer to real images. *2) Single-Module Ablation Study:* The experimental results in Tab. 5 using only the CRM module show that manipulating semantic evidence alone can effectively degrade the detector's performance, indicating that the semantic evidence from VFM that the detector relies on lacks invariance to carefully crafted transformations.

Further Analysis of ASM. *1) Residual Analysis:* Fig. 4(b) presents the residual analysis of the ASM, where the real image and its paired reconstructed image serve as the training input. The ideal correction is defined as the pixel-wise difference between the real image and the paired reconstructed image, while the actual correction is defined as the pixel-wise difference between the ASM output image and the reconstructed image. The remaining error is defined as the pixel-wise difference between the real image and the ASM output image. It can be observed that the ideal and actual correction maps are consistent, and that the remaining error map is uniformly close to gray, indicating that the ASM output image has effectively suppressed the artifacts introduced by the SD reconstruction. *2) Single-Module Ablation Study:* The experimental results in Tab. 5 using only the ASM module show that manipulating artifact evidence alone can partially degrade the detector's performance, indicating that the artifact evidence that the detector relies on lacks invariance to carefully crafted transformations.

4.5 Analysis and Ablation Study

Ablation Study. The ablation results in Tab. 5 confirm the complementary contributions of CRM and ASM. Without either module, i.e., on clean fake images, detectors achieve an average accuracy of 0.62. Applying CRM or ASM alone reduces this to 0.33 and 0.50, respectively, showing that both latent-space adversarial optimization and artifact suppression can independently weaken detection. Combining both modules further lowers the accuracy to 0.18, demonstrating their complementary and consistently effective anti-forensic contribution. Please see Appendix A.1.3 for ablation results regarding parameters in the experiment.



(a) t-SNE Visualization of CRM

(b) Residual Analysis of ASM

Figure 4: (a) Feature Space Visualization (t-SNE). Projections of CLIP embeddings from the GenImage and AIGCDetectBench datasets, showing clean real and fake images before attack, and adversarial images after attack. (b) Residual Analysis of ASM. Consistency between the ideal and actual correction maps indicates successful suppression of SD-induced artifacts.

Table 5: Ablation study of different components on the GenImage dataset. ✓ and × indicate usage.

Module		CLIP-based Detectors					Conventional Detectors					Avg.
CRM	ASM	UniFD	DGS-Net	Effort	CLIP+Linear	CLIP+LoRA	CNNSpot	GramNet	ResNet50	SAFE	NPR	
×	×	0.70	0.61	0.87	0.64	0.69	0.43	0.54	0.33	0.70	0.69	0.62
✓	×	0.41	0.34	0.54	0.20	0.12	0.33	0.32	0.00	0.29	0.38	0.33
×	✓	0.64	0.58	0.77	0.40	0.54	0.16	0.38	0.31	0.60	0.67	0.50
✓	✓	**0.33**	**0.33**	**0.22**	**0.05**	**0.03**	**0.08**	**0.22**	**0.00**	**0.20**	**0.37**	**0.18**

Computational Efficiency Analysis. Our method operates in a detector-agnostic black-box setting and relies only on a fixed VFM. It requires no access to a surrogate model during generation, and the resulting adversarial images can be used to attack multiple unseen detection models. On 2700 images, the early-stopping mechanism reduces the average number of latent optimization iterations to 64.9, with an observed end-to-end generation time of approximately 13 seconds per image.

Frequency Domain Analysis. Fig. 5 shows the spectral maps of real and fake images from the ADM and StarGAN subsets of AIGCDetectBench, along with the adversarial images generated by our method. It can be clearly observed that the spectral characteristics of our method closely resemble those of real images. Please refer to Appendix A.1.4 for spectral comparisons with other methods.

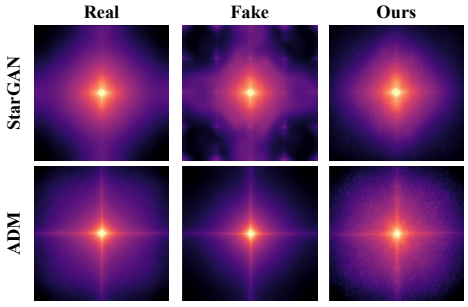


Figure 5: Spectral Analysis.

5 Conclusion and Future Work

In this paper, we categorize the decision cues exploited by existing detectors into two broad types: semantic representations derived from VFMs and synthesis-specific artifacts. We find that independently manipulating either type of evidence leads to substantial degradation in detection performance. Building on this observation, we propose CRAS, a detector-agnostic, black-box anti-forensic framework in which a Cross-Modal Reconstruction Module and an Artifact Suppression Module are cascaded to jointly attack semantic and artifact evidence. Extensive experiments demonstrate that CRAS achieves superior anti-forensic performance over existing methods while preserving high visual fidelity. Our findings reveal that current AI-generated image detectors rely heavily on semantic and synthesis-specific cues that are vulnerable to carefully controlled transformations, providing new insights into their decision-making mechanisms.

In future work, we aim to leverage the insights gained from anti-forensic probing to guide the development of next-generation detectors, and to explore manipulation-invariant forensic evidence that remains stable under adversarial transformations, which could serve as a more reliable foundation for robust detector design.

AI USE STATEMENT

In this work, we used generative AI tools to assist with and polish writing, support research ideation or execution, and draft sections of the paper. We have not used generative AI tools to generate synthetic datasets, and proving mathematical propositions is not applicable to this work. We have reviewed all AI-assisted work. The authors carefully reviewed all sentences polished by large language models and verified and tested LLM-generated code to ensure its correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the proposed method, implementation settings, and experimental protocols in the main paper and Appendix. We will release the complete source code, including training and evaluation scripts, together with representative adversarial image samples upon publication to support independent verification and future research.

REFERENCES

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

Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. Diffusion models for imperceptible and transferable adversarial attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):961–977, 2024b.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, Jim Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *Advances in Neural Information Processing Systems*, 38:48009–48036, 2026.

Feng Ding, Wenhui Yi, Xinan He, Mengyao Xiao, Jianfeng Xu, and Jianqiang Du. Difffaceedit: A diffusion-based facial dataset for forgery-semantic driven deepfake detection analysis. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 14117–14121. IEEE, 2026a.

Feng Ding, Wenhui Yi, Yunpeng Zhou, Xinan He, Hong Rao, and Shu Hu. Decoupling bias, aligning distributions: Synergistic fairness optimization for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32704–32713, 2026b.

Bing Fan, Zihan Jiang, Shu Hu, and Feng Ding. Attacking identity semantics in deepfakes via deep feature fusion. In *2023 IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 114–119. IEEE, 2023.

Hao Fang, Jiawei Kong, Wenbo Yu, Bin Chen, Jiawei Li, Hao Wu, Shu-Tao Xia, and Ke Xu. One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4090–4100. IEEE, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Xinan He, Yue Zhou, Bing Fan, Bin Li, Guopu Zhu, and Feng Ding. Vlforgery face triad: Detection, localization and attribution via multimodal large language models. *arXiv preprint arXiv:2503.06142*, 2025.

Jiewei Lai, Lan Zhang, Chen Tang, Pengcheng Sun, Xinming Wang, and Yunhao Wang. Untraceable deepfakes via traceable fingerprint elimination. In *International Conference on Learning Representations*, volume 2026, pp. 88489–88513, 2026.

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.

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, 2026.

Kaisheng Liang, Xuelong Dai, Yanjie Li, Dong Wang, and Bin Xiao. Improving transferable targeted attacks with feature tuning mixup. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25802–25811. IEEE, 2025.

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

Zhengzhe Liu, Xiaojuan Qi, and Philip H.S. Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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 (CVPR)*, pp. 24480–24489, June 2023.

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

Yuchen Ren, Zhengyu Zhao, Chenhao Lin, Bo Yang, Lu Zhou, Zhe Liu, and Chao Shen. Improving adversarial transferability on vision transformers via forward propagation refinement. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25071–25080. IEEE, 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 (CVPR)*, pp. 28130–28139, June 2024.

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 (CVPR)*, June 2020.

Jiazhen Yan, Ziqiang Li, Fan Wang, Boyu Wang, Ziwen He, and Zhangjie Fu. DGS-net: Distillation-guided gradient surgery for CLIP fine-tuning in AI-generated image detection. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=G5XGej7wNt.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=GFpjO8S8Po.

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

Yue Zhou, Xinan He, KaiQing Lin, Bin Fan, Feng Ding, and Bin Li. Breaking latent prior bias in detectors for generalizable aigc image detection. *arXiv preprint arXiv:2506.00874*, 2025a.

Ziyin Zhou, Ke Sun, Zhongxi Chen, Huafeng Kuang, Xiaoshuai Sun, and Rongrong Ji. Stealthdiffusion: Towards evading diffusion forensic detection through diffusion model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3627–3636, 2024.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18746–18758. IEEE, 2025b.

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.

Table 6: Additional Cross-Generator Generalization. Detection accuracy of two types of detectors on specific subsets of the AIGCDetectBench dataset, covering both Diffusion and GAN architectures.

Diffusion Models								GAN Models							
Generator	Metrics	CLIP-based Detectors			Conventional Detectors			Generator	Metrics	CLIP-based Detectors			Conventional Detectors		
		UniFD	DGS-Net	CLIP+LoRA	CNNSpot	ResNet50	NPR			UniFD	DGS-Net	CLIP+LoRA	CNNSpot	ResNet50	NPR
ADM	$\text{Acc}_{\text{clean}}$	0.21	0.46	0.16	0.02	0.16	0.25	BigGAN	$\text{Acc}_{\text{clean}}$	0.68	1.00	0.51	0.02	0.06	0.17
	$\text{Acc}_{\text{adv}} \downarrow$	0.16	0.30	0.00	0.00	0.00	0.10		$\text{Acc}_{\text{adv}} \downarrow$	0.38	0.36	0.06	0.00	0.00	0.12
DALL·E 2	$\text{Acc}_{\text{clean}}$	0.62	0.45	0.85	0.05	0.36	0.89	CycleGAN	$\text{Acc}_{\text{clean}}$	0.80	1.00	0.85	0.18	0.05	0.27
	$\text{Acc}_{\text{adv}} \downarrow$	0.08	0.28	0.00	0.00	0.00	0.28		$\text{Acc}_{\text{adv}} \downarrow$	0.58	0.44	0.08	0.02	0.00	0.14
GLIDE	$\text{Acc}_{\text{clean}}$	0.75	0.60	0.83	0.05	0.45	0.70	GauGAN	$\text{Acc}_{\text{clean}}$	0.76	1.00	0.44	0.03	0.01	0.02
	$\text{Acc}_{\text{adv}} \downarrow$	0.26	0.48	0.06	0.00	0.00	0.06		$\text{Acc}_{\text{adv}} \downarrow$	0.48	0.42	0.04	0.00	0.00	0.14
Midjourney	$\text{Acc}_{\text{clean}}$	0.65	0.19	0.41	0.22	0.22	0.71	ProGAN	$\text{Acc}_{\text{clean}}$	0.97	0.99	0.98	0.95	0.45	0.98
	$\text{Acc}_{\text{adv}} \downarrow$	0.22	0.32	0.02	0.00	0.00	0.42		$\text{Acc}_{\text{adv}} \downarrow$	0.68	0.49	0.08	0.09	0.00	0.19
VQ-Diffusion	$\text{Acc}_{\text{clean}}$	0.35	0.83	0.49	0.01	0.11	0.14	StarGAN	$\text{Acc}_{\text{clean}}$	0.86	1.00	0.95	0.23	0.83	0.06
	$\text{Acc}_{\text{adv}} \downarrow$	0.14	0.26	0.00	0.00	0.00	0.20		$\text{Acc}_{\text{adv}} \downarrow$	0.58	0.44	0.08	0.00	0.00	0.00
SD v1.4	$\text{Acc}_{\text{clean}}$	0.98	0.64	0.99	1.00	0.47	1.00	StyleGAN	$\text{Acc}_{\text{clean}}$	0.51	0.98	0.55	0.27	0.26	0.67
	$\text{Acc}_{\text{adv}} \downarrow$	0.50	0.30	0.02	0.30	0.00	0.64		$\text{Acc}_{\text{adv}} \downarrow$	0.12	0.24	0.00	0.00	0.00	0.16
SD v1.5	$\text{Acc}_{\text{clean}}$	0.98	0.63	0.99	1.00	0.49	1.00	StyleGAN2	$\text{Acc}_{\text{clean}}$	0.41	0.94	0.56	0.17	0.29	0.58
	$\text{Acc}_{\text{adv}} \downarrow$	0.54	0.40	0.08	0.22	0.00	0.68		$\text{Acc}_{\text{adv}} \downarrow$	0.14	0.30	0.00	0.00	0.00	0.08
Avg.	$\text{Acc}_{\text{clean}}$	0.65	0.54	0.67	0.34	0.32	0.67	**Avg.**	$\text{Acc}_{\text{clean}}$	0.71	0.99	0.69	0.26	0.28	0.39
	$\text{Acc}_{\text{adv}} \downarrow$	**0.27**	**0.33**	**0.02**	**0.07**	**0.00**	**0.34**		$\text{Acc}_{\text{adv}} \downarrow$	**0.42**	**0.38**	**0.04**	**0.02**	**0.00**	**0.12**

Table 7: Anti-forensic performance of the proposed method against CLIP2-based detectors across datasets. Negative Δ_{acc} indicates reduced detection accuracy and attack success.

Method	Metrics	Dataset				Avg.
		GenImage	AIGCDetectBench	AIGIBench	AIGI-Holmes	
CLIP2+Linear	$\text{Acc}_{\text{clean}}$	0.76	0.90	0.84	0.86	0.84
	$\text{Acc}_{\text{adv}} \downarrow$	0.05	0.03	0.07	0.09	0.06
	$\boldsymbol{\Delta}_{\text{acc}} \downarrow$	-0.71	-0.87	-0.77	-0.77	-0.78
CLIP2-ZeroShot	$\text{Acc}_{\text{clean}}$	0.79	0.86	0.86	0.87	0.84
	$\text{Acc}_{\text{adv}} \downarrow$	0.03	0.02	0.04	0.04	0.03
	$\boldsymbol{\Delta}_{\text{acc}} \downarrow$	-0.76	-0.84	-0.82	-0.83	-0.81

A Appendix

A.1 Additional Experimental Results

A.1.1 Cross-Generator Generalization

Tab. 6 reports the detection results of clip-based detectors and conventional detectors on different subsets within the AIGCDetectBench dataset, where different subsets are generated by different generators. The experimental results demonstrate that our method can consistently cause a significant drop in detector performance on adversarial images created by different generators.

A.1.2 Experimental Results Targeting CLIP2

Tab. 7 reports the anti-forensic performance of adversarial images crafted against CLIP2 as the attack target, evaluated on CLIP2-based detectors. CLIP2+Linear freezes the pretrained CLIP2 visual backbone and trains only a linear classification head for real/fake discrimination, while CLIP2-ZeroShot uses the frozen backbone's features directly, performing detection via a KNN classifier built from real and fake text anchors. The results demonstrate that our method achieves consistent attack effectiveness against different VFMs.

A.1.3 Parameter Ablation

The parameter ablation in Tab. 8 shows that the optimal configuration is $\lambda_{\text{mse}} = 0.1$ and $\lambda_{\text{align}} = 0.9$. Since diffusion reconstruction relies on a deterministic DDIM inversion process, which already strongly constrains the quality of the reconstructed image, increasing λ_{mse} yields only marginal improvements in image quality. In contrast, increasing λ_{align} strengthens the semantic-alignment constraint, leading to improved anti-forensic performance.

Table 8: Ablation study on the weighting parameters λ_{mse} and λ_{align} of the Cross-Modal Reconstruction Module against the NPR detector on the GenImage dataset.

Params.		$\text{ACC}_{\text{adv}} \downarrow$	SSIM ↑	PSNR ↑
λ_{mse}	λ_{align}			
0.9	0.1	0.48	**0.88**	**32.44**
0.7	0.3	0.47	0.87	32.38
0.5	0.5	0.45	0.86	32.31
0.3	0.7	0.44	0.86	32.28
0.1	**0.9**	**0.37**	0.86	32.25

A.1.4 Additional Frequency Domain Analysis

To further analyze the spectral characteristics of different attack methods, we conduct a qualitative analysis of the spectral properties of adversarial images produced by baseline attack methods and our proposed method. As clearly shown in Fig. 6, among the various attack strategies, our method most closely approximates the spectral signature of genuine images, whereas the other methods introduce unnatural spectral features.

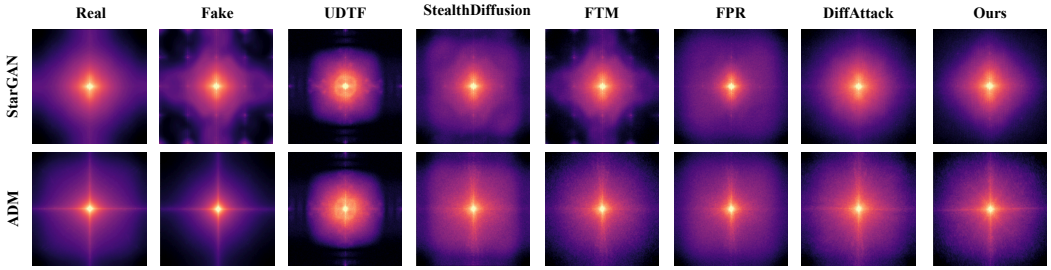


Figure 6: Complete Comparison of Spectral Characteristics.

A.2 Complete Experimental Setup

Victim Models. To comprehensively evaluate the adversarial effectiveness of the proposed framework, we select CLIP as the target VFM, given its widespread use as a feature extraction backbone in detectors. The attacked detectors span CLIP-based detectors and conventional detectors. The CLIP-based baselines include UniFD (Ojha et al., 2023), DGS-Net (Yan et al., 2026), Effort (Yan et al., 2025), CLIP+Linear, and CLIP+LoRA. CLIP+Linear freezes the pretrained CLIP visual encoder and trains only a linear head for real/fake classification, while CLIP+LoRA fine-tunes the CLIP visual backbone using LoRA with a rank of 4, an alpha of 8, and a dropout rate of 0.1. Conventional detectors include CNNSpot (Wang et al., 2020), GramNet (Liu et al., 2020), ResNet50 (He et al., 2016), SAFE (Li et al., 2025), and NPR (Tan et al., 2024), covering diverse forensic paradigms based on spatial-domain cues, texture statistics, and frequency-domain artifacts. CLIP+Linear and CLIP+LoRA are trained on the SD1.4 subset of GenImage, while all other models use officially released pretrained weights. For comparison with other anti-forensic methods, we select DiffAttack (Chen et al., 2024b), UDTF (Lai et al., 2026), FTM (Liang et al., 2025), FPR (Ren et al., 2025) and StealthDiffusion (Zhou et al., 2024).

Dataset. To comprehensively evaluate the consistent effectiveness of the proposed attack method, we conduct experiments on four representative datasets, including GenImage (Zhu et al., 2023), AIGCDetectBench (Zhong et al., 2023), AIGI-Holmes (Zhou et al., 2025b), and AIGIBench (Li et al., 2026). Tab. 9 reports the specific generative models for all datasets.

Evaluation Metrics. For evaluation metrics, we adopt the classification accuracy on clean fake images (Acc_{clean}) as the performance baseline of the target detector. We then evaluate attack effectiveness using the classification accuracy on adversarial images (Acc_{adv}) and the relative accuracy

Table 9: Overview of the datasets used in our experiments.

Dataset	Generative Models	Total
GenImage	ADM, BigGAN, Glide, Midjourney, SD v1.4, SD v1.5, VQDM, Wukong	8
AIGCDetectBench	ADM, DALL·E2, Glide, Midjourney, VQDM, BigGAN, CycleGAN, GauGAN, ProGAN, SD v1.4, SD v1.5, StarGAN, StyleGAN, StyleGAN2, WhichFaceIsReal, Wukong	16
AIGIBench	BLIP, BlendFace, CommunityAI, DALL·E-3, E4S, FLUX1-dev, FaceSwap, GLIDE, IP-Adapter, Imagen3, InSwap, Infinite-ID, InstantID, Midjourney, PhotoMaker, ProGAN, R3GAN, SD3, SDXL, SimSwap, SocialRF, StyleGAN-XL, StyleGAN3, StyleSwim, WFIR	25
AIGI-Holmes	FLUX, Infinity, Janus, Janus-Pro-1B, Janus-Pro-7B, LlamaGen, PixArt-XL, SD35-L, Show-o, VAR	10

change $\Delta_{\text{acc}} = Acc_{\text{adv}} - Acc_{\text{clean}}$, where a more negative Δ_{acc} indicates a larger degradation in detection accuracy and thus a more effective attack.

Implementation Details. In the Cross-Modal Reconstruction Module, to preserve the reconstruction quality, the input clean images are first resized to 512×512 and then fed into the pre-trained diffusion model. Specifically, we adopt Stable Diffusion v1.4 as the reconstruction model and apply 2 diffusion steps. The latent optimization is performed using the Adam optimizer with a learning rate of 1×10^{-3}, where the temperature coefficient is set to $\tau = 0.1$, and the loss weights are configured as $\lambda_{mse} = 0.1$ and $\lambda_{align} = 0.9$. The maximum optimization iterations are set to 100, and the optimization is terminated early when the generated image achieves the highest similarity with the real-text anchor. In the Artifact Suppression Module, the suppression network is trained for 20 epochs with the loss weight $\lambda_{L1} = 100$.

A.3 ASM Network Architecture

The architectural details of the ASM are as follows: 1) **Autoencoder**: The autoencoder adopts a U-Net architecture with a symmetric 8-layer encoder-decoder structure and skip connections. Each layer uses a 4×4 kernel with stride 2 and padding 1. The encoder progressively downsamples the input while increasing the channel depth up to 512, with each layer followed by batch normalization and LeakyReLU(0.2), except the first layer, which omits normalization. The decoder mirrors this structure via transposed convolutions, upsampling back to the original resolution while reducing channel depth, with skip connections preserving fine-grained details at each level. Each decoder layer is followed by batch normalization and ReLU, with dropout in the first three layers, except the final layer, which uses Tanh to constrain outputs to $[-1, 1]$. 2) **Discriminator**: The discriminator adopts a PatchGAN architecture that classifies overlapping 70×70 patches as real or fake. It consists of five 4×4 convolutional layers: the first three use stride 2 to downsample while increasing channel depth to 256, the fourth uses stride 1 to further increase the channel depth to 512, and the final layer uses stride 1 to produce a single-channel patch-wise prediction with no activation, yielding raw logits for the adversarial loss. We summarize the architecture of both models in Table 10 and Table 11.

A.4 Text Anchors Used in the CRM

To guide the reconstructed images toward the real image manifold in the CLIP embedding space, we construct two sets of textual anchors: a real-related text set and a fake-related text set. The real-related set comprises a diverse vocabulary describing authentic photographic attributes, spanning capture sources (e.g., *camera*, *smartphone*, *film*), photographic genres (e.g., *portrait*, *landscape*, *street*, *wildlife*), visual properties (e.g., *natural*, *authentic*, *realistic*, *unedited*, *raw*), and contextual descriptors (e.g., *documentary*, *candid*, *news*, *personal*, *travel*), covering 38 terms in total. The fake-related set instead consists of prompts directly denoting synthetic origin, namely *"AI generated"* and *"text to image"*. Each text anchor is independently encoded using the frozen CLIP text encoder, and the resulting embeddings are normalized before optimization. Given an image reconstructed

Table 10: Autoencoder Architecture

Component	Layer Type	Kernel Size	Stride	Output Shape	Activation/Normalization
Encoder	Conv2d	4×4	2	$64 \times \frac{H}{2} \times \frac{W}{2}$	ReLU
	Conv2d	4×4	2	$128 \times \frac{H}{4} \times \frac{W}{4}$	ReLU + BatchNorm
	Conv2d	4×4	2	$256 \times \frac{H}{8} \times \frac{W}{8}$	ReLU + BatchNorm
	Conv2d	4×4	2	$512 \times \frac{H}{16} \times \frac{W}{16}$	ReLU + BatchNorm
	Conv2d	4×4	2	$512 \times \frac{H}{32} \times \frac{W}{32}$	ReLU + BatchNorm
	Conv2d	4×4	2	$512 \times \frac{H}{64} \times \frac{W}{64}$	ReLU + BatchNorm
	Conv2d	4×4	2	$512 \times \frac{H}{128} \times \frac{W}{128}$	ReLU + BatchNorm
	Bottleneck	4×4	2	$512 \times \frac{H}{256} \times \frac{W}{256}$	ReLU
Decoder	ConvTranspose2d	4×4	2	$512 \times \frac{H}{128} \times \frac{W}{128}$	ReLU + BatchNorm
	ConvTranspose2d	4×4	2	$512 \times \frac{H}{64} \times \frac{W}{64}$	ReLU + BatchNorm + Dropout
	ConvTranspose2d	4×4	2	$512 \times \frac{H}{32} \times \frac{W}{32}$	ReLU + BatchNorm + Dropout
	ConvTranspose2d	4×4	2	$512 \times \frac{H}{16} \times \frac{W}{16}$	ReLU + BatchNorm + Dropout
	ConvTranspose2d	4×4	2	$256 \times \frac{H}{8} \times \frac{W}{8}$	ReLU + BatchNorm
	ConvTranspose2d	4×4	2	$128 \times \frac{H}{4} \times \frac{W}{4}$	ReLU + BatchNorm
	ConvTranspose2d	4×4	2	$64 \times \frac{H}{2} \times \frac{W}{2}$	ReLU + BatchNorm
	Output Layer	4×4	2	$C_{\text{out}} \times H \times W$	Tanh

Table 11: Discriminator Architecture

Component	Layer Type	Kernel Size	Stride	Output Shape	Activation/Normalization
PatchGAN	Conv2d	4×4	2	$64 \times \frac{H}{2} \times \frac{W}{2}$	ReLU
	Conv2d	4×4	2	$128 \times \frac{H}{4} \times \frac{W}{4}$	ReLU + BatchNorm
	Conv2d	4×4	2	$256 \times \frac{H}{8} \times \frac{W}{8}$	ReLU + BatchNorm
	Conv2d	4×4	1	$512 \times (\frac{H}{8} - 1) \times (\frac{W}{8} - 1)$	ReLU + BatchNorm
	Output Layer	4×4	1	$1 \times (\frac{H}{8} - 2) \times (\frac{W}{8} - 2)$	—

by Stable Diffusion, its visual representation is extracted using the frozen CLIP image encoder. We then optimize the reconstructed image to increase its cosine similarity to the real-related text anchors while decreasing its similarity to the fake-related anchors. Tab. 12 records the complete text pool.

Table 12: Complete list of text anchors used to construct the real-text and fake-text anchors in the CRM.

Category	Text Anchors	Total
Real-related	natural, authentic, genuine, original, real, camera, professional, documentary, candid, film, smartphone, landscape, portrait, street, nature, urban, indoor, outdoor, wildlife, architectural, macro, sharp, lighting, exposure, unedited, raw, human, texture, realistic, organic, physical, environment, news, social, stock, personal, family, travel	38
Fake-related	AI generated, text to image	2