TIGA: Trajectory-Injected Generative Attack against Black-box AIGC Detectors

Xia Du, Zhuosen Bao, Zheng Lin, Jizhe Zhou, Jiawei Lian, Chi-man Pun, *Senior Member, IEEE*, Jun Luo, *Fellow, IEEE*, Wei Ni, *Fellow, IEEE*, and Symeon Chatzinotas, *Fellow, IEEE*

***Abstract*—Recent diffusion models have achieved remarkable realism in facial image synthesis, posing growing challenges to artificial intelligence-generated content (AIGC) forensic detectors. Existing evasion methods typically perturb pre-generated images or require detector-aware training, which may introduce visible or statistical artifacts and limit applicability when the diffusion model must remain frozen and the target detector is accessible only through black-box queries. We propose Trajectory-Injected Generative Attack (TIGA), a source-image-free and training-free framework that generates detector-evasive images within a single diffusion sampling trajectory. TIGA steers the latent Denoising Diffusion Implicit Model (DDIM) trajectory so that adversarial properties emerge during generation rather than being added afterward. TIGA first aggregates gradients from multiple white-box surrogate detectors to form a transferable, sign-aware prior, and then performs anisotropic directional search with symmetric finite-difference queries to estimate the black-box target response. The estimated directions are stabilized by decayed momentum and injected according to the DDIM noise schedule, with frequency-domain reshaping to suppress high-frequency artifacts. Experiments on surrogate and unseen specialized forensic detectors show that TIGA achieves strong black-box attack performance, transferability, and high robustness under common post-processing operations without source images or diffusion-model retraining, while preserving high perceptual quality.**

***Index Terms*—Adversarial attack, AIGC detection, Diffusion Models, Trajectory Injection.**

I. Introduction

RECENT diffusion models [1]–[4] have achieved remarkable progress in image synthesis, enabling the generation of highly realistic artificial intelligence (AI)-generated images that are increasingly difficult to distinguish from real images by human observers [5]–[9]. Benefiting from their powerful generative capability, diffusion-based Artificial Intelligence Generated Content (AIGC) technologies have been widely adopted in digital content creation, virtual avatars, image editing, and multimedia generation. However, the rapid proliferation of realistic AIGC has raised serious concerns regarding misinformation, malicious forgery, identity manipulation, and media credibility. To mifigate these risks, a wide range of AIGC forensic detectors have been proposed, including Convolutional Neural Network (CNN)-based detectors that learn manipulation-sensitive spatial or residual cues [10]–[15], vision-transformer and foundation-model-based detectors that exploit pretrained visual representations [16]–[20], frequency-domain methods that capture spectral artifacts introduced by image synthesis pipelines [21], [22], and diffusion-reconstruction-based approaches that distinguish real and generated images through reconstruction discrepancies [23], [24].

Recent studies have shown that AIGC detectors remain vulnerable to adversarial evasion. Existing adversarial evasion methods can be generally divided into two categories. The first category performs post-hoc perturbation on already generated images, where adversarial signals are optimized in the pixel domain, frequency domain, or through realistic post-processing operations after image synthesis [25]–[27]. Although such perturbations are typically bounded by an ℓ_p budget or a trade-off parameter, these controls only limit their magnitude without keeping them on the diffusion manifold: tightening them trades attack strength for quality along a fixed frontier, and the residual may still cause visible artifacts, abnormal frequency patterns, or degraded quality within the budget, while remaining sensitive to image transformations and post-processing.

The second category moves beyond direct post-hoc perturbation by leveraging diffusion models for detector-evasive image generation or refinement [28], [29]. By optimizing detector-evasive objectives within diffusion-based pipelines, these methods can alleviate the artifact problem caused by post-hoc perturbation and improve the visual fidelity of adversarial images. However, existing methods typically require detector-aware retraining, partial finetuning of diffusion components, or post-generation optimization initialized from pre-existing source images, making them difficult to apply in source-free black-box scenarios where no original image is available, and the diffusion model must remain frozen.

In practice, a more realistic threat model arises when

Xia Du and Zhuosen Bao are with the School of Computer and Information Engineering, Xiamen University of Technology, Xiamen, 361000, China (email: baozhuosen@stu.xmut.edu.cn; duxia@xmut.edu.cn; szzhu@xmut.edu.cn).

Zheng Lin and Symeon Chatzinotas are with the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, L-1855 Luxembourg, Luxembourg (e-mail: zhenglin@ieee.org; symeon.chatzinotas@uni.lu).

Jizhe Zhou is with the School of Computer Science, Engineering Research Center of Machine Learning and Industry Intelligence, Sichuan University, Chengdu, China, 610020, China (email: jzzhou@scu.edu.cn).

Jiawei Lian is with the PCA Laboratory, Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: lianjw@njust.edu.cn).

Chi-man Pun is with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, 999078, China (email: cmpun@umac.mo).

Wei Ni is with the School of Engineering, Edith Cowan University, Perth, WA 6027, Australia (email: wei.ni@ieee.org).

Jun Luo is with the College of Computing and Data Science, Nanyang Technological University, Singapore (e-mail: junluo@ntu.edu.sg).

Corresponding authors: Zheng Lin (zhenglin@ieee.org).

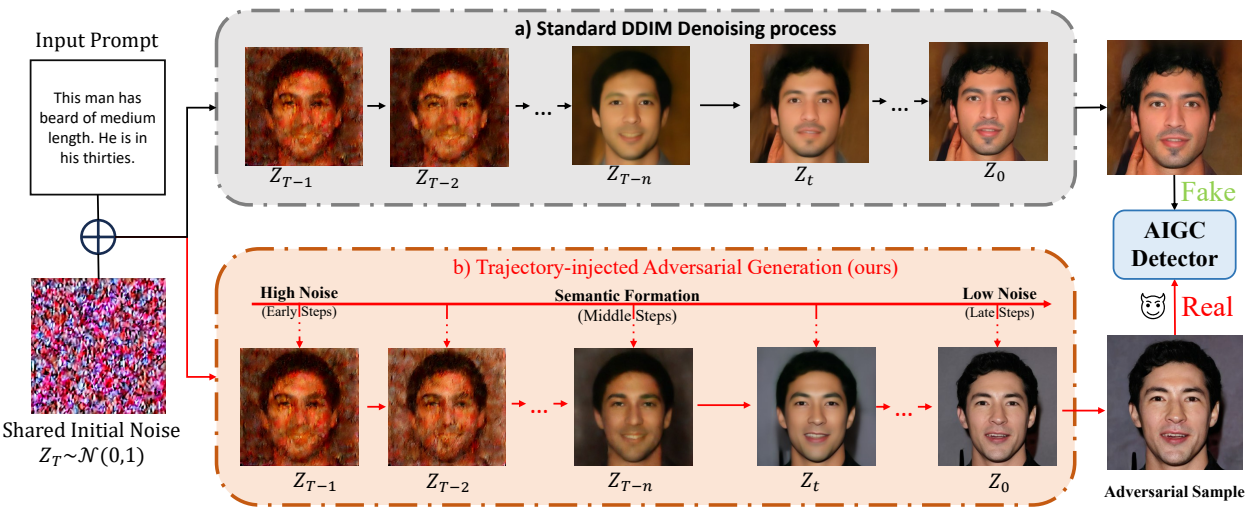


Fig. 1: Conceptual principle of TIGA. Standard DDIM sampling generates visually realistic AIGC images that can still be identified as fake by forensic detectors. In contrast, TIGA introduces trajectory-level adversarial guidance during diffusion sampling. Surrogate detectors first provide a transferable prior direction, while black-box symmetric probing refines the trajectory direction according to the target detector response. The resulting schedule-aware and frequency-shaped injection shifts detector decisions from fake to real while preserving visual quality and perceptual naturalness. EMA denotes exponential moving average.

attackers can query the output confidence of a target AIGC detector but cannot access its internal architecture, parameters, or gradients. Meanwhile, several accessible surrogate detectors may provide transferable forensic priors. Under this setting, a key question remains underexplored: can detector-evasive images be generated directly within a single diffusion sampling trajectory, starting only from random noise and conditional inputs, without relying on pre-existing source images or retraining the diffusion model? This source-image-free black-box setting is challenging for two reasons. On the one hand, direct zeroth-order search in the high-dimensional latent space incurs prohibitive query costs and often yields unstable direction estimates. On the other hand, naively injecting adversarial guidance into the denoising trajectory may distort the generated content or introduce unnatural high-frequency artifacts, particularly when the guidance is inconsistent with the diffusion noise schedule.

To address these challenges, we propose Trajectory-Injected Generative Attack (TIGA), a training-free trajectory-injected adversarial generation framework against black-box AIGC detectors via surrogate-guided directional search. Instead of perturbing generated images after synthesis, TIGA directly steers the Denoising Diffusion Implicit Model (DDIM) denoising trajectory in the latent space, allowing detector-evasive properties to emerge during the generation process. As illustrated in Fig. 1, TIGA shifts the detector decision by modifying the intermediate sampling trajectory rather than attaching external perturbations to the final image, thereby keeping the result a naturally generated sample rather than an artifact-laden image. The diffusion model remains completely frozen throughout sampling, and the target detector is only queried in a black-box manner.

The key idea of TIGA is to combine transferable white-box surrogate guidance with prior-guided black-box directional search, thereby avoiding detector-aware retraining and undirected isotropic zero-order exploration. Specifically, TIGA consists of three tightly coupled modules. First, a Surrogate-Guided Prior (SGP) module constructs a transferable adversarial direction by aggregating latent-space gradients from multiple white-box surrogate detectors. This prior direction provides both transferable detector-evasive knowledge and a sign-aware search orientation, enabling the subsequent black-box search to start from a more informative region rather than from random isotropic exploration. Second, a Surrogate-Guided Directional Search (SGDS) module performs black-box trajectory guidance under an anisotropic direction distribution centered around the surrogate prior. By combining the surrogate-guided direction with orthogonal exploration components, SGDS estimates the response of the black-box detector through symmetric finite-difference queries and aggregates the response-weighted probing directions into a stable trajectory guidance vector. This design generalizes conventional isotropic zero-order search as a boundary case while significantly reducing estimation variance under a fixed query budget. Third, a Schedule-Aware Trajectory Injection (SATI) module injects the accumulated momentum into the DDIM trajectory according to the diffusion noise schedule. The injected update is further reshaped in the frequency domain to suppress high-frequency artifacts and remove channel-wise bias, improving visual fidelity while weakening detector-sensitive forensic traces.

Different from post-hoc perturbation methods, TIGA does not attach an external adversarial pattern to the final image. Instead, it modifies the intermediate denoising trajectory so that the final sample remains naturally generated by the diffusion model, gaining evasiveness without the pixel-space residual that forces post-hoc methods to trade attack strength for quality. Unlike retraining-based adversarial generation

methods, TIGA does not require optimizing or finetuning any component of the diffusion model. Moreover, compared with purely random black-box search, the proposed surrogate-guided anisotropic search yields more stable and reliable direction estimates by constraining the exploration space around transferable forensic directions.

The main contributions of this paper are summarized as follows:

- We propose TIGA, a training-free trajectory-injected adversarial generation framework against black-box AIGC detectors. Without retraining diffusion models or accessing gradients of the target detector, TIGA directly manipulates DDIM denoising trajectories during sampling, allowing detector-evasive properties to emerge as part of the generation process.
- We integrate prior-guided black-box search into the diffusion sampling trajectory: gradients from multiple white-box surrogate detectors are aggregated into a transferable, sign-aware prior that only biases the search direction, while the black-box queries determine the actual update; the symmetric finite-difference query estimates are accumulated across steps into a stable trajectory guidance signal, coupling the per-step search with the generation process rather than attacking a fixed image.
- We develop a schedule-aware trajectory injection strategy with frequency-domain artifact suppression. The proposed injection adaptively scales adversarial guidance according to the DDIM noise schedule and reshapes the injected update in the frequency domain, thereby balancing attack effectiveness, perceptual naturalness, and visual fidelity.

II. Related Work

A. AIGC Image Detection

With the rapid development of generative models, especially generative adversarial networks (GANs) [30]–[32] and diffusion models [2], [4], [33]–[37], detecting AI-generated images has become an important problem in multimedia forensics. Early studies mainly focused on detecting images synthesized by GAN-based generators, where CNNs were trained to capture spatial artifacts and statistical inconsistencies left by image synthesis models [10], [13], [38]–[40]. These methods demonstrated that generated images often contain model-specific or generator-agnostic forensic traces, making it possible to distinguish real and fake images even when the visual appearance is highly realistic.

Beyond spatial-domain artifacts, many works have explored frequency-domain cues for AIGC image detection. Since generative models may introduce abnormal spectral distributions, periodic patterns, or high-frequency inconsistencies during image synthesis, frequency-aware detectors attempt to identify fake images by analyzing Fourier spectra, local frequency statistics, or texture-level forensic traces [21], [22], [41], [42]. These methods are particularly relevant to diffusion-generated images, as subtle frequency fingerprints may remain even when pixel-level visual artifacts are difficult to observe. However, frequency-based detectors may be sensitive to image compression, resizing, and adversarial frequency manipulation.

Recently, transformer-based and foundation-model-based detectors have been introduced to improve the generalization ability of AIGC detection. Compared with conventional CNN-based detectors, vision transformers and Contrastive Language–Image Pre-training (CLIP)-based representations can capture more global semantic and structural cues, which may help detect images generated by unseen models or under cross-domain settings [16], [18], [43], [44]. In addition, recent studies and benchmarks have systematically examined fake-image detection and localization across different generator families, image domains, post-processing conditions, and increasingly realistic manipulation scenarios [45]–[49]. These studies show that although existing forensic models achieve promising performance on in-distribution test data, their robustness and generalization can degrade significantly under unseen generators, cross-domain data, image transformations, or complex manipulation scenarios.

Another line of research detects diffusion-generated images based on reconstruction behavior. For example, diffusion-reconstruction-based methods observe that images generated by diffusion models can often be reconstructed more faithfully by a pretrained diffusion model than real images, and therefore use reconstruction errors as forensic representations [23], [24], [50], [51]. Such methods provide a new perspective for identifying diffusion-generated content and have shown strong detection performance in several settings. Since these detectors rely on specific reconstruction discrepancies or forensic traces, they may still be vulnerable when the generation trajectory is intentionally steered to suppress detector-sensitive cues.

These existing AIGC detectors exploit diverse forensic evidence, including spatial artifacts, frequency statistics, semantic representations, and diffusion reconstruction errors. However, the increasing realism of diffusion-generated images and the emergence of adversarial evasion methods raise new concerns about the reliability of these detectors. This motivates the study of adversarial generation methods that can directly challenge AIGC detectors under realistic black-box settings.

B. Adversarial Attacks against AIGC Image Detectors

Adversarial attacks were initially studied in conventional image classification, where carefully optimized perturbations were introduced to cause erroneous model predictions. Representative white-box methods include Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and the C&W attack [52]–[54]. When the target model is inaccessible, transfer-based attacks craft adversarial examples using surrogate models and exploit cross-model transferability [5], [55], [56], whereas query-based attacks estimate adversarial directions directly from model outputs through zeroth-order optimization, evolutionary strategies, or random search [57]–[60]. Although query-based attacks avoid access to target gradients, estimating reliable directions in high-dimensional image spaces often requires a substantial number of model queries. To improve query efficiency, a line of prior-guided black-box attacks combines the two paradigms using surrogate

gradients as a search prior: Guided ES augments random search with surrogate-gradient directions to shape the sampling distribution [61], and P-RGF further fuses a transfer-based prior with symmetric finite-difference queries to the target model, substantially reducing the query cost of gradient estimation [62]. Our surrogate-guided directional search builds on this prior-guided paradigm and extends it from perturbing a fixed image to steering a diffusion sampling trajectory.

Recent studies have extended adversarial evasion to DeepFake and AIGC image detectors. Early work demonstrated that synthetic-image detectors can be bypassed under both white-box and black-box settings using carefully optimized additive perturbations [25], [63]. Subsequent methods increasingly exploit forensic properties specific to generated images. StatAttack minimizes the statistical discrepancy between real and fake images by adversarially optimizing natural degradations such as exposure, blur, and noise [64]. FPBA introduces frequency-domain perturbations and a post-train Bayesian surrogate strategy to improve transferability across heterogeneous AIGC detectors [26]. R^2BA performs score-based black-box optimization over realistic post-processing operations, including Gaussian blur, JPEG compression, Gaussian noise, and illumination effects [27].

Most of these methods follow a post-hoc paradigm: they require a pre-generated fake image and optimize pixel-domain, frequency-domain, or post-processing transformations to change the detector prediction. Since the adversarial modification is performed after image synthesis, the resulting signal is not inherently produced by the original generative trajectory. Conventional perturbation-based attacks may therefore exhibit limited visual invisibility, cross-detector transferability, or robustness to practical image processing, motivating attacks that explicitly optimize natural degradations and low-level forensic statistics [26], [27], [64].

StealthDiffusion represents an intermediate direction between post-hoc perturbation and generation-time adversarial synthesis. Given an initially generated image, it performs latent adversarial optimization and employs a control-oriented variational autoencoder (VAE) to reduce visual and spectral discrepancies, thereby producing higher-quality detector-evasive images [28]. Nevertheless, it remains a diffusion-assisted post-generation refinement pipeline that depends on a pre-existing generated image. More recently, the Adversarial Diffusion Model (ADM) directly generates detector-evasive images from scratch by introducing an adversarial denoising U-Net to search for detector-sensitive latent representations [29]. Although ADM moves adversariality into the generation process, it requires additional detector-aware optimization of diffusion components, coupling the resulting generator to the detectors used during model construction.

Different from these approaches, TIGA considers a source-image-free, training-free, and score-based black-box adversarial generation setting. It does not require an externally provided source image, post-generation refinement, or parameter updates to the diffusion model. TIGA uses accessible surrogate detectors only to construct a transferable directional prior and employs confidence queries to the black-box target detector to refine the adversarial direction online. The resulting guidance is injected directly into the DDIM denoising trajectory, allowing detector-evasive properties to emerge during inference while keeping the diffusion model frozen.

III. Method

A. Overview and Threat Model

Overview. As illustrated in Fig. 2, TIGA performs adversarial generation entirely within the DDIM sampling process, without modifying or finetuning the diffusion model. Given a condition $c = (\text{mask}, \text{text})$ and an initial noise z_T, the frozen diffusion model produces a denoising trajectory $z_T \to z_{T-1} \to \cdots \to z_0$. At each step, the standard DDIM update first predicts a clean latent $\hat{z}_0$ and the next trajectory state z_{t-1}. TIGA then augments this step with three tightly coupled modules: (i) a *Surrogate-Guided Prior* (SGP) module that aggregates white-box surrogate gradients at $\hat{z}_0$ into a transferable prior direction $\hat{g}_{\text{dir}}$; (ii) a *Surrogate-Guided Directional Search* (SGDS) module that, guided by $\hat{g}_{\text{dir}}$, queries the black-box detector with symmetric finite differences and accumulates a stable momentum direction m_t; and (iii) a *Schedule-Aware Trajectory Injection* (SATI) module that scales m_t by the noise schedule, reshapes it in the frequency domain, and injects the result into the trajectory state z_{t-1}. The adversarial property therefore emerges as part of generation rather than being attached afterwards.

Notation. Let $\mathcal{D}(\cdot)$ denote the VAE decoder and $\mathcal{T}(\cdot)$ denote the differentiable detector preprocessing (resizing and normalization). We write the composite latent-to-detector mapping as $\mathcal{G} = \mathcal{T} \circ \mathcal{D}$. Following the DDIM formulation, the clean latent predicted at step t is

$$\hat{z}_0 = \frac{z_t - \sqrt{1-\alpha_t}\,\epsilon_\theta(z_t, t)}{\sqrt{\alpha_t}}, \tag{1}$$

where ϵ_θ is the frozen noise predictor, and α_t is the cumulative noise coefficient. The next trajectory state is obtained by the standard DDIM step, i.e.,

$$z_{t-1} = \sqrt{\alpha_{t-1}}\,\hat{z}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\,\epsilon_\theta + \sigma_t \epsilon, \tag{2}$$

with σ_t the DDIM stochasticity level, $\epsilon \sim \mathcal{N}(0, I)$, and I the identity matrix.

Threat model. We consider a realistic deployment setting in which the diffusion model is frozen and the attacker controls only the sampling process. The target AIGC detector is accessed in a strict black-box manner: only its output confidence can be queried, while its architecture, parameters, and gradients are unavailable. We denote the black-box target by its real-probability response $D_b(x) \in (0, 1)$, where $D_b(x) > 0.5$ indicates that x is judged *real*. In addition, the attacker holds M white-box surrogate detectors $\{f_j\}_{j=1}^{M}$, each producing a real-class logit $s_j^{\text{sur}}(z) = f_j(\mathcal{G}(z))$ from which gradients can be back-propagated.

We note that surrogate guidance is computed from logits, whereas the black-box objective is defined on the sigmoid probability D_b; this distinction is kept explicit throughout. The primary attack objective is to maximize the real-class

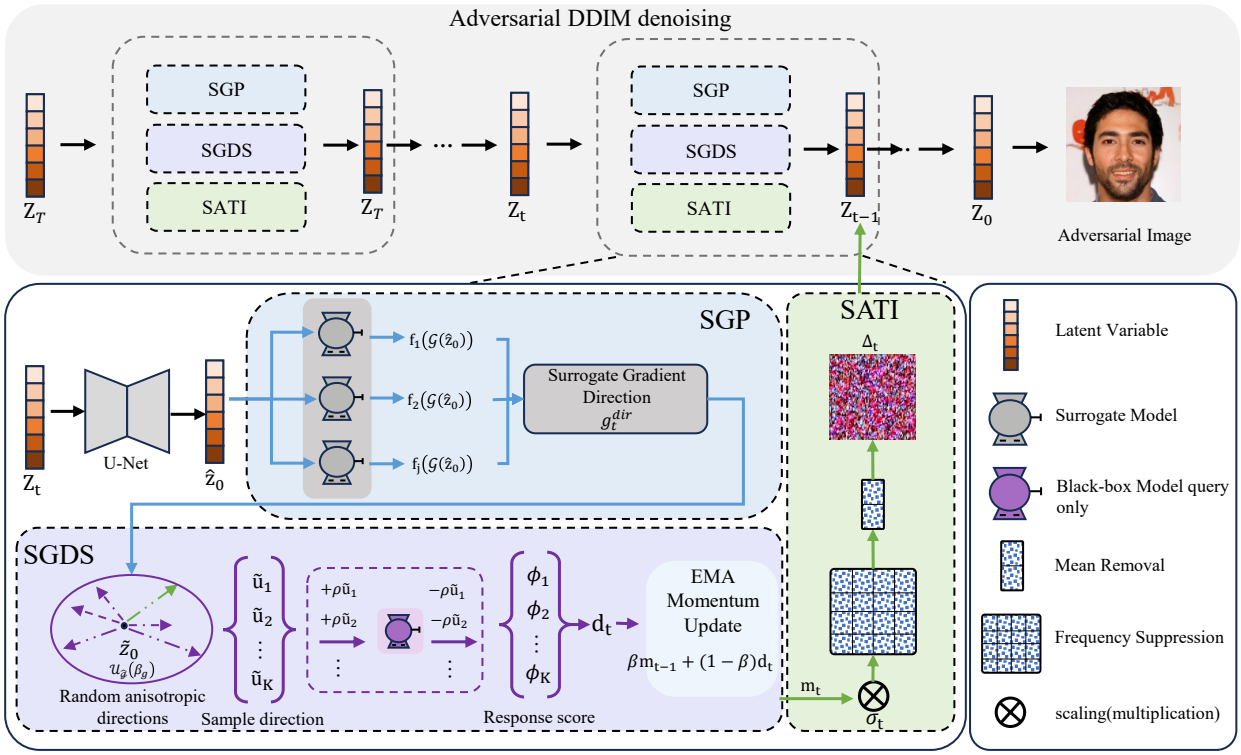


Fig. 2: Overall framework of TIGA. Given the shared initial noise z_T, standard DDIM sampling first predicts the clean latent $\hat{z}_0$ at each denoising step. The SGP module aggregates gradients from multiple surrogate detectors to construct a transferable prior direction. Guided by this prior, the SGDS module samples anisotropic probing directions, queries the black-box detector with symmetric finite differences, and accumulates the response-weighted directions into momentum. Finally, the SATI module performs schedule-aware trajectory injection with frequency-domain artifact suppression, producing detector-evasive images without retraining the diffusion model.

confidence of the target detector, subject only to the image being produced by the injected DDIM trajectory:

$$\max_{\{\Delta_t\}} \; D_b(x_0), \quad \text{s.t. } x_0 = \mathcal{D}(z_0),\; z_0 = \mathrm{DDIM}_{\{\Delta_t\}}(z_T, c), \tag{3}$$

where $\{\Delta_t\}$ are the per-step trajectory injections defined below. We intentionally keep Eq. (3) as a single-objective formulation rather than attaching explicit fidelity, semantic, or frequency penalties: perceptual fidelity is not optimized as a loss, term but instead is encouraged implicitly by the design of the injection itself, namely the schedule-aware bounded scaling and the frequency-domain reshaping introduced in Section III-D.

B. Surrogate-Guided Prior

The first module constructs a transferable adversarial direction from the white-box surrogate detectors, which serves as an informative prior for the subsequent black-box search. At each denoising step, the search and guidance are performed at the predicted clean latent $\hat{z}_0$ in Eq. (1), since the black-box and surrogate detectors only operate on clean images: the decoded preview $\mathcal{D}(\hat{z}_0)$ approximates the final synthesized image, whereas the noisy state z_t has no meaningful detector response.

Given the M surrogate detectors, we define their ensemble real-class score in the latent space as

$$\bar{s}^{\mathrm{sur}}(z) = \frac{1}{M}\sum_{j=1}^{M} s_j^{\mathrm{sur}}(z) = \frac{1}{M}\sum_{j=1}^{M} f_j\big(\mathcal{G}(z)\big), \tag{4}$$

and obtain the surrogate-guided prior direction by back-propagating the ensemble score to the latent and normalizing it:

$$\hat{g}_{\mathrm{dir}} = \mathrm{normalize}\left(\nabla_z\, \bar{s}^{\mathrm{sur}}(z)\Big|_{z=\hat{z}_0}\right). \tag{5}$$

The prior $\hat{g}_{\mathrm{dir}}$ has two desirable properties. First, it is *transferable*: by aggregating gradients across heterogeneous surrogate architectures, it captures detector-agnostic forensic directions rather than overfitting to a single model. Second, it is *sign-aware*: since $\hat{g}_{\mathrm{dir}}$ is the ascent direction of the real-class score, it already points toward the detector-evasive region, so the subsequent search requires neither a reference sample nor an explicit sign-calibration step to start stably. To balance guidance quality against the cost of back-propagation, the prior is refreshed every τ_g steps and reused in-between. Notably, $\hat{g}_{\mathrm{dir}}$ only specifies a search *orientation*, as opposed to the attack itself: as the surrogates differ from the target, directly injecting it reduces to a pure transfer attack, and the actual detector-evasive direction is determined by the black-box response in the next module.

C. Surrogate-Guided Directional Search

Given the prior $\hat{g}_{\text{dir}}$, the second module estimates a reliable trajectory guidance direction by querying the black-box target. Instead of performing isotropic zero-order exploration, which is notoriously high in variance and unstable in the high-dimensional latent space, we constrain the search to an anisotropic distribution centered around the surrogate prior.

Anisotropic guided sampling. Let $\beta_g \in [0,1]$ be a fusion coefficient controlling how strongly the search is biased toward the prior. To draw a probing direction, we first sample an isotropic Gaussian $r \sim \mathcal{N}(0, I)$ and extract its unit component orthogonal to the prior,

$$r_\perp = \frac{r - \langle r, \hat{g}_{\text{dir}} \rangle \, \hat{g}_{\text{dir}}}{\left\| r - \langle r, \hat{g}_{\text{dir}} \rangle \, \hat{g}_{\text{dir}} \right\|}, \tag{6}$$

and then mix the prior with the orthogonal exploration component to form the probing direction

$$\tilde{u} = \text{normalize}\big(\beta_g \, \hat{g}_{\text{dir}} + (1 - \beta_g) \, r_\perp\big), \qquad \tilde{u} \sim \mathcal{U}_{\hat{g}}(\beta_g). \tag{7}$$

Here, $\mathcal{U}_{\hat{g}}(\beta_g)$ denotes the resulting anisotropic direction distribution: it concentrates probing energy along the transferable prior while retaining orthogonal exploration to correct for surrogate-target mismatch.

Symmetric finite-difference response. For each probing direction $\tilde{u}_k$, we measure how fast the black-box real-probability changes along it using a symmetric (central) finite difference with probing radius ρ:

$$\phi_k = \frac{D_b\big(\mathcal{D}(\hat{z}_0 + \rho \, \tilde{u}_k)\big) - D_b\big(\mathcal{D}(\hat{z}_0 - \rho \, \tilde{u}_k)\big)}{2\rho}. \tag{8}$$

The scalar ϕ_k acts as a directional response weight: a large positive ϕ_k indicates that moving along $+\tilde{u}_k$ increases the real-probability, while $\phi_k < 0$ indicates the opposite. The central difference yields an $O(\rho^2)$ truncation error, giving a more accurate estimate than one-sided differencing at the same query cost.

Direction estimation. The per-step guided direction estimate is the response-weighted expectation of the probing directions over the guided distribution, approximated with N samples:

$$d_t = \mathbb{E}_{\tilde{u} \sim \mathcal{U}_{\hat{g}}(\beta_g)}\big[\phi(\hat{z}_0, \tilde{u}) \, \tilde{u}\big] \approx \frac{1}{N} \sum_{k=1}^{N} \phi_k \, \tilde{u}_k. \tag{9}$$

Directions that raise the real-probability dominate the weighted sum, while uninformative directions are suppressed toward zero. In the spirit of prior-guided black-box estimators such as P-RGF and Guided ES [61], [62], this per-step estimate interpolates between transfer and query-based search. At the boundary $\beta_g = 0$, the probing direction reduces to $\tilde{u} = r_\perp$, i.e., isotropic zero-order exploration restricted to the subspace orthogonal to the prior; in the high-dimensional latent space a random direction is almost surely nearly orthogonal to $\hat{g}_{\text{dir}}$, so this closely approximates fully isotropic search. At the opposite boundary $\beta_g = 1$, all probes align with $\hat{g}_{\text{dir}}$ and Eq. (9) reduces to a one-dimensional black-box line search along the surrogate prior, where the direction is fixed to the prior but its sign and magnitude are still determined by the target queries ϕ_k, rather than to a query-free transfer attack.

The intermediate regime $\beta_g \in (0,1)$ constrains the search subspace with the white-box prior, substantially reducing estimation variance under a fixed query budget. Unlike these methods, which estimate a single direction on a fixed image, here the estimate d_t is one link in a trajectory-level chain: it is accumulated into momentum across denoising steps and injected back into the DDIM trajectory through the schedule-aware, frequency-shaped update of the next module, so that the per-step search and the diffusion sampling process are tightly coupled, rather than applied to a static input.

Decayed momentum accumulation. To stabilize the trajectory-level signal across steps, we accumulate the normalized direction estimates using a step-decayed exponential moving average. We index the denoising steps in sampling order by $i = 0, 1, \ldots, S-1$, so that step i corresponds to DDIM time $t = T - i$, and $i = 0$ corresponds to the first, highest-noise step. Writing d_i for the per-step estimate d_t of Eq. (9) at that step, we define $\hat{d}_i = \text{normalize}(d_i)$ and set $\beta_i = \beta_{\text{momentum}}\big(1 - \frac{i}{S-1}\big)$. The momentum is bootstrapped from the first estimate and updated as

$$m_0 = \hat{d}_0, \qquad m_i = \beta_i \, m_{i-1} + (1 - \beta_i) \, \hat{d}_i. \tag{10}$$

Since β_i is the largest in the early, high-noise steps and decays toward zero as sampling proceeds, the momentum relies more on the accumulated direction early on, where individual estimates are noisier, and becomes more responsive to the current estimate in the later, low-noise steps.

To control the query cost, the black-box directional search is executed only once every τ_q denoising steps: on a query step the N finite-difference probes are issued and the momentum m is updated as above; on the intervening steps no query is made and the most recent momentum is reused for injection, with no update applied before the first query step. The default configuration uses $\tau_q = 1$, i.e., a guided search at every step; a larger τ_q trades attack strength for fewer queries, and this query interval is the parameter varied in the corresponding ablation.

D. Schedule-Aware Trajectory Injection

The third module converts the momentum direction m_t into an actual trajectory update. A key design choice is *where* and *how strongly* to inject. We inject into the trajectory state z_{t-1} from Eq. (2) rather than into the clean prediction $\hat{z}_0$: $\hat{z}_0$ is a transient preview that is recomputed and discarded at every step, whereas z_{t-1} is the actual variable propagated along the trajectory. Perturbing z_{t-1} allows the adversarial signal to accumulate across steps and be progressively absorbed by the remaining denoising, which is the essence of trajectory-level injection.

Schedule-aware scaling. TIGA operates under the stochastic DDIM sampler ($\eta = 1$ in all experiments), for which the stochasticity level σ_t in Eq. (2) is strictly positive and decays monotonically along the denoising process. Since the accumulated momentum m_t ($\equiv m_i$ at the corresponding step) is a unit direction, we scale it by this stochasticity level σ_t

Algorithm 1 TIGA: Trajectory-Injected Generative Attack

Require: frozen diffusion model, condition c, noise z_T, black-box D_b, surrogates $\{f_j\}$, coefficients $\beta_g, \beta_{\text{momentum}}, \lambda_0$, samples N, radius ρ, prior-refresh interval τ_g, query interval τ_q

$m \leftarrow \varnothing$
for $t = T$ **to** 1 **do**
  $\epsilon_\theta \leftarrow \texttt{UNet}(z_t, t, c)$
  compute $\hat{z}_0$ by Eq. (1), z_{t-1} by Eq. (2)
  if $t \bmod \tau_g = 0$ **then**
    $\hat{g}_{\text{dir}} \leftarrow \text{normalize}(\nabla_z \bar{s}^{\text{sur}}(z)|_{\hat{z}_0})$ // SGP
  end if
  if $t \bmod \tau_q = 0$ **then**
    for $k = 1$ **to** N **do**
      sample $\tilde{u}_k \sim \mathcal{U}_{\hat{g}}(\beta_g)$ by Eq. (7)
      $\phi_k \leftarrow$ symmetric finite difference by Eq. (8)
    end for
    $d_t \leftarrow \frac{1}{N}\sum_k \phi_k \tilde{u}_k$; update m by Eq. (10) // SGDS
  end if
  if $m \neq \varnothing$ **then**
    $\Delta_t \leftarrow$ scale and frequency-reshape m by Eqs. (11)–(12)
    $z_{t-1} \leftarrow \text{clip}(z_{t-1} + \Delta_t, -4, 4)$ // SATI
  end if
end for
return $x_0 = \mathcal{D}(z_0)$

and a strength coefficient λ_0, with n the latent dimension, and clip the magnitude to bound per-element distortion:

$$\Delta_t = \sigma_t \, \lambda_0 \, m_t \sqrt{n}, \qquad \Delta_t \leftarrow \text{clip}\big(\Delta_t, -\tfrac{\lambda_0}{2}, \tfrac{\lambda_0}{2}\big). \tag{11}$$

Because σ_t decays monotonically along the denoising process, the injection is strong in early high-noise steps, where it can reshape the trajectory toward the detector-evasive region, and vanishes in late low-noise steps, where the diffusion model recovers clean high-frequency details. This schedule-aware coupling ties the injection to the sampler's own stochasticity so that the perturbation is naturally assimilated rather than appearing as an external noise layer.

Frequency-domain artifact suppression. To further improve fidelity and weaken detector-sensitive forensic traces, we reshape Δ_t in the frequency domain. We attenuate its high-frequency component via low-pass energy preservation (with Down/Up denoting bilinear resampling), and remove the per-channel spatial mean to eliminate the zero-frequency bias that would otherwise cause color shift or over-exposure:

$$\begin{aligned} \Delta_t^{\text{low}} &= \text{Up}\big(\text{Down}(\Delta_t)\big), \\ \Delta_t &\leftarrow \Delta_t^{\text{low}} + \tfrac{1}{2}\big(\Delta_t - \Delta_t^{\text{low}}\big), \\ \Delta_t &\leftarrow \Delta_t - \overline{\Delta_t}^{(H,W)}. \end{aligned} \tag{12}$$

Attenuating the high-frequency residual suppresses visible texture artifacts and, at the same time, reduces the abnormal spectral fingerprints that frequency-domain forensic detectors rely on. Finally, the reshaped update is injected and the trajectory state is clamped to the valid VAE range:

$$z_{t-1} \leftarrow \text{clip}\big(z_{t-1} + \Delta_t, \, -4, \, 4\big). \tag{13}$$

The updated z_{t-1} becomes the input of the next denoising step, propagating the adversarial signal forward until the final detector-evasive image $x_0 = \mathcal{D}(z_0)$ is obtained. The complete procedure, with the three modules embedded inside the standard DDIM denoising loop over a frozen diffusion model, is summarized in Algorithm 1.

IV. Experiments

A. Experimental Setup

Implementation details. TIGA is implemented in PyTorch 1.7 with CUDA 11.0 on top of a frozen, pretrained Collaborative Diffusion backbone, and all experiments are run on an NVIDIA RTX 3090 GPU. The backbone synthesizes 512×512 face images conditioned jointly on a segmentation mask and a free-form text description, operating in a $3{\times}64{\times}64$ latent space; for every test instance, we pair a 20-class CelebAMask-HQ-style segmentation mask [65] (resized to 32×32 and one-hot encoded) with a textual attribute description (e.g., age, hairstyle, or beard length) to form the condition $c = (\text{mask}, \text{text})$ used throughout Section III. DDIM sampling runs for $S = 50$ steps with stochasticity $\eta = 1$, and for each instance the reference (unattacked) and adversarial trajectories share the same initial noise z_T and condition c, so that any difference between them is attributable only to the trajectory injection rather than to sampling variance. Unless otherwise noted, every reported TIGA result uses the full configuration that also serves as the reference point of the ablation study (Section IV-C): guidance strength $\lambda_0 = 0.35$, surrogate-prior fusion coefficient $\beta_{\text{guide}} = 0.5$, $N = 10$ probing directions per guided step with finite-difference radius $\rho = 0.05$, query interval $\tau_q = 1$, so that guided search is performed at *every* denoising step, momentum decay $\beta_{\text{momentum}} = 0.8$, and surrogate-prior refresh interval $\tau_g = 5$. Images are generated in batches under a fixed random seed so that results are reproducible across methods.

Generic AIGC detectors. We use four architecturally diverse, binary real/fake AIGC detectors as the target/surrogate pool, jointly denoted **REDS**: ResNet-50 (**R**, CNN) [66], EfficientNet-B0 (**E**, CNN) [67], DeiT-Base (**D**, plain Vision Transformer) [68], and Swin-Base (**S**, hierarchical Transformer) [69]. Each detector is fine-tuned in-domain on $12{,}000$ images synthesized by our diffusion backbone (balanced real/fake) as a single real/fake logit ($\text{sigmoid}(\cdot) > 0.5$ indicates "real"), and all four reach over 95% accuracy on a held-out test set, so that the evaluated targets are reliable rather than weak detectors. Following the threat model in Section III, we construct four black-box settings by cycling through REDS: in each setting, one model plays the role of the strict black-box target D_b, queried only through its output probability, while the remaining three serve as the white-box surrogates $\{f_j\}_{j=1}^{3}$ used by the SGP module. All four settings share the same test set of (mask, text) pairs.

Transfer detectors. To measure whether the induced evasiveness generalizes beyond the REDS pool used during optimization, we additionally consider five specialized forensic detectors that are never used as either a white-box surrogate or a black-box target: a CNN-based spatial-artifact classifier (**CNN**) [13], a diffusion-reconstruction-error

TABLE I: Black-box attack performance and visual quality under four black-box target settings on REDS (R: ResNet-50, E: EfficientNet-B0, D: DeiT, S: Swin-T). ASR↑ denotes the attack success rate (higher is stronger); BRIS↓ denotes the BRISQUE no-reference quality score (lower is better). The best result in each column is shown in **bold**, and our method is highlighted in gray.

Attacks	ResNet-50		EfficientNet-B0		DeiT		Swin-T		Average	
	ASR↑	BRIS↓	ASR↑	BRIS↓	ASR↑	BRIS↓	ASR↑	BRIS↓	ASR↑	BRIS↓
FGSM	86.64	87.37	99.92	106.63	91.13	106.63	96.37	106.63	93.52	101.82
PGD	28.15	36.43	83.27	36.05	84.98	36.05	94.43	36.06	72.71	36.15
SimBA	78.63	26.79	84.29	27.19	68.71	28.33	51.86	27.53	70.87	27.46
Square	99.14	69.20	**100.00**	78.11	**100.00**	57.64	99.86	52.39	99.89	64.34
BruSLe	99.11	35.40	**100.00**	35.05	**100.00**	35.14	89.14	35.93	97.06	35.38
R^2BA	**100.00**	21.76	**100.00**	27.39	**100.00**	25.83	98.43	35.93	99.61	24.89
Ours	**100.00**	**20.43**	**100.00**	**21.28**	**100.00**	**19.18**	**100.00**	**20.05**	**100.00**	**20.24**

TABLE II: Transfer-attack evasion rate (%, percentage judged "real") against five unseen detectors (CNN, DIRE, Uni, Effort, PGC), none used as a white-box surrogate or black-box target during optimization, evaluated separately for each of the four REDS black-box target settings used to craft the images. "Clean": unattacked images. Bold: best per column; gray row: our method.

	Target Detector: DeiT						Target Detector: EfficientNet-B0				
Methods	**CNN**	**DIRE**	**Uni**	**Effort**	**PGC**	**Methods**	**CNN**	**DIRE**	**Uni**	**Effort**	**PGC**
Clean	77.86	90.57	21.29	22.14	4.29	Clean	77.86	90.57	21.29	22.14	4.29
FGSM	**100.00**	82.86	35.14	0.86	7.86	FGSM	**100.00**	82.86	35.14	0.86	7.86
PGD	**100.00**	83.43	5.57	0.00	0.00	PGD	**100.00**	83.43	5.57	0.00	0.00
SimBA	98.57	89.97	14.86	4.43	1.57	SimBA	99.86	89.72	11.43	1.86	1.00
Square	**100.00**	92.43	6.86	0.71	0.86	Square	**100.00**	92.43	9.43	1.14	1.14
BruSLe	**100.00**	95.86	33.58	56.86	0.43	BruSLe	**100.0**	96.43	31.71	54.14	0.43
R^2BA	**100.00**	95.72	20.86	15.86	9.29	R^2BA	98.71	94.28	18.14	11.43	12.57
Ours	99.53	**100.00**	**68.40**	**86.32**	**49.06**	**Ours**	**100.00**	**99.53**	**47.17**	**66.98**	**22.17**
	Target Detector: ResNet-50						**Target Detector: Swin-T**				
Methods	**CNN**	**DIRE**	**Uni**	**Effort**	**PGC**	**Methods**	**CNN**	**DIRE**	**Uni**	**Effort**	**PGC**
Clean	77.86	90.57	21.29	22.14	4.29	Clean	77.86	90.57	21.29	22.14	4.29
FGSM	**100.00**	84.79	41.14	7.71	5.00	FGSM	**100.00**	82.86	35.14	0.86	7.86
PGD	**100.00**	84.15	15.43	0.00	0.57	PGD	**100.00**	83.43	5.57	0.00	0.00
SimBA	98.71	90.29	11.14	6.29	0.86	SimBA	99.71	89.86	12.29	2.57	0.86
Square	**100.00**	92.58	7.86	0.71	0.86	Square	**100.00**	92.15	7.43	0.14	0.86
BruSLe	**100.00**	95.67	31.43	55.71	0.43	BruSLe	**100.00**	97.86	34.29	59.71	0.43
R^2BA	**100.00**	95.29	19.86	21.86	12.86	R^2BA	**100.00**	94.65	22.14	35.57	12.73
Ours	**100.00**	**100.00**	**62.74**	**75.94**	**40.62**	**Ours**	**100.00**	**99.53**	**59.43**	**77.36**	**38.21**

detector (**DIRE**) [23], a CLIP-feature-based universal detector (**Uni**) [16], and two further generalizable detectors (**Effort** [70] and **PGC**) [71]. All five are trained on 144K images generated by Stable Diffusion v1.4 [2] and ProGAN covering the same four object categories, with the ProGAN data drawn from ForenSynths [13] and the SD-v1.4 data from GenImage [48]. We report their evasion behavior in Section IV-B, and use the abbreviations CNN/DIRE/Uni/Effort/PGC in all tables.

Baselines. We compare TIGA against six representative adversarial attacks that are adapted to AIGC-detector evasion: two white-box gradient attacks, FGSM [52] and PGD [53]; three query-based black-box attacks, SimBA [57], Square Attack [58], and BruSLe [59]; and the most recent AIGC-oriented black-box attack, R^2BA [27], which optimizes over realistic post-processing operations. All baselines are applied as *post-hoc* pixel-space perturbations on top of the identical clean image produced by the same DDIM trajectory as TIGA, sharing the same initial noise z_T, mask, and text, so that the underlying generative content is held fixed across methods and only the perturbation mechanisms differ. For fair comparisons, all attacks share the same maximum perturbation budget $\epsilon = 16/255$. The four query-based black-box attacks, SimBA, Square, BruSLe, and R^2BA, each query the black-box target under a fixed budget of 1,000 queries with at most 20 iterations, are launched from a random initialization, and never access the target gradients. As FGSM and PGD require gradients that are unavailable under the black-box target setting, we run them as transfer attacks: their perturbations are crafted on a single white-box surrogate detector and then transferred to the black-box target. Neither method issues any query to the target or accesses its gradients.

Evaluation metrics. We report: (i) the Attack Success Rate

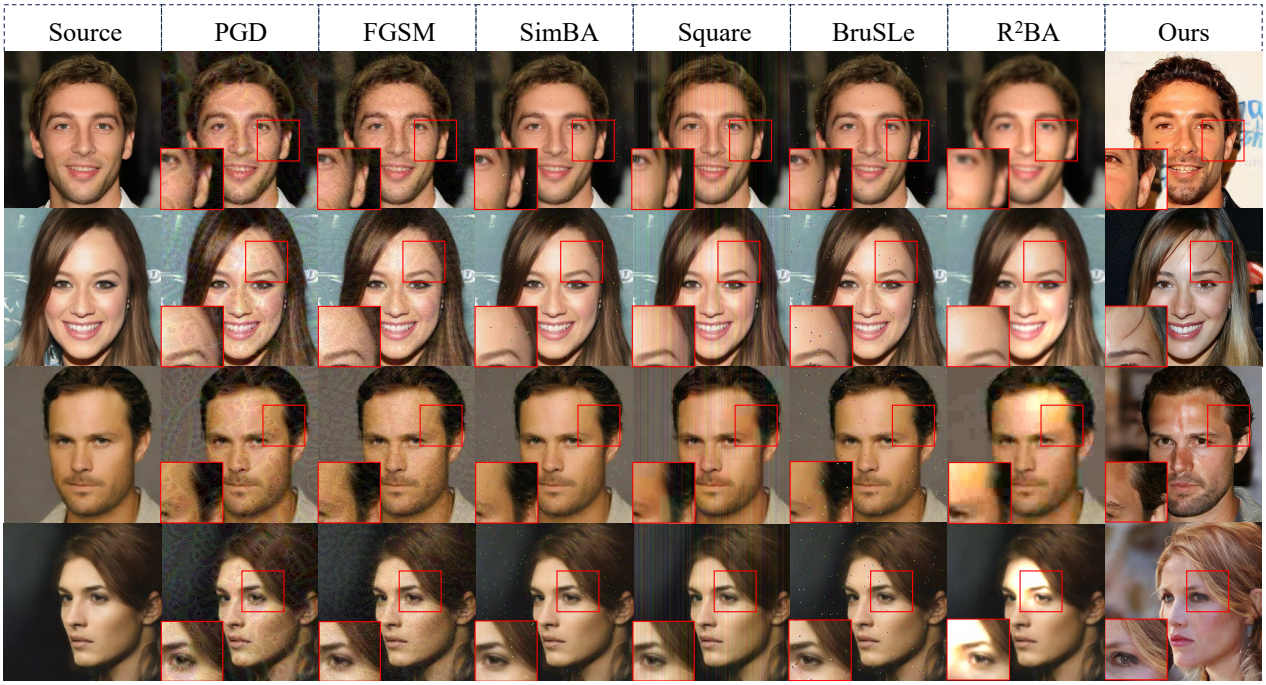


Fig. 3: Qualitative comparison of adversarial faces produced by each attack from the same DDIM-generated source image (leftmost column); red boxes zoom into the forehead, cheek, and periocular skin.

(ASR↑), the percentage of images that the target detector correctly labels "fake" before the attack and flips to "real" afterward; (ii) Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE↓), a no-reference perceptual quality score computed on the final adversarial image, where a lower value indicates fewer visible artifacts; (iii) the **detector-evasion rate** on fully unseen detectors, used to quantify transferability; and (iv) **robust ASR** under common post-processing operations.

B. Comparison Study

1) Attack Effectiveness and Visual Quality on REDS: Table I compares TIGA against the six baselines on all four REDS targets. TIGA attains a perfect 100% ASR on every target while obtaining the lowest BRISQUE in every column (average 20.24), i.e., it Pareto-dominates every baseline on both axes at once. Among the baselines, R^2BA is the strongest overall, nearly saturating ASR (average 99.61%) at the best perceptual quality of any baseline (average BRISQUE 24.89), yet it still trails TIGA on BRISQUE at every target. Square and BruSLe also saturate ASR (average 99.89%/97.06%) but at a much higher visual cost (BRISQUE 64.34/35.38), SimBA is moderate in quality (27.46) but weak in ASR (70.87%), and the gradient attacks are the least consistent (FGSM: 93.52% ASR at the worst BRISQUE 101.82; PGD collapses to 28.15% on ResNet-50). Once a perturbation is externally superimposed in pixel space, an attacker is forced to trade success rate against visible artifacts; TIGA, by shaping the sample *during* generation rather than after it, avoids this trade-off and dominates all six baselines on both axes.

2) Qualitative Comparison: Fig. 3 visualizes this trade-off. The six baselines all perturb pixels on top of the identical clean generation, so they preserve the coarse appearance of the source but leave visible signatures in the zoomed patches: uniform speckle noise for FGSM/PGD, sparse local changes for SimBA, colored blotches for Square/BruSLe, and a mild global blur/illumination shift for R^2BA. TIGA (rightmost) shows none of these superimposed patterns and keeps skin and hair smooth, consistent with its lowest BRISQUE. This follows from *where* the update is applied: injected into the DDIM trajectory (Eq. (11)) rather than onto a finished image, it is smoothed and regularized by the remaining denoising steps. The trade-off is that, since the schedule-aware injection is the strongest at early, high-noise steps, TIGA samples can show slightly larger coarse, low-frequency changes (e.g., hair color or skin tone) than the pixel-perturbation baselines. This is inherent to the source-image-free setting: TIGA only aims to keep the output a plausible, naturally generated face under the given condition, not pixel-level fidelity to a source; for detector evasion the perceptual naturalness of the final image is the operative criterion.

3) Transferability to Unseen Detectors: To test whether the evasiveness induced by TIGA generalizes beyond the four REDS models used during optimization, we re-score all attacked images with the five unseen transfer detectors described in Section IV-A (CNN, DIRE, Uni, Effort, and PGC), none of which is used as a white-box surrogate or a black-box target during attack optimization. Any evasion observed reflects genuine cross-architecture, cross-paradigm transfer, rather than optimization against a known model.

As reported in Table II, two trends stand out. On CNN and DIRE, which respond to low-level, spatially localized statistics, most methods including TIGA reach or nearly reach 100% evasion; the high Clean rate on CNN (77.86%) indicates it is weak even against unattacked samples. More importantly,

TABLE III: ASR↑ (%) of seven attack methods under different post-processing operations and parameter settings. For each target detector and each post-processing setting, the **bold** value denotes the highest ASR among all attack methods. Gray row denotes our method.

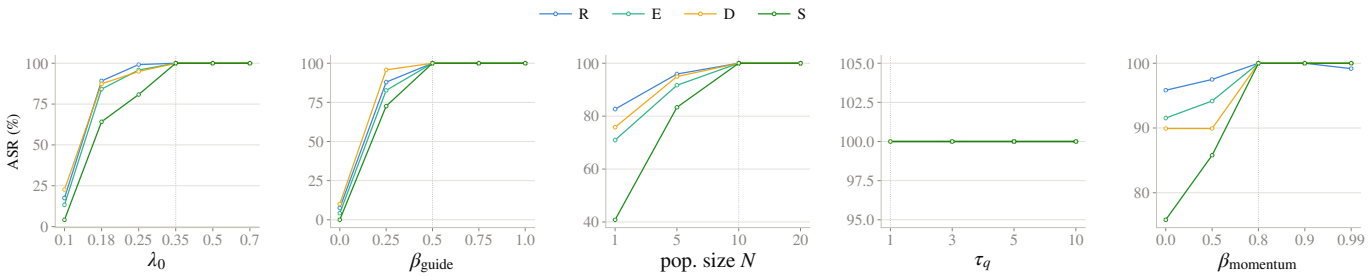
Detector	Attack	Gaussian Blur					Gaussian Noise					JPEG Compression					Resize				
		0.5	1	1.5	2	3	0.01	0.03	0.05	0.08	0.1	10	30	50	75	95	0.25	0.5	0.75	1.25	1.5
R	FGSM	86.18	90.54	87.57	73.86	44.35	85.57	85.43	92.62	99.43	**100.00**	93.71	84.14	85.71	84.43	85.79	38.29	87.43	89.86	87.29	89.97
	PGD	37.14	42.57	37.29	25.14	15.23	27.18	33.29	62.71	97.86	**100.00**	88.43	29.43	31.29	29.14	27.14	6.14	35.43	40.87	36.43	38.43
	Square	43.86	40.28	40.43	32.43	25.86	60.29	56.57	77.57	98.57	**100.00**	93.43	69.57	86.57	87.86	82.43	9.14	31.29	43.71	54.57	49.43
	SimBA	2.86	5.57	7.14	6.86	9.14	0.86	8.86	41.86	94.71	**100.00**	86.00	2.43	1.29	1.00	1.86	1.57	2.71	3.86	5.17	4.86
	BrusLe	89.86	41.29	17.29	11.14	8.29	97.43	80.71	73.43	96.71	99.86	94.52	67.86	77.42	86.86	97.71	3.96	22.29	61.33	83.71	83.71
	R^2BA	92.86	85.43	82.71	73.14	52.67	95.14	87.52	91.72	97.29	**100.00**	**98.86**	61.86	50.14	67.43	94.29	33.43	77.68	82.43	84.17	85.14
	Ours	**94.35**	**93.82**	**91.46**	**84.73**	**73.68**	**98.72**	**92.14**	**95.63**	**99.71**	**100.00**	97.29	**90.64**	**92.37**	**93.18**	**98.52**	**78.42**	**91.68**	**94.15**	**92.83**	**94.36**
E	FGSM	**100.00**	**100.00**	99.71	98.29	76.87	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	98.14	99.71	99.86	99.43	**100.00**	85.14	99.43	**100.00**	**100.00**	**100.00**
	PGD	**100.00**	**100.00**	98.29	90.71	52.86	**100.00**	99.86	99.23	99.86	**100.00**	82.86	98.57	99.86	99.86	**100.00**	65.43	99.12	**100.00**	**100.00**	**100.00**
	Square	60.71	30.00	35.71	25.43	15.00	78.71	82.14	91.43	99.00	99.86	60.00	58.71	77.29	73.29	83.00	7.29	31.00	47.86	58.43	58.43
	SimBA	11.71	5.43	6.43	6.43	7.43	22.43	49.17	83.86	98.29	99.86	31.14	3.43	1.71	1.00	12.43	3.00	1.29	3.57	7.29	9.57
	BrusLe	52.43	50.86	21.57	14.86	9.71	97.43	97.43	98.43	99.86	**100.00**	54.14	58.57	80.21	88.86	82.86	4.29	19.86	53.29	60.86	65.29
	R^2BA	89.86	30.14	27.86	23.57	22.32	92.57	97.29	99.14	99.86	**100.00**	68.43	19.29	14.71	16.47	49.57	9.43	23.22	34.86	47.79	42.14
	Ours	**100.00**	**100.00**	**99.86**	**99.14**	**83.46**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**99.26**	**99.86**	**99.93**	**99.94**	**100.00**	**90.72**	**99.71**	**100.00**	**100.00**	**100.00**
D	FGSM	90.17	87.14	83.14	71.92	45.86	91.14	92.29	94.14	95.57	96.57	92.29	91.65	89.86	91.14	91.57	60.86	85.14	89.12	90.63	90.88
	PGD	84.29	78.57	73.29	58.43	35.29	85.43	83.86	84.29	90.86	93.29	87.14	77.71	79.86	83.86	86.43	45.29	74.71	79.86	81.29	80.57
	Square	79.43	74.86	65.57	46.14	29.29	75.71	71.28	76.71	85.57	91.14	82.71	54.86	64.71	65.86	74.00	32.29	66.29	76.71	78.86	76.14
	SimBA	30.43	27.57	23.71	12.86	10.86	27.00	37.71	56.86	77.14	85.57	77.00	22.00	18.86	19.14	26.14	9.86	20.57	25.00	24.43	22.57
	BrusLe	55.86	38.71	28.71	14.57	8.14	70.29	63.43	71.57	83.57	88.86	79.29	38.43	40.14	43.79	60.71	9.12	31.14	41.29	48.57	47.14
	R^2BA	91.14	74.29	66.17	48.43	33.65	93.43	93.57	95.75	95.26	95.19	96.71	84.52	83.86	90.57	96.29	35.86	64.31	77.39	84.95	83.57
	Ours	**93.74**	**91.53**	**88.96**	**82.35**	**72.84**	**94.21**	**95.18**	**96.32**	**97.46**	**98.37**	**95.46**	**94.38**	**93.21**	**94.63**	**97.28**	**76.49**	**89.82**	**92.47**	**93.58**	**94.11**
S	FGSM	96.43	97.39	95.57	91.26	64.29	96.29	93.86	91.29	87.86	86.43	84.71	98.71	97.86	97.43	97.43	83.21	96.92	98.29	98.67	98.67
	PGD	93.57	92.57	90.14	80.57	48.69	90.43	75.29	64.43	62.57	68.57	64.57	88.86	90.29	90.57	96.14	68.86	90.14	94.57	94.86	95.86
	Square	90.14	93.14	87.57	68.29	40.29	14.86	24.71	27.43	34.29	40.71	54.71	71.14	81.71	88.57	88.86	44.86	87.71	96.86	94.29	92.57
	SimBA	7.00	7.00	9.29	7.86	8.43	0.14	3.71	9.86	21.57	31.86	35.14	9.71	9.43	9.71	5.43	8.14	7.57	11.71	6.71	6.29
	BrusLe	94.57	97.86	83.26	37.29	11.71	9.12	53.71	62.43	65.43	67.29	62.57	72.57	84.14	87.29	97.11	18.71	94.25	99.15	97.71	97.57
	R^2BA	76.15	82.43	83.57	79.29	64.93	89.31	90.43	90.29	89.57	89.86	89.91	88.71	97.43	95.14	70.96	81.43	84.14	84.14	80.29	80.71
	Ours	**97.82**	**98.64**	**97.16**	**93.87**	**74.56**	**97.38**	**95.27**	**93.14**	**90.72**	**93.36**	**88.94**	**99.18**	**98.42**	**98.11**	**98.36**	**97.92**	**98.16**	**99.52**	**99.14**	**99.05**

on the higher-level detectors Uni, Effort, and PGC the baselines stay close to the Clean rate (21.29%/22.14%/4.29%), whereas TIGA improves substantially and consistently across all four settings (Uni: 47.17–68.40%; Effort: 66.98–86.32%; PGC: 22.17–49.06%). Even R^2BA, the strongest baseline on REDS, stays near Clean here (e.g., PGC 12.86% vs. our 40.62% under the ResNet-50 target). This indicates that the baselines' post-hoc perturbations overfit to low-level statistics of their own target and do not transfer to higher-level detectors, while TIGA's trajectory-level modification shifts properties of the generated content itself and transfers more consistently.

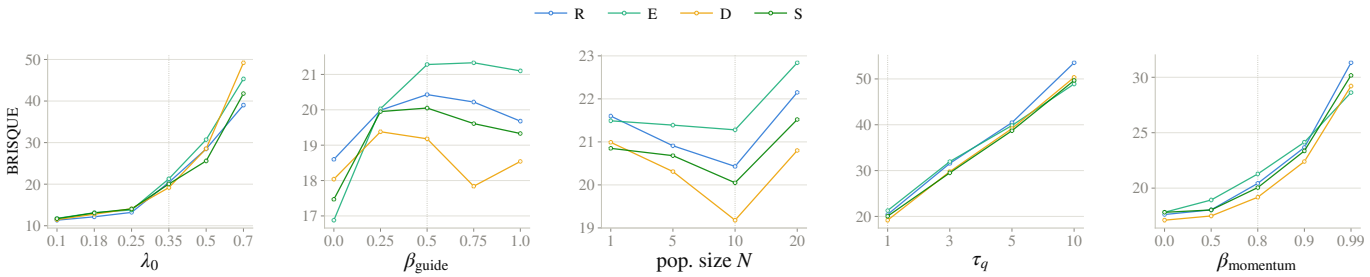
Transfer strength also depends on the surrogate set: the DeiT-target setting (surrogates {R,E,S}) transfers the best to Uni/Effort/PGC (68.40%/86.32%/49.06%) and the EfficientNet-target setting (surrogates {R,D,S}) transfers the worst (47.17%/66.98%/22.17%), even though both share R and S. Since the two differ only in EfficientNet vs. DeiT, including a CNN-style surrogate with a different inductive bias from ResNet-50 contributes disproportionately to the prior's transferability.

4) Robustness Evaluation: To assess whether the evasiveness of each method survives common image post-processing that an AIGC platform might apply before running detection, we take the adversarial images produced by each method against all four REDS targets and re-score them, after Gaussian blur, Gaussian noise, JPEG compression, or resizing, using all four REDS detectors; Table III reports, for each (perturbation family, parameter) setting and each scoring detector, the percentage of processed images still classified as "real."

Several observations follow from Table III, one per post-processing family. Under **Gaussian blur**, TIGA degrades gracefully (ASR-R $94.35\% \rightarrow 73.68\%$ from radius 0.5 to 3), whereas SimBA collapses to single digits already at radius 0.5 and Square/BruSLe fall 40–60 points; R^2BA is the most blur-robust baseline ($92.86\% \rightarrow 52.67\%$) but still trails TIGA at every level. Under **JPEG compression**, TIGA stays close to its uncompressed level (e.g., 97.29% at quality 10), while the baselines are the weakest at intermediate qualities (30–75): SimBA and PGD fall to 2.43% and 29.43% at quality 30, and even R^2BA drops to 50.14% at quality 50. Under **Resize**, at the largest downsampling (0.25) SimBA/BruSLe/Square/PGD fall to single digits while TIGA retains 78.42%. Under **Gaussian noise**, TIGA is the strongest at $\sigma \leq 0.05$, but at $\sigma = 0.10$ essentially all methods converge to near 100%; we read this as a metric ceiling effect, since the noise itself pushes images out of the detectors' training distribution regardless of any perturbation. Across the three families that remove or attenuate structure (blur, compression, resize), TIGA is the only method whose evasiveness survives realistic, moderate-strength post-processing, consistent with its adversarial signal being a property of the generated content rather than an added high-frequency pattern.



(a) Attack success rate.



(b) Perceptual quality measured by BRISQUE.

Fig. 4: Sensitivity of (a) ASR and (b) BRISQUE to each hyperparameter (λ_0, β_{guide}, population size N, query interval τ_q, and β_{momentum}). Each hyperparameter is swept independently while the remaining four are fixed at their full-configuration values. Results are evaluated against four REDS target detectors (R, E, D, and S). The dotted vertical line indicates the value adopted in the full configuration.

TABLE IV: Module ablation on REDS. Each cell reports ASR↑ (%) / BRISQUE↓ for the full method and four ablated variants (defined in the text). Bold: best per column. Full (Ours) is the same configuration reported in Table I.

Config	R		E		D		S		*Avg*	
	ASR	BRISQUE	ASR	BRISQUE	ASR	BRISQUE	ASR	BRISQUE	ASR	BRISQUE
Full (Ours)	**100.00**	20.43	**100.00**	21.28	**100.00**	19.18	**100.00**	20.05	**100.00**	20.24
w/o SGP	97.50	19.18	96.67	19.95	97.48	18.38	82.50	18.18	93.54	18.92
w/o SGDS	98.33	53.27	73.33	52.22	88.03	52.34	71.67	52.16	82.84	52.50
w/o Momentum	95.83	**17.62**	91.54	**17.82**	89.92	**17.11**	75.83	**17.82**	88.28	**17.59**
w/o Freq.	99.17	28.89	100.00	25.54	99.16	24.59	99.17	22.61	99.38	25.41

C. Ablation Study

1) Hyperparameter Sensitivity: Fig. 4 sweeps each of the five key hyperparameters independently, holding the other four at the full configuration ($\lambda_0 = 0.35$, $\beta_{\text{guide}} = 0.5$, $N = 10$, $\tau_q = 1$, $\beta_{\text{momentum}} = 0.8$). Across all five, ASR rises (near-)monotonically and then saturates while BRISQUE degrades gradually in the same direction, confirming that each parameter trades attack strength against perceptual quality as expected: λ_0 scales the injection magnitude (Eq. (11)) and has the most direct effect on both curves; β_{guide} biases probing toward the prior (Eq. (7)), raising ASR at little quality cost until orthogonal exploration is lost near 1; larger N reduces estimation variance (Eq. (9)) with diminishing returns; smaller τ_q accumulates stronger momentum at a higher query budget; and β_{momentum} stabilizes the direction up to a point beyond which it over-smooths. In every case the adopted operating point (dotted line) sits just past the knee where ASR has saturated but before BRISQUE degrades sharply, which is why it is chosen as the full configuration.

2) Module Ablation: To isolate the contribution of each of the three modules described in Section III, Table IV compares the full method against four ablated variants, all evaluated on the same four REDS targets under the same full-configuration hyperparameters: **w/o SGP** disables the surrogate-guided prior used to bias the anisotropic probing distribution described in Section III, so the guided search falls back to using only orthogonal, isotropic exploration; **w/o SGDS** disables the zero-order directional search entirely, so the trajectory is driven only by the schedule-aware injection of whatever direction the surrogate prior alone provides; **w/o Momentum** removes momentum accumulation across guided steps by setting $\beta_{\text{momentum}} = 0$ in Eq. (10), so each step's direction estimate is used without carrying information from previous steps; and **w/o Freq.** removes the frequency-domain artifact-suppression term of Eq. (12) from the schedule-aware injection, leaving the raw, unfiltered update.

Table IV quantifies each component and Fig. 5 shows the

Fig. 5: Qualitative module ablation. Each row shows the same (mask, text) condition attacked with one component removed, alongside the full method.

qualitative effect. Removing **SGDS** causes by far the largest degradation, in ASR (average 82.84%) and especially quality (BRISQUE $20.24 \rightarrow 52.50$), appearing as broad low-frequency color shifts: the zero-order search is the primary mechanism for a strong attack without a large quality-degrading injection. Removing **SGP** gives the second-largest ASR drop (average 93.54%, Swin-T 82.50%) while slightly improving BRISQUE (18.92), so the prior mainly contributes cross-target reliability rather than quality. Removing **Momentum** costs ASR (average 88.28%) at the best BRISQUE of any row (17.59), i.e., it trades a little quality for a more reliable attack. Removing **Freq.** barely affects ASR (99.38%) but worsens BRISQUE (25.41), confirming that it suppresses high-frequency artifacts rather than driving attack strength. The four components are thus complementary: SGDS and SGP drive effectiveness, Momentum and Freq. govern the quality/reliability trade-off, and the full combination is the only configuration attaining 100% ASR on all four targets while keeping BRISQUE competitive (the two variants with marginally lower BRISQUE reach it only by sacrificing ASR).

V. Conclusion

This paper proposed TIGA, a training-free, trajectory-injected framework that evades black-box AIGC detectors by steering the DDIM denoising trajectory through a surrogate-guided prior, an anisotropic black-box directional search, and a schedule-aware, frequency-reshaped injection. On general-purpose AIGC detectors, TIGA attains 100% black-box ASR with the lowest BRISQUE among all compared methods, transfers markedly better to unseen detectors, and remains robust under common post-processing, with the module ablation confirming that its three components are complementary. Future work will include broadening the surrogate/detector pools and extending TIGA beyond face generation and a single diffusion backbone.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[3] Z. Lin, W. Wei, Z. Chen, C.-T. Lam, X. Chen, Y. Gao, and J. Luo, "Hierarchical Split Federated Learning: Convergence Analysis and System Optimization," *IEEE Trans. Mobile Comput.*, 2025.

[4] Z. Huang, K. C. Chan, Y. Jiang, and Z. Liu, "Collaborative diffusion for multi-modal face generation and editing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6080–6090.

[5] X. Du, X. Liu, J. Zhou, Z. Lin, C.-m. Pun, C. Wu, T. Li, Z. Chen, W. Ni, and J. Luo, "Defensive adversarial captcha: A semantics-driven framework for natural adversarial example generation," *IEEE Transactions on Dependable and Secure Computing*, 2025.

[6] H. Zhang, X. Mao, G. Dong, Z. Li, X. Su, K. Chen, J. Yang, and Z. Lin, "Memmark: State-evolution attribution watermarking for agent long-term memory systems," *arXiv preprint arXiv:2605.25002*, 2026.

[7] T. Duan, Z. Zhang, S. Guo, D. Huang, Y. Zhao, Z. Lin, Z. Fang, D. Luan, H. Cui, and Y. Cui, "Leed: A highly efficient and scalable llm-empowered expert demonstrations framework for multi-agent reinforcement learning," *arXiv preprint arXiv:2509.14680*, 2025.

[8] X. Guan, Z. Zhang, Z. Lin, Z. Sun, T. Duan, Z. Fang, R. Wang, H. Cui, W. Ni, J. Luo *et al.*, "Fluxshard: Motion-aware feature cache reuse for collaborative video analytics in mobile edge computing," *arXiv preprint arXiv:2605.06027*, 2026.

[9] J. Zhan, H. Shen, Z. Lin, and T. He, "Prism: Privacy-aware routing for adaptive cloud–edge llm inference via semantic sketch collaboration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 33, 2026, pp. 28 150–28 158.

[10] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proceedings of the 4th ACM workshop on information hiding and multimedia security*, 2016, pp. 5–10.

[11] Z. Lin, Y. Zhang, Z. Chen, Z. Fang, Y. Yang, G. Zhang, H. Yang, C. Wu, X. Chen, and Y. Gao, "ESL-LEO: An Efficient Split Learning Framework over LEO Satellite Networks," in *Proc. WASA*, 2025, pp. 344–357.

[12] Z. Fang, Z. Lin, S. Hu, Y. Tao, Y. Deng, X. Chen, and Y. Fang, "Dynamic uncertainty-aware multimodal fusion for outdoor health monitoring," *arXiv preprint arXiv:2508.09085*, 2025.

[13] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. 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*, 2020, pp. 8695–8704.

[14] H. Yuan, Z. Chen, Z. Lin, J. Peng, Z. Fang, Y. Zhong, Z. Song, X. Wang, and Y. Gao, "Graph Learning for Multi-Satellite Based Spectrum Sensing," in *Proc. IEEE Int. Conf. Commun. Technol. (ICCT)*, 2023, pp. 1112–1116.

[15] Z. Lin, Z. Chen, X. Chen, W. Ni, and Y. Gao, "HASFL: Heterogeneity-aware Split Federated Learning over Edge Computing Systems," *IEEE Trans. Mobile Comput.*, 2026.

[16] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 24 480–24 489.

[17] Z. Lin, Y. Zhang, Z. Chen, Z. Fang, X. Chen, P. Vepakomma, W. Ni, J. Luo, and Y. Gao, "HSplitLoRA: A Heterogeneous Split Parameter-Efficient Fine-Tuning Framework for Large Language Models," *arXiv preprint arXiv:2505.02795*, 2025.

[18] S. Li, W. Ma, J. Guo, S. Xu, B. Li, and X. Zhang, "Unionformer: Unified-learning transformer with multi-view representation for im-

age manipulation detection and localization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 12 523–12 533.

[19] Z. Fang, Z. Lin, S. Hu, H. Cao, Y. Deng, X. Chen, and Y. Fang, "IC3M: In-Car Multimodal Multi-Object Monitoring for Abnormal Status of Both Driver and Passengers," *arXiv preprint arXiv:2410.02592*, 2024.

[20] Z. Lin, X. Hu, Y. Zhang, Z. Chen, Z. Fang, X. Chen, A. Li, P. Vepakomma, and Y. Gao, "SplitLoRA: A Split Parameter-Efficient Fine-Tuning Framework for Large Language Models," *arXiv preprint arXiv:2407.00952*, 2024.

[21] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *European conference on computer vision*. Springer, 2020, pp. 86–103.

[22] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. 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*, 2024, pp. 28 130–28 139.

[23] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "Dire for diffusion-generated image detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 445–22 455.

[24] Y. Luo, J. Du, K. Yan, and S. Ding, "LaRE^2: Latent reconstruction error based method for diffusion-generated image detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 006–17 015.

[25] N. Carlini and H. Farid, "Evading Deepfake-image detectors with White- and Black-Box Attacks," *arXiv preprint arXiv:2004.00622*, 2020.

[26] Y. Diao, N. Zhai, C. Miao, Z. Yu, X. Wei, X. Yang, and M. Wang, "Vulnerabilities in AI-generated image detection: The challenge of adversarial attacks," *IEEE Transactions on Multimedia*, pp. 1–13, 2026.

[27] C. Xie, D. Ye, Y. Zhang, Y. Shang, L. Tang, Y. Lv, J. Deng, and J. Song, "Take fake as real: Realistic-like robust black-box adversarial attack to evade AIGC detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[28] Z. Zhou, K. Sun, Z. Chen, H. Kuang, X. Sun, and R. Ji, "StealthDiffusion: Towards evading diffusion forensic detection through diffusion model," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 3627–3636.

[29] H. Wang, S. Li, Z. Qian, and X. Zhang, "Adversarial diffusion model: Generating high quality and undetectable images from scratch," *IEEE Transactions on Information Forensics and Security*, vol. 21, pp. 1725–1736, 2026.

[30] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[31] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[32] Z. Lin, L. Wang, J. Ding, B. Tan, and S. Jin, "Channel Power Gain Estimation for Terahertz Vehicle-to-Infrastructure Networks," *IEEE Commun. Lett.*, vol. 27, no. 1, pp. 155–159, 2022.

[33] Z. Lin, G. Qu, X. Chen, and K. Huang, "Split Learning in 6G Edge Networks," *IEEE Wirel. Commun.*, 2024.

[34] S. Wang, S. Chen, D.-H. Wang, Y. Hua, and Y. Yan, "CAMD: Context-aware masked distillation for general self-supervised facial representation pre-training," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[35] Z. Sun, X. Guan, Z. Lin, Y. Qing, H. Song, Z. Fang, Z. Chen, F. Liu, H. Cui, W. Ni *et al.*, "Rrto: A high-performance transparent offloading system for model inference in mobile edge computing," *arXiv preprint arXiv:2507.21739*, 2025.

[36] Y. Zhang, M. Hu, Z. Lin, X. Fan, F. Xie, Z. Fang, J. Yang, W. Zhu, Z. Chen, C. Lv *et al.*, "Hera: Learning long-horizon coordination for device-cloud collaborative llm agents," *arXiv preprint arXiv:2605.24598*, 2026.

[37] Y. Lu, Z. Fang, P. Qiao, Z. Lin, J. Yang, Y. Zhang, P. L. Yee, Z. Chen, and J. Luo, "Conflict-aware federated fine-tuning of large language models with mixture-of-experts," *arXiv preprint arXiv:2606.15625*, 2026.

[38] Z. Lin, Y. Zhang, Z. Chen, Z. Fang, C. Wu, X. Chen, Y. Gao, and J. Luo, "LEO-Split: A Semi-Supervised Split Learning Framework over LEO Satellite Networks," *IEEE Trans. Mobile Comput.*, 2025.

[39] Z. Liu, X. Qi, and P. H. Torr, "Global texture enhancement for fake face detection in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8060–8069.

[40] X. Chen, C. Dong, J. Ji, J. Cao, and X. Li, "Image manipulation detection by multi-view multi-scale supervision," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14 185–14 193.

[41] J. Yan, Z. Li, F. Wang, Z. He, and Z. Fu, "Dual frequency branch framework with reconstructed sliding windows attention for AI-generated image detection," *IEEE Transactions on Information Forensics and Security*, vol. 21, pp. 2313–2325, 2026.

[42] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Frequency-aware deepfake detection: Improving generalizability through frequency space learning," 2024.

[43] S. Yan, O. Li, J. Cai, Y. Hao, X. Jiang, Y. Hu, and W. Xie, "A sanity check for AI-generated image detection," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 70 702–70 720.

[44] B. Du, X. Ma, X. Zhu, Z. Yang, C. Niu, C. Qu, M. Fang, Z. Wang, J. Liu, J. Liu *et al.*, "Can we build a monolithic model for fake image detection? SICA: Semantic-induced constrained adaptation for unified-yet-discriminative artifact feature space reconstruction," *arXiv preprint arXiv:2602.06676*, 2026.

[45] X. Ma, X. Zhu, L. Su, B. Du, Z. Jiang, B. Tong, Z. Lei, X. Yang, C.-M. Pun, J. Lv *et al.*, "IMDL-BenCo: A comprehensive benchmark and codebase for image manipulation detection & localization," *Advances in Neural Information Processing Systems*, vol. 37, pp. 134 591–134 613, 2024.

[46] B. Du, X. Zhu, X. Ma, C. Qu, K. Feng, Z. Yang, C.-M. Pun, J.-Z. Zhou *et al.*, "ForensicHub: A unified benchmark & codebase for all-domain fake image detection and localization," *Advances in neural information processing systems*, vol. 38, 2026.

[47] Z. Li, J. Yan, Z. He, K. Zeng, W. Jiang, L. Xiong, and Z. Fu, "Is artificial intelligence generated image detection a solved problem?" *Advances in neural information processing systems*, vol. 38, 2026.

[48] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang, "Genimage: A million-scale benchmark for detecting AI-generated image," *Advances in neural information processing systems*, vol. 36, pp. 77 771–77 782, 2023.

[49] X. Zhu, J.-Z. Zhou, K. Feng, C. Qu, X. Wang, Y. Wang, L. Zhou, and J. Liu, "Revisiting image manipulation localization under realistic manipulation scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 7198–7207.

[50] J. Ricker, D. Lukovnikov, and A. Fischer, "AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9130–9140.

[51] B. Chen, J. Zeng, J. Yang, and R. Yang, "DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images," in *Forty-first International Conference on Machine Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=oRLwyayrh1

[52] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[53] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[54] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[55] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.

[56] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[57] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple black-box adversarial attacks," in *International conference on machine learning*. PMLR, 2019, pp. 2484–2493.

[58] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *European conference on computer vision*. Springer, 2020, pp. 484–501.

[59] Q. V. Vo, E. Abbasnejad, and D. Ranasinghe, "BRUSLEATTACK: A query-efficient score-based black-box sparse adversarial attack," in *International Conference on Learning Representations*, vol. 2024, 2024, pp. 54 462–54 499.

[60] J. Zhu, X. Du, J. Zhou, C.-M. Pun, Q. Xu, and X. Liu, "DP-RAE: A dual-phase merging reversible adversarial example for image privacy protection," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 671–680.

[61] N. Maheswaranathan, L. Metz, G. Tucker, D. Choi, and J. Sohl-Dickstein, "Guided evolutionary strategies: Augmenting random search with surrogate gradients," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4264–4273.

[62] S. Cheng, Y. Dong, T. Pang, H. Su, and J. Zhu, "Improving black-box adversarial attacks with a transfer-based prior," *Advances in neural information processing systems*, vol. 32, 2019.

[63] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley, "Adversarial Deepfakes: Evaluating vulnerability of Deepfake detectors to adversarial examples," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3348–3357.

[64] Y. Hou, Q. Guo, Y. Huang, X. Xie, L. Ma, and J. Zhao, "Evading Deepfake detectors via adversarial statistical consistency," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 12 271–12 280.

[65] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[67] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[68] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[69] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[70] Z. Yan, J. Wang, P. Jin, K.-Y. Zhang, C. Liu, S. Chen, T. Yao, S. Ding, B. Wu, and L. Yuan, "Orthogonal subspace decomposition for generalizable AI-generated image detection," in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, Eds., vol. 267. PMLR, 13–19 Jul 2025, pp. 70 268–70 288. [Online]. Available: https://proceedings.mlr.press/v267/yan25b.html

[71] X. Zhou, J. Fei, P. Yu, J. Xie, C. Cheng, and Z. Xia, "PGC: Peak-guided calibration for generalizable AI-generated image detection," 2026.