

USENIX

THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

VOID: Defeating Unauthorized Mimicry in Latent Diffusion Models

Chunlin Qiu, Ang Li, Tianxiao Huang, Ruilin Gan, Yunjie Ge, and Shenyi Zhang, *Wuhan University;* Huayi Duan, *The Hong Kong University of Science and Technology (Guangzhou);* Lingchen Zhao, *Wuhan University;* Chao Shen, *Xi'an Jiaotong University;* Qian Wang, *Wuhan University*

https://www.usenix.org/conference/usenixsecurity26/presentation/qiu-chunlin

This paper is included in the Proceedings of the 35th USENIX Security Symposium.

August 12–14, 2026 • Baltimore, MD, USA

ISBN 978-1-939133-58-8

Open access to the Proceedings of the 35th USENIX Security Symposium is sponsored by



جامعة الملك عبدالله للعلوم والتقنية

King Abdullah University of Science and Technology

ARTIFACT EVALUATED
USENIX
AVAILABLE

VOID: Defeating Unauthorized Mimicry in Latent Diffusion Models

Chunlin Qiu[1], Ang Li[1], Tianxiao Huang[1], Ruilin Gan[2], Yunjie Ge[3], Shenyi Zhang[3], Huayi Duan[4], Lingchen Zhao[1], Chao Shen[5], Qian Wang[1]*
[1] *School of Cyber Science and Engineering, Wuhan University,*
[2] *School of Computer Science, Wuhan University,*
[3] *Institute for Math&AI, Wuhan University,*
[4] *The Hong Kong University of Science and Technology (Guangzhou),*
[5] *School of Cyber Science and Engineering, Xi'an Jiaotong University*

Abstract

While Latent Diffusion Models (LDMs) have revolutionized visual synthesis, they are increasingly exploited for unauthorized mimicry of individuals. Existing defenses inject deceptive perturbations to steer the generated images toward irrelevant targets. However, this approach hinges on an ungrounded assumption: subtle perturbations can maintain their deceptive efficacy throughout an LDM's extensive generation process. In reality, the model's innate restoration mechanism will remove such perturbations and cause individual identities to re-emerge in the images generated.

We propose VOID, a defense framework that overcomes this conundrum by manipulating an LDM's intrinsic stochasticity. VOID perturbs the diffusion pipeline in two novel ways: 1) amplifying the latent encoding errors to shatter an image's semantic structure, and 2) counteracting the target guidance signals to suppress the model's restoration capabilities. This results in a semantic corruption that thwarts any unauthorized mimicry. Notably, the security gain does not come at the price of visual utility, as VOID simultaneously manages to confine perturbations to human-imperceptible regions of protected images. Our comprehensive evaluation of 24 state-of-the-art defenses against 10 mimicry attacks on 5 datasets demonstrates VOID's unprecedented protection power: it increases the average Fréchet Inception Distance (FID) from 113 to 365, a 223% improvement over the strongest defense to date.

1 Introduction

Latent Diffusion Models (LDMs) [10, 42, 50] have emerged as the dominant framework for high-fidelity visual synthesis. Compared to traditional generative models, LDMs demonstrate superior capabilities in zero-shot image editing [45, 72] and few-shot model personalization [24, 52]. These large-scale pre-trained models function as highly personalized creation tools for generating specific contents or artistic styles across diverse scenarios. Furthermore, their computational efficiency significantly reduces hardware requirements, enabling individual users to leverage sophisticated synthesis capabilities on consumer-grade devices. However, the barriers are equally low for malicious actors to perform unauthorized mimicry, e.g., cloning sensitive facial identities or distinctive artistic styles with only a handful of reference images. This poses severe threats to personal privacy [41, 61] and intellectual property rights [38, 56], raising critical concerns about the abuse of generative AI.

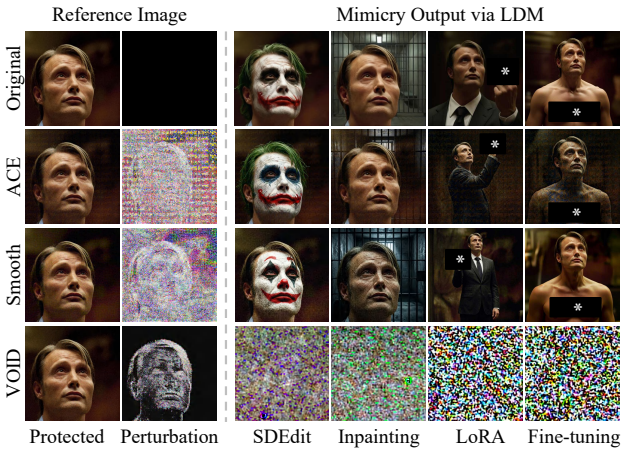


Figure 1: Visual comparison of defenses against unauthorized mimicry. While SOTA methods fail to prevent high-fidelity identity reconstruction, VOID fundamentally collapses the generation trajectory, rendering synthesized outputs semantically void. *Sensitive content is redacted for presentation.*

To counter such threats, a list of defense mechanisms [37, 38, 56, 61] have been proposed. All of them follow a *semantic-steering* paradigm: injecting subtle perturbations into an image that needs protection prior to its release, in the hope of deceiving LDMs and steer their denoising trajectory away from the image's original semantics. The steering process can apply to either the initial latent encoding [37, 56] or the subsequent denoising process [38, 61].

*Corresponding author.

However, our extensive empirical analysis reveals a critical vulnerability in this approach, where defensive perturbations fail to sustain their deceptive efficacy along an LDM's generation process. As illustrated by the visual comparisons in Fig. 1, even the most advanced defense mechanisms, ACE [76] and Smooth [8], fail to prevent the reconstruction of identities they intend to protect. Evidently, superficial visual perturbations are insufficient to prevent the synthesis of semantic content.

We further reveal that the failure of semantic-steering defenses is inevitable due to the fundamental design of LDMs. Specifically, a protected identity effectively acts as a robust semantic anchor that allows LDMs to discard protective perturbations as additive noise. Moreover, defensive signals will undergo structural suppression through the VAE [29] component's encoding and the subsequent diffusion noise injection process. Consequently, the weakened perturbations fail to override the model's strong generative restoration priors.

To overcome these systemic limitations, we propose VOID, a defense framework that realizes an alternative *semantic-corruption* paradigm. As shown in Fig. 2, rather than contending with an LDM's denoising instincts, VOID embraces and weaponizes the model's intrinsic stochasticity. Our core insight is that, while the model always purifies external protective perturbations, it cannot distinguish or suppress its own *internal probabilistic errors*. By amplifying these errors, we can trigger a cascading trajectory divergence that shatters semantic structure and ensure that mimicry of protected images degenerates into meaningless noise.

Specifically, VOID perturbs the LDM pipeline in two novel ways. First, the variational autoencoder (VAE) [29] always introduces probabilistic noise when encoding an image into a latent space, and we amplify such encoding uncertainty to an extent that it becomes irreducible. Second, after the U-Net diffusion process the latent feature vector will be decoded back to an image steered by the Classifier-Free Guidance (CFG) mechanism [21], yet we manage to neutralize its conditional steering effect so that the restoration trajectory collapses. We realize both techniques as perturbations on the image under protection, without modifying any part of the LDM itself.

A distorted image will be of little use. We thereby incorporate a Just Noticeable Difference (JND) [27] model to estimate perceptual sensitivity, effectively confining perturbations to imperceptible regions. Since this restricts our security budget, we further develop a customized perturbation selection strategy to ensure robust defense. Orchestrating all these techniques into a joint optimization process, VOID offers strong resilience to mimicry while maintaining the utility of images.

We validate VOID through an extensive evaluation involving 10 mimicry attacks and 24 state-of-the-art defenses across 5 diverse datasets. Quantitative results show that VOID raises the average FID (Fréchet Inception Distance, the standard quality metric for generative models) from 113 to 365, a 223% improvement over the second best defense, while reducing latent identity correlation to near-zero. Meanwhile,

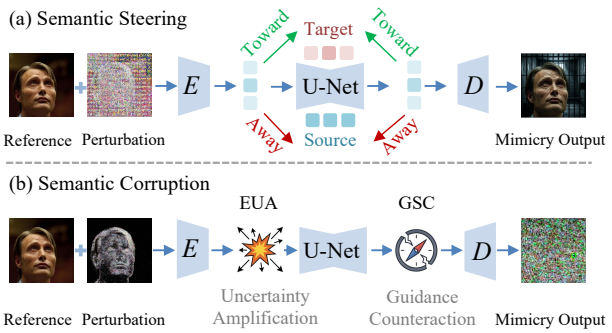


Figure 2: Comparison between the existing semantic steering paradigm and our proposed semantic corruption paradigm.

VOID maintains superior visual utility with a PSNR [26] of 34.41 dB and an SSIM [64] of 0.89; it stands as the only defense in our benchmarks to achieve a perceptual LPIPS [73] below 0.1. Moreover, our stress tests demonstrate its transferability across diverse LDM variants like SD v2.1 [1] and LCM [42] as well as its resilience against advanced adaptive countermeasures including prior-guided purification and mechanism-aware adaptation. These comprehensive results establish VOID as a powerful and highly effective tool for proactive protection against generative mimicry.

In summary, we make the following contributions:

- We provide the first systematic study to dissect the semantic-steering paradigm for countering LDM-based mimicry threats and reveal its inherent limitations.
- We propose VOID, a novel framework that shifts the defensive paradigm to semantic corruption. It leverages the inherent stochasticity of LDMs to defeat mimicry attacks without degrading the utility of protected images.
- We establish comprehensive benchmarks with state-of-the-art defenses against attacks over diverse datasets, demonstrating the superior performance of VOID.

2 Background and Related Work

2.1 Unauthorized Mimicry in LDMs

Latent Diffusion Models (LDMs) [50] enable efficient image editing and personalization, but the same capabilities also lower the cost of mimicking identities, objects, and artistic styles from reference images. We defer technical details of LDMs to Section A. Unauthorized mimicry generally follows two routes: training-based model personalization and inference-based image editing.

Training-based model personalization: This attack family uses the victim's images as unauthorized training data to bind an identity, object, or style to a model or a learnable

token [17, 50, 52]. The adversary first uses the VAE [29] encoder $\mathcal{E}$ to map each image x into the latent space, obtaining $z_0 = \mathcal{E}(x)$. Personalization is then performed by minimizing the diffusion training loss on these unauthorized latents:

$$\min_{\theta'} \mathbb{E}_{t,z_0,\varepsilon,c}[\|\varepsilon - \varepsilon_{\theta'}(z_t,t,c)\|_2^2], \tag{1}$$

where $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon$ is the noisy latent at timestep t.

Here, $\varepsilon_{\theta'}$ denotes the noise predictor after updating the trainable variables θ', and c denotes the conditioning signal, such as a text prompt. These attacks mainly differ in what variables are updated. *Full Fine-tuning (FT)* updates all model parameters, with *DreamBooth (DB)* [52] further using prior preservation to improve fidelity. *Parameter-Efficient Fine-Tuning (PEFT)* [23], such as *LoRA* [24] and *SVDiff* [4], updates only a small set of low-rank or singular-value parameters. *Embedding Optimization*, such as *Textual Inversion* [17], freezes the model and optimizes a textual embedding that serves as a trigger for the unauthorized concept.

Inference-based image editing: This attack family uses off-the-shelf LDMs to mimic a protected concept without updating model parameters. It typically consists of inversion and guided generation. First, inversion maps a reference image into a latent state z_t, either by adding noise to the image, as in *SDEdit* [45] and *Inpainting* [50], or by reversing the diffusion trajectory, as in *DDIM Inversion* [59] and *DiffEdit* [11]. This step preserves source information that can later be reused during generation. The subsequent generation is then steered toward a target concept using conditioning signals. Classifier-Free Guidance (CFG) [21] provides semantic control, while *ControlNet* [72] adds structural constraints such as edges or poses to preserve spatial alignment with the source image. Compared with training-based personalization, these attacks avoid fine-tuning and therefore reduce the cost of unauthorized mimicry, but they still rely on the reference image to provide semantic or structural guidance.

Together, training-based personalization and inference-based editing cover two major unauthorized mimicry surfaces. The former learns the protected concept through parameter or embedding updates, while the latter reuses a protected image through inversion and guidance. These attacks differ in cost, access assumptions, and affected pipeline stages, but both can reproduce identities, objects, or styles from reference images. Therefore, a practical defense should be evaluated against both training-time and inference-time mimicry.

2.2 Defenses against LDM-based Mimicry

To safeguard visual privacy against unauthorized mimicry, existing defenses protect images by adding subtle perturbations δ ($\|\delta\|_p \leq \alpha$) before release. Most methods follow semantic steering, where the perturbation redirects the LDM [50] away from the original semantics or toward an irrelevant target. These defenses can be described by their adversarial setting, such as evasion against frozen models [38, 55] or poisoning against future personalization [61, 63], and by their main target in the LDM [50] pipeline, such as VAE [29] encoding, U-Net [51] denoising, or full-pipeline optimization.

Defenses that target encoding mainly perturb the representation before denoising. Methods such as *Glaze* [56] and *NightShade* [57] add pixel-space perturbations so that the VAE [29] encoder $\mathcal{E}$ maps the protected image to a misleading latent representation. Methods that operate in the latent space, such as LDS [47], further optimize the steering objective in the latent domain, but they still rely on shifting the protected representation toward an external adversarial direction. The objective can be written as

$$\min_{\delta} L(\mathcal{E}(x+\delta), z_{ref}) \quad \text{s.t.} \quad \|\delta\|_p \leq \alpha, \tag{2}$$

where L is a distance metric and z_{ref} specifies either a decoy latent target or a direction away from the original latent.

Defenses that target the denoising trajectory optimize perturbations against the U-Net [51] noise predictor ε_θ at selected diffusion timesteps under the conditioning signal. In evasion settings, methods such as *AdvDM* [38] and *Smooth* [8] optimize perturbations against frozen LDMs by changing their noise estimates across timesteps, with Smooth [8] emphasizing later timesteps where editing noise is stronger to improve robustness for inference-based editing. Other methods use different timestep ranges. For example, *SimAC* [63] and *GAPDiff* [78] focus on earlier timesteps during diffusion training and report stronger protection against personalization-based mimicry, while *AntiDB* [61] incorporates denoising objectives into a training-aware optimization for personalization. The objective can be written as

$$\min_{\delta} \mathbb{E}_t [L(\varepsilon_\theta(z_t(x+\delta), t, c), y_{ref})] \quad \text{s.t.} \quad \|\delta\|_p \leq \alpha, \tag{3}$$

where y_{ref} specifies the steering target, either a decoy noise pattern or a signal that increases prediction error.

Beyond the steering target, defenses also differ in how the perturbation is optimized. Evasion-based methods keep the LDM parameters fixed and update only the perturbation δ, often with PGD [44] optimization, to construct protected samples. This setting may use VAE losses, U-Net losses, or their combination, as in *Mist* [37] and *PhotoGuard* [55]. Poisoning-based methods further account for model updates during personalization, using training-aware or bi-level objectives to optimize protected samples, as in *ACE* [76], *MetaCloak* [41] and *Pretender* [60]. Despite these different optimization regimes, they still rely on steering losses that shift latent representations or denoising trajectories away from protected semantics.

Overall, existing defenses differ in where and how they optimize perturbations, but share the same semantic-steering assumption. A small external signal is expected to persist through the LDM pipeline and redirect generation away from protected semantics. However, LDMs restore coherent semantics from noisy latents, suppressing such external signals while retaining the original semantic anchor. As a result, these defenses often provide limited and attack-specific protection.

3 Threat Model

The threat model involves two parties: a *Data Owner* (defender) and an *Adversary* (attacker).

Adversary. The adversary aims to perform successful mimicry using the data owner's images without authorization.

- **Capabilities:** The adversary's capabilities are defined by their knowledge of the defense. Black-box attackers possess no specific knowledge of the protective perturbations but are capable of employing general pre-processing schemes [14], adversarial purification [75], or extensive hyper-parameter tuning [22] to evade potential protection. White-box attackers are assumed to have full access to a defense's internals, including the specific LDM used to craft the defense and the underlying algorithm. This comprehensive knowledge allows them to devise specialized adaptive attacks [6, 22] or modify the mimicry pipeline [40] to directly circumvent the protection.

- **Objectives:** The primary goal of an adversary is to produce high-fidelity unauthorized mimicry that aligns with adversarial instructions. This involves two competing requirements, which are the successful replication of identity or style features from the source images and the accurate execution of desired modifications, such as changes in background or context.

Data Owner. The data owner seeks to protect personal images by applying protective perturbations to render the aforementioned mimicry attempts ineffective.

- **Capabilities:** We assume the data owner relies solely on a locally available LDM to construct defensive perturbations, with no prior knowledge of the adversary's specific LDM variants, mimicry regimes, or configurations across adversarial settings.

- **Objectives:** The data owner adheres to three primary goals for the protection: (1) Effectiveness: The defense should fundamentally disrupt the adversary's generation process ensuring that any mimicry attempt fails both identity preservation and intent alignment; (2) Utility: The defense should preserve the high visual quality of the protected images, ensuring they retain their utility for legitimate public display with satisfactory perceptual quality; (3) Robustness: The defense should demonstrate transferability across heterogeneous mimicry pipelines and maintain strong resilience against diverse adversarial countermeasures, ranging from generic countermeasures to white-box adaptive attacks.

4 Analysis of Defense Vulnerability

Existing semantic-steering defenses hinge on the assumption that subtle perturbations can maintain their deceptive efficacy throughout the extensive denoising process of LDMs. In reality, models effectively purify these signals to recover the protected identity. We trace this vulnerability to two systemic root causes: intrinsic feature persistence (Section 4.1) and structural signal degradation (Section 4.2).

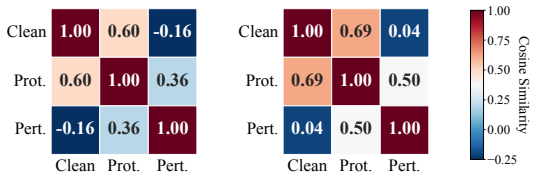
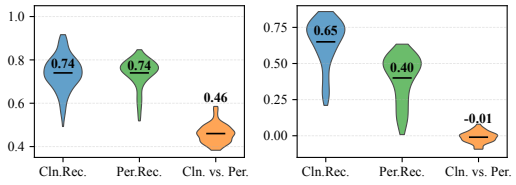


Figure 3: Pairwise latent cosine similarity between clean, protected, and perturbation samples, comparing evasion-based [38] (left) and poisoning-based [61] (right) defenses.



(a) CLIP similarity (b) DINO similarity

Figure 4: Quantitative reconstruction quality, showing high fidelity for both reconstructed images and perturbations.

4.1 Persistence of Intrinsic Features

We first investigate the semantic integrity of latent representations across 24 state-of-the-art defenses, all of which adhere to the semantic-steering paradigm. Specifically, we project original images, protected samples, and isolated perturbations from the TI-Dataset [17] into the VAE [29] latent space of Stable Diffusion v1.5 [53]. By computing pairwise cosine similarities among these components, we quantify the semantic correlations between protected outputs and original identities to evaluate the actual degree of identity erasure achieved by current methods.

> ***Observation 1:*** *Existing defenses fail to decouple the latent representation from the original identity, leading to semantic persistence.*

Ideally, an effective defense should displace the original semantic features so that the resulting latent representation aligns more closely with a deceptive target than with the clean identity. However, Fig. 3 reveals that protected features consistently maintain a disproportionately high similarity to the original samples, showing minimal correlation with the intended deceptive targets. This phenomenon indicates that current methods merely overlay additive noise onto the images

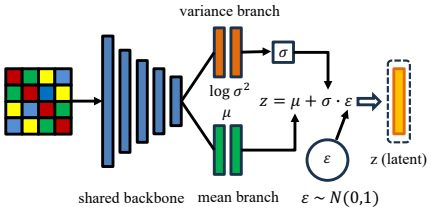


Figure 5: Dual-branch structure of the VAE encoder.

rather than fundamentally redirecting or corrupting the underlying semantic anchor. Consequently, the original identity remains the dominant signal in the latent space, allowing the generative restoration mechanism of the LDM to effectively ignore the defensive signals during the denoising process.

> ***Observation 2:*** *Diffusion models can naturally separate distinct representations from mixed features.*

We further evaluate whether LDMs can decompose protected latent representations into their constituent parts: the clean original identity and the additive perturbation. To verify this separability, we apply DDIM inversion [59] to protected latents and perform denoising guided by neutral prompts (e.g., "image") to minimize semantic leakage. Quantitative evaluations across 24 defenses (Fig. 4) reveal high CLIP [49] and DINO [46] similarities between the reconstructed components and their ground truths. We provide qualitative evidence of this separation in Section B. These results demonstrate the model's capacity to isolate and recover original semantics from superficial protective overlays, explaining the re-emergence of protected identities.

4.2 Degradation of Defensive Signals

> ***Observation 3:*** *The VAE mean branch structurally attenuates defensive perturbations during latent mapping.*

In addition to semantic persistence, we investigate how defensive perturbations survive through the LDM pipeline. As illustrated in Fig. 5, the latent representation is jointly determined by the mean (μ) and variance (σ) branches of the VAE encoder. We evaluate the sensitivity of these branches by computing their Lipschitz constants (K) to quantify the survival of defensive perturbations passing through the encoder bottleneck. Since semantic-steering defenses aim to alter deterministic identity features, they fundamentally rely on manipulating the VAE's mean (μ) branch. However, results in Table 1 reveal a structural bias across five mainstream VAEs. Specifically, the mean branch exhibits a contractive mapping, effectively attenuating these steering signals. In contrast, the variance branch remains approximately eight times more sensitive to input changes. This disparity exposes a fundamental bottleneck for defenses that rely primarily on the manipulation of the mean latent vector.

Table 1: Lipschitz constants of five VAE encoders.

VAE Encoder	K_μ	K_σ	Ratio (K_σ/K_μ)
SDv1-VAE [53]	0.6747	5.5859	8.2791
SDv2-D16-VAE [1]	0.8759	12.7562	14.5635
SDv3-VAE [3]	0.6334	6.1018	9.6334
SDXL-VAE [2]	0.8590	3.9507	4.5992
SDXL-EQ-VAE [30]	0.8708	3.9087	4.4886
AVG	0.7828	6.4607	8.2533

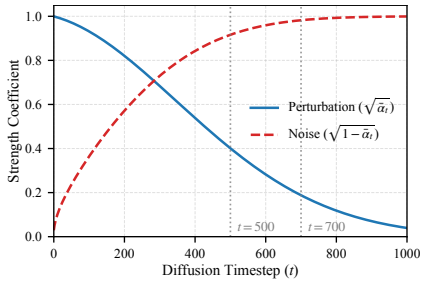


Figure 6: Decay of protective signals during Stable Diffusion v1.5 [53] forward diffusion process.

Following the VAE bottleneck, we further examine the impact of systematic noise injection on the surviving protective signals. In LDM-based mimicry, the perturbed latent $z_0 + \delta_z$ is transformed into a corresponding noisy state z_t according to the forward process:

$$z_t = \sqrt{\bar{\alpha}_t}(z_0 + \delta_z) + \sqrt{1-\bar{\alpha}_t}\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \mathbf{I}), \qquad (4)$$

where the scaling coefficient $\sqrt{\bar{\alpha}_t}$ is governed by the LDM noise schedule. As the timestep t increases to facilitate substantial content modification, $\sqrt{\bar{\alpha}_t}$ decreases monotonically according to the schedule, such as the scaled-linear schedule in Stable Diffusion v1.5 [53]. As shown in Fig. 6, the perturbation weakens because the residual signal decays while noise becomes dominant. For example, at $t = 700$, which is a standard setting for robust mimicry tasks, the protective signal is attenuated to about 0.2 of its initial magnitude. This indicates that the noise injection inherent to the editing pipeline effectively washes out surviving defensive signals, making steering-based defenses fragile in practical attacks.

> ***Observation 4:*** *Initial noise in the editing process further degrades the remaining protective signals.*

These observations expose that the semantic-steering paradigm is fundamentally vulnerable due to its reliance on external deceptive perturbations. Being extrinsic to the generative process, these signals are inevitably identified as additive noise and purified by the model's restoration mechanisms, including VAE structural compression and diffusion stochasticity, rather than preserved as stable semantic changes. Consequently, the efficacy of such defenses is inherently upper-bounded by the model's restoration capability, rendering them structurally insufficient for robust protection.

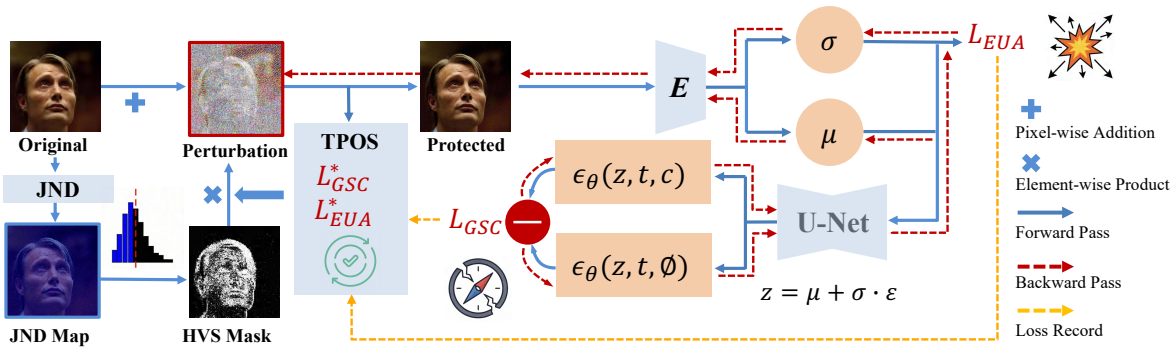


Figure 7: Overview of the proposed VOID framework. The pipeline jointly optimizes the protective perturbation by maximizing encoding uncertainty (L_{EUA}) during VAE encoding and counteracting guidance signals (L_{GSC}) during U-Net denoising.

5 Design of VOID

5.1 Defense Overview

Our analysis in Section 4 establishes that the LDM's restoration priors act as a robust semantic filter, effectively stripping deceptive perturbations introduced by semantic-steering defenses as additive noise. Consequently, a robust defense cannot rely on extrinsic confrontation, as the model is trained to rectify such deviations.

To overcome this, we exploit a critical structural vulnerability. Specifically, while the model effectively purifies external noise, it is strictly compelled to preserve its own intrinsic probabilistic errors to maintain the continuity of the latent manifold. Driven by this insight, we propose a paradigm shift to semantic-corruption. Instead of fighting the model's restoration mechanism, VOID weaponizes this necessity, amplifying internal deviations to trigger a cascading collapse of the generative trajectory.

VOID operationalizes this strategy by targeting two distinct forms of system uncertainty: the encoding deviation inherent in VAE sampling and the stochastic randomness injected throughout the diffusion chain. To exploit the former, we introduce encoding uncertainty amplification (EUA), which penetrates the VAE bottleneck by escalating latent encoding errors into irreducible stochasticity. This ensures that the initial semantic structure is shattered at the very source of the generation process. Simultaneously, to manipulate the latter, we employ guidance signal counteraction (GSC), which leverages diffusion stochasticity to collapse the Classifier-Free Guidance [21] vector. By neutralizing the steering force required for model restoration, GSC prevents the LDM from rectifying the chaos seeded by EUA.

Realizing this paradigm, however, requires reconciling aggressive corruption with high visual utility. We resolve this tension via HVS-guided perturbation masking (HGPM), which restricts artifacts to human-imperceptible regions. To stabilize optimization under these strict spatial constraints, we further integrate timestep-partitioned optimal selection (TPOS). As illustrated in Fig. 7, VOID synthesizes these components to achieve a robust equilibrium between unprecedented mimicry prevention and superior image fidelity.

5.2 Encoding Uncertainty Amplification

To initiate intrinsic corruption at the source of the diffusion chain, we first target the VAE bottleneck. Standard LDMs employ a VAE [29] to map images into a latent space distribution $\mathcal{N}(\mu, \sigma^2)$. Structurally, the encoder $\mathcal{E}$ consists of two parallel branches: a mean branch (μ) that captures deterministic semantic features, and a variance branch (σ) that represents encoding uncertainty.

Our analysis in Section 4 reveals that the mean branch acts as a contractive mapping, effectively attenuating external perturbations to restore original semantics. Instead of engaging in a futile struggle against this robust mean branch, VOID pivots to exploit the highly sensitive variance branch. Our Lipschitz evaluation confirms this strategic shift: the variance branch exhibits a sensitivity coefficient ($K_\sigma \approx 6.5$) nearly an order of magnitude higher than that of the mean branch ($K_\mu \approx 0.8$). By targeting this sensitive component, we can aggressively amplify inherent encoding uncertainty to penetrate the VAE bottleneck without being suppressed.

In the latent diffusion framework, this stochasticity is not mere noise but a structural necessity defined by the reparameterization trick: $\mathbf{z} = \mu + \sigma \odot \mathbf{n}$. Unlike deterministic features, the variance σ cannot be filtered by the model's restoration mechanism without destroying the continuity of the latent manifold. Consequently, maximizing this variance fundamentally alters the topology of the latent representation, creating a probabilistic shield that forces the model to interpret the perturbation as inherent ambiguity within the source data. We operationalize this by formulating the EUA loss to maximize the aggregate predicted uncertainty:

$$L_{\text{EUA}} = \sum_i \sigma_{\mathcal{E},i}(\mathbf{x} + \delta). \tag{5}$$

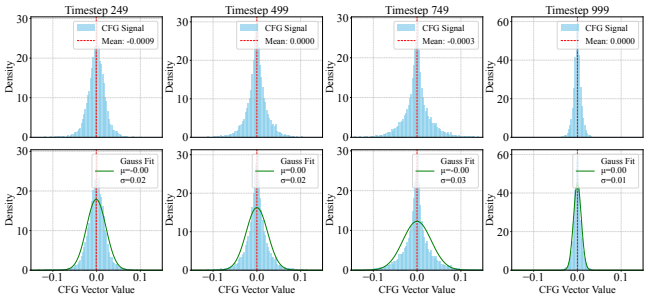


Figure 8: Distribution analysis of the CFG vector across different timestep ranges t. The distributions exhibit near-zero mean and small variance, resembling Gaussian distributions.

By forcing the encoder to map the image into a region of maximal uncertainty, this objective converts the defense into irreducible stochasticity. This seeds the diffusion pipeline with high-entropy chaos, ensuring that subsequent mimicry collapses into meaningless noise.

5.3 Guidance Signal Counteraction

While EUA initiates latent chaos at the source, the LDM inherently employs Classifier-Free Guidance (CFG) [21] to rectify such deviations during the denoising process. CFG functions as the primary restoration engine by extrapolating a guidance vector $\Delta\varepsilon_t$, which represents the semantic discrepancy between conditional and unconditional noise predictions. This vector acts as the steering force that aligns the generation trajectory with the target prompt c as follows:

$$\Delta\varepsilon_t(\delta) = \varepsilon_\theta(\mathbf{z}_t, t, c) - \varepsilon_\theta(\mathbf{z}_t, t, \emptyset). \tag{6}$$

To prevent the model from recovering protected features via this mechanism, guidance signal counteraction (GSC) aims to fundamentally collapse the LDM's restoration capability. Our key insight is that the effectiveness of CFG depends not merely on the magnitude of $\Delta\varepsilon_t$ but on its directional variance, which provides the necessary gradient information for semantic alignment. As visualized in Fig. 8, the values of $\Delta\varepsilon_t$ follow an approximately Gaussian distribution with near-zero mean. Leveraging this stochastic nature, GSC neutralizes the restoration force by collapsing the signal distribution rather than simply suppressing its magnitude. We achieve this by minimizing the standard deviation of the guidance vector to compress the signal into a non-informative state:

$$L_{\text{GSC}}(\delta) = \mathbb{E}_{t,\varepsilon}\Big[\sigma\left(\Delta\varepsilon_t(\delta)\right)\Big]. \tag{7}$$

Minimizing this objective effectively decouples the generation from its semantic anchor. By inducing a non-informative guidance signal, we ensure that the model receives no consistent directional information to recover original identities, thereby forcing the denoising process to propagate the chaos seeded by EUA throughout the entire diffusion chain.

Algorithm 1: VOID Protection Framework

Input : Original image $\mathbf{x}$ and proxy prompt $\hat{c}$, VAE Encoder E and LDM ε_θ, Perturbation budget α, Step size η, Total iterations N, JND masking ratio ρ, TPOS partition number K, Loss weight λ.

Output : Protected image $\mathbf{x}'$

Initialize $\delta \leftarrow 0, \delta^* \leftarrow 0, L_{\text{best}} \in \mathbb{R}^K \leftarrow -\infty$

$M_{\text{HVS}} \leftarrow \mathbb{I}(\text{TD-JND}(\mathbf{x}) \in \text{Top}_\rho)$

for $n = 1$ ***to*** N

  $t \sim U[1,T], k \leftarrow \lfloor t \cdot K/T \rfloor$

  $z_0' \leftarrow E(\mathbf{x}+\delta), \varepsilon \sim N(0,\mathbf{I})$

  // EUA

  $L_{\text{EUA}} \leftarrow \sum \sigma(z_0')$

  $z_t' \leftarrow \sqrt{\bar{\alpha}_t} z_0' + \sqrt{1-\bar{\alpha}_t}\varepsilon$

  // GSC

  $L_{\text{GSC}} \leftarrow \sigma(\varepsilon_\theta(z_t', t, \hat{c}) - \varepsilon_\theta(z_t', t, \emptyset))$

  $L \leftarrow \lambda L_{\text{EUA}} - L_{\text{GSC}}$

  // HGPM

  $\delta \leftarrow \text{Proj}_{\ell_\infty \le \alpha}(\delta + \eta \cdot \text{sign}(\nabla_\delta L)) \odot M_{\text{HVS}}$

  // TPOS

  if $L > L_{best}[k]$ **then**

    $L_{\text{best}}[k] \leftarrow L, \delta^* \leftarrow \delta$

  end

end

return $\mathbf{x}' \leftarrow \mathbf{x} + \delta^*$

To handle practical black-box scenarios where the adversary's attack prompt c is unknown, we employ a shadow condition strategy. We use blind captioning (e.g., BLIP [36]) to generate a proxy prompt $\hat{c}$ that captures the inherent visual attributes of the reference image. Because unauthorized mimicry fundamentally targets these visual features, $\hat{c}$ represents a critical semantic intersection with potential attack prompts. This approach allows VOID to disrupt the fundamental visual features targeted by mimicry, ensuring the defense remains effective even when attack prompts are unknown.

5.4 Joint Optimization Framework

To operationalize semantic corruption, we integrate the EUA and GSC into a joint optimization objective:

$$\max_\delta J(\delta) = \lambda L_{\text{EUA}}(\mathbf{x}+\delta) - L_{\text{GSC}}(\mathbf{x}+\delta), \tag{8}$$

where λ balances the trade-off between latent chaos and guidance suppression ($\lambda = 10$ by default). To preserve visual utility, we frame this as a constrained optimization problem using two mechanisms to ensure stealthiness and stability.

HVS-Guided Perturbation Masking. Standard defenses typically impose a spatially uniform ℓ_∞-norm budget, which disregards local texture sensitivity and introduces visible artifacts

in smooth regions. To rectify this, VOID employs HVS-guided perturbation masking (HGPM) to adaptively allocate the security budget. Leveraging the Top-Down Just Noticeable Difference (TD-JND) [27] model, we compute a redundancy map that quantifies the perceptual tolerance of each pixel. We then construct a binary mask M_{HVS} by retaining only the top $\rho\%$ regions with the highest JND values, effectively confining the perturbation to a texture-aware subspace:

$$\delta_{k+1} \leftarrow \text{Proj}_{\ell_\infty \le \alpha}(\delta_{k+1} \odot M_{\text{HVS}}). \tag{9}$$

This approach ensures that aggressive corruption signals are embedded exclusively within human-insensitive regions, maintaining high visual fidelity.

Timestep-Partitioned Optimal Selection. While HGPM ensures stealthiness, strictly enforcing spatial constraints can limit the search space and cause optimization oscillation. Furthermore, because timesteps t are randomly sampled in Eq. (8), a high loss value often reflects sampling bias rather than actual perturbation quality. To recover performance, we introduce timestep-partitioned optimal selection (TPOS). We divide diffusion timelines into K disjoint intervals and maintain separate optima for each partition. This allows the optimization to decouple sampling variance and independently recover the optimal perturbation for each timestep partition.

Overall Algorithm. The complete execution flow is detailed in Algorithm 1. By synthesizing static perceptual constraints from HGPM with the dynamic joint optimization of EUA and GSC, VOID achieves a robust equilibrium between unprecedented mimicry prevention and superior image fidelity.

6 Evaluation

6.1 Evaluation Setup

Datasets. We evaluate unauthorized mimicry on 5 datasets across 3 tasks: *identity mimicry* on CelebA-HQ [28] and VGGFace2 [7], *object mimicry* on TI-Dataset [17] and DB-Dataset [52], and *style mimicry* on WikiArt [54].

Models. We evaluate 4 LDMs: Stable Diffusion (SD) v1.5 [53], SD v2.1 [1], DreamShaper [43], and Latent Consistency Models [42]. We also evaluate two transformer-based LDMs, FLUX.1 Kontext [33] and FLUX.2 Klein [32]. Following prior works [60, 78], we use SD v1.5 [53] as the surrogate.

Mimicry Attacks. We evaluate 10 mimicry attacks spanning *training-based personalization* and *inference-based editing*. The former includes Fine-tuning (FT) [50], DreamBooth (DB) [52], LoRA (LR) [24], SVDiff (SV) [4], and Textual Inversion (TI) [17]; the latter includes SDEdit (SE) [45], Inpainting (IP) [50], ControlNet (CN) [72], DDIM Inversion (DI) [16], and DiffEdit (DF) [11].

Mimicry Defenses. We compare against 24 defenses, including 12 evasion-based methods, LDMR [70], AdvDM [38], Mist [37], PhotoGuard [55], Glaze [56], Smooth [8], SDST [68], PID [35], DiffGuard [9], Nightshade [57], GAPDiff [78], and LDS [47], and 12 poisoning-based methods, EUDP [74], AntiDB [61], ACE [76], SimAC [63], Metacloak [41], DisDiff [39], CAAT [65], PAP [62], AntiDiff [77], GoodAC [66], Pretender [60], and HAAD [67].

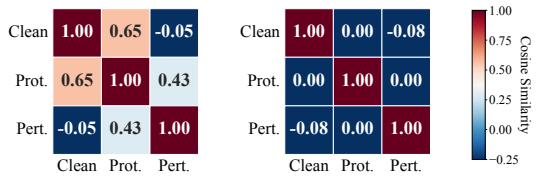


Figure 9: Latent semantic erasure analysis.

Table 2: Resilience against intense diffusion. Comparison of protection effectiveness (FID ↑) across varying timesteps (t).

Defense	Diffusion Timestep (t)				
	100	300	500	700	900
Smooth	127	148	158	124	7
ACE	168	142	57	21	10
VOID (Ours)	**286**	**293**	**296**	**301**	**306**

Metrics. We evaluate two dimensions: *Effectiveness*, measured by FID [20], CLIP-S [19], and CLIP-I [49], and *Utility*, using visual quality metrics, including PSNR [26], SSIM [64], FSIM [71], VIF [58], GMSD [69], LPIPS [73], DeepIQA [5], PieAPP [48], DISTS [15], and ContentLoss [18].

Implementation. All defenses use the same perturbation budget, $\varepsilon = 8/255$ under the ℓ_∞-norm. Detailed configurations are provided in Section C.

6.2 Validation of Defense Mechanisms

We first validate whether semantic corruption achieves latent semantic erasure and persistence under diffusion-induced signal degradation. For latent erasure, we evaluate VAE [29] latent identity correlation. As shown in Fig. 9, steering-based defenses preserve correlation after VAE compression (Avg. Sim ≈ 0.65), whereas VOID reduces it to nearly zero (0.00), decoupling protected representations from semantics. Section E further shows that this corruption is concept-specific rather than global pipeline failure.

We then evaluate whether the protection remains stable under diffusion noise injection. As shown in Table 2, Smooth [8] and ACE [76] collapse at large timesteps (FID ≈ 10), while VOID maintains strong disruption (Avg. FID ≈ 296) and even strengthens as the timestep increases. Based on this result, we adopt adversary-favored settings in subsequent experiments. For inference-based image editing, we use a large timestep of $t = 700$ to increase noise injection and editing strength, as in prior works [8, 70]. For training-based personalization, we use $t = 300$ as the starting point of the diffusion training range. These high-step settings impose stronger signal degradation and provide a rigorous stress test for defense persistence.

Table 3: Quantitative evaluation of defense efficacy. Results are averaged across 5 datasets (TI-Dataset [17], DB-Dataset [52], CelebA-HQ [28], VGGFace2 [7] and WikiArt [54]) and 6 mimicry methods: FT [50], SV [4], LR [24], SE [45], IP [50] and CN [72]). Arrows indicate the direction of better protection performance. **Bold** and underlined values denote the best and second-best results, respectively.

Type	Defense	Training-based mimicry			Inference-based mimicry			Average		
		FID [20] ↑	CLIP-S [19] ↓	CLIP-I [49] ↓	FID [20] ↑	CLIP-S [19] ↓	CLIP-I [49] ↓	FID [20] ↑	CLIP-S [19] ↓	CLIP-I [49] ↓
Poisoning	ACE [76]	146.22	24.52	0.69	47.87	27.04	0.80	97.04	25.78	0.74
	AntiDB [61]	69.32	28.49	0.77	65.86	26.24	0.84	67.59	27.36	0.81
	AntiDiff [77]	55.37	28.97	0.79	48.55	26.46	0.86	51.96	27.71	0.83
	CAAT [65]	67.90	28.73	0.77	45.40	26.74	0.86	56.65	27.74	0.81
	DisDiff [39]	71.97	28.75	0.77	28.11	26.79	0.88	50.04	27.77	0.82
	EUDP [74]	76.59	28.62	0.76	29.69	27.12	0.87	53.14	27.87	0.82
	GoodAC [66]	70.63	28.38	0.77	74.73	26.11	0.84	72.68	27.25	0.80
	MetaCloak [41]	44.03	29.09	0.81	38.48	26.65	0.87	41.26	27.87	0.84
	PAP [62]	72.26	28.36	0.76	87.86	26.33	0.83	80.06	27.35	0.80
	Pretender [60]	56.95	28.98	0.79	41.79	26.79	0.86	49.37	27.88	0.82
	SimAC [63]	44.75	29.19	0.82	22.67	26.60	0.89	33.71	27.90	0.85
	HAAD [67]	70.95	28.56	0.76	78.26	26.89	0.83	74.60	27.72	0.80
-	RandomNoise	22.15	29.16	0.88	5.64	26.55	0.93	13.90	27.85	0.90
Evasion	AdvDM [38]	53.27	29.18	0.80	39.85	27.07	0.85	46.56	28.13	0.82
	DiffGuard [9]	31.29	29.65	0.84	14.85	26.56	0.89	23.07	28.11	0.87
	GAPDiff [78]	103.96	26.47	0.71	33.48	26.85	0.86	68.72	26.66	0.78
	Glaze [56]	138.80	24.57	0.69	48.96	27.31	0.81	93.88	25.94	0.75
	LDMR [70]	46.70	29.03	0.79	43.66	26.45	0.82	45.18	27.74	0.81
	MIST [37]	143.17	24.53	0.69	48.32	27.39	0.81	95.74	25.96	0.75
	NightShade [57]	140.97	24.58	0.69	49.98	27.09	0.80	95.48	25.83	0.75
	PID [35]	34.14	29.26	0.82	27.68	27.28	0.84	30.91	28.27	0.83
	PhotoGuard [55]	32.18	29.20	0.83	23.19	27.01	0.85	27.69	28.11	0.84
	SDST [68]	48.09	28.20	0.79	35.61	27.39	0.82	41.85	27.79	0.81
	Smooth [8]	65.96	28.68	0.78	160.92	25.50	0.79	113.44	27.09	0.78
	LDS [47]	23.55	28.77	0.84	11.42	27.21	0.91	17.48	27.99	0.87
	VOID (Ours)	**404.94**	**20.08**	**0.53**	**325.80**	**18.82**	**0.49**	**365.37**	**19.45**	**0.51**

6.3 Effectiveness against Mimicry Attacks

Quantitative Results. Table 3 reports the mimicry defense effectiveness averaged across 5 datasets and 6 mimicry attacks. To benchmark performance against the most capable counterparts, we identify ACE [76] and Smooth [8] as the leading representatives of poisoning and evasion defenses, respectively. Despite their relative effectiveness, VOID demonstrates overwhelming superiority, outperforming these baselines by an average factor of **3.2**× in FID [20] (elevating the average disruption level from 113.44 to **365.37**). Beyond substantial visual degradation, the CLIP-S [19] metric reveals a fundamental mechanistic distinction: even the strongest baselines retain high semantic alignment (25.83) perilously close to unprotected images, indicating persistent identity leakage, whereas VOID drastically reduces CLIP-S [19] to **19.45**. Such semantic persistence in existing defenses explains their specialized yet fragile nature, as the latent identity remains accessible to diverse attack vectors. Furthermore, Fig. 10 reveals that existing semantic-steering paradigms fail to provide uniform protection across diverse mimicry threats. Specifically, leading defenses such as ACE [76] and Smooth [8] exhibit significant performance disparities across different mimicry paradigms, as they often provide effective defense against limited attack vectors while remaining vulnerable to others. VOID overcomes this limitation by maintaining a comprehensive protection coverage, consistently achieving superior mimicry disruption in both training-based model personalization and inference-based image editing scenarios.

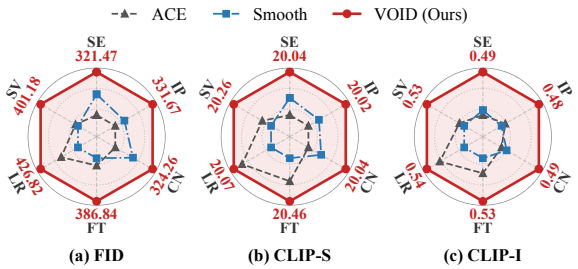


Figure 10: Defense effectiveness across diverse mimicry scenarios. Unlike baselines that exhibit jagged profiles across different attacks, VOID consistently encompasses others, demonstrating comprehensive protection.

Qualitative Results. Qualitative results in Fig. 11 confirm that although steering-based baselines like ACE [76] and Smooth [8] introduce distortion into mimicry results, original features frequently re-emerge. In contrast, VOID induces a robust semantic collapse that consistently renders mimicry outputs as semantically void noise. Visual inspection reveals that these outputs manifest as pure random noise with no perceptible trace of the original identity or style, confirming that VOID prevents the model from reconstructing any meaningful content from the corrupted latents. By corrupting the denoising trajectory to neutralize the model's restoration capability, VOID effectively severs the semantic link between original features and outputs to render mimicry attempts infeasible.

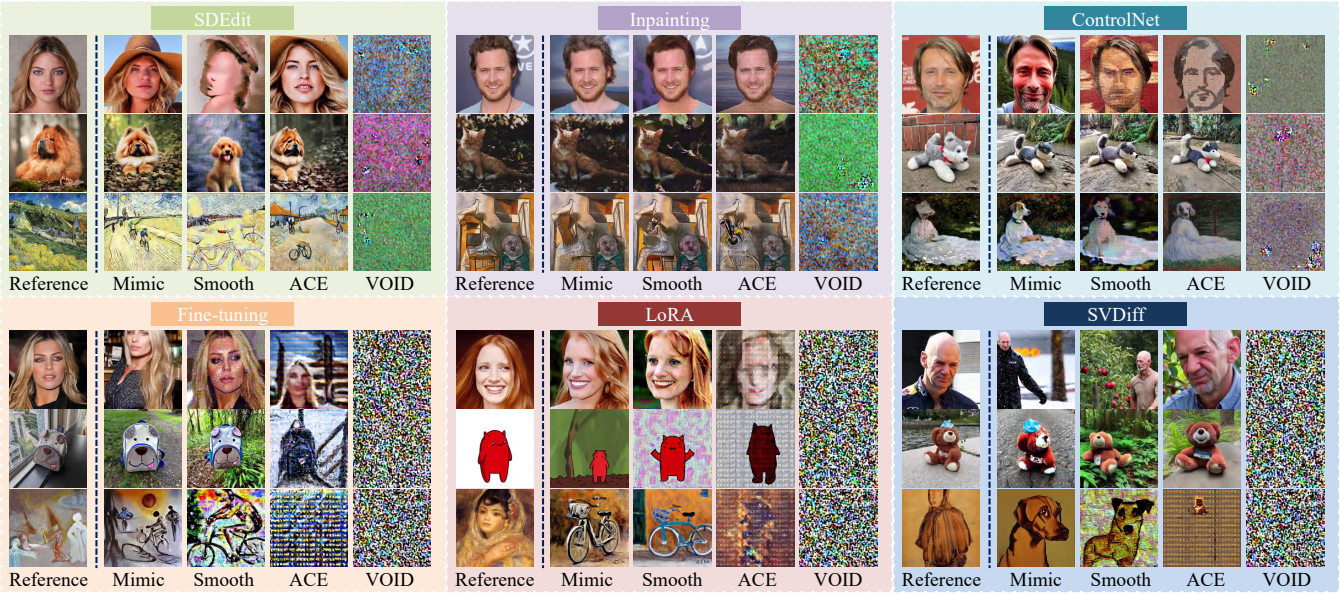


Figure 11: Qualitative comparison of defense efficacy. Rows 1-3 correspond to *inference-based image mimicry* (SE [45], IP [50], CN [72]), and Rows 4-6 correspond to *training-based model personalization* (FT [50], LoRA [24], SVDiff [4]). Columns compare the unprotected baseline (Mimic), the top-performing specialized defenses (Smooth [8] and ACE [76]), and our VOID.

Table 4: Imperceptibility evaluation. VOID is compared against the top-4 visual quality baselines among 24 methods.

Defense	Signal-level Metrics					Neural-based Metrics				
	PSNR [26]↑	SSIM [64]↑	FSIM [71]↑	VIF [58]↑	GMSD [69]↓	LPIPS [73]↓	DISTS [15]↓	DeepIQA [5]↑	PieAPP [48]↓	ContentLoss [18]↓
AntiDiff [77]	33.708	0.870	0.965	0.598	0.048	0.162	0.199	-0.026	1.330	1297.6
DiffGuard [9]	33.841	0.854	0.975	0.611	0.036	0.135	0.192	-0.019	0.912	1247.8
PhotoGuard [55]	33.367	0.843	0.979	0.584	0.028	0.122	0.219	-0.014	0.728	1589.5
SDST [68]	32.903	0.823	0.977	0.534	0.028	0.116	0.291	-0.017	0.716	2182.5
VOID (Ours)	**34.408**	**0.888**	**0.985**	**0.624**	**0.019**	**0.096**	**0.177**	**-0.010**	**0.679**	**1204.9**

6.4 Practicality for Real-world Deployment

Visual Quality. We evaluate the visual quality of the protected samples produced by VOID using 10 visual quality metrics. This evaluation encompasses both traditional signal-level measures, such as PSNR [26] and SSIM [64], and neural-based perceptual models, including LPIPS [73] and DeepIQA [5]. To ensure a rigorous assessment of utility, we benchmark VOID against 24 state-of-the-art defenses across diverse image domains. Table 4 compares VOID against the top-4 baselines. As reported, VOID consistently outperforms existing defenses across all evaluated metrics, achieving **34.408** dB PSNR and **0.888** SSIM. Notably, VOID is the only method achieving an LPIPS below 0.1 (**0.096**), confirming high visual quality. While leading baselines such as PhotoGuard [55] and SDST [68] apply indiscriminate global perturbations, the protective noise from VOID is remarkably sparse (Fig. 12). This superior image utility stems from our HVS-guided perturbation masking strategy, which leverages the Just Noticeable Difference (JND) [27] model to adaptively allocate perturbations to human-insensitive regions.

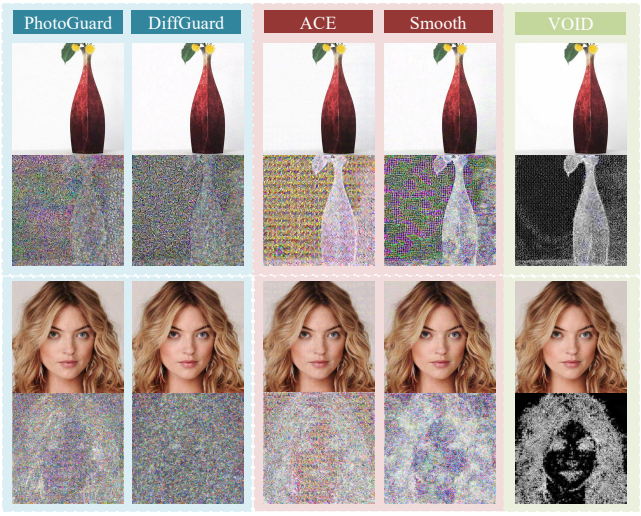


Figure 12: Visual quality comparison. Protected samples and their corresponding perturbation maps are presented, with perturbations magnified by 30× for visibility.

Table 5: Cross-model transferability evaluation on TI-Dataset [17]. Bold indicates the best performance among defenses.

Defense	Attack	SD v1.5* [53]			SD v2.1 [1]			DreamShaper [43]			LCM [42]		
		FID↑	CLIP-S↓	CLIP-I↓	FID↑	CLIP-S↓	CLIP-I↓	FID↑	CLIP-S↓	CLIP-I↓	FID↑	CLIP-S↓	CLIP-I↓
Smooth [8]	SDEdit [45]	121.34	30.20	0.79	108.15	27.00	0.77	63.72	33.55	0.84	86.73	29.50	0.85
	Inpainting [50]	119.15	28.86	0.85	141.21	29.27	0.84	77.87	30.63	0.87	66.05	30.35	0.90
	ControlNet [72]	110.32	30.09	0.82	103.77	27.45	0.80	55.46	32.75	0.85	113.88	30.64	0.84
ACE [76]	Fine-tuning [50]	219.69	24.26	0.64	202.78	23.33	0.64	214.95	25.15	0.66	139.47	22.26	0.76
	LoRA [24]	208.58	23.30	0.67	253.61	20.97	0.65	197.52	23.83	0.66	221.76	19.93	0.79
	SVDiff [4]	84.60	29.42	0.74	51.74	28.66	0.80	103.88	29.10	0.72	96.49	21.51	0.84
VOID (Ours)	SDEdit [45]	299.18	19.67	0.52	301.15	19.58	0.58	310.11	19.57	0.52	304.00	19.67	0.56
	Inpainting [50]	319.47	19.68	0.55	319.13	19.54	0.54	323.30	19.57	0.56	331.84	19.69	0.57
	ControlNet [72]	300.53	19.67	0.53	300.93	19.58	0.59	305.93	19.57	0.53	316.77	19.68	0.60
	Fine-tuning [50]	416.51	19.71	0.56	368.45	18.86	0.53	377.75	20.40	0.54	455.36	19.42	0.65
	LoRA [24]	384.01	19.20	0.57	351.80	19.07	0.57	445.43	19.86	0.56	498.19	19.32	0.69
	SVDiff [4]	343.43	19.60	0.54	365.25	18.82	0.52	334.13	19.96	0.53	414.78	19.17	0.70

Table 6: Practicality comparison across defenses.

Defense	Time	Memory	Effectiveness		Visual Quality	
	min↓	GB↓	CLIP-S↓	CLIP-I↓	SSIM↑	LPIPS↓
Smooth [8]	3.52	**8.63**	27.09	0.78	0.817	0.186
PhotoGuard [55]	14.40	33.09	28.11	0.84	0.843	0.122
SDST [68]	**2.55**	9.80	27.79	0.81	0.823	0.116
ACE [76]	6.13	13.17	25.78	0.74	0.813	0.158
AntiDiff [77]	7.71	16.92	27.71	0.83	0.870	0.162
Pretender [60]	5.86	13.34	27.88	0.82	0.842	0.183
VOID (Ours)	3.79	9.95	**19.45**	**0.51**	**0.888**	**0.096**

Table 7: Cross-model transferability on transformer-based latent diffusion models. Comparison on TI-Dataset [17].

Defense	FLUX.1 Kontext [33]			FLUX.2 Klein [32]		
	FID↑	CLIP-S↓	CLIP-I↓	FID↑	CLIP-S↓	CLIP-I↓
RandomNoise	7.64	28.42	0.93	11.13	28.11	0.91
Smooth [8]	27.63	27.42	0.84	19.87	27.64	0.86
ACE [76]	12.44	28.01	0.90	7.30	28.36	0.92
VOID (Ours)	**105.27**	**24.58**	**0.68**	**86.61**	**25.15**	**0.71**

Computational Overhead. As shown in Table 6, we compare VOID with six representative defenses covering the main cost profiles of prior work, including Smooth [8] with vanilla evasion, SDST [68] with SDS acceleration, PhotoGuard [55] with full-pipeline gradients, ACE [76] with vanilla poisoning, AntiDiff [77] with prompt tuning, and Pretender [60] with SGBP acceleration. Runtime and memory are measured on a single NVIDIA A100 GPU. VOID requires 3.79 minutes and 9.95 GB memory, which is comparable to lightweight methods such as Smooth [8] and SDST [68] and substantially lower than PhotoGuard [55]. VOID achieves the best effectiveness and visual quality, with the lowest CLIP-S/CLIP-I scores and the best SSIM/LPIPS results. These results demonstrate that VOID provides a favorable trade-off between deployment efficiency, protection strength, and image utility.

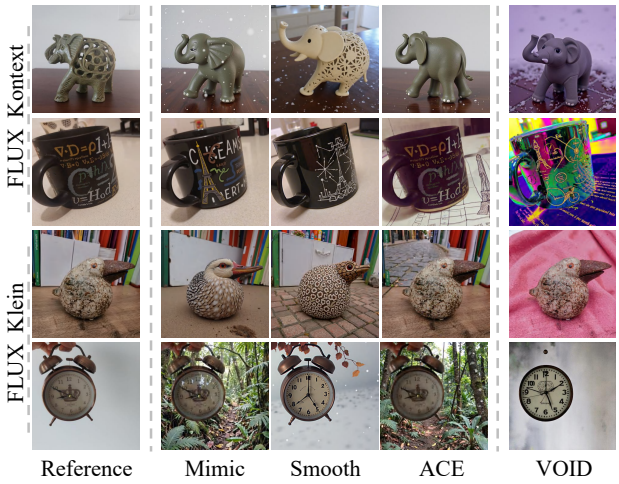


Figure 13: Comparison of defense efficacy on FLUX models.

6.5 Transferability across Diverse Scenarios

Cross-model transferability. We assess the transferability of VOID by crafting perturbations on SD v1.5 [53] and evaluating them on SD v2.1 [1], DreamShaper [43], and LCM [42], which differ in text encoders, community fine-tuning, and distillation mechanisms. Table 5 reports quantitative results. Unlike baselines that suffer significant performance degradation on unseen models, VOID exhibits superior robustness. Notably, while ACE [76]'s protection collapses on the distilled LCM [42] during fine-tuning (FID 139.47), VOID maintains a dominant FID of **455.36** (a 3.2× improvement). This shows that VOID targets the intrinsic generative consensus shared across diffusion models rather than model-specific surrogate parameters. We further evaluate VOID on transformer-based LDMs, including FLUX.1 Kontext [33] and FLUX.2 Klein [32]. As shown in Table 7, VOID achieves the best results on both models, with the highest FID and lowest CLIP-S/CLIP-I. Figure 13 shows that FLUX outputs mainly exhibit severe color drift and semantic distortion rather than pure-noise collapse, still making the edited images unusable.

Table 8: Robustness evaluation of VOID on TI-Dataset [17] against various countermeasures.

Defense	Attack	Model-Agnostic Transformation			Prior-Guided Purification			Mechanism-Aware Adaptation	
		GNoise	GSmooth	JPEG Comp.	Impress [6]	Impress++ [22]	Upscale [13]	DiffShortcut [40]	TimestepSel.
Smooth [8]	SDEdit [45]	111.34	25.50	101.90	62.43	50.20	36.26	76.97	7.81
	Inpaint [50]	109.15	72.39	109.15	109.15	97.75	109.15	86.74	5.27
	ControlNet [72]	100.32	29.88	92.46	61.62	47.98	83.33	61.06	5.88
ACE [76]	Fine-tuning [50]	209.69	73.33	134.53	197.91	152.96	192.34	87.99	40.99
	LoRA [24]	198.58	66.78	114.88	177.25	143.08	140.94	75.72	38.89
	SVDiff [39]	74.60	49.04	51.19	74.60	66.01	72.32	37.22	32.24
VOID (Ours)	SDEdit [45]	205.26	162.34	200.54	180.81	175.70	167.72	113.24	316.67
	Inpaint [50]	204.49	195.93	207.55	224.96	206.78	254.78	122.16	339.21
	ControlNet [72]	200.42	165.20	196.49	181.08	173.08	191.93	131.43	316.11
	Fine-tuning [50]	313.10	244.92	275.52	307.21	295.50	304.42	163.32	417.38
	LoRA [24]	291.30	225.40	249.45	280.63	267.76	262.47	172.33	377.56
	SVDiff [39]	209.01	196.23	197.31	209.01	203.09	207.87	123.51	349.42

Table 9: Defense effectiveness across heterogeneous mimicry Defenses. Comparison on TI-Dataset [17].

Defense	DB [52]			TI [17]		
	FID ↑	CLIP-S ↓	CLIP-I ↓	FID ↑	CLIP-S ↓	CLIP-I ↓
Smooth [8]	51.78	31.60	0.81	113.69	24.01	0.64
ACE [76]	87.64	28.98	0.74	149.28	23.39	0.66
VOID (Ours)	**222.32**	**28.05**	**0.69**	**247.47**	**19.85**	**0.60**

Defense	DI [16]			DF [11]		
	FID ↑	CLIP-S ↓	CLIP-I ↓	FID ↑	CLIP-S ↓	CLIP-I ↓
Smooth [8]	152.13	28.76	0.78	120.08	27.00	0.79
ACE [76]	124.22	26.12	0.72	159.28	26.19	0.77
VOID (Ours)	**244.35**	**20.32**	**0.58**	**627.33**	**20.96**	**0.57**

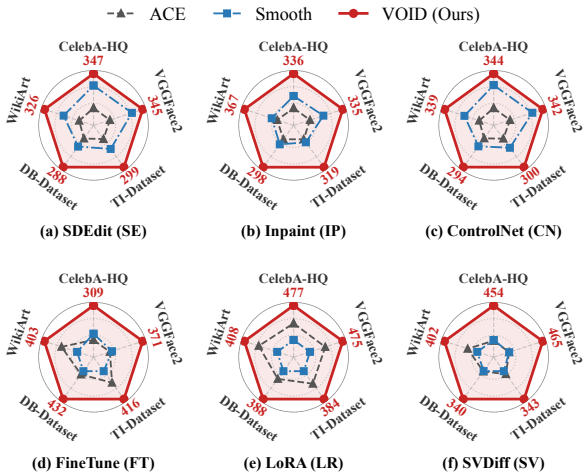


Figure 14: Evaluation of defense consistency across semantic domains. Radar charts illustrate FID scores (higher is better) across 5 diverse datasets under 6 mimicry attacks.

Cross-method transferability. Fig. 14 verifies the defense consistency of VOID across 6 mimicry attacks; we further extend the scope to 4 heterogeneous paradigms, spanning *training-based personalization* (DreamBooth (DB) [52], Textual Inversion (TI) [17]) and *inference-based editing* (DDIM Inversion (DI) [16], DiffEdit (DF) [11]). As detailed in Table 9, specialized baselines like ACE [76] and Smooth [8] exhibit distinct performance polarity. In contrast, VOID demonstrates universal superiority, notably achieving a dominant FID [20] of **627.33** in the DiffEdit [11] setting, surpassing the runner-up ACE [76] by nearly **4**×. This confirms that the semantic-corruption paradigm serves as a highly transferable defense capable of neutralizing diverse mimicry mechanisms.

Cross-domain transferability. We evaluate VOID across 5 distinct domains, including facial identities, general objects, and artistic styles. The results, visualized in Fig. 14, demonstrate that VOID maintains consistent protection stability across all categories. Quantitatively, VOID achieves high disruption levels with FID scores consistently exceeding 300 (peaking at **477** in LoRA [24]) and exhibits negligible variance. In contrast, baselines like ACE and Smooth display significant domain bias, evidenced by their inconsistent radar profiles. This confirms that VOID is content-agnostic, ensuring robust protection regardless of the target image semantics.

6.6 Robustness against Countermeasures

We evaluate VOID against three countermeasure categories on TI-Dataset [17]. As shown in Table 8, VOID consistently outperforms baselines across all settings.

Model-Agnostic Transformations. Against standard transformations (Gaussian noise, smoothing, JPEG [14]), VOID retains substantial efficacy even under aggressive smoothing. In contrast, baselines like ACE [76] collapse to non-disruptive levels under identical conditions. This disparity confirms that VOID relies on resilient structural corruption rather than fragile high-frequency perturbations.

Prior-Guided Purification. VOID exhibits exceptional stability under advanced purification methods such as Impress [6]. Specifically, for SDEdit [45], it retains a dominant FID of **180.81**, significantly outperforming Smooth [8] (62.43) and ACE [76] (14.78). This indicates that the semantic-corruption is deeply entangled with generative features, rendering it highly resistant to gradient-based inversion.

Table 10: Ablation study of VOID components. We evaluate Effectiveness, Utility, Transferability, and Robustness.

Defense Component	**Effectiveness** (FID ↑)			**Utility** (Visual Quality)			**Transferability** (FID ↑)		**Robustness** (FID ↑)	
	Training-based	Inference-based	Avg.	SSIM (↑)	PSNR (↑)	LPIPS (↓)	SD v2.1 [1]	LCM [42]	Impress [6]	TimestepSel.
EUA (Base)	**413.66**	**317.91**	**365.79**	0.789	32.766	0.182	365.67	370.24	263.44	354.78
+ GSC	392.30	308.06	350.18	0.789	32.775	0.180	**384.64**	**398.23**	**284.13**	**378.90**
+ HGPM	376.64	292.43	334.54	**0.894**	**35.327**	**0.081**	328.03	325.75	215.70	330.27
+ TPOS (VOID)	381.31	306.39	343.85	0.875	34.405	0.097	334.45	331.78	230.62	346.06

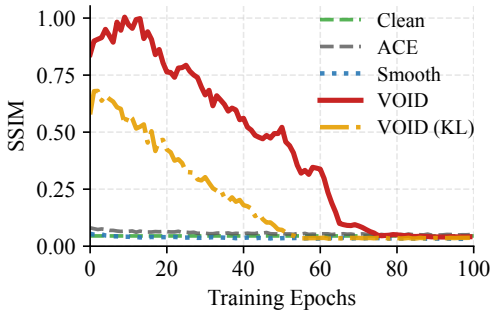
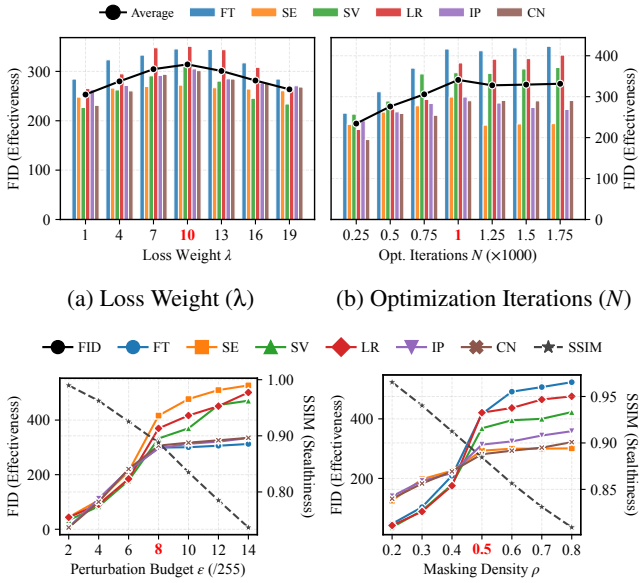


Figure 15: Optimization barrier under strong adaptive attacks.

Table 11: Strong adaptive attack results on TI-Dataset [17]. Perturbations are crafted on SD v1.5 and evaluated on different target models under standard and KL-regularized training.

Target Model	FID↑	CLIP-S↓	CLIP-I↓	Memory(GB)↓	Time(H)↓
SD v1.5 [53]*	247.23	25.11	0.66	78.53	2.13
SD v1.5 [53]* + KL	218.64	25.78	0.69	78.74	1.68
SD v2.1 [1]	198.71	26.83	0.70	79.12	2.44
SD v2.1 [1] + KL	176.32	27.21	0.72	79.35	1.92



(a) Loss Weight (λ) (b) Optimization Iterations (*N*)

(c) Perturbation Budget (α) (d) Masking Density (ρ)

Figure 16: Parameters sensitivity analysis.

Mechanism-Aware Adaptation. Motivated by our decay analysis (Section 6.2), we employ *Timestep Selection* to maximize diffusion intensity (inference strength 0.9, training start $t = 500$). Unlike baselines that falter under such high-noise regimes, VOID exhibits exceptional resilience, achieving FIDs of **316.67** for SDEdit [45] and **417.38** for Fine-tuning [50]. Moreover, its sustained effectiveness against DiffShortcut [40] (FID $>$ **113**) further validates the defense's robustness against targeted, mechanism-specific countermeasures.

Strong Adaptive Adversary. Existing adaptive evaluations [60] of defenses against LDM-based mimicry mainly fine-tune the U-Net, which largely overlaps with the FT attack [50] evaluated above. We implement a stronger adaptive attack that fine-tunes the full editing pipeline, including the VAE and U-Net, by editing the protected input with its own caption and forcing the edited output to reconstruct the input image in pixel space. We also test VOID (KL), which adds a KL regularizer to pull the VAE encoder's variance output toward the standard Gaussian, thereby reducing the EUA-induced variance shift. As shown in Fig. 15, VOID still creates a strong optimization barrier, with the attack loss stagnating for over 70 steps. We further report adaptive results in Table 11, where protections crafted on SD v1.5 [53] are evaluated on both SD v1.5 [53] and SD v2.1 [1] under standard and KL-regularized training. Although the adapted models improve the quantitative editing metrics, they require about 79GB memory and over 2 hours on a single A100 GPU, while the outputs remain severely blurred as shown in Fig. 20. This shows that adaptive training is costly and still fails to recover usable mimicry. Detailed settings are provided in Section D.

6.7 Ablation and Sensitivity Analysis

Component Ablation. We isolate the contribution of each VOID component in Table 10. Encoding Uncertainty Amplification (EUA) is the primary source of trajectory collapse, yielding the highest baseline effectiveness (Avg. FID **365.79**) by maximizing latent variance, though it compromises visual utility. Guidance Signal Counteraction (GSC) improves generalization, achieving the best transferability (FID **398.23**) and robustness (FID **284.13**), confirming that neutralizing semantic guidance prevents surrogate overfitting. HVS-Guided Perturbation Masking (HGPM) confines perturbations to texture-rich regions and boosts PSNR to **35.327**, with a trade-off in

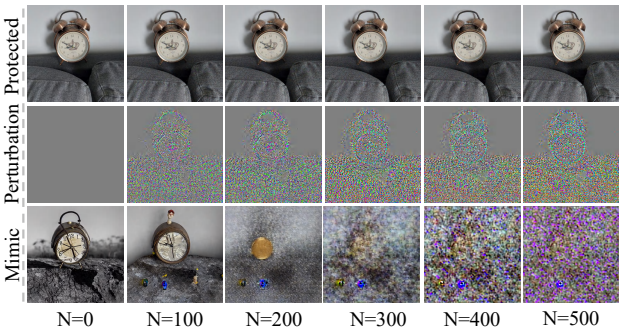


Figure 17: Visualization of optimization evolution.

Table 12: Sensitivity to perturbation budget and surrogate model setting on TI-Dataset [17].

Budget	Surrogate	Defense	SDEdit [45]-FID↑	FT [50]-FID↑
		Perturbation budget		
8/255	SDv1.5 [53]	Smooth [8]	121.34	81.48
		ACE [76]	24.78	219.69
		VOID (Ours)	**299.18**	**416.51**
16/255	SDv1.5 [53]	Smooth [8]	187.66	154.21
		ACE [76]	40.05	301.34
		VOID (Ours)	**337.53**	**481.40**
		Surrogate setting		
8/255	SDv1.5 [53]	VOID (Ours)	307.07	415.56
	SDv2.1 [1]		300.28	387.30
	SDv1.5 [53]+v2.1 [1]		**362.61**	**460.34**
16/255	SDv1.5 [53]+2.1 [1]	VOID (Ours)	**429.87**	**499.51**

raw attack strength. Timestep-Partitioned Optimal Selection (TPOS) mitigates stochastic sampling variance, recovering effectiveness (Avg. FID **343.85**) while preserving high fidelity. Fig. 17 further shows the evolution across optimization steps, where the protected image remains stable while mimicry outputs progressively degrade into global noise.

Parameter Sensitivity. We assess the sensitivity of VOID to four key hyperparameters on TI-Dataset [17], as shown in Fig. 16. Increasing the perturbation budget α improves protection but degrades visual quality, and $\alpha = 8/255$ balances strong disruption (FID > 300) with high fidelity (SSIM ≈ 0.88). Effectiveness plateaus after 1000 iterations, so we set $N = 1000$. The loss weight λ balances EUA and GSC: moderate weighting improves transferability on SD v2.1 [1], while excessive weighting suppresses EUA-driven corruption; we use $\lambda = 10$. The masking density ρ controls the trade-off between search space and stealthiness, and $\rho = 0.5$ restricts perturbations to the top 50% texture-rich regions. Table 12 further quantifies perturbation-budget and surrogate-model sensitivity, with results averaged over LCM [42] and DreamShaper [43] as target models. Increasing the budget to $16/255$ improves protection, and using both SD v1.5 [53] and SD v2.1 [1] as surrogates boosts transferability, showing the potential of ensemble-based optimization.

Table 13: Robustness evaluation (Accuracy %) of LDM-based purifiers on ImageNet using ResNet-50. The perturbation budget is $\varepsilon = 4/255$, and PGD attack runs for 100 iterations.

Defense	Clean	Evaluation Methods		
		PGD [44]	AutoAttack [12]	VOID (Ours)
W/O Purifier	80.55	0.01	0.00	—
OSCP [34]	77.63	73.89	74.19	**0.11**
DBLP [25]	78.00	75.60	74.80	**0.10**

7 Discussion

7.1 VOID as a Red-Teaming Tool

Beyond privacy protection against mimicry, VOID can also evaluate LDM-based adversarial purification defenses. In this setting, generative purifiers such as OSCP [34] and DBLP [25] use LDM priors to remove adversarial perturbations and restore clean data manifolds before classification. However, their reliance on stable generation trajectories makes them vulnerable to semantic corruption. To test this, we apply the core EUA and GSC objectives without perceptual masking under a standard ℓ_∞ budget ($\varepsilon = 4/255$). As shown in Table 13, purifier robust accuracy drops to nearly 0%, even though these models remain robust against PGD [44] and AutoAttack [12]. This gap shows that standard adversarial attacks treat LDMs as generic differentiable functions, while VOID directly targets their intrinsic generative dynamics. Therefore, VOID provides a complementary benchmark for exposing hidden fragility in LDM-based security modules.

7.2 Limitations

Generalizability across Modalities. Our current evaluation focuses exclusively on image-domain LDMs. While the generative capabilities of LDMs have expanded to other modalities such as audio, video, and 3D meshes, the applicability of VOID's semantic corruption paradigm in these domains remains unverified. Different modalities may exhibit distinct feature persistence dynamics and latent structures, potentially requiring domain-specific adaptations of our EUA and GSC modules. We leave the exploration of extending VOID to multimodal protection as future work.

Computational Overhead. The GSC module requires gradients from the guidance vector and thus performs both conditional and unconditional forward passes during optimization. This introduces a modest overhead compared to single-pass baselines such as AdvDM [38]. Given the significant performance advantage, we argue that this cost is acceptable for offline privacy protection and remains considerably lower than complex bilevel-optimization-based poisoning attacks. It is also a one-time cost incurred only during the protection phase, with no impact on subsequent image storage or usage.

8 Conclusion

The great potential of LDMs also brings unprecedented privacy threats. Existing semantic-steering mimicry defense mechanisms overlook the model's internal machinery that can render them futile, as we have elaborated in this paper. We explore an alternative principled semantic-corruption approach, which leverages the intrinsic stochastic nature of LDMs, and demonstrate its effectiveness through the design, implementation, and extensive evaluation of the VOID framework.

While this paper focuses primarily on image synthesis, we believe that the proposed semantic-corruption paradigm has unveiled a fundamental trait of diffusion priors, in particular their reliance on continuous latent evolution. This suggests transferability to other modalities such as voice cloning and 3D synthesis. An interesting research direction is to turn VOID into a unified framework that can support non-image domains and thwart unforeseen mimicry attacks.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This work was supported by National Key Research and Development Program of China (2023YFB3107400), the NSFC under Grants U2441240 ("Ye Qisun" Science Foundation), 62441238, 62302344, 62441237, 62502350, 62521002, 62132011, U24B20185, the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM114). Thanks to the New Cornerstone Science Foundation and the Xplorer Prize.

Ethical Considerations

Since VOID studies adversarial perturbations against generative models, it may create misuse risks. We discuss its impacts on four stakeholder groups.

Protected data and content owners. This group includes individuals, creators, artists, and IP holders who may face unauthorized identity or style mimicry. VOID provides a proactive opt-out defense, but its protection is not absolute and may be weakened by future adaptive attacks. It should therefore complement consent-based data protection, platform safeguards, and policy interventions.

Adversaries and malicious actors. VOID is designed to raise the cost of unauthorized mimicry. However, perturbation-based techniques may be repurposed for data poisoning, model disruption, or adaptive circumvention. We reduce these risks by framing VOID as a defensive tool and releasing artifacts with restrictive licensing and careful documentation.

Benign AI ecosystem actors. Developers, dataset maintainers, and model providers may use VOID for robustness evaluation and non-consensual data-use analysis. However, widespread perturbation-based protection may increase filtering costs, introduce false positives, or interfere with legitimate data use. Such defenses should complement transparent data policies and platform-level consent mechanisms.

Research and publication stakeholders. Our experiments use public datasets and collect no new private information. Still, represented people or works may raise consent or redistribution concerns, especially in face-related settings. We obscure sensitive facial attributes and gestures in mimicry results and conduct no human-subject study. Although publication may inform circumvention attempts, we mitigate this risk through restrictive licensing.

Overall, publication is ethically warranted because unauthorized mimicry causes concrete privacy and intellectual-property harms. VOID should be viewed as one defensive layer, not a replacement for broader consent, provenance, and governance mechanisms.

Open Science

We make the artifacts supporting our core contribution publicly available. The release includes: (1) the source code of VOID; (2) the mimicry attack implementations used in our evaluation; and (3) evaluation scripts for measuring defense effectiveness, including FID, CLIP-S, and CLIP-I. The artifact package and setup instructions are available at:

https://doi.org/10.5281/zenodo.20233998

References

[1] Stability AI. Stable diffusion version 2.1. https://huggingface.co/stabilityai/stable-diffusion-2-1, 2022.

[2] Stability AI. Stable diffusion xl base 0.9. https://huggingface.co/stabilityai/stable-diffusion-xl-base-0.9, 2023.

[3] Stability AI. Stable diffusion 3 medium. https://huggingface.co/stabilityai/stable-diffusion-3-medium, 2024.

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[5] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017.

[6] Bochuan Cao, Changjiang Li, Ting Wang, Jinyuan Jia, Bo Li, and Jinghui Chen. Impress: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative ai. *Advances in Neural Information Processing Systems*, 36:10657–10677, 2023.

[7] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[8] Junxi Chen, Junhao Dong, and Xiaohua Xie. Exploring adversarial attacks against latent diffusion model from the perspective of adversarial transferability. *arXiv preprint arXiv:2401.07087*, 2024.

[9] June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. Diffusionguard: A robust defense against malicious diffusion-based image editing. *arXiv preprint arXiv:2410.05694*, 2024.

[10] CompVis. Stable diffusion repository. https://github.com/CompVis/stable-diffusion, 2022.

[11] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

[12] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 2206–2216. PMLR, 2020.

[13] Katherine Crowson and Stability AI. Stable diffusion x2 latent upscaler. https://huggingface.co/stabilityai/sd-x2-latent-upscaler, 2023. Model accessed on Hugging Face.

[14] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017.

[15] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

[16] Hugging Face. The hugging face diffusion course - unit 4: Exploring the training script. https://huggingface.co/learn/diffusion-course/en/unit4/2, 2023. Accessed: 2025-10-21.

[17] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2022.

[18] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[19] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[22] Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. Adversarial perturbations cannot reliably protect artists from generative ai. *arXiv preprint arXiv:2406.12027*, 2024.

[23] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

[25] Chihan Huang, Belal Alsinglawi, and Islam Al-Qudah. DBLP: noise bridge consistency distillation for efficient and reliable adversarial purification. *CoRR*, abs/2508.00552, 2025.

[26] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

[27] Qiuping Jiang, Zhentao Liu, Shiqi Wang, Feng Shao, and Weisi Lin. Toward top-down just noticeable difference

estimation of natural images. *IEEE Transactions on Image Processing*, 31:3697–3712, 2022.

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

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

[30] Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Eq-vae: Equivariance regularized latent space for improved generative image modeling. *arXiv preprint arXiv:2502.09509*, 2025.

[31] Black Forest Labs. Flux.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024.

[32] Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

[33] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025.

[34] Chun Tong Lei, Hon Ming Yam, Zhongliang Guo, Yifei Qian, and Chun Pong Lau. Instant adversarial purification with adversarial consistency distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24331–24340, 2025.

[35] Ang Li, Yichuan Mo, Mingjie Li, and Yisen Wang. Pid: Prompt-independent data protection against latent diffusion models. *arXiv preprint arXiv:2406.15305*, 2024.

[36] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[37] Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*, 2023.

[38] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. *arXiv preprint arXiv:2302.04578*, 2023.

[39] Yisu Liu, Jinyang An, Wanqian Zhang, Dayan Wu, Jingzi Gu, Zheng Lin, and Weiping Wang. Disrupting diffusion: Token-level attention erasure attack against diffusion-based customization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3587–3596, 2024.

[40] Yixin Liu, Ruoxi Chen, Xun Chen, and Lichao Sun. Rethinking and defending protective perturbation in personalized diffusion models. *arXiv preprint arXiv:2406.18944*, 2024.

[41] Yixin Liu, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, and Lichao Sun. Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24219–24228, 2024.

[42] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

[43] Lykon. Dreamshaper. https://huggingface.co/Lykon/DreamShaper, 2023. Accessed: 2024-05-20.

[44] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[45] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[47] Huy Phan, Boshi Huang, Ayush Jaiswal, Ekraam Sabir, Prateek Singhal, and Bo Yuan. Latent diffusion shield-mitigating malicious use of diffusion models through latent space adversarial perturbations. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1440–1448, 2025.

[48] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018.

[49] 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*, pages 8748–8763. PmLR, 2021.

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[52] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine-tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[53] RunwayML. Stable diffusion v1.5. https://huggingface.co/runwayml/stable-diffusion-v1-5, 2024.

[54] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.

[55] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023.

[56] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023.

[57] Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 807–825. IEEE, 2024.

[58] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006.

[59] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[60] Zekun Sun, Zijian Liu, Shouling Ji, Chenhao Lin, and Na Ruan. Pretender: Universal active defense against diffusion finetuning attacks. In *The 34th USENIX Security Symposium*, 2025.

[61] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2116–2127, 2023.

[62] Cong Wan, Yuhang He, Xiang Song, and Yihong Gong. Prompt-agnostic adversarial perturbation for customized diffusion models. *Advances in Neural Information Processing Systems*, 37:136576–136619, 2024.

[63] Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, and Qidong Huang. Simac: A simple anti-customization method for protecting face privacy against text-to-image synthesis of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12047–12056, 2024.

[64] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[65] Jingyao Xu, Yuetong Lu, Yandong Li, Siyang Lu, Dongdong Wang, and Xiang Wei. Perturbing attention gives you more bang for the buck: Subtle imaging perturbations that efficiently fool customized diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24534–24543, 2024.

[66] Long Xu, Jiakai Wang, Haojie Hao, Haotong Qin, Jiejie Zhao, and Xianglong Liu. Harnessing global-local collaborative adversarial perturbation for anti-customization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13414–13423, 2025.

[67] Xide Xu, Sandesh Kamath, Muhammad Atif Butt, and Bogdan Raducanu. An h-space based adversarial attack for protection against few-shot personalization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4904–4913, 2025.

[68] Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion based mimicry through score distillation, 2024.

[69] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE transactions on image processing*, 23(2):684–695, 2013.

[70] Jianping Zhang, Zhuoer Xu, Shiwen Cui, Changhua Meng, Weibin Wu, and Michael R Lyu. On the robustness of latent diffusion models. *arXiv preprint arXiv:2306.08257*, 2023.

[71] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

[72] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[73] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[74] Zhengyue Zhao, Jinhao Duan, Xing Hu, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Yunji Chen. Unlearnable examples for diffusion models: Protect data from unauthorized exploitation. *arXiv preprint arXiv:2306.01902*, 2023.

[75] Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by stable diffusion? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24398–24407, 2024.

[76] Boyang Zheng, Chumeng Liang, and Xiaoyu Wu. Targeted attack improves protection against unauthorized diffusion customization. *arXiv preprint arXiv:2310.04687*, 2023.

[77] Li Zheng, Liangbin Xie, Jiantao Zhou, Xintao Wang, Haiwei Wu, and Jinyu Tian. Anti-diffusion: Preventing abuse of modifications of diffusion-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10582–10590, 2025.

[78] Haotian Zhu, Shuchao Pang, Zhigang Lu, Yongbin Zhou, and Minhui Xue. Gap-diff: protecting jpeg-compressed images from diffusion-based facial customization. In *NDSS*, 2025.

A Latent Diffusion Models

Latent Diffusion Models (LDMs) [31, 50] have become the de facto standard for high-fidelity image synthesis. Structurally, LDMs decouple the generative process into two distinct stages: perceptual compression and semantic generation.

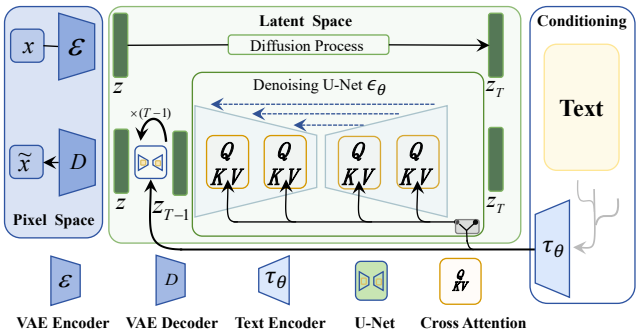


Figure 18: Overall architecture of Latent Diffusion Models.

As shown in Fig. 18, a VAE encoder $\mathcal{E}$ [29] first maps the input x to a low-dimensional latent $z_0 = \mathcal{E}(x)$, and a decoder $\mathcal{D}$ reconstructs the image from the latent.

The core generation occurs in the latent space via a probabilistic diffusion process. Formally, the *forward process* is a fixed Markov chain that gradually destroys the data structure by injecting Gaussian noise $n \sim \mathcal{N}(0, \mathbf{I})$ over T steps. At any timestep t, the noisy latent z_t can be expressed as:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t} n, \tag{10}$$

where $\{\bar{\alpha}_t\}$ follows a fixed variance schedule. The *reverse process* employs a time-conditional U-Net ε_θ [51] to reverse this corruption. Trained by minimizing the mean squared error between the predicted and actual noise, the objective function is:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,x_0,n}\left[\|\varepsilon - \varepsilon_\theta(z_t, t, c)\|^2\right], \tag{11}$$

where c denotes the conditioning signal (e.g., text prompts). Minimizing this objective allows the model to learn the data distribution conditioned on c. During inference, to strictly enforce semantic alignment with c, LDMs utilize Classifier-Free Guidance (CFG) [21]. The final noise prediction $\hat{\varepsilon}$ is extrapolated by scaling a *guidance vector*:

$$\hat{\varepsilon}_\theta(z_t, t, c) = \varepsilon_\theta(z_t, t, \emptyset) + w \cdot \underbrace{(\varepsilon_\theta(z_t, t, c) - \varepsilon_\theta(z_t, t, \emptyset))}_{\text{guidance vector}}, \tag{12}$$

where w is the guidance scale. This vector dictates the direction towards the conditional semantics c, effectively serving as the primary control mechanism for generation.

The computational efficiency and open availability of LDM weights have fundamentally lowered the barrier for high-fidelity synthesis [24, 50]. Unlike API-based models restricted by centralized safeguards, these models are typically deployed on personal devices in uncontrolled environments. This grants adversaries full control over the inference pipeline, allowing them to bypass safety checkers entirely and weaponize community-developed tools (e.g., LoRA [24], ControlNet [72]) for unauthorized mimicry of identities and artistic styles [37, 55]. Consequently, passive restrictions are insufficient, creating a compelling need for proactive defenses embedded directly within the data.

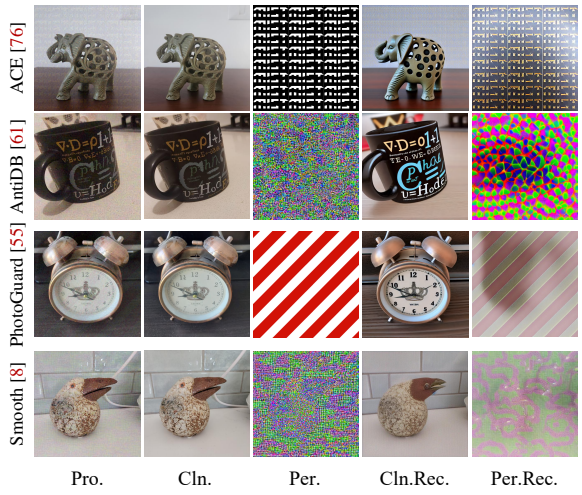


Figure 19: Visual reconstruction of original images and protective perturbations from protected samples.

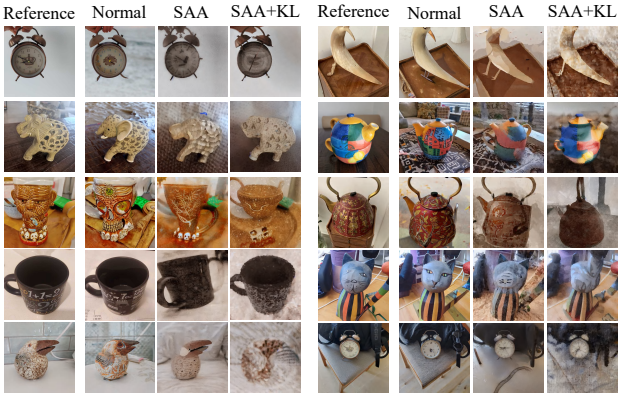


Figure 20: Functional collapse under adaptive adversary.

B Visual Reconstruction of Mixed Features

This section provides qualitative evidence for Observation 2 in Section 4.1. We apply DDIM inversion [59] to protected latents and denoise them with neutral prompts. As shown in Fig. 19, the model separates a protected sample into a reconstructed original identity and an isolated protective perturbation. This shows that steering-based defenses preserve semantic links, allowing the model's restoration mechanism to recover protected content.

C Implementation Details

For *inference-based editing*, we set the editing strength to 0.7. For *training-based personalization*, we use $t = 300$ as the starting point of the diffusion training range and sample timesteps from 300 to 1000. All defenses use an L_∞ budget of $\alpha = 8/255$ and 1000 optimization steps. Following PhotoGuard [55], protection generation uses `float32` to reduce quantization errors, while attacks use mixed precision. For VOID, we set $\lambda = 10$ and $\rho = 50$ to select the top 50% JND regions. Experiments mainly run on NVIDIA RTX 4090 GPUs, except PhotoGuard generation and the holistic adaptive attack, which run on NVIDIA A100 (80GB) GPUs.

Table 14: Semantic specificity analysis on SD v1.5 [53].

Prompt Group	FID↑		CLIP-S↓	
	w/o	VOID	w/o	VOID
Protected-concept	23.75	415.35	28.89	19.12
Near-concept	20.37	37.42	28.03	24.39
Unrelated-concept	19.52	19.93	27.88	27.08

D Details of Strong Adaptive Adversary

We design a holistic adaptive adversary that exploits $x_{prot} \approx x_{clean}$ by using the protected image as the reconstruction target. The attack fine-tunes the full editing pipeline, including the VAE Encoder/Decoder and U-Net. It injects noise into x_{prot} at timestep $t = 600$, performs 15-step unrolled denoising with the image caption, and backpropagates the pixel-space reconstruction loss through the loop. This attack requires about 79GB VRAM on a single A100 GPU. Despite this cost, VOID creates a strong optimization barrier. As shown in Fig. 15, optimization stagnates for over 70 steps due to unstable gradients from EUA-amplified latent variance. We also evaluate VOID (KL), where the adversary adds a KL regularizer to pull the VAE posterior toward the standard Gaussian and reduce the EUA-induced variance shift. Although KL regularization accelerates convergence, it does not restore usable mimicry. As shown in Fig. 20, the adapted model still produces severely blurred outputs, indicating functional degradation. Thus, even a holistic adaptive adversary improves reconstruction metrics only at high cost and fails to restore visually usable editing.

E Semantic Specificity Analysis

VOID protects images by corrupting the semantics of the protected concept, but this should not imply global generation failure. To verify semantic specificity and separate concept-specific corruption from global instability, we finetune SD v1.5 [53] on VOID-protected images and evaluate protected-concept, near-concept, and unrelated-concept prompts. If VOID destabilizes the whole diffusion pipeline, all prompt groups should degrade similarly. As shown in Table 14, degradation is strongest for protected concepts, where FID [20] increases from 23.75 to 415.35 and CLIP-S [19] drops from 28.89 to 19.12. The effect weakens on near concepts and nearly disappears on unrelated concepts, confirming that VOID mainly corrupts protected semantics rather than the overall diffusion generation process.