VOID: Defeating Unauthorized Mimicry in Latent Diffusion Models

Anonymous authors

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 22 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, 40, 47] 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 [43, 68] and few-shot model personalization [24, 49]. 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 [39, 58] and intellectual property rights [36, 53], raising critical concerns about the abuse of generative AI.

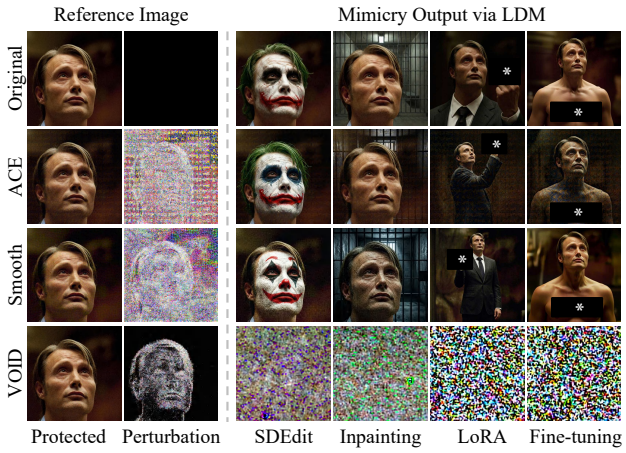

Figure 1: Visual comparison of defenses against unauthorized mimicry. While SOTA methods fail to prevent high-fidelity identity reconstruction, VOID fundamentally collapses the generation trajectory, rendering synthesized outputs semantically void. *Sensitive content is redacted for presentation.*

To counter such threats, a list of defense mechanisms [35, 36, 53, 58] 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 [35, 53] or the subsequent denoising process [36, 58].

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 [72] 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 22 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, VOID maintains superior visual utility with a PSNR [26] of 34.41 dB and an SSIM [61] of 0.89; it stands as the only defense in our benchmarks to achieve a perceptual LPIPS [69] below 0.1. Moreover, our stress tests demonstrate its transferability across diverse LDM variants like SD v2.1 [1] and LCM [40] 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.

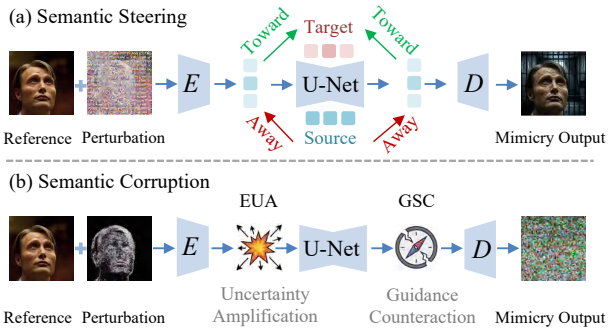

Figure 2: Comparison between the existing semantic steering paradigm and our proposed semantic corruption paradigm.

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.

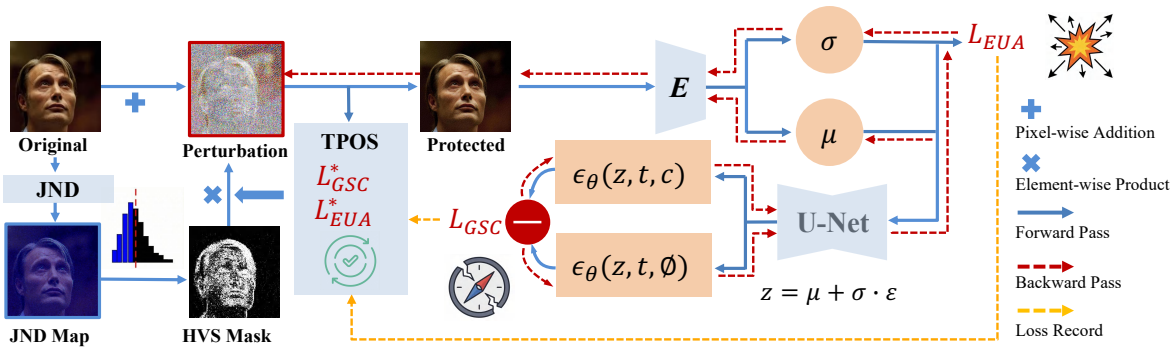

Figure 7: Overview of the proposed VOID framework. The pipeline jointly optimizes the protective perturbation by maximizing encoding uncertainty (L_{EUA}) during VAE encoding and counteracting guidance signals (L_{GSC}) during U-Net denoising.

Table 3: Quantitative evaluation of defense efficacy. Results are averaged across 5 datasets (TI-Dataset [17], DB-Dataset [49], CelebA-HQ [28], VGGFace2 [7] and WikiArt [51]) and 6 mimicry methods: FT [47], SD [4], LR [24], SE [43], IP [47] and CN [68]). Arrows indicate the direction of better protection performance. **Bold** and underlined values denote the best and second-best results, respectively. Detailed breakdowns per mimicry method and dataset are provided in Section G.

Type	Defense	Training-based mimicry			Inference-based mimicry			Average		
		FID [20] ↑	CLIP-S [19] ↓	CLIP-I [46] ↓	FID [20] ↑	CLIP-S [19] ↓	CLIP-I [46] ↓	FID [20] ↑	CLIP-S [19] ↓	CLIP-I [46] ↓
Poisoning	ACE [72]	146.22	24.52	0.69	47.87	27.04	0.80	97.04	25.78	0.74
	AntiDB [58]	69.32	28.49	0.77	65.86	26.24	0.84	67.59	27.36	0.81
	AntiDiff [73]	55.37	28.97	0.79	48.55	26.46	0.86	51.96	27.71	0.83
	CAAT [62]	67.90	28.73	0.77	45.40	26.74	0.86	56.65	27.74	0.81
	DisDiff [37]	71.97	28.75	0.77	28.11	26.79	0.88	50.04	27.77	0.82
	EUDP [70]	76.59	28.62	0.76	29.69	27.12	0.87	53.14	27.87	0.82
	GoodAC [63]	70.63	28.38	0.77	74.73	26.11	0.84	72.68	27.25	0.80
	MetaCloak [39]	44.03	29.09	0.81	38.48	26.65	0.87	41.26	27.87	0.84
	PAP [59]	72.26	28.36	0.76	87.86	26.33	0.83	80.06	27.35	0.80
	Pretender [57]	56.95	28.98	0.79	41.79	26.79	0.86	49.37	27.88	0.82
	SimAC [60]	44.75	29.19	0.82	22.67	26.60	0.89	33.71	27.90	0.85
-	RandomNoise	22.15	29.16	0.88	5.64	26.55	0.93	13.90	27.85	0.90
Evasion	AdvDM [36]	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 [74]	103.96	26.47	0.71	33.48	26.85	0.86	68.72	26.66	0.78
	Glaze [53]	138.80	24.57	0.69	48.96	27.31	0.81	93.88	25.94	0.75
	LDMR [66]	46.70	29.03	0.79	43.66	26.45	0.82	45.18	27.74	0.81
	MIST [35]	143.17	24.53	0.69	48.32	27.39	0.81	95.74	25.96	0.75
	NightShade [54]	140.97	24.58	0.69	49.98	27.09	0.80	95.48	25.83	0.75
	PID [33]	34.14	29.26	0.82	27.68	27.28	0.84	30.91	28.27	0.83
	PhotoGuard [52]	32.18	29.20	0.83	23.19	27.01	0.85	27.69	28.11	0.84
	SDST [64]	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
	VOID (Ours)	**404.94**	**20.08**	**0.53**	**325.80**	**18.82**	**0.49**	**365.37**	**19.45**	**0.51**

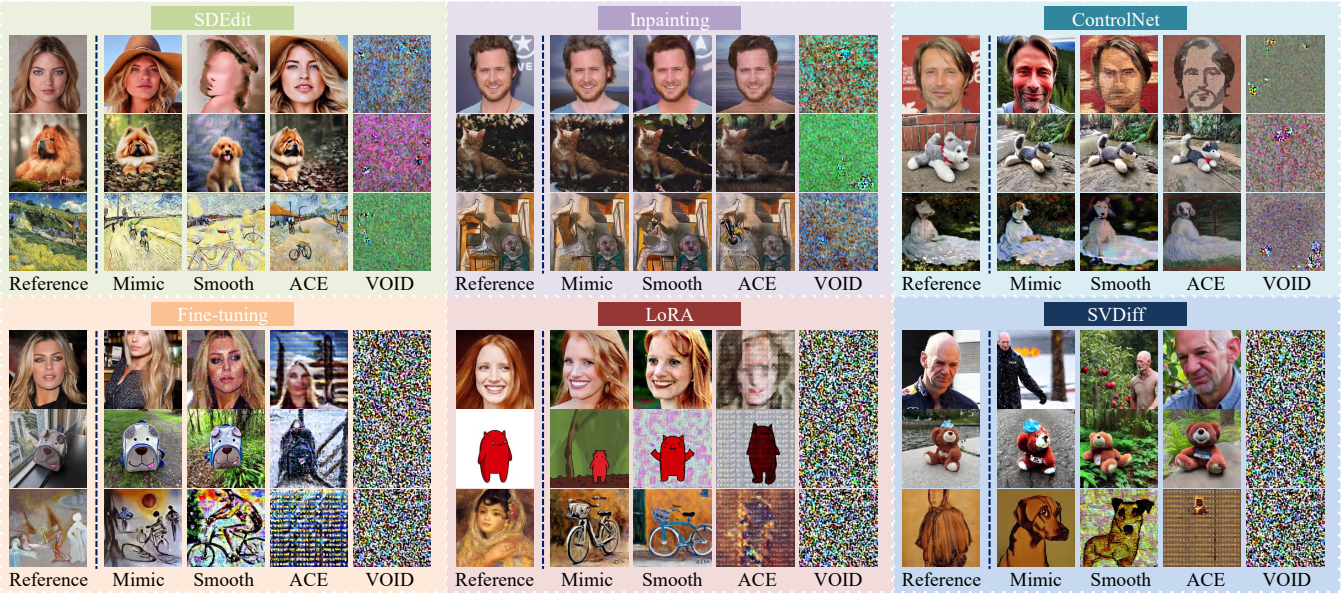

Figure 11: Qualitative comparison of defense efficacy. Rows 1-3 correspond to *inference-based image mimicry* (SE [43], IP [47], CN [68]), and Rows 4-6 correspond to *training-based model personalization* (FT [47], LoRA [24], SVDiff [4]). Columns compare the unprotected baseline (Mimic), the top-performing specialized defenses (Smooth [8] and ACE [72]), and our VOID.

Table 4: Imperceptibility evaluation. VOID is compared against the top-4 visual quality baselines among 22 methods.

Method	**Signal-level Metrics**					**Neural-based Metrics**				
	PSNR [26]↑	SSIM [61]↑	FSIM [67]↑	VIF [55]↑	GMSD [65]↓	LPIPS [69]↓	DISTS [15]↓	DeepIQA [5]↑	PieAPP [45]↓	ContentLoss [18]↓
AntiDiff [73]	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 [52]	33.367	0.843	0.979	0.584	0.028	0.122	0.219	-0.014	0.728	1589.5
SDST [64]	32.903	0.823	0.977	0.534	0.028	0.116	0.291	-0.017	0.716	2182.5
VOID	**34.408**	**0.888**	**0.985**	**0.624**	**0.019**	**0.096**	**0.177**	**-0.010**	**0.679**	**1204.9**

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.