DSSA: Dynamic Sparse Semantic Anchoring for Purifying Protective Perturbations against Diffusion Models

Weiwei Tan[1,2], **Rui Wang**[1,2*], **Lihua Jing**[1,2], **Yanjun Zhang**[3], **Runbo Li**[1,2], **Zhishen Wang**[1,2], **Leo Yu Zhang**[3]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]Griffith University

{tanweiwei,wangrui,jinglihua,lirunbo,wangzhishen}@iie.ac.cn
{yanjun.zhang,leo.zhang}@griffith.edu.au

Abstract

The rapid proliferation of personalization techniques in Stable Diffusion has driven the widespread deployment of adversarial perturbations as a protective measure against unauthorized data mimicry. Existing diffusion-based purification methods primarily rely on the noising and denoising paradigm and have proven somewhat effective in neutralizing these protections. However, they suffer from a severe trade-off between purification efficacy and perceptual fidelity, especially under large perturbation budgets. To mitigate this problem, in this paper, we propose Dynamic Sparse Semantic Anchoring (DSSA), a novel purification framework. By starting the reverse denoising process from pure Gaussian noise, DSSA removes the initial adversarial perturbations and circumvents the dilemma of selecting an optimal noising level. To prevent the loss of original semantics caused by pure noise generation, we propose a dynamic sparse anchoring strategy that directly integrates sparse pixels from the adversarial sample to guide the reconstruction. Additionally, we design a perception-guided early-stopping mechanism to prevent the perturbation resurgence caused by this direct pixel integration. Extensive experiments across diverse datasets and protection schemes demonstrate that our method consistently outperforms existing baselines, achieving state-of-the-art performance in efficacy and fidelity, successfully neutralizing even advanced adaptive protections.

1 Introduction

In recent years, diffusion models have significantly advanced the field of image generation [1, 2, 3, 4, 5]. Notably, Stable Diffusion, based on the Latent Diffusion Model (LDM) framework [6], has demonstrated exceptional generative capabilities and achieved widespread adoption in the open-source community. Personalized generation techniques [7, 8, 9] enable diffusion models to learn a specific subject or style from only a few reference images, allowing the synthesis of highly realistic portraits or artworks. However, this powerful mimicry capability raises severe copyright and privacy concerns. Malicious users can exploit minimal data, such as a few portraits or paintings, to fabricate deepfakes or art forgeries, leading to substantial copyright infringement and privacy risks [10, 11, 12, 13].

To mitigate these risks, researchers have introduced adversarial perturbations [14, 15] to protect data copyright [16, 17, 18, 19, 20, 21, 22, 23, 24]. These methods inject protective perturbations into images that require protection to disrupt feature extraction or noise prediction during model fine-tuning. Models fine-tuned on such protected data ultimately generate degraded or semantically

*Corresponding author.

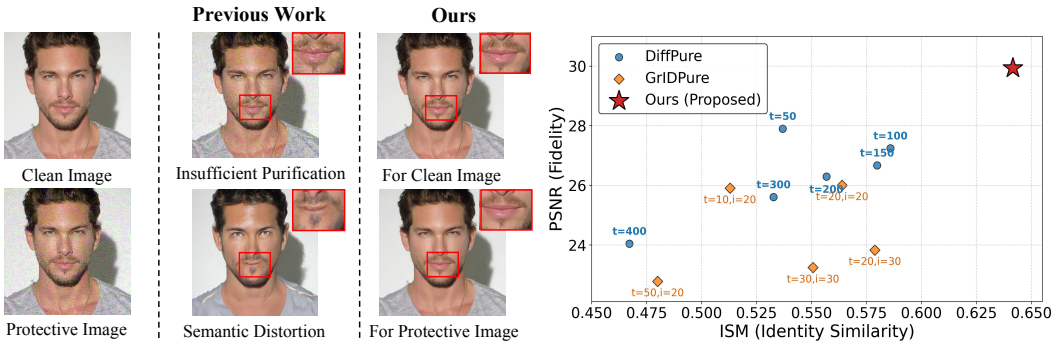


Figure 1: The trade-off in existing diffusion-based purification methods. **Left:** Visual comparison demonstrating that previous works either fail to remove protective perturbations (at small t) or destroy original semantics (at large t). In contrast, our proposed DSSA faithfully recovers the clean image. **Right:** Quantitative evaluation on the VGGFace2 dataset. Baseline methods exhibit a severe trade-off between purification efficacy and fidelity across different noise scales, whereas our approach successfully breaks this bottleneck, achieving state-of-the-art performance in both dimensions.

corrupted images, effectively preventing unauthorized personalization. Although early works primarily focus on designing protective perturbations to safeguard data copyright, recent studies have begun to investigate the robustness of these protections under image purification. In practical scenarios, malicious users may attempt to remove protective perturbations through purification techniques before personalization [25, 26, 27, 28]. Therefore, evaluating the resilience of protection methods against purification has become an important step toward understanding their real-world security.

Recently, noising and denoising mechanisms via diffusion models have become the mainstream paradigm for purifying protective perturbations. For instance, methods such as GrIDPure [26] build upon the DiffPure [25] framework by introducing patch partitioning and multi-step denoising with small noise scales to better preserve structural details. These methods explicitly project adversarial examples onto the clean data manifold through a forward noising and reverse denoising process.

Despite these advancements, purification methods relying on the noising and denoising paradigm suffer from a severe trade-off between perceptual fidelity and purification efficacy, particularly when confronted with high-intensity perturbations. As illustrated by the quantitative analysis in Figure 1, methods such as DiffPure and GrIDPure exhibit a distinctive inflection point of efficacy: while minimal noise maintains fidelity, it fails to neutralize the protection; conversely, as the noise scale increases, both fidelity and efficacy eventually collapse due to severe semantic distortions and reconstruction hallucinations. Consequently, these methods struggle to faithfully approximate the true clean image manifold, making it difficult to reliably evaluate the robustness of existing copyright protection techniques and potentially leading to a false sense of security. To understand the root cause of this trade-off, we investigate the characteristics of protective perturbations in the frequency domain. Our analysis reveals that these perturbations are not merely confined to high-frequency pixel fluctuations but are strategically distributed across mid-frequency and low-frequency semantic bands. This exposure highlights a fundamental flaw in the SDEdit paradigm: a small noise budget affects only high-frequency components, leaving low-frequency adversarial structures intact. Conversely, adding sufficient noise to eliminate low-frequency perturbations inevitably destroys the core content of the original image. Without reliable low-frequency guidance during the subsequent denoising process, the model is forced to guess the missing details, leading to the observed structural hallucinations.

To circumvent the dilemma of selecting an optimal noising level while effectively neutralizing adversarial perturbations, we depart from the conventional noising and denoising paradigm. Instead, we start the denoising process directly from pure Gaussian noise. To mitigate the structural hallucinations and semantic distortions inherent in such unconditional denoising, we propose a Dynamic Sparse Semantic Anchoring (DSSA) strategy. DSSA provides the necessary semantic guidance by selectively anchoring structural signals derived from the original protected image. Motivated by the observation that the reverse denoising process of diffusion models has different preferences at different stages, our designed guidance mask favors the overall low-frequency semantic framework of the image in the early stages. In the later stages, the denoising process tends to restore the high-frequency details of the image, and the dynamic mask subsequently adapts to retain these parts. However, directly an-

choring all high-frequency signals poses a significant risk. During this restoration, the strong gradient signals representing natural semantic edges become heavily entangled with the protective adversarial perturbations. Leveraging the disparity in their gradient distributions, we introduce a non-linear Cumulative Distribution Function (CDF) mapping to anchor high-amplitude natural edges while filtering out low-amplitude adversarial patterns. During the final denoising stages, as the injected noise scale decays, adversarial perturbations lead to a perturbation resurgence phenomenon, causing the model to inadvertently reconstruct the original protective patterns. To address this, we reveal a cross-domain alignment between reference-free perceptual quality and latent feature purification, based on which we design a dynamic perception-guided early-stopping mechanism to precisely stop the trajectory before the adversarial perturbations begin to recover.

In conclusion, our main contributions are threefold. (1) We empirically and theoretically reveal that existing diffusion-based purification methods, which predominantly rely on the noising and denoising paradigm, face an intrinsic trade-off between purification efficacy and semantic fidelity. (2) We propose a novel purification framework called DSSA. By initiating the denoising trajectory directly from pure Gaussian noise and employing a dynamic sparse semantic anchoring strategy, our method eradicates adversarial patterns while maximizing semantic fidelity. Additionally, we introduce a perception-guided early-stopping mechanism that precisely halts the trajectory to prevent the recovery of these perturbations during the late stages of denoising. (3) Extensive experiments demonstrate that our approach achieves state-of-the-art performance in both purification efficacy and perceptual quality, successfully neutralizing various advanced adaptive protections.

2 Background and Related Work

Text-to-Image (T2I) Diffusion Models. Diffusion models are generative models that learn to synthesize data by reversing a progressive noise-adding process. Given a clean image x_0, the forward process gradually adds Gaussian noise ϵ to produce a noisy state x_t at timestep t. The model is trained to denoise x_t using a neural network ϵ_θ, optimized by a simple mean squared error loss:

$$\mathcal{L}_{\text{DM}}(\theta, x_t) = \mathbb{E}_{\epsilon,t,(x,p)}[\|\hat{\epsilon}_\theta(x_t, p, t) - \epsilon\|_2^2] \tag{1}$$

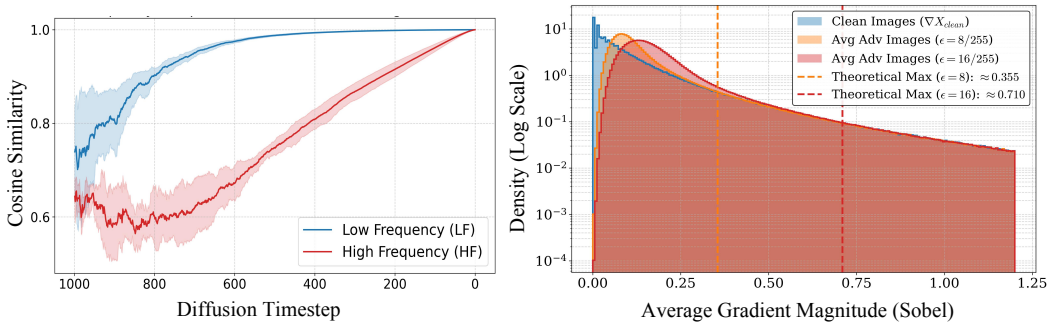
Early models perform this diffusion process directly in the pixel space. Pixel space diffusion requires heavy computational resources. To improve efficiency, Latent Diffusion Models (LDMs), most notably Stable Diffusion (SD), compress images into a low-dimensional space using a Variational Autoencoder (VAE) [29, 30] and perform the diffusion process within this latent space. Due to their efficiency and quality of generation, LDMs have become the standard framework for T2I generation.

Protective Perturbations against Unauthorized Personalization. Latent Diffusion Models (LDMs) have demonstrated exceptional performance in image generation. Personalization techniques allow models to capture specific concepts from minimal data. However, this capability raises significant concerns regarding data copyright and privacy. To mitigate these risks, adversarial examples have been proposed to safeguard data copyright. Existing methods can be categorized into three main paradigms based on their primary targets: (1) Encoder-targeted protections, which perturb the VAE to severely distort latent features [10, 20]; (2) Denoising-targeted protections, which disrupt the UNet's noise predictions to hinder the model from learning personalized concepts [16, 11]; and (3) Hybrid approaches, which combine both encoder and UNet disruptions to generate more comprehensive and robust perturbations [17]. Furthermore, recent advancements have focused on addressing the practical limitations of these early protections. PID [18] proposes a prompt-independent defense by directly manipulating the visual encoder, decoupling the protection efficacy from textual conditions. AIF [19] introduces an alternating iterative framework with an angle shifting mechanism and ensemble learning, enabling the generation of more stable and durable adversarial perturbations.

Image Purification Breaks Protection. Despite their promise, these protections remain vulnerable to image purification. DiffPure [25], a pioneering diffusion-based purification method, primarily eliminates perturbations through a process of forward noising and reverse denoising. GrIDPure [26] improves this approach with a patch-based iterative strategy to better preserve structural details. Similarly, Noisy Upscaling [28] disrupts adversarial structures with Gaussian noise before reconstructing images via more powerful super-resolution models. However, these noising and denoising methods suffer from a fundamental trade-off between fidelity and purification efficacy, especially against high-intensity perturbations. Specifically, adding minimal noise fails to effectively eliminate adversarial effects, whereas applying excessive noise induces semantic distortion and hallucinations. Alternatively, IMPRESS [31] identifies the reconstruction inconsistency between adversarial

and clean samples, proposing an optimization-based strategy guided by similarity and consistency losses. Nevertheless, IMPRESS remains limited by its high computational overhead and suboptimal performance against high-intensity perturbations.

3 Methodology



(a) Frequency-Wise Semantic Recovery **(b)** Gradient Magnitude Distributions

Figure 2: **(a)** Cosine similarity of low-frequency (LF) and high-frequency (HF) components across diffusion timesteps. **(b)** Density distributions of gradient magnitudes for clean and adversarial images.

3.1 Dynamic Sparse Semantic Anchoring

To completely eliminate adversarial perturbations at the physical level, we bypass the fidelity-purification trade-off of traditional methods by starting the reverse denoising process directly from pure Gaussian noise ($t = T$). However, unconditional full denoising often leads to severe semantic loss. To address this, we propose a dynamic semantic anchoring strategy based on sparse observations.

Spatial Blending via Bernoulli Sampling. Instead of a simple concatenation, we perform a precise spatial blend between the noisy structural anchors and the model's unconditional predictions. At each timestep t, we obtain the noisy adversarial image x_{adv}^{t-1} via forward diffusion, while the model generates an unconditional prediction x_{t-1}^{pred}. The fused state x_{t-1}^{fused} is formulated as:

$$x_{t-1}^{fused} = \mathbf{M}_t \odot x_{adv}^{t-1} + (1 - \mathbf{M}_t) \odot x_{t-1}^{pred} \tag{2}$$

where $\odot$ denotes the Hadamard product. The binary mask $\mathbf{M}_t$ is not deterministic; it is sampled from a dynamic retention probability map $\mathbf{P}_t$ using Bernoulli sampling: $\mathbf{M}_{t,(i,j)} \sim \text{Bernoulli}(\mathbf{P}_{t,(i,j)})$. The design of $\mathbf{P}_t$ evolves according to the spectral preferences of the diffusion denoising process.

Dynamic Scheduling Based on Spectral Bias. Diffusion models exhibit a significant spectral bias: they prioritize reconstructing the low-frequency global framework in early stages and shift towards high-frequency details in later stages (Figure 2 Left). Adapting to this dynamic, we design the probability map $\mathbf{P}_t(i, j)$ to smoothly transition via a time-dependent cosine weight $w(t)$:

$$\mathbf{P}_t(i, j) = w(t) \cdot P_{uni} + (1 - w(t)) \cdot P_{sobel}(i, j) \tag{3}$$

In the early denoising phase (when t is large and $w(t) \to 1$), the map is dominated by a uniform baseline probability P_{uni}. This acts as a random sparse sampling of the image's global layout, providing enough low-frequency information to anchor the object's position while the high-variance Gaussian noise easily neutralizes adversarial perturbations. As denoising progresses (when $t \to 0$ and $w(t) \to 0$), the guidance smoothly shifts towards P_{sobel} to anchor high-frequency real textures.

Sparse Semantic Anchoring via CDF Mapping. During the later stages, directly anchoring all high-frequency signals is risky, as the high-frequency components of adversarial perturbations become deeply entangled with natural edges. However, they exhibit a distinctive disparity in gradient distributions: adversarial fluctuations are strictly bounded by a small budget ϵ, whereas natural semantic edges exhibit unbounded, strong local contrast, as shown in Figure 2 (Right). To selectively anchor natural edges while filtering out toxicity, we calculate the gradient magnitude $g(i, j)$ of the

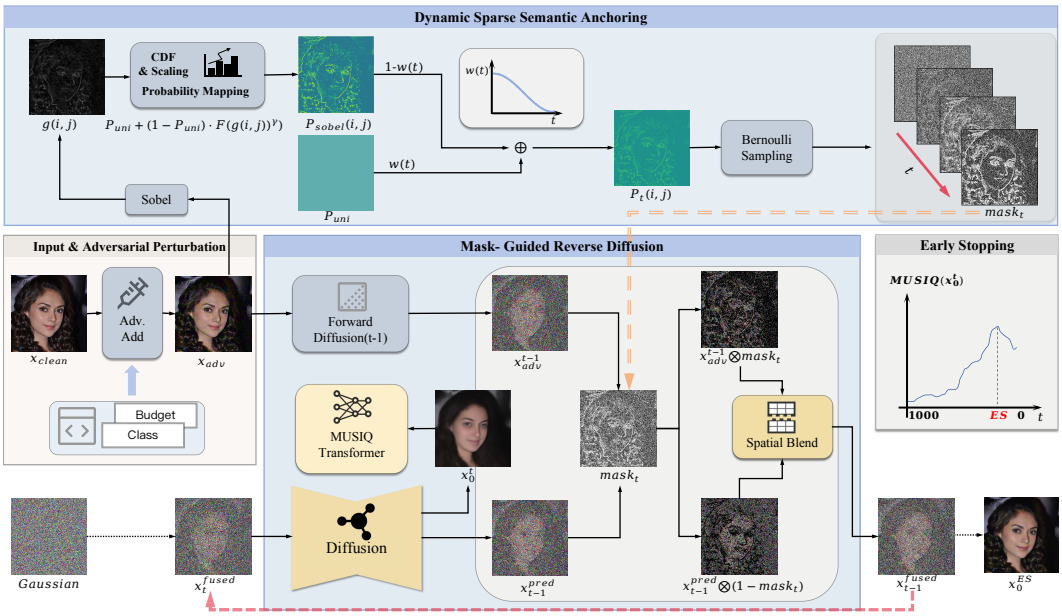


Figure 3: Overview of the proposed DSSA framework. (1) Initialization: To blindly neutralize protective perturbations of unknown intensities, the reverse diffusion process is initiated directly from pure Gaussian noise. (2) Mask-Guided Generation: At each timestep, the DSSA strategy computes a sparse retention mask to anchor semantic structures from the adversarial image, guiding the unconditional denoising trajectory. (3) Early-Stopping: Early-stopping mechanism halts the generation process at the optimal timestep, effectively preventing the perturbation resurgence.

adversarial image and introduce its empirical Cumulative Distribution Function (CDF), denoted as $F(\cdot)$. The structure-aware retention probability P_{sobel} is formulated as:

$$P_{sobel}(i,j) = P_{uni} + (1 - P_{uni}) \cdot [F(g(i,j))]^{\gamma} \tag{4}$$

Through this CDF mapping, we implement a soft-thresholding quantile anchoring strategy. The non-linear decay factor $\gamma(\gamma > 1)$ exponentially compresses the retention probability of pixels in the low-to-mid quantiles, effectively suppressing both the bounded adversarial noise and the overlapping weak textures. Simultaneously, it almost losslessly preserves the genuine semantic anchors located at the extreme high quantiles ($F(g) \to 1$).

3.2 Perception-Guided Early-Stopping

Late-Stage Perturbation Resurgence. In the early and middle denoising stages, the scale of the injected Gaussian noise is sufficiently large to mask the adversarial perturbation preserved within the structural anchors. This allows the model to safely traverse the natural image manifold. However, as the timestep $t \to 0$, the noise variance σ_t^2 gradually decays. Given the robust recovery capability of diffusion models, the diminishing noise can no longer suppress the inherent frequencies of the anchored adversarial patterns. Consequently, the final denoising stages inevitably lead to perturbation resurgence, wherein the model reconstructs the original adversarial perturbation. Therefore, a theoretical optimal early-stopping point, t^*, must exist to stop the trajectory before this resurgence occurs.

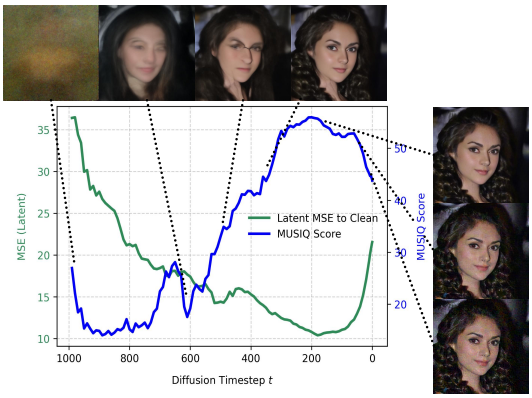


Figure 4: The peak of the reference-free perceptual metric aligns with the minimum VAE latent feature distance to the clean image.

Cross-Domain Alignment for Blind Purification and Early-Stopping Mechanism. Protective adversarial examples against Latent Diffu-

sion Models primarily function by inducing severe feature shifts within the VAE encoder [21]. Such shifts distort the latent representations of protected images, thereby disrupting the subsequent diffusion process and degrading generation quality. Thus, the theoretical optimal stopping point t^* should be the moment when the purified image $\hat{x}_0^t$ is closest to the original clean sample x_{clean} in the VAE latent space:$t^* = \arg\min_t ||\mathcal{E}(\hat{x}_0^t) - \mathcal{E}(x_{clean})||_2$. However, in real-world blind purification scenarios, x_{clean} is strictly unavailable, rendering this supervised objective intractable. To bridge this gap, we reveal a striking cross-domain alignment between unsupervised pixel-space visual quality and supervised latent-space detoxification. As illustrated in Figure 4, the peak values of the reference-free perceptual metric (MUSIQ [32]) along the denoising timeline meticulously align with the minimum points of the VAE latent feature distance. This observation provides a critical insight: the recovery of natural perceptual quality practically strictly equates to the neutralization of adversarial shifts in the latent space. Based on this alignment, we reformulate the intractable supervised purification into an unsupervised optimization guided by the MUSIQ score. At a given timestep t, we first project the current noisy state x_t directly back to the pixel space to obtain the clean image estimation $\hat{x}_0^t$ using Tweedie's formula. To ensure computational efficiency and avoid executing the MUSIQ Transformer at every single step, we evaluate the score at a regular stride K. We employ a sliding window strategy of size W to monitor the trend of MUSIQ($\hat{x}_0^t$). If the score passes its peak and exhibits a continuous monotonic decline within the window, it indicates that the trajectory has crossed the optimal point and the model has begun reconstructing adversarial patterns. We then immediately trigger early stopping and backtrack to output the $\hat{x}_0^t$ associated with the highest MUSIQ score, yielding a highly purified and structurally faithful final image.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We conduct experiments on two image domains: human faces and artistic styles. For human faces, we select 20 identities from the VGGFace2 dataset [33]. For artistic styles, we choose collections from 10 different artists in the WikiArt dataset. To comprehensively evaluate the purification methods, we establish an evaluation framework based on two dimensions: Fidelity and Purification Efficacy. For fidelity, we use PSNR, SSIM, and LPIPS [34] to quantify the visual discrepancy between the purified and original clean samples. This rigorously assesses whether the purification process degrades the quality of images or induces severe hallucinations. For purification efficacy regarding artistic style protection, we primarily use FID [35] and CLIP Score [36]. These metrics evaluate whether models fine-tuned on purified samples can approximate the ideal performance of those fine-tuned on clean data. Lower FID and higher CLIP Scores indicate that the purified samples successfully recover the artistic style features obscured by perturbations. For the human face dataset, we additionally introduce the Identity Similarity Metric (ISM) [11]. By extracting and matching facial feature vectors between generated and real images, a higher ISM demonstrates more thorough perturbation removal and better identity preservation.

BaseLine and Diffusion Models. For purification baselines, we compare three state-of-the-art strategies: DiffPure [25], GrIDPure [26], and Noisy Upscaling [28]. For fair comparison, we strictly adhere to the default hyperparameters from their official open-source repositories. Specifically, we use the 512×512 unconditional diffusion model for DiffPure. To accommodate GrIDPure's patch-partitioning strategy, we utilize the 256×256 unconditional diffusion model. For Noisy Upscaling, we employ the SD-x4-upscaler [6]. To evaluate purification performance, our main experiments are conducted using DreamBooth. Additional fine-tuning settings are described in the Appendix.

Implementation Details. For constructing protected data, we consistently use Stable Diffusion v1.5 (SD v1.5) as the target architecture. We cover three mainstream protection paradigms: VAE encoder-targeted perturbations (PhotoGuard, PID), denoising network-targeted perturbations (AdvDM, AIF), and hybrid perturbations (Mist). Furthermore, to validate the boundary capabilities and robustness of our purification scheme, we conduct stress tests with perturbation budgets set to $\epsilon = 8/255$ and $\epsilon = 16/255$. To address adaptive anti-purification defenses, we select StyleGuard and AntiPure, strictly following the experimental scenarios established in their original papers under a perturbation budget of 8/255. For our DSSA purification method, we primarily employ a 512×512 unconditional diffusion model. The uniform retention probability $P_{\text{uni}} = 0.5$ is chosen mainly to leverage the generative capability of the diffusion model rather than being treated as a hyperparameter; a detailed

analysis is provided in Appendix F.2. For the decay factor γ after mapping through the CDF function, we set $\gamma = 10$ in the main experiments. See Appendix A for more experimental details.

4.2 Experimental Results

Table 1: Quantitative evaluation of DreamBooth fine-tuning under different perturbation budgets.

Budget	Purification	AdvDM		PhotoGuard		Mist		PID		AIF	
		ISM ↑	FID ↓	ISM ↑	FID ↓	ISM ↑	FID ↓	ISM ↑	FID ↓	ISM ↑	FID ↓
Dataset: VGGFace2 (Human Faces)											
$8/255$	None *(Unpurified)*	0.4913	306.85	0.5638	176.72	0.5184	198.25	0.4837	210.32	0.5228	282.94
	IMPRESS	0.4778	298.41	0.5687	170.24	0.5101	187.67	0.4906	207.11	0.5263	267.44
	DiffPure	**0.6566**	129.29	0.6214	141.28	0.6393	129.32	**0.6552**	127.37	0.6503	132.10
	GrIDPure	0.584	156.84	0.6196	152.51	0.5910	148.90	0.6145	151.79	0.5923	170.37
	Noisy Upscaling	0.5848	150.11	0.5846	144.18	0.6001	152.69	0.4626	174.79	0.5936	161.48
	DSSA (Ours)	0.6374	**123.13**	**0.6512**	**131.38**	**0.6410**	**126.31**	0.6300	**126.55**	**0.6600**	**126.50**
$16/255$	None *(Unpurified)*	0.2750	383.49	0.1466	478.57	0.3934	281.96	0.0982	408.23	0.4039	346.45
	IMPRESS	0.3121	349.54	0.2181	427.15	0.3374	311.09	0.1743	398.72	0.4226	321.06
	DiffPure	0.5549	171.16	0.5754	166.33	0.5355	180.06	0.5423	162.52	0.5444	166.47
	GrIDPure	0.5129	253.57	0.5592	214.43	0.5976	163.22	0.6137	152.81	0.5072	270.96
	Noisy Upscaling	0.5732	164.71	0.5639	170.66	0.5500	178.50	0.5756	162.75	0.5704	164.44
	DSSA (Ours)	**0.6299**	**135.60**	**0.6154**	**140.98**	**0.6011**	**150.78**	**0.6356**	**141.19**	**0.6211**	**145.46**
Dataset: WikiArt (Artistic Styles)											
Metrics Reference		CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓
$8/255$	None *(Unpurified)*	0.7122	163.16	0.6546	162.83	0.6719	160.88	0.7135	156.17	0.6323	169.02
	IMPRESS	0.7921	131.01	0.7632	145.88	0.8021	108.13	0.7816	120.10	0.7727	124.06
	DiffPure	0.8069	101.41	0.8174	103.29	0.7911	110.13	0.7945	114.63	0.8247	108.80
	GrIDPure	0.8214	115.26	0.8303	101.84	0.8391	**81.52**	0.8360	97.36	0.8118	104.77
	Noisy Upscaling	**0.8379**	82.77	**0.8408**	95.11	0.8136	104.77	0.8351	111.49	**0.8329**	98.42
	DSSA (Ours)	0.8249	**78.29**	0.8206	**85.85**	**0.8461**	92.07	**0.8484**	**73.61**	0.8299	**87.36**
$16/255$	None *(Unpurified)*	0.6766	221.06	0.6426	273.11	0.6993	249.71	0.7033	199.56	0.6951	206.97
	IMPRESS	0.6851	214.79	0.6524	263.45	0.7103	223.64	0.7113	190.43	0.7199	199.35
	DiffPure	0.8008	105.17	0.7894	127.86	0.7916	110.92	0.7865	121.99	0.8033	115.81
	GrIDPure	0.8027	159.73	0.8048	160.61	0.7823	110.28	**0.8269**	99.02	0.8135	132.55
	Noisy Upscaling	0.8315	96.63	0.8318	104.67	**0.8339**	99.82	0.7975	97.43	0.8183	112.24
	DSSA (Ours)	**0.8371**	**92.39**	**0.8337**	**81.24**	0.8305	**94.79**	0.8208	**92.87**	**0.8218**	**108.39**

Main Results. We begin by evaluating the purification efficacy of different methods under downstream DreamBooth fine-tuning. As shown in Table 1, under a low-budget perturbation setting (e.g., $\epsilon = 8/255$), most SDEdit-based baseline methods are able to partially recover the clean semantic features. However, when faced with high-budget adversarial perturbations ($\epsilon = 16/255$), existing methods expose a fundamental vulnerability. They universally fail to completely remove the injected adversarial perturbation, resulting in downstream models that cannot learn the correct facial identities or artistic styles. This fragility stems from the inherent mechanism of prior purification methods, which struggle to find a feasible solution space between corrupting adversarial signals and preserving the original image semantics. In contrast, our proposed method achieves the highest performance in the majority of evaluated settings across both the VGGFace2 and WikiArt datasets. Beyond purification efficacy, we quantify the image fidelity by measuring the perceptual distance between the purified samples and the original clean images (Table 2). By leveraging a frequency-aware dynamic masking strategy, our approach anchors the denoising trajectory to the natural manifold of the original image using high-confidence sparse gradients as semantic priors. These design choices enable our method to achieve the highest image fidelity among the evaluated baselines. We provide qualitative visualizations in Figure 5 and Figure 6 to demonstrate the superior perceptual fidelity and purification efficacy of our framework. Consequently, our purified samples most closely resemble the real clean images while stripping away high-budget adversarial perturbations.

Ablation Study. We further conduct an ablation study on the face domain against the high-budget AdvDM attack ($\epsilon = 16/255$). As presented in Table 3, each component plays an indispensable role in the overall purification performance. Specifically, compared to a naive static uniform mask, our Dynamic Scheduled Masking strategy consistently improves image fidelity regardless of the early stopping mechanism, as evidenced by an SSIM increase from 0.6383 to 0.6435 without early stopping, and a further boost from 0.7896 to 0.8338 when combined with early stopping. This verifies its capability to accurately recover high-frequency structural details without sacrificing the underlying semantic manifold. Furthermore, incorporating the MUSIQ-based early stopping mechanism explicitly prevents perturbation resurgence in late stage. By stopping the denoising process right before the adversarial perturbations are reconstructed, this mechanism not only successfully

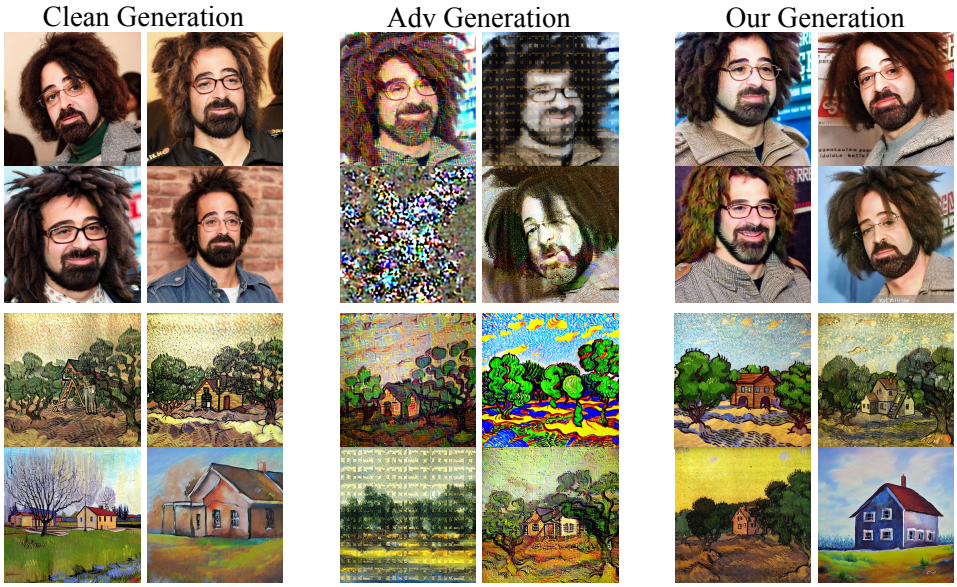


Figure 5: Qualitative comparison of DreamBooth generation results. While models fine-tuned on adversarial examples suffer from severe artifacts and semantic corruption, our method successfully neutralizes the protection to generate high-fidelity images that accurately preserve target identities and artistic styles.

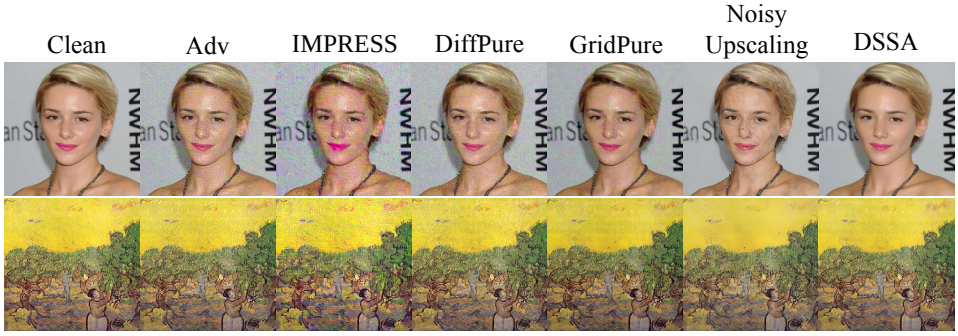
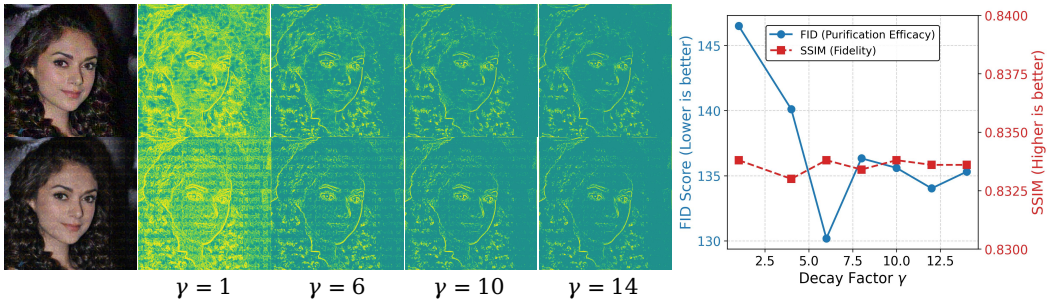


Figure 6: Visual comparison of purified samples. Existing baseline methods often struggle to completely remove adversarial artifacts without compromising image quality, whereas our DSSA method achieves superior perceptual fidelity and effectively restores the original clean semantics.



(a) Retention Probability Maps **(b)** Purification Efficacy and Fidelity

Figure 7: Analysis of the decay factor γ. **(a)** Retention probability maps P_{sobel} for different values of γ. Increasing γ emphasizes prominent facial contours while reducing retention probabilities in low-gradient regions. **(b)** Quantitative impact of γ on purification efficacy (FID) and image fidelity (SSIM). SSIM remains relatively stable across the evaluated values, while FID is lowest at $\gamma = 6$.

Table 2: Quantitative comparison of image fidelity across different purification methods under perturbation budgets $\epsilon = 8/255$ and $\epsilon = 16/255$.

Method	**Perturbation Budget:** $\epsilon = 8/255$			**Perturbation Budget:** $\epsilon = 16/255$		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
IMPRESS	27.40	0.7176	0.3781	23.61	0.5254	0.6229
DiffPure	29.17	0.7988	0.1861	27.09	0.6733	0.3707
GrIDPure	26.55	0.8478	0.1798	25.49	0.8079	0.2759
Noisy Upscaling	27.12	0.7408	0.2197	26.87	0.7396	0.2301
DSSA (Ours)	**32.48**	**0.8911**	**0.1224**	**29.05**	**0.8264**	**0.2117**

Table 3: **Ablation study on the face domain** ($\epsilon = 16/255$). We investigate the contribution of the Dynamic Masking strategy (Sobel-based) and the MUSIQ-based Early Stopping mechanism.

Component		Fidelity ↑		Purification Efficacy	
Masking Strategy	Early Stopping	PSNR ↑	SSIM ↑	FID ↓	ISM ↑
Static Uniform	×	27.06	0.6383	221.16	0.5531
Dynamic Scheduled	×	27.44	0.6435	370.35	0.2799
Static Uniform	✓	27.89	0.7896	148.08	0.5846
Dynamic Scheduled (Ours)	✓	**29.94**	**0.8338**	**135.6**	**0.6299**

evades toxicity recovery but also further stabilizes the fidelity metrics. The synergistic combination of both components ultimately achieves the optimal balance between high-fidelity reconstruction (PSNR: 29.94) and superior purification efficacy (ISM: 0.6299).

Hyperparameter Analyses. We investigate the impact of the decay factor γ. As shown in Figure 7 (Right), the fidelity metric (SSIM) remains strictly stable across all γ values, demonstrating that our dynamic masking framework intrinsically guarantees structural preservation. Regarding purification efficacy, while $\gamma = 6$ achieves a local optimum, the FID scores maintain highly competitive performance across a broad intermediate range. To provide an intuitive physical understanding of how γ acts as a spatial toxicity filter, we visualize the evolution of the retention probability maps P_{sobel}. As demonstrated in the heatmaps, a small γ fails to adequately suppress the background adversarial noise. As γ increases to an optimal range, the mask achieves extreme sparse semantic anchoring—successfully isolating the core facial contours while safely compressing the gradient overlap regions. Conversely, an excessively large γ (e.g., $\gamma = 14$) begins to fragment natural textures.

Robustness against Anti-Purification Defenses. Recently, advanced protective perturbations specifically designed to resist diffusion-based purification have emerged. To rigorously evaluate the robustness of our framework against such adaptive protections, we select two state-of-the-art anti-purification methods: StyleGuard [37] and AntiPure [38]. To strictly respect their original threat models and optimal operational domains, we evaluate StyleGuard primarily on the WikiArt dataset (targeting style mimicry) and AntiPure on the VGGFace2 dataset (targeting structural identity). As shown in Table 4, while these advanced protections successfully break their targeted baselines (e.g., StyleGuard effectively bypasses Noisy Upscaling, and AntiPure circumvents GrIDPure), our proposed framework maintains consistently high purification efficacy. The intrinsic robustness of DSSA stems from its fundamental departure from the partial-noising paradigm; by initiating the reconstruction trajectory from pure Gaussian noise, our method prioritizes the restoration of low-frequency global semantics before progressively refining high-frequency details. This progressive reconstruction effectively circumvents the core assumptions of adaptive protection like AntiPure, which rely on preserving adversarial high-frequency structures during shallow denoising cycles. Since our globally sparse anchoring stochastically avoids the local patch dependencies and fixed frequency modulations exploited by these defenses, DSSA reliably decouples intrinsic genuine image contours from adversarial artifacts across different semantic scales.

Table 4: Robustness against advanced anti-purification defenses.

StyleGuard	**FID ↓**	**CLIP ↑**
None (Unpurified)	380.61	0.6131
DiffPure	343.77	0.6822
Noisy Upscaling	298.98	0.7256
DSSA (Ours)	**101.86**	**0.7938**
AntiPure	**FID ↓**	**ISM ↑**
None (Unpurified)	167.01	0.5478
GrIDPure	140.47	0.5823
DSSA (Ours)	**121.78**	**0.6744**

Table 5: Purification performance on zero-shot personalization methods using VGGFace2 ($\epsilon = 16/255$). Best results are in bold.

		IP-Adapter		InstantID	
Attack	Purification	FID ↓	ISM ↑	FID ↓	ISM ↑
AdvDM	DiffPure	110.02	0.2895	93.84	0.3819
	GrIDPure	82.75	0.3194	82.01	0.4277
	Noisy Upscaling	85.86	0.2840	102.12	0.3306
	DSSA (Ours)	**61.17**	**0.3216**	**69.01**	**0.4504**
Mist	DiffPure	86.13	0.2972	89.86	0.3701
	GrIDPure	68.43	**0.3348**	76.37	0.4637
	Noisy Upscaling	84.16	0.2829	110.93	0.3120
	DSSA (Ours)	**65.79**	0.3290	**72.88**	**0.4708**

Generalization to Zero-Shot Personalization. To further evaluate the generalizability of DSSA, we conduct additional experiments on two zero-shot personalization methods, IP-Adapter [8] and InstantID [39], using the VGGFace2 dataset. We consider AdvDM and Mist with a perturbation budget of 16/255 and compare DSSA against three purification baselines. As shown in Table 5, DSSA achieves the lowest FID across all four evaluated settings and the highest ISM in three of them. In particular, DSSA achieves FID scores of 61.17 and 65.79 with IP-Adapter, and 69.01 and 72.88 with InstantID, under AdvDM and Mist, respectively. These results demonstrate that DSSA's purification effectiveness extends beyond fine-tuning-based personalization to zero-shot personalization methods.

Computational Efficiency. We compare the computational cost of different purification methods using the same input resolution on a single NVIDIA RTX 3090 GPU. As shown in Table 6, DSSA requires 189.76 seconds per image, comparable to GrIDPure (179.45 s) but slower than DiffPure and Noisy Upscaling. This overhead mainly stems from the longer denoising trajectory initialized from pure Gaussian noise. Nevertheless, DSSA requires only 4.27 GB of peak GPU memory, nearly identical to DiffPure (4.28 GB). Therefore, the primary computational limitation of DSSA lies in inference latency rather than memory consumption. Reducing the number of denoising steps is a promising direction for future work.

Table 6: Computational efficiency comparison on a single NVIDIA RTX 3090 GPU.

Purification	Runtime (s) ↓	Memory (GB) ↓
IMPRESS	1216.04	9.68
DiffPure	**15.24**	4.28
GrIDPure	179.45	**2.88**
Noisy Upscaling	52.17	18.74
DSSA (Ours)	189.76	4.27

5 Conclusion

In this paper, we propose Dynamic Sparse Semantic Anchoring (DSSA), a novel framework that fundamentally mitigate the purification efficacy and fidelity trade-off inherent in existing diffusion-based purification methods. By initiating the denoising trajectory from pure Gaussian noise and leveraging a perception-guided sparse anchoring strategy, DSSA effectively neutralizes high-budget protective perturbations while meticulously preserving semantic fidelity. Our experiments show that current protective methods have serious vulnerabilities. They are not robust against real-world purification attempts by motivated Image Exploiters. We hope these findings urge the Image Protector community to develop next-generation, robust defense algorithms to reliably safeguard data copyright.

References

[1] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020.

[2] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020.

[3] Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023.

[4] William S. Peebles and Saining Xie. Scalable diffusion models with transformers. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, 2022.

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

[6] 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 (CVPR)*, pages 10684–10695, June 2022.

[7] 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. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2022.

[8] Hu Ye, Jun Zhang, Siyi Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *ArXiv*, abs/2308.06721, 2023.

[9] 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. *ArXiv*, abs/2208.01618, 2022.

[10] 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 *USENIX Security Symposium*, 2023.

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

[12] Zhenting Wang, Chen Chen, Lingjuan Lyu, Dimitris N Metaxas, and Shiqing Ma. Diagnosis: Detecting unauthorized data usages in text-to-image diffusion models. *arXiv preprint arXiv:2307.03108*, 2023.

[13] Boheng Li, Yanhao Wei, Yankai Fu, Zhenting Wang, Yiming Li, Jie Zhang, Run Wang, and Tianwei Zhang. Towards reliable verification of unauthorized data usage in personalized text-to-image diffusion models. *2025 IEEE Symposium on Security and Privacy (SP)*, pages 2564–2582, 2024.

[14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

[15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014.

[16] 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*, abs/2302.04578, 2023.

[17] Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *ArXiv*, abs/2305.12683, 2023.

[18] Ang Li, Yichuan Mo, Mingjie Li, and Yisen Wang. Pid: Prompt-independent data protection against latent diffusion models. *ArXiv*, abs/2406.15305, 2024.

[19] Minghao Li, Rui Wang, Mingze Sun, and Lihua Jing. Preventing latent diffusion model-based image mimicry via angle shifting and ensemble learning. In *International Joint Conference on Artificial Intelligence*, 2025.

[20] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *International Conference on Machine Learning*, 2023.

[21] Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion based mimicry through score distillation. *ArXiv*, abs/2311.12832, 2023.

[22] Jiang Liu, Chun Pong Lau, and Ramalingam Chellappa. Diffprotect: Generate adversarial examples with diffusion models for facial privacy protection. *ArXiv*, abs/2305.13625, 2023.

[23] 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. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12047–12056, 2023.

[24] 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. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24219–24228, 2023.

[25] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, 2022.

[26] Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by stable diffusion? *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24398–24407, 2023.

[27] Haotian Xue and Yongxin Chen. Pixel is a barrier: Diffusion models are more adversarially robust than we think. *ArXiv*, abs/2404.13320, 2024.

[28] Robert Honig, Javier Rando, Nicholas Carlini, and Florian Tramèr. Adversarial perturbations cannot reliably protect artists from generative ai. *ArXiv*, abs/2406.12027, 2024.

[29] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[30] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[31] 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. *ArXiv*, abs/2310.19248, 2023.

[32] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5128–5137, 2021.

[33] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74, 2017.

[34] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[35] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *ArXiv*, abs/1706.08500, 2017.

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

[37] Yanjie Li, Wenxuan Zhang, Xinqi Lyu, Yihao Liu, and Bin Xiao. Styleguard: Preventing text-to-image-model-based style mimicry attacks by style perturbations. *ArXiv*, abs/2505.18766, 2025.

[38] Wenkui Yang, Jie Cao, Junxian Duan, and Ran He. Towards robust defense against customization via protective perturbation resistant to diffusion-based purification. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19290–19300, 2025.

[39] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *ArXiv*, abs/2401.07519, 2024.

[40] Jiankang Deng, J. Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *ArXiv*, abs/1905.00641, 2019.

[41] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI Conference on Artificial Intelligence*, 2022.

A Detailed Experimental Settings

A.1 DSSA Purification and Protective Sample Generation Details

In our experiments, the DSSA purification process is conducted using a 512×512 unconditional diffusion model. All input images are processed as 3-channel RGB images and strictly normalized to a $[0, 1]$ range. All experiments are executed on 4 NVIDIA RTX 3090 GPUs. The hyper-parameters for our dynamic sparse masking are set to $P_{uni} = 0.5$ and $\gamma = 10$, which we empirically found to generalize exceptionally well across both the VGGFace2 and WikiArt datasets. Purifying a single image consumes 4.2 GB of VRAM on a single RTX 3090 GPU and takes approximately 3 minutes and 10 seconds.

For the generation of protective adversarial samples, we strictly adhere to the default configurations specified in their respective original papers. Specifically, the optimization for PID is conducted for 1000 iterations, whereas the other baseline attacks (e.g., AdvDM, PhotoGuard, Mist, AIF) are optimized for 100 iterations. The adversarial perturbation budgets are uniformly set to $\epsilon = 8/255$ and $\epsilon = 16/255$ under the L_∞ norm.

A.2 DreamBooth Fine-Tuning Details

For personalization via DreamBooth, we employ specific configurations tailored to the respective data domains. In the human face domain (VGGFace2), the model is fine-tuned for 500 steps per identity. We utilize the prompt "a photo of a <sks> person" for the subsequent image generation. In the artistic style domain (WikiArt), we randomly select 10 images per artist as the training set, and the model is fine-tuned for 1000 steps. We use "an <sks> painting" as the training prompt, and evaluate the style mimicry capability using the target generation prompt "a <sks> painting of a house".

A.3 LoRA Fine-Tuning Details

To evaluate the purification efficacy under parameter-efficient fine-tuning paradigms, we utilize Low-Rank Adaptation (LoRA). For the VGGFace2 dataset, the LoRA modules are trained for 40 epochs per identity. We then generate 640 images per identity using the prompt "a photo of a <sks> person". Similarly, for the WikiArt dataset, the models are fine-tuned for 40 epochs per artistic style, yielding 640 generated images per artist using the prompt "a painting in the style of <sks>".

A.4 Details of Purification Baselines

We compare our framework against three representative diffusion-based purification baselines. For DiffPure, we set the forward noising timestep to $t = 100$ and employ the exact same 512×512 unconditional diffusion model used in our method to ensure a fair comparison. For GrIDPure, we configure the parameters according to the optimal settings reported in their study, setting the noise scale to $t = 10$, the number of iterations to $iter = 20$, and the blending parameter to $\gamma = 0.1$. For Noisy Upscaling, we strictly adhere to the default parameters provided by the authors, utilizing the SD-x4-upscaler architecture with the noise level set to 320.

A.5 Evaluation Metrics

We adopt three widely used metrics to comprehensively assess the purification efficacy and generation quality:

- **FID:** The Fréchet Inception Distance (FID) evaluates the generative fidelity by measuring the discrepancy between the feature distributions of the generated images and the original, clean images.
- **CLIP Score:** We adapt the CLIP Score to evaluate the semantic disruption induced by the protective perturbations. Specifically, we compute the image-to-image CLIP similarity between two sets of generated images: (1) images generated by a model personalized on clean, unprotected samples, and (2) images generated by a model personalized on the purified adversarial samples.

- **ISM:** The Identity Similarity Metric (ISM) is explicitly used to evaluate performance in the facial domain. We leverage the RetinaFace detector [40] to robustly locate and crop facial regions, from which we extract identity-specific feature embeddings for comparison.

B Theoretical Upper Bound of Adversarial Gradient Magnitude

In Section 3.2, we introduce the Dynamic Sparse Semantic Anchoring (DSSA) strategy, which anchors high-frequency structural signals based on the disparity in gradient distributions. We state that the gradient magnitudes of adversarial perturbations are strictly bounded. Here, we provide the formal derivation for the theoretical maximum upper bound of the magnitude of adversarial gradient under the L_∞-norm threat model, as illustrated in Figure 2. Let the adversarial perturbation be δ, which is strictly bounded by the perturbation budget ϵ under the L_∞-norm. Thus, for any pixel (i, j) in the perturbation map, its amplitude satisfies:

$$|\delta(i, j)| \leq \epsilon \tag{5}$$

In our framework, the gradient magnitude $g(i, j)$ is computed using the Sobel operator, which consists of two 3×3 convolution kernels for the horizontal (G_x) and vertical (G_y) derivatives:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{6}$$

The horizontal gradient response $\nabla_x \delta$ at a given pixel is obtained by the convolution $G_x * \delta$. To find the maximum possible response, the pixels corresponding to positive weights in G_x must take the maximum positive perturbation ϵ, while the pixels corresponding to negative weights must take the minimum negative perturbation $-\epsilon$. Therefore, the maximum absolute horizontal gradient is:

$$\max |\nabla_x \delta| = \big(1 \cdot \epsilon + 2 \cdot \epsilon + 1 \cdot \epsilon\big) - \big(-1 \cdot (-\epsilon) - 2 \cdot (-\epsilon) - 1 \cdot (-\epsilon)\big) = 4\epsilon + 4\epsilon = 8\epsilon \tag{7}$$

By symmetry, the maximum possible vertical gradient response is:

$$\max |\nabla_y \delta| = 8\epsilon$$

The overall gradient magnitude is defined as $g = \sqrt{(\nabla_x \delta)^2 + (\nabla_y \delta)^2}$. To establish a strict upper bound, we consider the extreme case where both horizontal and vertical gradients reach their theoretical maximums simultaneously. Thus, the theoretical maximum gradient magnitude g_{max} for the adversarial perturbation is bounded by:

$$g_{max} \leq \sqrt{(8\epsilon)^2 + (8\epsilon)^2} = \sqrt{128\epsilon^2} = 8\sqrt{2}\epsilon \tag{8}$$

By substituting the specific perturbation budgets commonly used in protective attacks into this bound, we obtain: For $\epsilon = 8/255$, the theoretical upper bound is $g_{max} = 8\sqrt{2} \times \frac{8}{255} \approx 0.355$. For $\epsilon = 16/255$, the theoretical upper bound is $g_{max} = 8\sqrt{2} \times \frac{16}{255} \approx 0.710$.

This theoretical bound demonstrates that under a strict L_∞-norm , the maximum contribution of adversarial perturbations to the local gradient magnitude is inherently limited by the properties of the Sobel operator. Consequently, for high-contrast structural edges whose gradient magnitudes significantly exceed this bound (e.g., $g > 0.8$), the gradient response is fundamentally dominated by the genuine image content rather than the adversarial noise. While this does not imply the absolute absence of adversarial signals in these regions, it provides a strong mathematical guarantee that the adversarial interference is quantitatively capped. This property serves as a robust prior for our CDF-based spatial probability mapping, enabling our framework to effectively prioritize the anchoring of genuine semantic contours while substantially suppressing the influence of adversarial perturbations.

C Effectiveness across More Personalization Methods

Beyond full-parameter adaptation, we further evaluate our framework under parameter-efficient fine-tuning paradigms, specifically LoRA. As demonstrated in Table 7, our method consistently maintains state-of-the-art purification efficacy. Notably, while Noisy Upscaling occasionally exhibits competitive performance, this phenomenon fundamentally stems from its reliance on the powerful

generative prior of the SD super-resolution model. Rather than strictly detoxifying the adversarial remnants, Noisy Upscaling tends to actively synthesize and hallucinate rich high-frequency details. While such generative hallucinations inadvertently benefit the subjective evaluation of artistic styles, they fail to guarantee true structural fidelity. In contrast, our method robustly eliminates perturbations and achieves superior overall performance without relying on generative illusions.

Table 7: Quantitative evaluation of LoRA under large perturbation budgets.

Budget	Purification	AdvDM		PhotoGuard		Mist		PID		AIF	
		CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓
Dataset: VGGFace2 (Human Faces)											
$16/255$	None *(Unpurified)*	0.4889	279.33	0.4793	350.29	0.5223	186.05	0.4521	281.35	0.4923	231.73
	DiffPure	0.5579	110.64	0.5177	114.69	0.5404	84.47	0.5650	94.27	0.5565	102.23
	GrIDPure	0.5784	120.70	0.5422	155.84	0.5799	88.59	0.5754	102.64	0.5653	123.2
	Noisy Upscaling	0.5971	52.09	0.5713	107.38	0.5927	59.83	0.5721	90.87	0.6135	48.37
	DSSA (Ours)	**0.6162**	**41.81**	**0.6021**	**63.73**	**0.6095**	**56.89**	**0.6155**	**76.41**	**0.6219**	**42.04**
Dataset: WikiArt (Artistic Styles)											
$16/255$	None *(Unpurified)*	0.6522	269.22	0.6514	377.31	0.6767	246.83	0.6855	236.58	0.6722	239.29
	DiffPure	0.7921	110.75	0.7886	110.44	0.7711	128.71	0.7703	113.71	0.7972	86.39
	GrIDPure	0.7910	127.76	0.7878	151.50	0.7653	194.01	0.7923	128.99	0.7893	140.69
	Noisy Upscaling	0.7733	93.46	**0.7945**	**93.53**	0.7247	125.54	0.7657	140.63	0.7630	102.01
	DSSA (Ours)	**0.8118**	**65.38**	0.7901	118.6	**0.7812**	**116.61**	**0.8022**	**89.32**	**0.8032**	**72.81**

D Consistency Across Identities and Artistic Styles

To further investigate the consistency of DSSA across different identities and artistic styles, we conduct identity-level and style-level evaluations on VGGFace2 and WikiArt under AdvDM and Mist attacks ($\epsilon = 16/255$). We report the mean and standard deviation of ISM across 20 facial identities and FID across 10 artistic styles. Furthermore, we perform matched comparisons against three purification baselines on each identity or style. As shown in Tables 8 and 9, DSSA outperforms the baselines in 80.0% of identity-level comparisons and 76.7% of style-level comparisons. Notably, DSSA surpasses GrIDPure across all 20 identities under AdvDM. These results indicate that the overall improvements are broadly distributed across identities and styles rather than dominated by a few favorable cases.

E Additional Analysis of Perception-Guided Early Stopping

E.1 Correlation Between MUSIQ and the Fidelity–Purification Trade-off

To further validate the reliability of MUSIQ as a reference-free stopping signal, we conduct a trajectory-level correlation analysis across different datasets, protection methods, and perturbation budgets. At each saved denoising checkpoint, we compute MUSIQ together with clean-referenced VAE-MSE, LPIPS, PSNR, and SSIM. Correlations are calculated along each image trajectory.

As shown in Table 10, MUSIQ exhibits positive correlations with the purification trade-off across all evaluated settings. Under AdvDM, the Spearman correlation on VGGFace2 decreases from 0.8215 to 0.5989 as the perturbation budget increases, while remaining positive. On WikiArt, strong correlations are maintained across both perturbation budgets and protection methods. These results suggest that MUSIQ provides a useful proxy for identifying favorable regions of the denoising trajectory, rather than necessarily coinciding with the oracle optimum of every individual metric.

E.2 Sensitivity to the Choice of Perceptual Quality Estimator

To investigate the sensitivity of our early-stopping mechanism to the perceptual quality estimator, we replace MUSIQ [32] with CLIP-IQA [41] while keeping the saved denoising trajectories and all other experimental settings unchanged.

As shown in Table 11, both estimators exhibit positive correlations with the fidelity–purification trade-off, indicating that the stopping mechanism is not exclusively dependent on MUSIQ. However, MUSIQ achieves higher Pearson and Spearman correlations and lower VAE and trade-off regrets in

Table 8: Identity- and style-level statistics under AdvDM and Mist ($\epsilon = 16/255$). Values denote mean $\pm$ standard deviation.

Dataset / Metric	AdvDM	Mist
VGGFace2 / ISM ↑	0.6299 ± 0.0466	0.6011 ± 0.0478
WikiArt / FID ↓	92.39 ± 31.12	94.79 ± 28.26

Table 9: Matched identity/style comparisons under AdvDM and Mist ($\epsilon = 16/255$). Entries indicate DSSA wins over each baseline.

	VGGFace2		WikiArt	
Baseline	AdvDM	Mist	AdvDM	Mist
DiffPure	16/20	13/20	7/10	8/10
GrIDPure	20/20	16/20	8/10	9/10
Noisy Upscaling	16/20	15/20	6/10	8/10
Total	**52/60**	**44/60**	**21/30**	**25/30**

Table 10: Trajectory-level correlations between MUSIQ and fidelity–purification metrics across different datasets, protection methods, and perturbation budgets.

			Spearman ρ			Trade-off	
Dataset	Protection	Budget	VAE	LPIPS	PSNR	Pearson r	Spearman ρ
VGGFace2	AdvDM	8/255	0.6762	0.8150	0.8164	0.7856	0.8215
VGGFace2	AdvDM	16/255	0.5250	0.6374	0.5645	0.5181	0.5989
WikiArt	AdvDM	8/255	0.8678	0.8016	0.8016	0.8583	0.8128
WikiArt	AdvDM	16/255	0.8430	0.7697	0.7670	0.8976	0.8379
WikiArt	Mist	8/255	0.8540	0.7833	0.7358	0.8767	0.8252
WikiArt	Mist	16/255	0.8372	0.7680	0.7637	0.8253	0.7825

both evaluated settings. Although CLIP-IQA achieves slightly lower PSNR regret, MUSIQ provides a more consistent balance between perceptual fidelity and perturbation removal. These findings justify our choice of MUSIQ as the default estimator while highlighting the influence of perceptual quality estimation on stopping performance.

F Design Rationale and Empirical Analysis of the Masking Mechanism

In this section, we provide the rationale for the design of employing random pixel masking and detail the empirical study used to determine the optimal uniform retention probability of P_{uni}.

F.1 Why Random Pixel Masking instead of Patch/Region Masking?

The primary objective of the early stage masking in DSSA is to anchor the low-frequency semantic framework. We abandon traditional patch-based or region-based masking strategies for a critical reason: these strategies will induce severe structural hallucinations. When using contiguous patch masks, large continuous areas of the original image are entirely obscured (as shown in Figure 8). During the mask-guided spatial blending process, the diffusion model is forced to independently hallucinate these expansive missing regions based on the surrounding context. Due to the high degrees of freedom in unconditional generation, this often leads to catastrophic semantic distortions, such as the generation of disjointed textures or anatomically incorrect structures. In contrast, our Bernoulli-based random pixel masking is different. Instead of continuous blank regions, the model receives uniformly distributed, extremely sparse discrete pixels. This strategy perfectly aligns with the spectral bias of diffusion models. The sparse pixels act as a low-frequency anchor, restricting the macroscopic generation trajectory without constraining the local high-frequency details. Consequently, the model smoothly reconstructs the image using its rich generative prior without suffering from regional hallucinations.

F.2 Empirical Determination of the Uniform Retention Probability P_{uni}

Instead of treating the uniform retention probability P_{uni} as a hyperparameter, we rigorously derive its optimal value through an empirical study on the trade-off between semantic reconstruction and toxicity retention. We randomly selected 100 clean images and performed mask-guided reverse diffusion starting directly from pure Gaussian noise, varying the uniform keep probability from 0.9

Table 11: Comparison of MUSIQ and CLIP-IQA as reference-free early-stopping estimators. Higher correlations and lower regrets indicate better performance.

		Trade-off Correlation ↑		Regret ↓		
Setting	Estimator	Pearson	Spearman	VAE	PSNR	Trade-off
VGGFace2 + AdvDM	MUSIQ	**0.518**	**0.599**	**0.130**	0.164	**0.146**
	CLIP-IQA	0.430	0.432	0.157	**0.137**	0.168
WikiArt + Mist	MUSIQ	**0.825**	**0.783**	**0.156**	0.538	**0.221**
	CLIP-IQA	0.601	0.542	0.255	**0.536**	0.297



Figure 8: Catastrophic semantic distortions caused by continuous patch and region masking.

down to 0.1. We quantitatively evaluated the perceptual distance and fidelity of the reconstructed images against the original clean images using LPIPS, SSIM. The reconstruction fidelity exhibits a predictable relationship with the keep probability. When the keep probability is high, the model achieves perfect reconstruction (Figure 9). However, the curves also reveal a critical capability of diffusion models: even when the retention probability drops to approximately 0.5, the diffusion model still demonstrates a robust capability to reconstruct the original image with highly acceptable fidelity. While a higher P_{uni} yields better reconstruction, our primary objective is adversarial purification. Retaining a higher percentage of initial pixels inevitably means preserving a larger proportion of the adversarial perturbations, which would directly undermine the physical eradication of the protection. Therefore, the principle of our design is to select the minimum possible retention probability that still guarantees semantic structural integrity. Based on our empirical observations, $P_{uni} = 0.5$ serves as this optimal inflection point. It provides just enough sparse low-frequency signals to prevent the model from deviating from the original semantic manifold, while maintaining a 50% physical destruction rate of the initial adversarial noise even before the dynamic high-frequency filtering (CDF mapping) takes effect. Consequently, P_{uni} is fixed as an empirically grounded constant in our framework rather than a tunable variable.

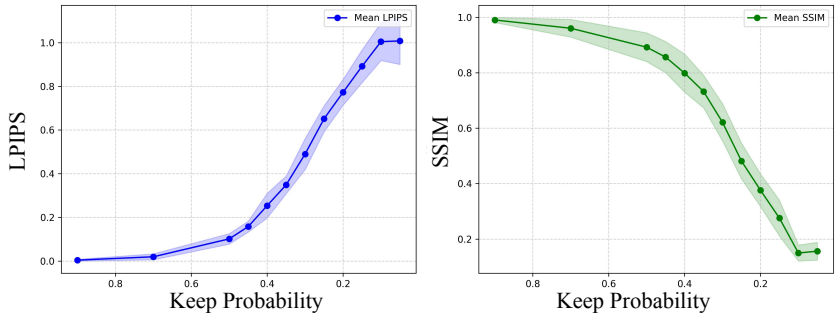


Figure 9: Empirical determination of the uniform retention probability P_{uni}. The plots illustrate the impact of varying keep probabilities on unconditional reconstruction fidelity (measured by LPIPS and SSIM). The results justify $P_{uni} = 0.5$ as the optimal inflection point, balancing robust semantic preservation with maximum adversarial noise destruction.

Table 12: High-frequency reconstruction analysis. Values for high-frequency metrics are reported as mean ± standard deviation.

Input	R_{HF} ($\rightarrow 1$)	$E_{\mathrm{HF}} \downarrow$	PSNR ↑	SSIM ↑
Adversarial	1.8313 ± 0.4541	2.4964 ± 0.8238	27.33	0.6348
DSSA (Ours)	0.6825 ± 0.0972	0.9436 ± 0.0758	**29.94**	**0.8338**

G High-Frequency Detail Preservation and Reconstruction Analysis

High-Frequency Trade-off. At large perturbation budgets, protected images contain substantial adversarial high-frequency components that are inconsistent with clean images. Although completing the denoising trajectory may reconstruct additional high-frequency content, such content does not necessarily improve fidelity. As shown in our ablation study, removing early stopping reduces PSNR from 29.94 to 27.44 and SSIM from 0.8338 to 0.6435. These results are consistent with the late-stage perturbation resurgence discussed in Section 3.2.

Quantitative High-Frequency Analysis. To quantify high-frequency preservation, we extract the high-frequency component of an image x as

$$\mathcal{H}_\sigma(x) = x - \mathcal{G}_\sigma(x), \tag{9}$$

where $\mathcal{G}_\sigma$ denotes Gaussian smoothing with standard deviation σ.

We evaluate two metrics: (1) the high-frequency magnitude retention ratio R_{HF}, and (2) the clean-referenced high-frequency reconstruction error E_{HF}:

$$R_{\mathrm{HF}} = \frac{\|\mathcal{H}_\sigma(x)\|_2}{\|\mathcal{H}_\sigma(x_{\mathrm{clean}})\|_2}, \tag{10}$$

$$E_{\mathrm{HF}} = \frac{\|\mathcal{H}_\sigma(x) - \mathcal{H}_\sigma(x_{\mathrm{clean}})\|_1}{\|\mathcal{H}_\sigma(x_{\mathrm{clean}})\|_1}. \tag{11}$$

Here, x denotes the adversarial or purified image, and x_{clean} is the corresponding clean reference. An ideal R_{HF} is 1, whereas a lower E_{HF} indicates better high-frequency reconstruction.

Results and Discussion. As shown in Table 12, adversarial inputs exhibit excessive high-frequency magnitude, with an average retention ratio of 1.8313 relative to clean images. DSSA reduces the clean-referenced high-frequency reconstruction error from 2.4964 to 0.9436, corresponding to a 62.2% relative reduction. The deviation of the mean retention ratio from the ideal value of 1 is also reduced by 61.8%.

Meanwhile, DSSA improves PSNR by 2.61 dB and SSIM by 0.1990. Although the purified images retain only 68.3% of the clean high-frequency magnitude on average, their substantially lower reconstruction error indicates that DSSA effectively suppresses incorrect adversarial components while sacrificing some genuine fine textures. These findings highlight the trade-off between high-frequency detail preservation and adversarial perturbation removal.

H Limitations

While our proposed DSSA framework achieves state-of-the-art performance in both purification efficacy and perceptual fidelity, we acknowledge certain limitations that warrant future investigation.

Inference Latency and Computational Overhead. To ensure the complete physical eradication of high-budget adversarial perturbations, DSSA deliberately initiates the reverse denoising trajectory directly from pure Gaussian noise ($t = T$). Consequently, compared to existing bounded-noise methods (e.g., DiffPure) that operate over limited noise scales, DSSA requires a step-by-step traversal of a significantly longer generation path. This full-trajectory denoising process inherently demands a higher number of Neural Function Evaluations (NFEs), inevitably introducing higher inference latency and computational overhead. While this computational trade-off is necessary to bypass the fundamental vulnerabilities of partial noising, it may constrain the immediate deployment of

DSSA in real-time systems or resource-limited edge devices. Furthermore, when employed as a preprocessing step for large-scale datasets prior to personalized fine-tuning (e.g., DreamBooth or LoRA), the per-image computational cost accumulates, leading to substantial overall processing times. Future work could explore integrating advanced acceleration techniques, such as higher-order ODE solvers (e.g., DPM-Solver) or consistency Models into our mask-guided generation framework. This would aim to significantly compress the sampling trajectory and mitigate the computational overhead without sacrificing the established purification efficacy.

Minor Compromise in High-Frequency Details. Furthermore, while the perception-guided early-stopping mechanism is highly effective at preventing late-stage perturbation resurgence, it intrinsically introduces a minor trade-off regarding extreme high-frequency details. Diffusion models typically synthesize the finest micro-structures (e.g., subtle skin pores or intricate canvas brushstrokes) during the final few denoising timesteps ($t \to 0$). By prematurely halting the trajectory, the mechanism can result in slightly over-smoothed visual outputs at the microscopic level. This compromise is noticeable in artistic style images that inherently rely on rich, high-frequency textural semantics.

I Broader Societal Impacts

As an advanced image purification technique, our proposed DSSA framework carries dual-use societal implications. On the negative side, this technology could potentially be misused by malicious **Image Exploiters** to bypass existing copyright and privacy safeguards. By effectively neutralizing protective adversarial perturbations, unauthorized users could employ our method to preprocess protected data (e.g., personal portraits or artists' original artworks) and subsequently train high-fidelity personalized models to generate deepfakes or art forgeries.

However, from a positive and defensive perspective, this research is crucial for the AI security community. By demonstrating that current noising-denoising protection paradigms can be effectively bypassed through our dynamic sparse anchoring strategy, we expose the fundamental vulnerabilities of existing defense mechanisms. We hope that these findings will alert the **Image Protector** community to the limitations of current local perturbations or simple frequency manipulations, thereby motivating the development of next-generation, robust active copyright protection technologies that can withstand unconditional generative priors.

J Additional Qualitative Results



Figure 10: Visual comparison of purified images across different methods under various adversarial perturbations.



Figure 11: Qualitative comparison of DreamBooth personalized generation results across different purification methods on the facial dataset.

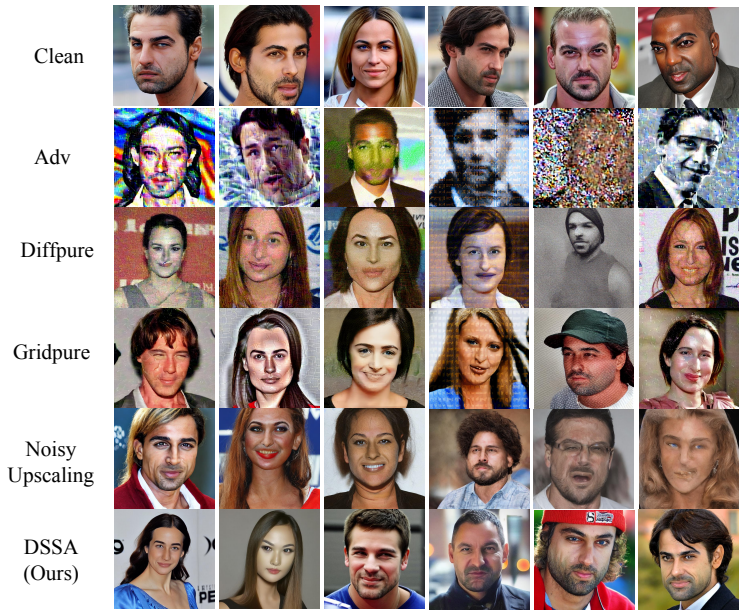


Figure 12: Qualitative comparison of LoRA personalized generation results across different purification methods on the VGGFace2 dataset. Each column displays generated samples corresponding to a specific facial identity.

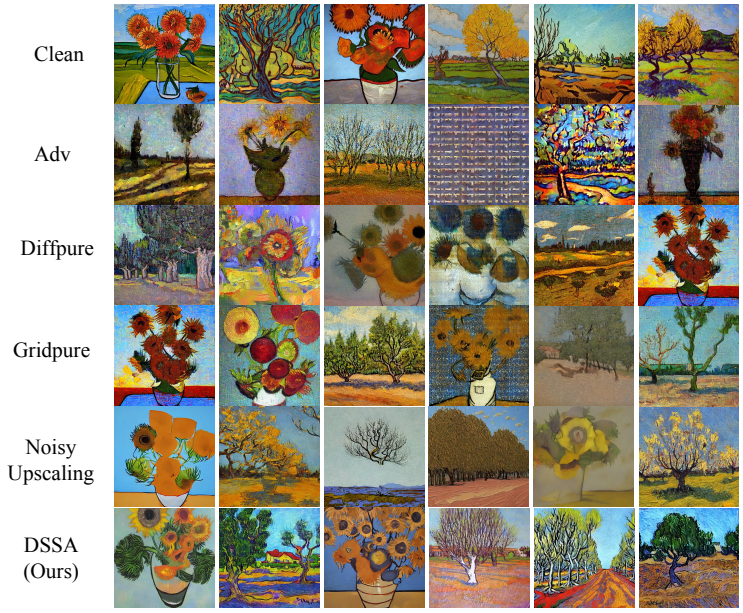


Figure 13: Qualitative comparison of LoRA personalized generation results across different purification methods on the WikiArt dataset. Each column displays generated samples corresponding to a specific artistic style. All images were generated using the prompt “a sks painting of a house”.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made, including the contributions made in the paper and scopes.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We claim the limitations of our work in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We have conducted detailed proofs and verifications for both the hypotheses and observations presented in the Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided detailed experimental descriptions and parameter specifications in the main text and appendices to facilitate the reproduction of the main experiments and the associated conclusions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will publicly release the detailed data and code once the paper has been accepted for publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have introduced the detailed experimental settings and implementation details in our main paper and Appendix A

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to the high computational cost of full-trajectory diffusion generation and downstream fine-tuning (DreamBooth/LoRA) across multiple baselines, running multiple random seeds to compute error bars is computationally prohibitive. However, the statistical significance and stability of our results are supported by the extensive scale of our evaluation. We tested across diverse domains (VGGFace2 and WikiArt) and generated hundreds of samples per setting (e.g., 640 images per identity/style for LoRA), ensuring that the reported metrics are robust, and representative without the need for multi-run averaging.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have detailed the experimental details in the Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm our results follow the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We thoroughly discuss both the positive and negative societal impacts of our proposed purification framework in the Appendix I.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper introduces a purification algorithm and an evaluation framework. We do not release any high-risk pre-trained foundation models.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We confirm all used assets are cited or credited, and the licenses are respected

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not introduce new monolithic assets such as new datasets or pre-trained foundation models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this paper does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.