MIDGUARD: FREQUENCY-AWARE ADVERSARIAL PROTECTION FOR PERSONALIZED DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have achieved remarkable success in high-quality image synthesis, while personalization further enable them to reproduce personal identities from only a few reference images. Despite enabling customized applications, this capability raises privacy concerns through unauthorized identity replication and manipulation. Existing adversarial defenses optimize diffusion-based objectives to construct imperceptible perturbations, yet they face a trade-off between disrupting unauthorized identity learning and preserving image quality. Addressing this trade-off calls for update directions that effectively disrupt identity learning while introducing minimal visual distortion. To investigate how this balance varies across spatial scales, we decompose adversarial gradients into different frequency bands and compare the resulting updates under the same perturbation budget. In our analysis, low-frequency updates suppress identity learning but cause greater distortion, whereas high-frequency updates better preserve appearance but offer weaker protection. Mid-frequency updates achieve comparable suppression to low-frequency updates with less distortion. We therefore propose **MidGuard**, a frequency-aware adversarial protection method for personalized diffusion models. It fuses Gaussian band-pass-filtered mid-frequency gradients with raw gradients to retain complementary signals, while progressively decays the contribution of mid-frequency guidance within each inner optimization loop to enable less spectrally constrained refinement. Extensive experiments demonstrate stronger overall identity protection than representative adversarial defenses while maintaining competitive visual fidelity. Code is available at `https://anonymous.4open.science/r/luck`.

1 INTRODUCTION

Diffusion models have become a foundation of modern text-to-image generation, enabling high-quality, diverse, and controllable image synthesis (Ho et al., 2020; Rombach et al., 2022; Yang et al., 2023). Building on these models, personalization techniques such as DreamBooth, Textual Inversion, LoRA, and Custom Diffusion can learn a subject from only a few reference images and reproduce its identity across diverse prompts and scenes (Ruiz et al., 2023; Gal et al., 2023; Hu et al., 2022; Kumari et al., 2023). While this capability enables personalized portrait synthesis and creative applications, it also introduces serious privacy risks. Publicly shared photos may be collected without consent and used for unauthorized identity replication or manipulation (Lee et al., 2025). Moreover, verifying whether personalized diffusion models have been trained on unauthorized data remains an important challenge (Li et al., 2025a). Studies of diffusion-model memorization have also shown that training examples can be reproduced or extracted, further exposing the risks of unauthorized data use (Carlini et al., 2023; Wen et al., 2024; Chen et al., 2024). Protecting personal images against unauthorized personalization is therefore an increasingly important problem.

In response, existing studies protect personal images by adding imperceptible perturbations before release, so that subsequent personalization fails to learn the intended identity. Early methods formulate protection as an adversarial optimization problem and construct image-specific perturbations against diffusion training (Liang et al., 2023; Van Le et al., 2023). Subsequent studies improve the effectiveness and generalization of this paradigm by exploiting the behavior of diffusion models across timesteps and feature layers (Wang et al., 2024) and by reducing the dependence of protective perturbations on specific text prompts (Wan et al., 2024). Although these methods differ in design, most improvements focus on the adversarial objective or optimization strategy. Yet under imperceptible-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tibility constraints, it remains difficult to achieve strong identity disruption without sacrificing the visual quality of protected images. Understanding how different update components affect identity protection and visual distortion may help improve this trade-off.

To further investigate this challenge, we first examine the spatial distribution of perturbation effects under conventional adversarial optimization. By tracking the optimization loss, diffusion latent shift, and region-wise latent shifts across facial and background areas, we observe that conventional optimization already induces substantial changes, with a considerable share occurring in background regions that are weakly associated with identity. This suggests that increasing perturbation strength alone may not efficiently target identity-related information. Since different frequency bands capture visual structures at different spatial scales and play distinct roles in diffusion denoising (Si et al., 2024; Yang et al., 2025), we further examine the frequency composition of adversarial gradients. Our analysis shows that mid-frequency updates achieve a more favorable balance between identity disruption and visual fidelity than low- and high-frequency updates.

We therefore propose **MidGuard**, a frequency-aware adversarial protection method for personalized diffusion models. MidGuard first applies Gaussian band-pass filtering to the raw diffusion gradient to emphasize mid-frequency components. Because emphasizing a selected frequency band may attenuate useful information from the remaining spectrum, MidGuard fuses the filtered gradient with the raw gradient to preserve complementary optimization signals. It further employs a reverse frequency schedule that gradually reduces mid-frequency guidance during each adversarial optimization loop, allowing the optimization to place greater emphasis on mid-frequency components in early updates while retaining more information from the original gradient in later steps.

This work makes three main contributions: (1) We provide a diagnostic analysis showing that stronger perturbation effects do not necessarily translate into proportionally stronger identity disruption, and further identify mid-frequency components as providing a favorable balance between identity disruption and visual fidelity. (2) Based on this insight, we propose **MidGuard**, a frequency-aware method that combines mid-frequency gradient guidance, full-spectrum gradient fusion, and stage-adaptive frequency scheduling. (3) Extensive evaluations across different diffusion models demonstrate that MidGuard achieves stronger identity protection than representative defenses while maintaining competitive visual fidelity.

2 Related Work

Personalized diffusion models adapt pretrained text-to-image models to a subject using only a few reference images, enabling subject-consistent generation across prompts and scenes (Gal et al., 2023; Ruiz et al., 2023; Hu et al., 2022; Kumari et al., 2023). This capability also introduces a serious privacy risk by facilitating unauthorized identity replication (Li et al., 2025a; Lee et al., 2025). Related proactive defenses therefore add imperceptible perturbations before image release to disrupt malicious editing or style imitation (Salman et al., 2023; Shan et al., 2023), and to prevent subject-driven personalization from learning the protected identity or concept (Liang et al., 2023; Van Le et al., 2023; Wang et al., 2024; Wan et al., 2024).

Early adversarial protections establish the use of image-specific perturbations to interfere with diffusion-model learning. MIST (Liang & Wu, 2023) employs diffusion-oriented adversarial objectives to prevent models from learning protected visual content, while Anti-DreamBooth (Van Le et al., 2023) extends this paradigm to subject-driven personalization through surrogate DreamBooth training. Subsequent work explores several complementary directions, including improving transferability and robustness through surrogate adaptation and transformation sampling (Liu et al., 2024b), selecting more effective timesteps and feature representations (Wang et al., 2024), and enhancing prompt generalization (Wan et al., 2024). Other studies redesign the optimization target through token-level attention manipulation, targeted objectives, or subject–class attention disruption (Liu et al., 2024a; Zheng et al., 2025; Xu et al., 2024). Frequency-domain perturbation learning has also been explored to jointly improve protection effectiveness and visual stealthiness (Yang et al., 2024).

Beyond improving adversarial perturbation optimization, recent studies also explore alternative protection settings and intervention mechanisms outside this standard formulation. DiffusionGuard focuses on diffusion-based image editing and improves resistance to masked editing through early-stage objectives and mask augmentation (Choi et al., 2025). To avoid costly per-image iterative

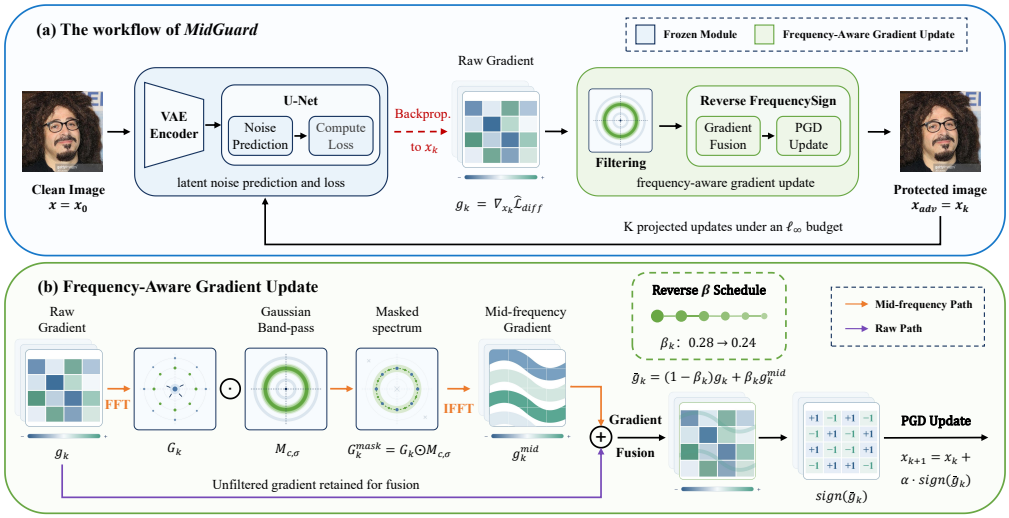


Figure 1: Overview of MidGuard. (a) A surrogate diffusion model, frozen during image updates, computes the raw gradient g_k at each optimization step, and the protected image x_{adv} is obtained through K projected steps under an ℓ_∞ perturbation budget. (b) A Gaussian band-pass mask extracts the mid-frequency gradient g_k^{mid}, which is fused with the raw gradient using β_k. A reverse schedule gradually reduces frequency guidance within each inner PGD loop before the signed PGD update.

optimization, DiffVax replaces repeated optimization with a learned immunization model for more efficient protection against diffusion-based editing in practice (Ozden et al., 2026). At a different intervention level, APDM protects against unauthorized personalization by modifying the diffusion model itself rather than the released images (Lee et al., 2025).

Frequency information has received some attention in adversarial robustness and diffusion-model research, revealing that different frequency ranges can exhibit distinct robustness properties and denoising behaviors (Yin et al., 2019; Guo et al., 2020; Benita et al., 2025). Frequency characteristics have also been introduced into anti-customization, for example through frequency-aware timestep selection or frequency-domain perturbation learning (Wang et al., 2024; Yang et al., 2024). These studies establish the importance of spectral information, but the frequency structure of diffusion-loss adversarial gradients themselves remains comparatively underexplored. In particular, how different gradient frequency components affect the balance between identity disruption and visual fidelity during adversarial optimization is still not well understood. MidGuard addresses this gap by directly reorganizing adversarial gradients in the frequency domain and introducing frequency-aware gradient guidance for personalized diffusion protection.

3 Method

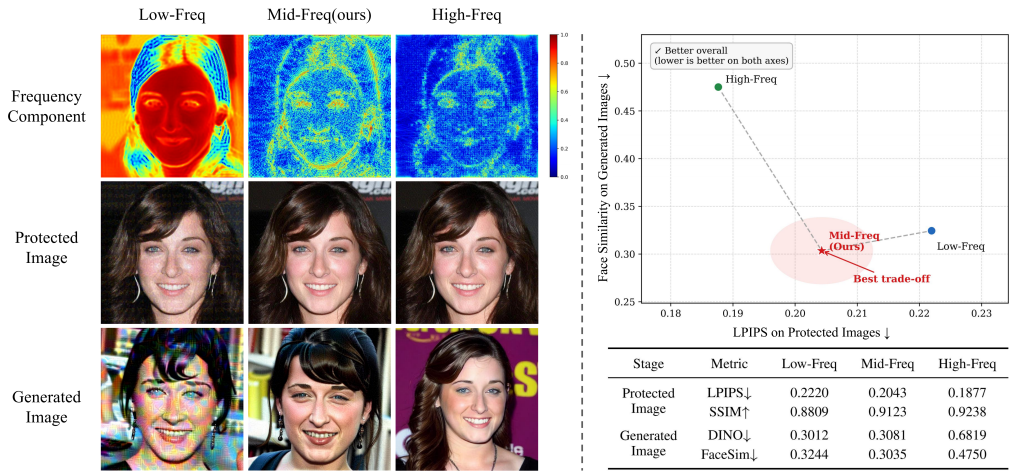
Given a clean reference image $x_0 \in [-1, 1]^{H \times W \times C}$, adversarial protection aims to construct a protected image $x_{\text{adv}} = x_0 + \delta$ with $\|\delta\|_\infty \leq \eta$. The perturbation is optimized to maximize the diffusion training loss $\mathcal{L}_{\text{diff}}(x_0 + \delta; \theta_s)$ of a surrogate model θ_s, thereby preventing subsequent personalization from learning the target identity while preserving the visual appearance of the reference image. Instead of directly updating the protected image using the raw diffusion gradient, MidGuard reorganizes the gradient in the frequency domain. It emphasizes mid-frequency components through Gaussian band-pass filtering, retains complementary information through raw-gradient fusion, and adjusts the contribution of frequency guidance using a reverse frequency schedule. Figure 1 presents the overall workflow of MidGuard.

3.1 Frequency Analysis

Adversarial protection aims to disrupt identity learning while preserving the visual quality of reference images. Our diagnostic analysis shows that conventional optimization already produces substantial

latent shifts, but a considerable share of these changes occurs in background regions that are weakly associated with identity, as detailed in Appendix A.1. However, spatial localization alone does not reveal which gradient components carry identity-relevant information, because updates combine signals across multiple scales. We examine the gradient spectrum to compare how components at different spatial scales affect identity disruption and visual distortion. Since different frequency bands capture visual structures at different spatial scales and interact differently with diffusion denoising (Si et al., 2024; Yang et al., 2025; Benita et al., 2025; Amerehi & Healy, 2025; Wang et al., 2026), their relative contributions may affect the balance between protection and visual fidelity. Although frequency-selective perturbations and frequency-domain masks have exhibited distinct adversarial properties (Yin et al., 2019; Guo et al., 2020; Li et al., 2025b), the most effective frequency range for protecting personalized images remains unclear.

To identify the most suitable frequency range, we isolate low-, mid-, and high-frequency components of the adversarial gradient and use each of them separately to construct protected images under the same perturbation budget. The Gaussian masks use centers $c = 0.05$, 0.45, and 0.75, respectively, with a shared bandwidth $\sigma = 0.18$. All other optimization settings are kept unchanged across all three frequency settings to ensure a fair comparison. The resulting protected images are then used for downstream personalized diffusion training under identical conditions, allowing us to evaluate both protected image fidelity and identity suppression after personalization.



Stage	Metric	Low-Freq	Mid-Freq	High-Freq
Protected Image	LPIPS↓	0.2220	0.2043	0.1877
	SSIM↑	0.8809	0.9123	0.9238
Generated Image	DINO↓	0.3012	0.3081	0.6819
	FaceSim↓	0.3244	0.3035	0.4750

(a) Qualitative comparison of different frequency perturbations

(b) Protection-fidelity trade-off

Figure 2: Frequency analysis of perturbations. (a) Qualitative effects of perturbations across frequency ranges on protected images and personalized generation. (b) Quantitative protection–fidelity trade-off, with LPIPS and face similarity visualized in the scatter plot and complementary metrics reported below. Mid-frequency-guided updates achieve the most favorable balance.

As shown in Figure 2(a), low-frequency perturbations provide stronger identity suppression than high-frequency perturbations, but at the cost of noticeable visual degradation in the protected images. High-frequency perturbations better preserve the original appearance, yet yield substantially weaker identity suppression after personalization. In contrast, mid-frequency perturbations achieve identity suppression comparable to that of low-frequency perturbations while introducing considerably less visual distortion. The quantitative results in Figure 2(b) further show that the mid-frequency range offers a more favorable balance between protected-image fidelity and identity suppression. Based on this observation, we emphasize mid-frequency components in the subsequent adversarial optimization. Appendix A.2 derives a sufficient condition for a local mid-frequency advantage at equal visual cost.

3.2 Gradient Filtering and Fusion

The preceding analysis identifies the mid-frequency range as a favorable region for balancing identity disruption and visual fidelity. Nevertheless, frequency selectivity alone does not guarantee an effective adversarial update. The raw diffusion gradient contains optimization signals distributed across the entire frequency spectrum, and completely removing components outside the selected range may

discard information that remains useful for maximizing the diffusion loss. MidGuard therefore first applies smooth frequency filtering to the raw gradient to extract its mid-frequency component. It then fuses the filtered component with the raw gradient to construct the fused adversarial gradient while retaining information from the remaining spectrum.

At the k-th optimization step, let x_k denote the current protected image. During the image-update stage, the parameters of the surrogate diffusion model θ_s remain fixed, and only x_k is updated. Following the standard diffusion training objective (Ho et al., 2020; Rombach et al., 2022), we randomly sample a timestep t and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, yielding the noisy latent $z_{k,t} = \sqrt{\bar{\alpha}_t}\mathcal{E}(x_k) + \sqrt{1-\bar{\alpha}_t}\epsilon$. Here, $\mathcal{E}(\cdot)$ denotes the pretrained VAE encoder, including its latent scaling factor, and $\bar{\alpha}_t$ is determined by the diffusion noise schedule. The diffusion training loss is then

$$\mathcal{L}_{\text{diff}}(x_k; \theta_s) = \mathbb{E}_{t,\epsilon}\left[\|\epsilon - \epsilon_{\theta_s}(z_{k,t}, t, p)\|_2^2\right], \tag{1}$$

where ϵ_{θ_s} is the predicted noise, p is the instance prompt containing a subject identifier and class noun, and $\widehat{\mathcal{L}}_{\text{diff}}(x_k; \theta_s, t, \epsilon)$ denotes the integrand evaluated at the sampled t and ϵ. Backpropagating this loss yields the raw stochastic gradient $g_k = \nabla_{x_k}\widehat{\mathcal{L}}_{\text{diff}}(x_k; \theta_s, t, \epsilon)$. Because g_k aggregates responses from the VAE encoder and denoising network across the spatial representation, it contains components distributed throughout the frequency spectrum.

To control the contribution of these components, we transform each channel of g_k into the frequency domain. Frequency-domain representations provide a convenient way to isolate and modulate diffusion-related signals (Li et al., 2025b; Wang et al., 2026). Specifically, we compute the shifted spectrum $G_k = \text{fftshift}(\mathcal{F}(g_k))$, where $\mathcal{F}(\cdot)$ denotes the two-dimensional Fourier transform. The spectrum shift moves the zero-frequency component to the center, allowing frequency ranges to be characterized by their radial distance. We define the normalized radial frequency as $D(u, v) = \sqrt{u^2 + v^2}/\sqrt{2}$, where $(u, v) \in [-1, 1]^2$ are uniformly sampled coordinates on the shifted Fourier grid. Accordingly, values near zero represent low-frequency components around the spectrum center, whereas larger values represent progressively higher frequencies. Based on this radial representation, we construct a Gaussian band-pass mask

$$M_{c,\sigma}(u, v) = \exp\left(-\frac{(D(u, v) - c)^2}{2\sigma^2}\right), \tag{2}$$

where c specifies the center of the emphasized frequency range and σ controls its bandwidth. Frequency components close to c receive weights near one, while components farther from the selected range are progressively attenuated. Because the mask depends only on the radial distance $D(u, v)$, it applies consistent frequency weighting across orientations without introducing a directional preference. The Gaussian formulation avoids abrupt changes at frequency boundaries and provides smooth weighting across neighboring components rather than imposing a hard frequency cutoff. Unless otherwise specified, we use $c = 0.45$ and $\sigma = 0.18$ in the main experiments.

We apply the mask to the shifted spectrum and transform the result back to the spatial domain:

$$g_k^{\text{mid}} = \text{Re}\left[\mathcal{F}^{-1}\left(\text{ifftshift}\left(M_{c,\sigma} \odot G_k\right)\right)\right]. \tag{3}$$

Here, $\odot$ denotes element-wise multiplication and $\text{Re}(\cdot)$ extracts the real component of the inverse transform. The resulting gradient g_k^{mid} retains components near the selected mid-frequency range while suppressing components farther away. The same radial mask is applied to each gradient channel, enabling frequency modulation without requiring an additional spatial identity mask.

Although g_k^{mid} concentrates the gradient around the selected frequency range, using it alone would attenuate complementary information from the remaining spectrum. We therefore fuse it with the raw gradient:

$$\bar{g}_k = (1 - \beta_k)g_k + \beta_k g_k^{\text{mid}}, \qquad 0 \le \beta_k \le 1. \tag{4}$$

The raw-gradient component preserves full-spectrum optimization signals, while the filtered component favors the selected mid-frequency range. Let $\bar{G}_k$ denote the shifted Fourier spectrum of $\bar{g}_k$, and $\widetilde{M}_{c,\sigma}$ the effective mask after taking the real part (Appendix A.3). The frequency-domain fusion is $\bar{G}_k(u, v) = [1 - \beta_k + \beta_k\widetilde{M}_{c,\sigma}(u, v)]G_k(u, v)$. Near the selected range, $\widetilde{M}_{c,\sigma}(u, v)$ approaches one, largely retaining the original gradient strength. Components outside this range receive smaller but nonzero weights through the raw-gradient residual. The fused gradient therefore emphasizes the mid-frequency range without completely discarding complementary optimization signals from the rest of the spectrum.

3.3 Reverse Frequency Schedule

The fusion coefficient β_k controls how strongly the update is guided toward the selected mid-frequency range. A larger value places more emphasis on g_k^{mid} and more strongly attenuates components outside the passband, whereas a smaller value preserves a larger contribution from the raw gradient g_k. A fixed coefficient imposes the same spectral preference at every optimization step, whereas our schedule varies this preference within each inner loop. Early steps emphasize the selected mid-frequency components to guide adversarial updates under the surrogate obtained after model adaptation. Later steps gradually upweight complementary components from the raw gradient to allow less frequency-selective refinement of the update direction. This design coordinates stronger initial guidance with subsequent relaxation within each loop, without assuming that restarting the schedule moves the image away from the projection boundary.

To coordinate these two stages, we introduce a reverse frequency schedule that gradually relaxes frequency guidance as the optimization progresses within each PGD loop. For a loop containing K adversarial update steps, the fusion coefficient is defined as

$$\beta_k = \beta_{\text{start}} + \frac{k}{K-1}\left(\beta_{\text{end}} - \beta_{\text{start}}\right), \qquad k = 0, \ldots, K-1, \tag{5}$$

where $\beta_{\text{start}} \geq \beta_{\text{end}}$ and $K > 1$. When $K = 1$, we set $\beta_0 = \beta_{\text{start}}$. We use $\beta_{\text{start}} = 0.28$ and $\beta_{\text{end}} = 0.24$. After each surrogate adaptation, the schedule restarts and decays within the next inner PGD loop, rather than globally across all TK updates. This renews mid-frequency guidance under the changed diffusion-loss landscape before increasing raw-gradient contributions for refinement. Here, "early" and "later" refer to steps within the same inner loop.

The effect of this schedule can be interpreted directly through the equivalent spectral weighting of the fused gradient. At frequency coordinate (u, v), the raw spectrum is multiplied by

$$W_k(u, v) = 1 - \beta_k + \beta_k \widetilde{M}_{c,\sigma}(u, v) = 1 - \beta_k \left[1 - \widetilde{M}_{c,\sigma}(u, v)\right]. \tag{6}$$

For frequencies near the passband center, $\widetilde{M}_{c,\sigma}(u, v) \approx 1$, and hence $W_k(u, v) \approx 1$ throughout optimization. For frequencies outside the passband, $\widetilde{M}_{c,\sigma}(u, v) < 1$, so decreasing β_k gradually increases their weights toward one. The reverse schedule therefore preserves the selected mid-frequency components while progressively upweighting complementary components from the raw gradient. This produces a smooth transition from frequency-guided direction establishment to full-spectrum refinement rather than switching abruptly between two update rules.

Finally, after obtaining the stage-adaptive fused gradient, we convert it into an ℓ_∞-compatible update direction and perform projected gradient ascent (Madry et al., 2017):

$$x_{k+1} = \Pi_{\mathcal{B}_\infty(x_0,\eta)\cap[-1,1]}\left(x_k + \alpha \ \text{sign}(\bar{g}_k)\right), \tag{7}$$

where α denotes the step size, and the projection constrains the updated image to both the valid normalized image range and the ℓ_∞ ball of radius η centered at the clean image x_0. This projection ensures that every intermediate image satisfies the same perturbation constraint. Although the sign operator fixes the element-wise update magnitude, frequency filtering and fusion reshape the spatial update direction before the sign operation. Consequently, MidGuard gradually shifts the balance from stronger mid-frequency guidance toward a greater contribution from the raw gradient during PGD optimization, while keeping the perturbation bound η fixed.

4 Experiments

4.1 Experimental Setup

Implementation Details. We conduct experiments on Stable Diffusion v1.5 (SD1.5), Stable Diffusion v2.1 (SD2.1) (Rombach et al., 2022), and Stable Diffusion XL 1.0 (SDXL) (Podell et al., 2023). Experiments are conducted on CelebA-HQ (Karras et al., 2018) and VGGFace2 (Cao et al., 2018), with 50 identities selected from each dataset and four reference images used for each identity. For SD1.5 and SD2.1, we fine-tune the UNet and text encoder with DreamBooth (Ruiz et al., 2023) at 512×512 resolution for 1,000 steps using a batch size of 1 and a learning rate of 5×10^{-7}.

Table 1: Comparison with adversarial protection methods after DreamBooth personalization using SD1.5. Results are reported on CelebA-HQ and VGGFace2 under the prompt "a photo of sks person"; arrows indicate the preferred metric directions.

Method	CelebA-HQ					VGGFace2				
	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓
Clean	0.8077	0.8008	0.5875	147.5516	0.6190	0.7942	0.7477	0.6745	194.7966	0.6055
MIST	0.6181	0.5639	0.6411	255.9407	0.3970	0.4048	0.2412	0.7648	429.9999	0.2774
Anti-DreamBooth	0.4591	0.3314	0.7011	354.9742	0.3712	0.2626	0.0884	0.8520	514.3143	0.2323
SimAC	0.4031	0.3142	0.7344	381.1371	0.3410	0.3083	0.0947	0.8438	488.0776	0.4175
PAP	0.5040	0.3720	0.7508	365.3730	0.4048	0.2740	0.0910	0.8260	513.5296	0.2570
DiffusionGuard	0.6655	0.7116	0.6400	199.0947	0.4344	0.6761	0.6882	0.7154	225.2115	0.4591
DiffVax	0.6273	0.6650	0.6457	210.0161	0.5594	0.7210	0.7003	0.6916	213.5720	0.5850
MidGuard (Ours)	**0.3661**	**0.2951**	**0.7968**	**386.7291**	**0.3110**	**0.2603**	**0.0770**	**0.8724**	**515.0376**	**0.2238**

For SDXL, we train a rank-16 DreamBooth LoRA (Hu et al., 2022) on the UNet at 1024×1024 resolution for 1,000 steps using a learning rate of 10^{-4}. All settings use 200 class images and prior preservation with weight 1.0. MidGuard uses 50 outer iterations with an ℓ_∞ budget of $\eta = 5 \times 10^{-2}$ and a PGD step size of $\alpha = 5 \times 10^{-3}$. We set $c = 0.45$, $\sigma = 0.18$, $\beta_{\text{start}} = 0.28$, and $\beta_{\text{end}} = 0.24$. All experiments are implemented in PyTorch and run on NVIDIA RTX 4090 GPUs.

Comparison Methods. We use the unprotected clean images as the no-defense baseline and compare MidGuard against MIST (Liang & Wu, 2023), Anti-DreamBooth (Van Le et al., 2023), SimAC (Wang et al., 2024), PAP (Wan et al., 2024), DiffusionGuard (Choi et al., 2025), and DiffVax (Ozden et al., 2026). All baseline methods are implemented using their official publicly available codebases and the best-performing configurations released in their respective repositories. Every method is evaluated under the same maximum ℓ_∞ perturbation budget after accounting for image normalization. The downstream personalization and generation settings are kept identical across methods for each diffusion backbone to ensure a fair comparison.

Evaluation. After personalization, we generate 64 images per identity using the prompt "a photo of sks person", where "sks" is the identity token associated with the reference identity. For each identity, all methods are evaluated using the same fixed set of 64 random seeds and the corresponding initial latents. Protected-image fidelity is evaluated using LPIPS (Zhang et al., 2018) (↓), SSIM (Wang et al., 2004) (↑), FSIM (Zhang et al., 2011) (↑), CLIP-I (Radford et al., 2021) (↑), and DINO (Caron et al., 2021) (↑). Identity protection after personalization is evaluated using CLIP-I (↓), DINO (↓), LPIPS (↑), FID (Heusel et al., 2017) (↑), and ArcFace similarity (Deng et al., 2019) (↓). Detailed metric computation and aggregation are provided in Appendix B.2.

4.2 Main Results

Protection Effectiveness. Table 1 presents the protection results on CelebA-HQ and VGGFace2 using SD1.5. MidGuard consistently outperforms the compared defenses across both datasets. Importantly, the reduction in face similarity is accompanied by improvements in the complementary evaluation metrics, indicating that the observed advantage is not limited to a single evaluator. Figure 3 provides corresponding visual evidence: compared with other methods, MidGuard produces personalized outputs with weaker resemblance to the reference identities. Together, these results support the effectiveness of MidGuard in interfering with identity learning during personalization.

Image Fidelity. Our results show that stronger identity protection can be achieved without increasing distortion in the released images. The SD1.5 fidelity results in Table 2 show that MidGuard preserves the reference images in terms of visual quality, achieving the best numerical results across all five criteria on both CelebA-HQ and VGGFace2. Viewed together with Table 1, these results indicate that the protection gains are not accompanied by a decline in protected-image fidelity. MidGuard offers a favorable balance between disrupting downstream identity learning and preserving the visual content of the images intended for release. Additional quantitative and qualitative results, including protection and fidelity evaluations on SD2.1 and SDXL, are provided in Appendix C.

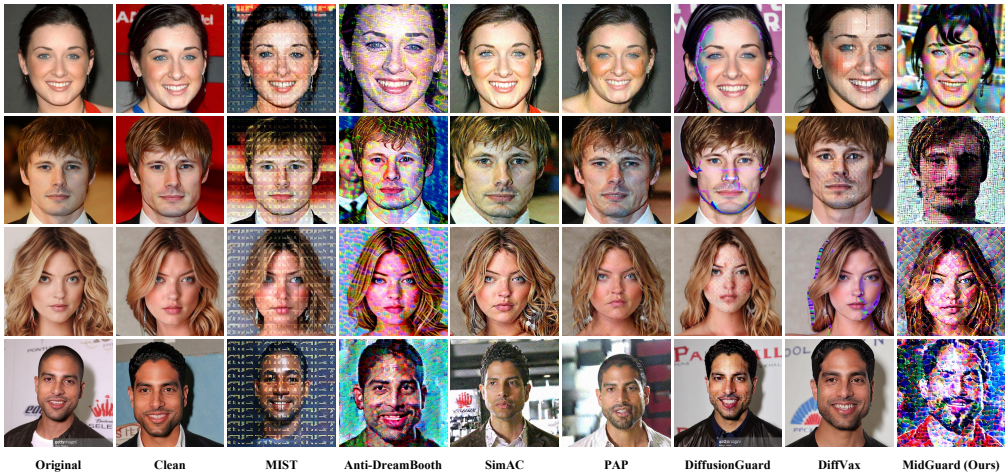


Figure 3: Qualitative comparison of personalized generation on CelebA-HQ and VGGFace2 under the prompt "a photo of sks person". "Original" shows the target identity reference, "Clean" denotes personalization without protection, and the remaining columns show outputs obtained after training on images protected by the corresponding methods.

Table 2: Fidelity of images protected by different methods on CelebA-HQ and VGGFace2 using SD1.5; arrows indicate the preferred metric directions.

Method	CelebA-HQ					VGGFace2				
	LPIPS↓	SSIM↑	FSIM↑	CLIP-I↑	DINO↑	LPIPS↓	SSIM↑	FSIM↑	CLIP-I↑	DINO↑
MIST	0.4109	0.8079	0.9176	0.8083	0.8458	0.4485	0.7609	0.9146	0.8703	0.8560
Anti-DreamBooth	0.1766	0.9300	0.9718	0.8449	0.9411	0.2102	0.9142	0.9713	0.9230	0.9456
SimAC	0.2096	0.9009	0.9634	0.8729	0.9126	0.2236	0.8959	0.9613	0.9064	0.9209
PAP	0.1794	0.9225	0.9720	0.8551	0.9411	0.2141	0.9170	0.9712	0.9166	0.9451
DiffusionGuard	0.1660	0.9294	0.9714	0.8757	0.9435	0.2109	0.9493	0.9729	0.9208	0.9476
DiffVax	0.1728	0.8861	0.9322	0.7501	0.8695	0.2172	0.9205	0.9751	0.9006	0.9460
MidGuard (Ours)	**0.1642**	**0.9305**	**0.9736**	**0.8821**	**0.9448**	**0.2093**	**0.9539**	**0.9796**	**0.9265**	**0.9479**

4.3 Robustness

We evaluate whether the protection remains effective after protected images undergo JPEG compression, resizing, Gaussian blur, center cropping, median filtering, or combined JPEG compression and resizing before personalization. Figure 4 presents the SD1.5 results. Across these common image transformations, MidGuard generally preserves its protection effectiveness and exhibits favorable robustness relative to the compared defenses. Notably, this advantage is not confined to a particular type of preprocessing, but remains evident across compression-, resampling-, and filtering-based transformations. For more challenging operations such as center cropping or combined transformations, the relative advantage of MidGuard becomes smaller in some cases, while the method remains competitive overall. These results suggest that the proposed protection is not limited to unprocessed images and remains effective under a range of common image manipulations before personalization. Additional robustness results under SD2.1 are provided in Appendix D.1, while evaluation under LoRA-based personalization is reported in Appendix D.2.

4.4 Ablation Study

We examine whether mid-frequency filtering is sufficient and how raw-gradient fusion and scheduling contribute to protection. Table 3 progressively introduces the components of MidGuard. Using only the filtered gradient weakens protection across all reported metrics relative to Raw PGD, indicating

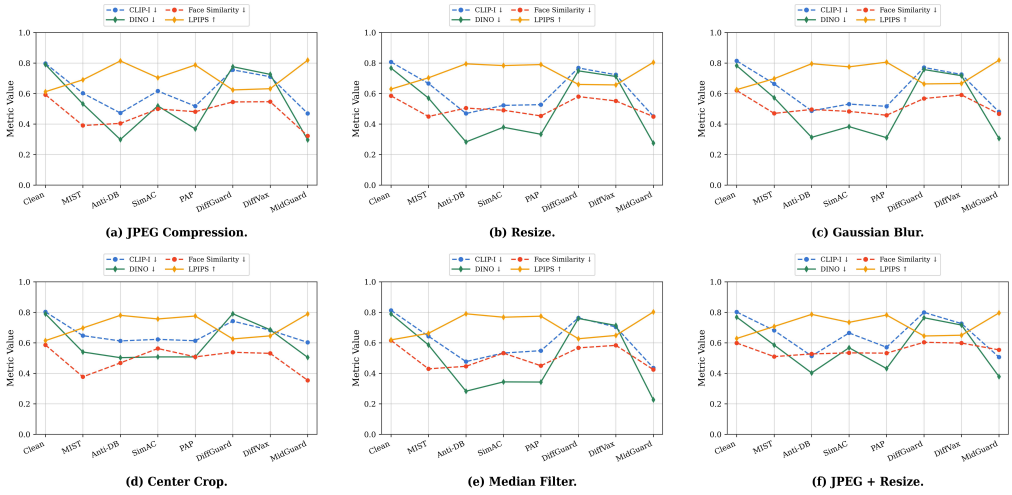


Figure 4: Identity protection after image preprocessing using SD1.5. Protected images undergo each operation before personalization. Panels compare JPEG compression, resizing, Gaussian blur, center cropping, median filtering, and combined JPEG compression and resizing.

that selecting a favorable frequency band does not by itself guarantee an effective adversarial update. Fusing the filtered and raw gradients recovers much of the lost performance, although it does not consistently outperform Raw PGD across all metrics. Introducing the reverse frequency schedule further improves all protection metrics over fixed fusion, showing the benefit of adapting the strength of frequency guidance during optimization. The complete MidGuard method therefore achieves the strongest protection among the evaluated configurations.

Beyond the component ablation, we examine the sensitivity of MidGuard to the center frequency, bandwidth, frequency-guidance strength, and scheduling strategy in Appendix E. With the other mask parameter fixed at its main setting, separate sweeps favor $c = 0.45$ and $\sigma = 0.18$, jointly supporting the configuration used in the main experiments. A moderate guidance strength better balances the filtered and raw gradients, while comparisons across scheduling strategies show that the reverse schedule outperforms fixed, increasing, and abruptly changing frequency guidance.

Table 3: Component ablation of MidGuard. Filtering, Fusion, and Schedule denote mid-frequency filtering, raw-gradient fusion, and the reverse frequency schedule, respectively. The configuration without these components corresponds to Raw PGD.

Filtering	Fusion	Schedule	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓
–	–	–	0.3481	0.1424	0.8309	476.70	0.3277
✓	–	–	0.4150	0.2061	0.8148	459.92	0.3671
✓	✓	–	0.3588	0.1451	0.8385	479.05	0.3226
✓	✓	✓	**0.3420**	**0.1377**	**0.8480**	**486.74**	**0.3154**

5 Conclusion

This paper introduces MidGuard, a frequency-aware adversarial protection method for personalized diffusion models. Our analysis showed that conventional optimization suffers from ineffective perturbation allocation and that mid-frequency components provide a favorable balance between identity disruption and visual fidelity. Based on this finding, MidGuard combines Gaussian band-pass filtering, raw-gradient fusion, and a reverse frequency schedule to guide adversarial optimization. Extensive experiments across datasets and diffusion backbones demonstrated stronger identity protection than representative defenses while maintaining competitive image quality, with ablation studies further validating the importance of frequency selection, guidance strength, and scheduling.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, translation, and presentation refinement; literature retrieval and verification; parts of method ideation and refinement; providing suggestions for developing and writing the local theoretical analysis and its derivation; and data analysis and consistency checking. The authors independently reviewed all AI-assisted suggestions and verified the experimental results, supporting data, cited references, and scientific claims.

ETHICS STATEMENT

This work is intended to protect individuals against unauthorized identity personalization and manipulation. Our experiments use publicly available face datasets solely for research purposes, and we do not collect or release new personal data. Although MidGuard is designed as a defensive privacy mechanism, adversarial perturbations could potentially be misused to disrupt authorized model training or benign data pipelines. We therefore emphasize that the method should be applied only to images for which the user has appropriate rights and solely for consent-aware privacy protection. Researchers and practitioners remain responsible for complying with applicable dataset licenses, privacy requirements, and relevant laws.

REPRODUCIBILITY STATEMENT

We provide the complete formulation of MidGuard and its optimization procedure in Section 3, and report the datasets, comparison methods, training configurations, and evaluation protocol in Section 4.1. Appendix B.2 specifies the computation and aggregation of all evaluation metrics. The main ablation study and additional analyses in Appendix E document the contribution of each module and the selection of the frequency guidance strength, schedule, center frequency, and bandwidth. Additional diagnostic, robustness, LoRA-based personalization, and cross-backbone results are also included in the appendix. An anonymous implementation is available at the URL provided in the abstract, including the core MidGuard implementation and a verified SD1.5 example configuration.

REFERENCES

Fatemeh Amerehi and Patrick Healy. Defending against frequency-based attacks with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3562–3572, 2025.

Roi Benita, Miki Elad, and Joseph Keshet. Spectral analysis of diffusion models with application to schedule design. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67–74. IEEE, 2018. doi: 10.1109/FG.2018.00020. URL https://github.com/ox-vgg/vgg_face2.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium*, pp. 5253–5270, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chen Chen, Daochang Liu, and Chang Xu. Towards memorization-free diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8425–8434, 2024.

William June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. DiffusionGuard: A robust defense against malicious diffusion-based image editing. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/43b77cef2a83a25aa27d3271d209e4fd-Abstract-Conference.html.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. Low frequency adversarial perturbation. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 1127–1137, 2020.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023.

Tae-Young Lee, Juwon Seo, Jong Hwan Ko, and Gyeong-Moon Park. Perturb a model, not an image: Towards robust privacy protection via anti-personalized diffusion models. *Advances in Neural Information Processing Systems*, 38:92410–92442, 2025.

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. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 2564–2582. IEEE, 2025a.

Zelin Li, Ruohan Zong, Yifan Liu, Ruichen Yao, Yaokun Liu, Yang Zhang, and Dong Wang. Anti-tamper protection for unauthorized individual image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15501–15510, 2025b.

Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*, 2023. URL https://arxiv.org/abs/2305.12683.

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. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20763–20786. PMLR, 2023. URL https://proceedings.mlr.press/v202/liang23g.html.

Yisu Liu, Jinyang An, Wangqian Zhang, Dayan Wu, Jingzi Gu, Zheng Lin, and Weiping Wang. Disrupting diffusion: Token-level attention erasure attack against diffusion-based customization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3587–3596, 2024a. doi: 10.1145/3664647.3681243.

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

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

Tarik Can Ozden, Ozgur Kara, Oguzhan Akcin, Kerem Zaman, Shashank Srivastava, Sandeep P. Chinchali, and James M. Rehg. DiffVax: Optimization-free image immunization against diffusion-based editing. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2411.17957.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. URL https://arxiv.org/abs/2307.01952.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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*, pp. 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29894–29918, 2023.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. In *32nd USENIX Security Symposium*, pp. 2187–2204, 2023.

Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. FreeU: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4733–4743, 2024.

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

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

Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, and Qidong Huang. SimAC: A simple anti-customization method for protecting face privacy against text-to-image synthesis of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12047–12056, 2024. URL https://arxiv.org/abs/2312.07865.

Qisen Wang, Yifan Zhao, and Jia Li. Beyond heat dissipation: Optimizing diffusion models in frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(7): 7534–7549, 2026. doi: 10.1109/TPAMI.2026.3666860.

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

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

Xide Xu, Muhammad Atif Butt, Sandesh Kamath, and Bogdan Raducanu. Privacy protection in personalized diffusion models via targeted cross-attention adversarial attack. *arXiv preprint arXiv:2411.16437*, 2024.

Haosen Yang, Adrian Bulat, Isma Hadji, Hai X Pham, Xiatian Zhu, Georgios Tzimiropoulos, and Brais Martinez. Fam diffusion: Frequency and attention modulation for high-resolution image generation with stable diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2459–2468, 2025.

Jing Yang, Runping Xi, Yingxin Lai, Xun Lin, and Zitong Yu. DDAP: Dual-domain anti-personalization against text-to-image diffusion models. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10. IEEE, 2024. doi: 10.1109/IJCB62174.2024.10744490.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. doi: 10.1109/TIP.2011.2109730.

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

Boyang Zheng, Chumeng Liang, and Xiaoyu Wu. Targeted attack improves protection against unauthorized diffusion customization. In *International Conference on Learning Representations*, 2025.

A Optimization Diagnostics and Local Frequency Analysis

A.1 Optimization Process and Spatial Perturbation Allocation

To examine the behavior of the baseline adversarial optimization, we record several diagnostic quantities at every outer iteration, including the diffusion optimization loss, the mean absolute shift from the clean image in the VAE latent space, the area-normalized latent shifts within the face and background regions, and the proportion of the regional shift assigned to the face region. The face region is determined by the facial mask, while its complement is treated as the background region. The mask is used only for this diagnostic measurement and does not participate in adversarial optimization.

As shown in Figure 5, the absolute latent shift increases rapidly and then remains at a consistently high level, indicating that the optimization already produces a substantial change in the diffusion latent space. However, the area-normalized background shift remains larger than the face-region shift throughout optimization, and the face-region share stabilizes below the equal-allocation reference. These observations suggest that the primary limitation is not insufficient perturbation magnitude, but the ineffective allocation of the perturbation budget to regions weakly associated with identity. This finding further motivates the frequency-aware optimization adopted by MidGuard.

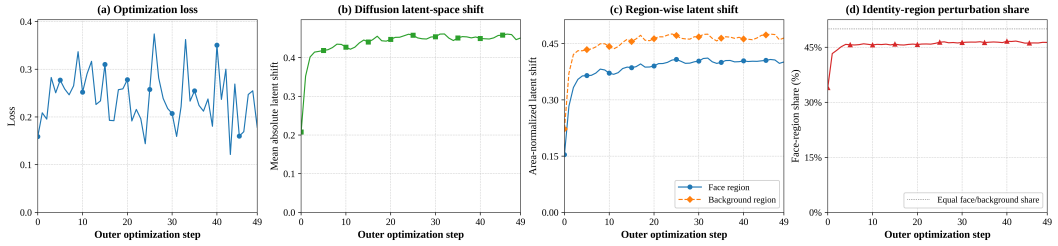


Figure 5: Analysis of the baseline adversarial optimization process and perturbation allocation across the outer iterations. We report (a) the diffusion optimization loss, (b) the mean absolute shift in the VAE latent space, (c) the area-normalized latent shifts in the face and background regions, and (d) the proportion of the regional shift assigned to the face region. The dotted line in (d) denotes equal allocation between the face and background regions.

A.2 Local Analysis of Frequency-Selective Directions

We derive a sufficient condition under which a mid-frequency direction offers a better local trade-off between increasing the surrogate diffusion loss and preserving visual features. The analysis compares frequency-selective directions without assuming that mid frequencies are universally optimal.

Local objective and visual cost. Fix the surrogate parameters and all sampled random variables, and write the sampled diffusion loss as $\mathcal{L}(x)$ with $g = \nabla_x \mathcal{L}(x)$, treating the image as a vector. Assume that $\mathcal{L}$ has an L-Lipschitz gradient in a neighborhood of x. For perturbations whose connecting segments remain in this neighborhood,

$$\mathcal{L}(x+\delta) - \mathcal{L}(x) = g^\top \delta + r(\delta), \qquad |r(\delta)| \le \tfrac{L}{2}\|\delta\|_2^2. \tag{8}$$

For a differentiable visual feature mapping ϕ with Jacobian J_ϕ at x, define

$$C_x(\delta) = \delta^\top Q_x \delta, \qquad Q_x = J_\phi^\top J_\phi + \lambda I, \quad \lambda > 0. \tag{9}$$

Differentiability gives $\|\phi(x+\delta)-\phi(x)\|_2^2+\lambda\|\delta\|_2^2 = C_x(\delta)+o(\|\delta\|_2^2)$. Thus, C_x is a positive-definite local approximation to regularized feature distortion, rather than a replacement for the algorithm's ℓ_∞ constraint or an exact expression for LPIPS.

Gain at equal local cost. Let $h_b = P_b g \neq 0$ be the real-valued filtered gradient for $b \in \{\text{low}, \text{mid}, \text{high}\}$. The filters may overlap, as with Gaussian masks. Set

$$\delta_b(\varepsilon) = \frac{\varepsilon h_b}{\sqrt{h_b^\top Q_x h_b}}, \qquad R_b = \frac{g^\top h_b}{\sqrt{h_b^\top Q_x h_b}}. \tag{10}$$

Then $C_x(\delta_b) = \varepsilon^2$ and $\mathcal{L}(x+\delta_b) - \mathcal{L}(x) = \varepsilon R_b + r_b$. Since $Q_x \succeq \lambda I$, Equation 8 implies $|r_b| \le L\varepsilon^2/(2\lambda)$. The ratio R_b therefore measures first-order loss gain per unit square root of local visual cost, balancing alignment with the diffusion gradient against the associated visual change.

Sufficient condition for a mid-frequency advantage. Suppose $\Delta = R_{\text{mid}} - \max\{R_{\text{low}}, R_{\text{high}}\} > 0$. For either $b \in \{\text{low}, \text{high}\}$,

$$\begin{aligned} \mathcal{L}(x+\delta_{\text{mid}}) - \mathcal{L}(x+\delta_b) &= \varepsilon(R_{\text{mid}} - R_b) + r_{\text{mid}} - r_b \\ &\ge \varepsilon\Delta - L\varepsilon^2/\lambda. \end{aligned} \tag{11}$$

For $L > 0$, the difference is strictly positive when $0 < \varepsilon < \lambda\Delta/L$ and all perturbation segments remain in the assumed neighborhood. For $L = 0$, the strict first-order advantage suffices. Thus, a larger R_{mid} guarantees a greater local loss increase than the compared low- and high-frequency directions at equal quadratic cost, for sufficiently small steps.

This analysis provides a local criterion for comparing frequency-selective directions through their loss gain relative to visual cost. The empirical results in Figure 2 show a favorable protection–fidelity trade-off for mid-frequency filtering, consistent with this interpretation.

A.3 Discrete Frequency Filtering

Our implementation samples each mask coordinate uniformly on $[-1, 1]$ and takes the real part after inverse transformation. On an even-sized grid, the sampled mask need not be symmetric under discrete frequency reversal. To describe the resulting operator exactly, let $A = \mathrm{ifftshift}(M_{c,\sigma})$ be the real-valued mask in unshifted FFT order, and define $A^{\sharp}_{i,j} = A_{(-i) \bmod H,(-j) \bmod W}$ for $0 \le i < H$ and $0 \le j < W$. The effective shifted mask is $\widetilde{M}_{c,\sigma} = \mathrm{fftshift}((A + A^{\sharp})/2)$.

For a real gradient g with unshifted spectrum $\widehat{g} = \mathcal{F}(g)$, conjugate symmetry gives $\overline{\widehat{g}_{(-i) \bmod H,(-j) \bmod W}} = \widehat{g}_{i,j}$. Using $\mathrm{Re}(z) = (z + \overline{z})/2$ therefore yields

$$\mathcal{F}\big(\mathrm{Re}[\mathcal{F}^{-1}(A \odot \widehat{g})]\big) = \tfrac{1}{2}(A + A^{\sharp}) \odot \widehat{g}. \tag{12}$$

Thus, the effective mask gives the exact spectral representation of the implemented real-valued filtering operation, without changing the algorithm. Since $0 < A \le 1$, its pairwise average also lies in $(0, 1]$, preserving the interpretation of fusion as frequency-dependent attenuation. For a conjugate-symmetric mask, $\widetilde{M}_{c,\sigma} = M_{c,\sigma}$. The implementation additionally uses 10^{-8} in the radial normalization denominator for numerical stability.

B Additional Experimental Details

B.1 Prompts and Perturbation Settings

We use `sks` as the identity token associated with each reference identity, while `person` denotes the corresponding class. The instance prompt is "a photo of sks person", and the class prompt used for prior preservation is "a photo of person". Each identity is represented by four reference images, and the same instance prompt is used for adversarial optimization and subsequent personalization. All adversarial perturbations are optimized in the normalized image space $[-1, 1]$. MidGuard uses an ℓ_∞ perturbation budget of $\eta = 5 \times 10^{-2}$ and a PGD step size of $\alpha = 5 \times 10^{-3}$. The same maximum perturbation budget is enforced for all compared methods after accounting for their image-normalization conventions. MidGuard runs for 50 outer iterations, with 3 clean-image surrogate adaptation steps, 6 adversarial image-update steps, and 3 protected-image main-model adaptation steps in each iteration. The complete procedure is given in Algorithm 1. Unless otherwise specified, the Gaussian band-pass parameters are set to $c = 0.45$ and $\sigma = 0.18$, while the reverse frequency schedule decreases the guidance coefficient from $\beta_{\mathrm{start}} = 0.28$ to $\beta_{\mathrm{end}} = 0.24$ within each PGD loop.

B.2 Metric Computation

For evaluation after personalization, we use a fixed set of 64 random seeds and generate 64 images for each identity using the prompt "a photo of sks person". The same seed set and corresponding initial latents are used across all compared methods to ensure consistent evaluation conditions.

Protected-image fidelity. For protected-image fidelity, each protected image is paired with its corresponding clean reference image. For a fidelity metric $m(\cdot, \cdot)$, the dataset-level score is computed as

$$\mathcal{M}_{\mathrm{fid}} = \frac{1}{N} \sum_{i=1}^{N} m\big(x_i^{\mathrm{clean}}, x_i^{\mathrm{prot}}\big), \tag{13}$$

where N denotes the total number of reference images in the dataset. LPIPS (Zhang et al., 2018) is computed using an AlexNet backbone on images resized to 512×512. SSIM (Wang et al., 2004) and FSIM (Zhang et al., 2011) are computed on RGB images in the $[0, 1]$ range at the same resolution. CLIP-I (Radford et al., 2021) is computed as the cosine similarity between the pooled image representations of CLIP ViT-L/14, while DINO (Caron et al., 2021) is computed as the cosine similarity between normalized class-token features from DINO ViT-S/16. CLIP and DINO inputs are resized to 224×224.

Identity Protection after Personalization. For identity protection after personalization, let $\hat{x}_{i,j}$ denote the j-th generated image of identity i and $x_{i,r}$ its r-th clean reference image. The identity-level

score is computed as

$$\mathcal{M}_i = \frac{1}{64} \sum_{j=1}^{64} \underset{r \in \{1,\dots,4\}}{\mathrm{Agg}} m(\hat{x}_{i,j}, x_{i,r}) \,, \tag{14}$$

where Agg is the maximum for CLIP-I and DINO, and the minimum for LPIPS. Dataset-level results are obtained by averaging $\mathcal{M}_i$ over identities.

Face similarity is measured using ArcFace (Deng et al., 2019). For images containing multiple detected faces, we use the face with the largest bounding-box area. Let $\mathcal{J}_i$ denote the generated images of identity i with valid face embeddings and $\mathcal{R}_i$ the corresponding reference images with valid face embeddings. For identities with both sets nonempty, we compute

$$\mathcal{M}_i^{\mathrm{face}} = \frac{1}{|\mathcal{J}_i|} \sum_{j \in \mathcal{J}_i} \max_{r \in \mathcal{R}_i} \langle e(\hat{x}_{i,j}), e(x_{i,r}) \rangle \,, \tag{15}$$

where $e(\cdot)$ denotes the normalized face embedding. Dataset-level face similarity is obtained by equally averaging $\mathcal{M}_i^{\mathrm{face}}$ over identities with valid face embeddings. Missing embeddings are excluded rather than assigned a similarity of zero. Thus, face similarity measures identity resemblance conditional on successful face-embedding extraction, rather than the probability of generating a detectable face.

FID (Heusel et al., 2017) is computed separately for each identity between its four clean reference images and the corresponding 64 generated images under the same evaluation protocol for all methods. The resulting identity-level FID scores are averaged across identities within each dataset. Given the limited number of reference images per identity, we use FID only as a supplementary measure for relative distributional discrepancy, rather than as a standalone measure of identity suppression.

B.3 Protected-Image Generation Algorithm

Algorithm 1 summarizes the alternating optimization used to generate protected images. The main model is an internal model used during protection generation, not the downstream model used for evaluation. At each outer iteration, a fresh surrogate is copied from the current main model and adapted on clean images. Its parameters are then held fixed while the protected images are updated. Finally, the main model is adapted on the updated protected images, and its parameters are retained for the next outer iteration. The VAE remains frozen throughout. Each adaptation call uses the instance diffusion loss and class prior-preservation loss with a fresh optimizer. We use $T = 50$, $S = 3$, and $K = 6$, giving 300 image-update steps. The frequency schedule restarts in every inner PGD loop, whereas the protected images retain the updates from previous outer iterations.

C Additional Experimental Results

We supplement the main SD1.5 evaluation with protection and fidelity results on SD2.1 and SDXL, followed by qualitative comparisons with and without the identity token.

C.1 Protection Effectiveness

SD2.1. Table 4 reports the protection results under SD2.1 (Rombach et al., 2022). MidGuard achieves the best performance across all reported metrics on both CelebA-HQ and VGGFace2. The improvement is observed consistently across semantic similarity, perceptual distance, distributional discrepancy, and face identity similarity, rather than being concentrated on a single evaluation criterion. These results demonstrate that MidGuard remains effective when optimized and evaluated on SD2.1, supporting the applicability of the proposed frequency-aware optimization across different diffusion backbones.

SDXL. Table 5 reports the protection performance under SDXL (Podell et al., 2023). MidGuard achieves the strongest overall protection on both datasets, obtaining the best result on four of the five metrics in each case. Specifically, MIST obtains a slightly higher FID on CelebA-HQ, while DiffusionGuard achieves a lower face similarity on VGGFace2; nevertheless, MidGuard performs best on the remaining metrics in both datasets. Compared with SD2.1, the differences among protection methods are generally smaller under the SDXL LoRA setting, suggesting that their

Algorithm 1 MidGuard protected-image generation

Require: Initial main-model parameters $\theta^{(0)}$; clean adaptation images $\mathcal{X}_{\text{train}}$; clean images to protect $\mathcal{X}_0$; class images $\mathcal{C}$; instance and class prompts; $T, S, K, \eta, \alpha, c, \sigma, \beta_{\text{start}}, \beta_{\text{end}}$
Ensure: Protected images $\mathcal{X}_{\text{adv}}$

$\mathcal{X}_{\text{adv}} \leftarrow \mathcal{X}_0$
Construct the radial Gaussian mask $M_{c,\sigma}$
for $i = 0, \ldots, T-1$ **do**
  $\theta_s \leftarrow \text{Copy}(\theta^{(i)})$
  Adapt θ_s for S steps on $\mathcal{X}_{\text{train}}$ and $\mathcal{C}$ using the instance and prior-preservation losses
  Hold θ_s fixed during image updates
  for $k = 0, \ldots, K-1$ **do**
    $\beta_k \leftarrow \beta_{\text{start}} + \dfrac{k}{\max\{K-1, 1\}}(\beta_{\text{end}} - \beta_{\text{start}})$
    Sample diffusion timestep t and Gaussian noise ϵ
    $g_k \leftarrow \nabla_{\mathcal{X}_{\text{adv}}} \widehat{\mathcal{L}}_{\text{diff}}(\mathcal{X}_{\text{adv}}; \theta_s, t, \epsilon)$
    $G_k \leftarrow \text{fftshift}(\mathcal{F}(g_k))$
    $g_k^{\text{mid}} \leftarrow \text{Re}\{\mathcal{F}^{-1}[\text{ifftshift}(M_{c,\sigma} \odot G_k)]\}$
    $\bar{g}_k \leftarrow (1-\beta_k) g_k + \beta_k g_k^{\text{mid}}$
    $\mathcal{X}_{\text{adv}} \leftarrow \Pi_{\mathcal{B}_\infty(\mathcal{X}_0, \eta) \cap [-1,1]}[\mathcal{X}_{\text{adv}} + \alpha \,\text{sign}(\bar{g}_k)]$
  end for
  $\theta^{(i+1)} \leftarrow$ adapt $\theta^{(i)}$ for S steps on $\mathcal{X}_{\text{adv}}$ and $\mathcal{C}$
end for
return $\mathcal{X}_{\text{adv}}$

Table 4: Comparison with adversarial protection methods after DreamBooth personalization using SD2.1. Results are reported on CelebA-HQ and VGGFace2 under the prompt "a photo of sks person"; arrows indicate the preferred metric directions.

Method	CelebA-HQ					VGGFace2				
	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓
Clean	0.8185	0.8135	0.5941	139.8875	0.6412	0.8098	0.7603	0.6638	187.8534	0.6360
MIST	0.5839	0.4495	0.7142	308.6899	0.4652	0.4632	0.2130	0.7341	446.2382	0.3562
Anti-DreamBooth	0.5757	0.4092	0.7769	339.3417	0.4715	0.4622	0.2275	0.8144	450.8116	0.3533
SimAC	0.5697	0.4631	0.7699	342.8505	0.4546	0.4711	0.2644	0.8007	422.3129	0.3760
PAP	0.5765	0.4309	0.7839	338.2414	0.4928	0.4650	0.2304	0.8300	441.5537	0.3557
DiffusionGuard	0.6414	0.6938	0.6431	205.0093	0.5059	0.6597	0.6893	0.6967	234.8279	0.5189
DiffVax	0.6284	0.6741	0.6365	209.3338	0.5915	0.7037	0.7169	0.6668	211.0841	0.6320
MidGuard (Ours)	**0.5270**	**0.4068**	**0.7949**	**357.2217**	**0.4353**	**0.4525**	**0.2028**	**0.8405**	**452.0813**	**0.3212**

relative performance depends on the backbone and personalization setup. Despite this reduced margin, MidGuard maintains the most consistent overall performance across semantic, perceptual, distributional, and identity-based evaluations, demonstrating that its effectiveness extends beyond the SD1.5 and SD2.1 settings.

C.2 Image Fidelity

SD2.1. Table 6 reports protected-image fidelity under SD2.1. MidGuard maintains strong fidelity across both datasets and across perceptual, structural, and representation-based evaluations. More importantly, when considered together with the protection results in Table 4, the stronger identity suppression achieved by MidGuard is not accompanied by a corresponding degradation in protected-image quality. This result further shows that the favorable trade-off between identity protection and visual fidelity observed under SD1.5 extends to SD2.1.

SDXL. Table 7 reports protected-image fidelity under SDXL. MidGuard remains competitive across both datasets, although no single method dominates all fidelity measures in this setting. When considered together with the protection results in Table 5, MidGuard achieves strong identity protection while preserving high visual fidelity. These results indicate that the proposed method remains

Table 5: Comparison with adversarial protection methods after DreamBooth LoRA personalization using SDXL. Results are reported on CelebA-HQ and VGGFace2 under the prompt "a photo of sks person"; arrows indicate the preferred metric directions.

Method	CelebA-HQ					VGGFace2				
	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓
Clean	0.7932	0.8469	0.5086	128.6839	0.6620	0.8047	0.7962	0.5902	183.4675	0.6789
MIST	0.7092	0.7459	0.5744	**184.8993**	0.5933	0.7051	0.7118	0.6606	208.8476	0.6291
Anti-DreamBooth	0.7551	0.8069	0.5520	153.8692	0.6290	0.7173	0.7304	0.6547	207.3125	0.6309
SimAC	0.7226	0.7851	0.5569	160.6391	0.6221	0.7199	0.7229	0.6383	210.4407	0.6299
PAP	0.7325	0.8005	0.5606	167.8841	0.6343	0.7082	0.7281	0.6588	210.0204	0.6221
DiffusionGuard	0.7046	0.8441	0.5283	143.6705	0.6017	0.7222	0.7875	0.6026	211.2345	**0.6102**
DiffVax	0.7845	0.8416	0.5130	129.3954	0.6440	0.7880	0.7808	0.5956	190.7409	0.6706
MidGuard (Ours)	**0.7013**	**0.7236**	**0.5791**	184.2893	**0.5912**	**0.7037**	**0.7106**	**0.6629**	**213.8681**	0.6281

Table 6: Fidelity of images protected by different methods on CelebA-HQ and VGGFace2 using SD2.1; arrows indicate the preferred metric directions.

Method	CelebA-HQ					VGGFace2				
	LPIPS↓	SSIM↑	FSIM↑	CLIP-I↑	DINO↑	LPIPS↓	SSIM↑	FSIM↑	CLIP-I↑	DINO↑
MIST	0.4109	0.8079	0.9114	0.8083	0.8458	0.4485	0.7609	0.9146	0.8703	0.8560
Anti-DreamBooth	0.1782	0.9302	0.9720	0.8466	0.9415	0.2116	0.9142	0.9714	0.9231	0.9453
SimAC	0.2041	0.9003	0.9636	0.8668	0.9111	0.2222	0.8948	0.9615	0.9056	0.9196
PAP	0.1810	0.9324	0.9720	0.8482	0.9407	0.2149	0.9170	0.9713	0.9253	0.9444
DiffusionGuard	0.1696	0.9348	0.9612	**0.8985**	0.9326	0.2011	**0.9716**	0.9746	0.9164	0.9639
DiffVax	0.1678	0.8861	0.9322	0.7501	0.8695	0.2072	0.9605	0.9721	0.9006	0.9485
MidGuard (Ours)	**0.1653**	**0.9377**	**0.9776**	0.8522	**0.9451**	**0.2004**	0.9642	**0.9752**	**0.9269**	**0.9661**

effective under the more challenging SDXL setting without relying on substantially increased image distortion.

C.3 Qualitative Comparisons

We distinguish between prompts without the identity token, which test generic portrait generation under different descriptions, and prompts containing `sks`, which explicitly request the personalized identity. The evaluated prompts include a standard photo, a DSLR portrait, and a photo with the additional attribute "wearing a hat". For all qualitative comparisons, we use the same identities, prompts, and generation seeds across methods to ensure directly comparable outputs.

Generation without the Identity Token. Figure 6 shows outputs under "a photo of person", "a dslr portrait of person", and "a photo of person wearing a hat", none of which contains `sks`. The model personalized on MidGuard-protected images continues to generate visually coherent portraits across all three prompts and follows the additional hat attribute without obvious degradation in the displayed examples. These observations suggest that disrupting the target identity association need not compromise generic portrait generation or basic prompt adherence.

Protection across Identity-Conditioned Prompts. Figure 7 presents results under "a photo of sks person", "a dslr portrait of sks person", and "a photo of sks person wearing a hat", all of which request the personalized identity through `sks`. Across the displayed examples, MidGuard produces outputs with weaker resemblance to the corresponding reference identities than the compared defenses. This effect remains evident under both the DSLR portrait description and the additional hat attribute, suggesting that MidGuard maintains its protection across the tested identity-conditioned prompts rather than relying on a single training-style prompt.

Table 7: Fidelity of images protected by different methods on CelebA-HQ and VGGFace2 using SDXL; arrows indicate the preferred metric directions.

Method	CelebA-HQ					VGGFace2				
	LPIPS↓	SSIM↑	FSIM↑	CLIP-I↑	DINO↑	LPIPS↓	SSIM↑	FSIM↑	CLIP-I↑	DINO↑
MIST	0.4099	0.8111	0.9205	0.7917	0.8156	0.4067	0.7923	0.9222	0.8468	0.8359
Anti-DreamBooth	0.1215	0.9510	0.9812	0.9501	0.9782	0.1844	0.9311	0.9815	0.9521	0.9717
SimAC	0.1863	0.9336	0.9793	0.9139	0.9616	0.2041	0.9299	0.9785	0.9431	0.9628
PAP	0.1391	0.9443	0.9772	0.9410	0.9732	0.2032	0.9292	0.9774	0.9469	0.9681
DiffusionGuard	**0.1186**	0.9663	0.9927	0.9208	**0.9953**	0.1859	**0.9496**	**0.9937**	0.9561	0.9753
DiffVax	0.1205	0.9590	0.9899	0.9589	0.9899	0.1867	0.9334	0.9899	0.9538	0.9799
MidGuard (Ours)	0.1207	**0.9669**	**0.9987**	**0.9602**	0.9878	**0.1805**	0.9386	0.9890	**0.9592**	**0.9809**

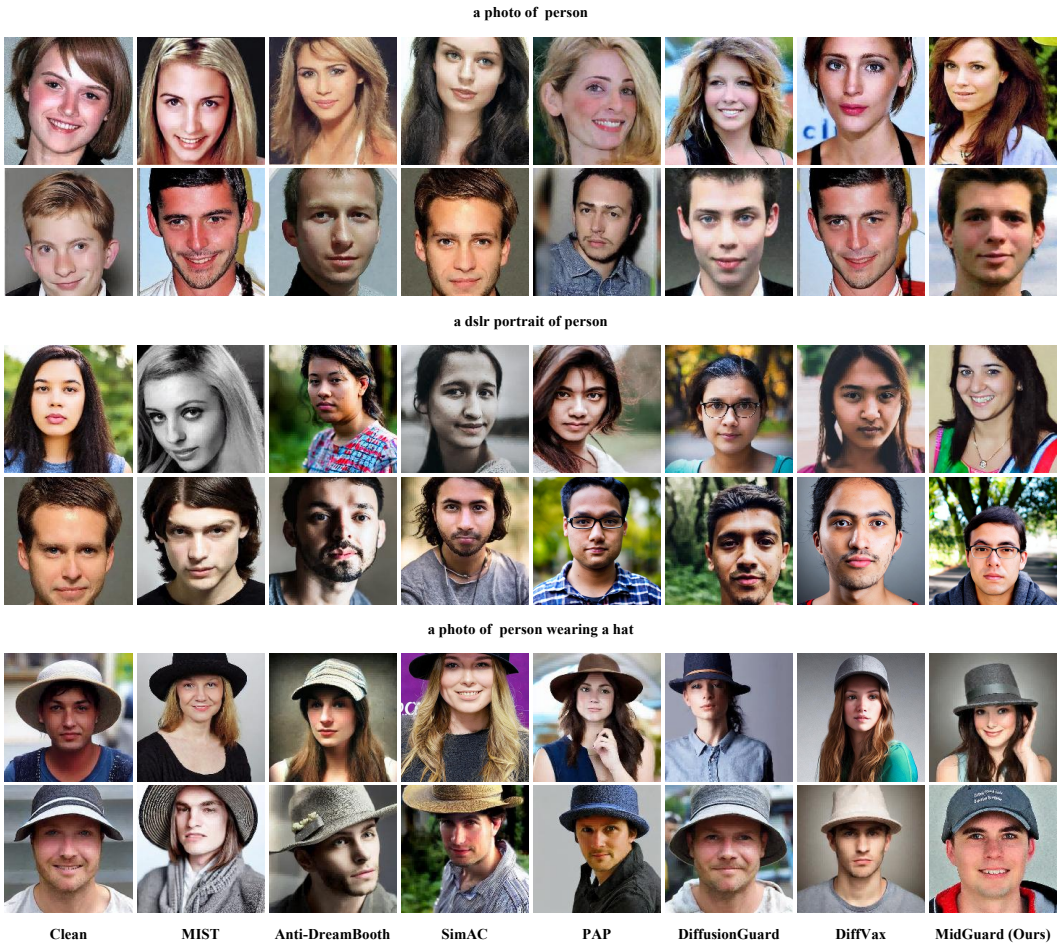


Figure 6: Generic portrait generation without the identity token sks. The three prompt groups correspond to "a photo of person", "a dslr portrait of person", and "a photo of person wearing a hat", with two identities shown for each prompt. Columns correspond to the clean baseline and different protection methods.

D Robustness and Generalization

We further evaluate whether the protection remains effective under common image transformations, an alternative personalization mechanism, and a change of target backbone. These experiments

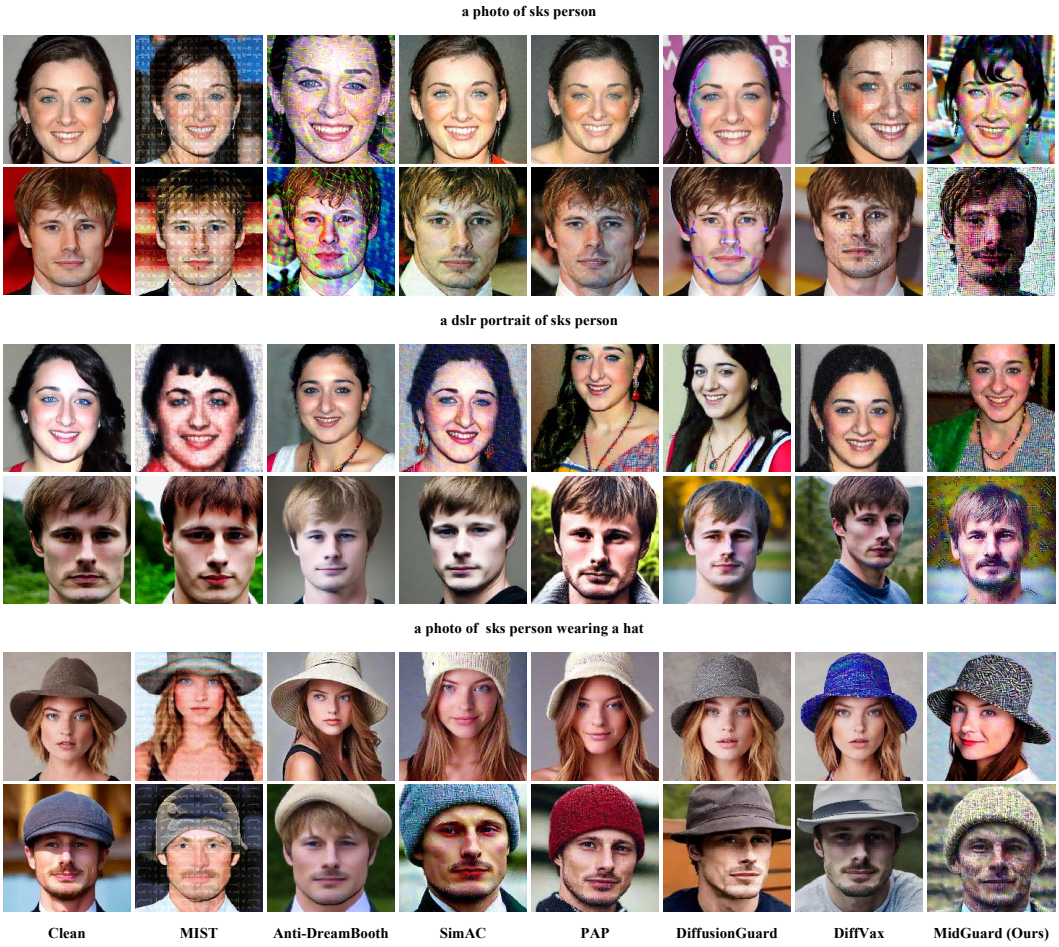


Figure 7: Identity-conditioned generation under "a photo of sks person", "a dslr portrait of sks person", and "a photo of sks person wearing a hat". Each prompt group shows two identities, and columns correspond to the clean baseline and different protection methods.

complement the main DreamBooth evaluation by examining input-level robustness, transfer across personalization strategies, and protection transfer from SD1.5 to SD2.1 and SDXL.

D.1 Image Preprocessing

Protection should remain effective when protected images are processed before personalization. We apply six common preprocessing operations to the protected images. JPEG compression uses quality 70. Resizing downsamples images to 256×256 and then resizes them back to 512×512 using bicubic interpolation. Gaussian blur uses a radius of 1.0. Center cropping retains 90% of the image width and height and resizes the cropped image back to 512×512 using bicubic interpolation. Median filtering uses a 3×3 kernel. The combined transformation applies JPEG compression with quality 70 followed by the same 256×256 resizing operation.

Tables 8 and 9 summarize performance averaged across the six operations. The Clean (unprotected) row is included only as a reference baseline and is not part of this six-operation average. MidGuard achieves the best averaged result on every reported metric for both backbones. This supports protection beyond the unprocessed setting, without implying that it leads in every metric under every operation.

Figure 8 presents the per-operation results under SD2.1, complementing the SD1.5 results in Figure 4. Although the relative differences among methods vary across transformations, MidGuard remains

Table 8: Average identity-protection performance after six image preprocessing operations using SD1.5. Clean (unprotected) denotes personalization using clean reference images without protection or preprocessing and is included as an unprotected baseline.

Method	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓
Clean (unprotected)	0.8061	0.7814	0.6215	187.52	0.5998
MIST	0.6505	0.5644	0.6930	280.01	0.4377
Anti-DreamBooth	0.5053	0.3471	0.7935	392.37	0.4744
SimAC	0.5818	0.4502	0.7536	340.38	0.5171
PAP	0.5491	0.3824	0.7859	378.55	0.4804
DiffusionGuard	0.7667	0.7662	0.6407	195.78	0.5669
DiffVax	0.7119	0.7118	0.6498	210.11	0.5670
MidGuard (Ours)	**0.4909**	**0.3313**	**0.8049**	**405.36**	**0.4286**

Table 9: Average identity-protection performance after six image preprocessing operations using SD2.1. Clean (unprotected) denotes personalization using clean reference images without protection or preprocessing and is included as an unprotected baseline.

Method	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓
Clean (unprotected)	0.8220	0.7982	0.6380	177.88	0.6536
MIST	0.6198	0.5045	0.7498	307.17	0.4650
Anti-DreamBooth	0.6162	0.4538	0.8045	356.89	0.5198
SimAC	0.6293	0.4741	0.7819	321.53	0.5666
PAP	0.6240	0.4854	0.8003	342.23	0.5594
DiffusionGuard	0.7476	0.7628	0.6690	200.33	0.6319
DiffVax	0.6975	0.7310	0.6527	199.32	0.6612
MidGuard (Ours)	**0.6120**	**0.4368**	**0.8125**	**360.92**	**0.4596**

competitive under compression, resizing, cropping, and filtering operations. Together with the averaged results above, these comparisons show that the protection remains effective across different forms of image preprocessing rather than depending on a particular transformation.

D.2 LoRA-Based Personalization

We additionally replace full DreamBooth fine-tuning with LoRA-based personalization (Hu et al., 2022) using SD1.5. Results are averaged across identities from CelebA-HQ (Karras et al., 2018) and VGGFace2 (Cao et al., 2018). As shown in Table 10, MidGuard consistently provides stronger identity protection than the compared methods under LoRA-based adaptation. This indicates that MidGuard is not tied to full-model DreamBooth fine-tuning and remains effective when personalization is performed through low-rank parameter updates.

D.3 Black-Box Cross-Model Transfer

We evaluate transfer from SD1.5 to SD2.1 and SDXL on CelebA-HQ and VGGFace2, each with 50 identities and four reference images per identity. For all protection methods, images protected using SD1.5 are directly used for target-model personalization without re-optimizing perturbations. We use DreamBooth for SD2.1 and DreamBooth LoRA for SDXL, following the corresponding main-experiment training, generation, and metric settings. For each target backbone, we generate 64 images per identity using the same prompt and a fixed seed set shared across methods.

SD1.5-to-SD2.1 Transfer. Table 11 reports results separately for both datasets. MidGuard achieves the best numerical result across all five protection metrics on each dataset. Face similarity decreases from 0.6488 to 0.3489 on CelebA-HQ and from 0.6497 to 0.3077 on VGGFace2, outperforming

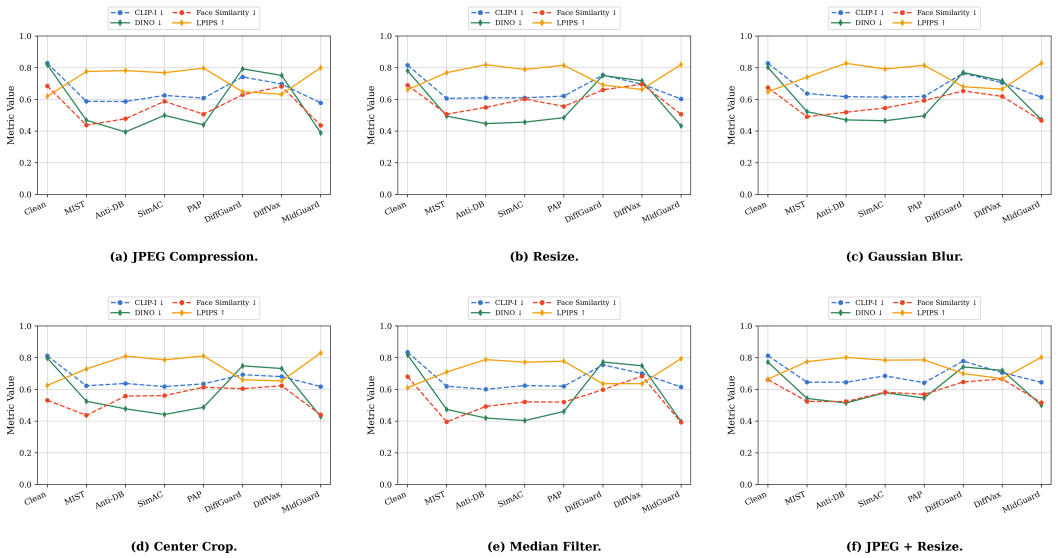


Figure 8: Identity protection after image preprocessing using SD2.1. Panels compare JPEG compression, resizing, Gaussian blur, center cropping, median filtering, and combined JPEG compression and resizing.

Table 10: Identity-protection performance under LoRA-based personalization using SD1.5, averaged across identities from CelebA-HQ and VGGFace2.

Method	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓
Clean	0.6910	0.7599	0.6400	184.78	0.3926
MIST	0.4712	0.5111	0.7160	391.24	0.1298
Anti-DreamBooth	0.5172	0.6096	0.6986	246.56	0.2048
SimAC	0.4955	0.5714	0.6959	260.32	0.2016
PAP	0.4950	0.5861	0.7395	276.91	0.2227
DiffusionGuard	0.6005	0.7439	0.6526	197.60	0.2539
DiffVax	0.6585	0.7316	0.6468	192.10	0.3725
MidGuard (Ours)	**0.4553**	**0.4986**	**0.7553**	**403.15**	**0.1142**

the strongest competing baseline, MIST, at 0.3609 and 0.3198, respectively. These results support protection transfer from SD1.5 to SD2.1.

SD1.5-to-SDXL Transfer. Table 12 reports transfer results under SDXL DreamBooth LoRA personalization. MidGuard achieves the best numerical result across all five metrics on both datasets. Face similarity decreases from 0.6490 to 0.5336 on CelebA-HQ and from 0.6803 to 0.5782 on VGGFace2, compared with 0.5527 and 0.6199 for MIST, the strongest competing baseline on this metric. Together with the SD2.1 results, these findings demonstrate that MidGuard's protection is not confined to the source backbone and remains effective across the tested target models without target-specific perturbation optimization, supporting its applicability when the personalization backbone is unknown.

E Additional Ablation Studies

This section supplements the component comparison in Section 4.4. We examine protected-image fidelity, the Gaussian mask parameters, the fixed guidance weight, and frequency schedules.

All ablation studies use the same 50 CelebA-HQ identities and four reference images per identity as the main evaluation. Within each ablation study, all variants use the same set of 64 generation seeds

Table 11: SD1.5-to-SD2.1 protection transfer after DreamBooth personalization on CelebA-HQ and VGGFace2. Clean denotes unprotected training images; arrows indicate the preferred metric directions.

Method	CelebA-HQ					VGGFace2				
	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓
Clean	0.8168	0.8133	0.5940	139.9518	0.6488	0.8102	0.7615	0.6647	187.6756	0.6497
MIST	0.5852	0.4522	0.7118	306.7028	0.3609	0.4583	0.2241	0.7360	449.6615	0.3198
Anti-DreamBooth	0.5801	0.4197	0.7772	334.2838	0.4732	0.4702	0.2329	0.8169	453.9172	0.3295
SimAC	0.5772	0.4095	0.7605	344.7021	0.4577	0.4892	0.2917	0.8035	414.3807	0.3545
PAP	0.5791	0.4379	0.7783	340.4669	0.4738	0.4708	0.2372	0.8160	449.6708	0.3216
DiffusionGuard	0.6529	0.6876	0.6571	199.5046	0.4884	0.6636	0.6858	0.7058	230.1363	0.4908
DiffVax	0.6270	0.6737	0.6361	208.2967	0.5885	0.7056	0.7149	0.6658	212.2711	0.6348
MidGuard (Ours)	**0.4965**	**0.3823**	**0.7863**	**372.7232**	**0.3489**	**0.3844**	**0.2059**	**0.8396**	**458.9082**	**0.3077**

Table 12: SD1.5-to-SDXL protection transfer after DreamBooth LoRA personalization on CelebA-HQ and VGGFace2. Clean denotes unprotected training images; arrows indicate the preferred metric directions.

Method	CelebA-HQ					VGGFace2				
	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓
Clean	0.7959	0.8479	0.5086	128.7808	0.6490	0.8061	0.7955	0.5922	182.6169	0.6803
MIST	0.6293	0.6565	0.6103	223.3560	0.5527	0.6335	0.6833	0.6955	266.4469	0.6199
Anti-DreamBooth	0.6690	0.7628	0.6043	194.8996	0.6295	0.6403	0.6853	0.6971	263.4664	0.6226
SimAC	0.6551	0.7285	0.5878	214.3703	0.6136	0.6670	0.6877	0.6621	253.4197	0.6254
PAP	0.6780	0.7645	0.6092	193.4023	0.6247	0.6450	0.7029	0.6906	250.6588	0.6216
DiffusionGuard	0.7248	0.8413	0.5169	147.2807	0.5932	0.7367	0.7836	0.5998	192.7328	0.6204
DiffVax	0.6669	0.7558	0.5647	203.4142	0.6228	0.6646	0.7415	0.6026	205.2150	0.6495
MidGuard (Ours)	**0.5938**	**0.6511**	**0.6161**	**242.5951**	**0.5336**	**0.6089**	**0.6392**	**0.7011**	**279.4497**	**0.5782**

for each identity, consistent with the main evaluation protocol. Metrics are computed and aggregated using the same evaluation procedure across variants. Seed sets may differ between other ablation studies, so we compare variants within each study. The center-frequency and bandwidth sweeps report identical aggregate results for their shared configuration $c = 0.45$ and $\sigma = 0.18$.

Table 13: Ablation of frequency filtering, gradient fusion, and scheduling on protected-image fidelity; arrows indicate the preferred metric directions.

Variant	LPIPS↓	SSIM↑	FSIM↑	CLIP-I↑	DINO↑
Raw PGD	0.2053	0.9204	0.9695	0.8732	0.9377
Mid-frequency only	0.2267	0.9128	0.9633	0.8812	0.9422
Fixed fusion	0.2049	0.9205	0.9695	0.8740	0.9399
MidGuard (Ours)	**0.2005**	**0.9222**	**0.9708**	**0.8980**	**0.9503**

Component-Wise Image Fidelity. Table 13 complements the protection comparison in Table 3. Filtering alone worsens LPIPS, SSIM, and FSIM relative to Raw PGD, despite higher CLIP-I and DINO similarities. Fixed fusion restores the former metrics to approximately the baseline level. The complete method achieves the best fidelity across all five criteria, indicating that its improved protection in the component experiment does not require a corresponding loss of reference-image quality.

Center Frequency. Table 15 varies c while fixing $\sigma = 0.18$. The main-experiment setting $c = 0.45$ gives the best result across the five reported protection metrics. Performance varies non-monotonically

Table 14: Sensitivity to the Gaussian band-pass bandwidth σ with $c = 0.45$; arrows indicate the preferred metric directions.

σ	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓
0.02	0.3343	0.1216	0.8527	473.7793	0.2416
0.04	0.3295	0.1241	0.8514	480.1001	0.2687
0.06	0.3412	0.1371	0.8439	470.6282	0.2653
0.08	0.3332	0.1241	0.8471	480.8199	0.3105
0.10	0.3441	0.1322	0.8455	475.3948	0.2557
0.12	0.3432	0.1348	0.8467	468.3806	0.2577
0.14	0.3576	0.1440	0.8422	464.9500	0.2557
0.16	0.3335	0.1239	0.8490	476.7723	0.2811
0.18	**0.3226**	**0.1055**	**0.8551**	**484.9839**	**0.2378**
0.20	0.3331	0.1246	0.8475	480.9460	0.2382

as the center moves toward higher frequencies, so the results do not support a general preference for either lower or higher center frequencies.

Table 15: Sensitivity to the Gaussian band-pass center frequency c with $\sigma = 0.18$; arrows indicate the preferred metric directions.

c	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓
0.05	0.3381	0.1231	0.8506	479.9103	0.2489
0.15	0.3509	0.1422	0.8449	469.8545	0.2712
0.25	0.3476	0.1266	0.8514	474.2416	0.2439
0.35	0.3484	0.1220	0.8500	477.1473	0.2613
0.45	**0.3226**	**0.1055**	**0.8551**	**484.9839**	**0.2378**
0.55	0.3498	0.1440	0.8461	471.0538	0.2884
0.65	0.3465	0.1421	0.8482	470.1085	0.2720
0.75	0.3600	0.1520	0.8401	457.6105	0.2539
0.85	0.3428	0.1211	0.8523	474.6691	0.2555
0.95	0.3325	0.1130	0.8504	475.3460	0.2403

Table 16: Comparison of frequency-guidance schedules within the six-step inner PGD loop; arrows indicate the preferred metric directions.

Schedule	β sequence	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓
Alternate	$0.24, 0.28, 0.24, 0.28, 0.24, 0.28$	0.2959	0.1047	0.8611	520.7134	0.1349
Ramp	$0.240, 0.248, 0.256, 0.264, 0.272, 0.280$	0.2970	0.1028	0.8575	518.6118	0.1537
Reverse	$0.280, 0.272, 0.264, 0.256, 0.248, 0.240$	**0.2856**	**0.1022**	**0.8715**	**535.6632**	**0.0382**
LateDrop1	$0.28, 0.28, 0.28, 0.28, 0.28, 0.24$	0.2954	0.1107	0.8574	521.7711	0.1800
LateDrop2	$0.28, 0.28, 0.28, 0.28, 0.24, 0.24$	0.2951	0.1093	0.8615	530.1945	0.1195
Fixed	$0.28, 0.28, 0.28, 0.28, 0.28, 0.28$	0.2868	0.1050	0.8705	533.7485	0.0612

Bandwidth. Table 14 varies σ while fixing $c = 0.45$. The main-experiment setting $\sigma = 0.18$ achieves the best result across all five reported protection metrics. The non-monotonic results show that protection depends on bandwidth. The $\sigma = 0.18$ row matches the $c = 0.45$ row in Table 15, as both evaluate the same Gaussian mask.

Frequency Guidance Schedule.

Table 16 compares fixed, increasing, reverse, alternating, and delayed-drop schedules. The fixed schedule uses $\beta = 0.28$ throughout the six-step inner PGD loop and therefore corresponds to the $\beta = 0.28$ setting in Table 17. The reverse schedule achieves the best result across all five metrics

Table 17: Effect of the fixed frequency-guidance weight β on identity protection after personalization; arrows indicate the preferred metric directions.

β	CLIP-I↓	DINO↓	LPIPS↑	FID↑	Face Sim.↓
0.10	0.2952	0.1068	0.8593	514.1674	0.1668
0.20	0.2921	0.1057	0.8608	526.2605	0.1410
0.24	0.2998	0.1201	0.8623	519.3822	0.1597
0.26	0.2996	0.1059	0.8607	517.3310	0.1425
0.27	0.2924	0.1068	0.8620	520.1210	0.1059
0.28	**0.2868**	**0.1050**	**0.8705**	**533.7485**	**0.0612**
0.30	0.2990	0.1092	0.8586	519.4559	0.0813
0.32	0.3041	0.1139	0.8587	521.6353	0.1182
0.35	0.2896	0.1082	0.8582	522.1154	0.1825
0.50	0.3217	0.1180	0.8562	514.8854	0.1924

and consistently improves upon fixed guidance. The other schedules do not match its performance, supporting gradual relaxation from stronger mid-frequency guidance to less constrained refinement. However, the final scores alone do not establish the underlying optimization mechanism.

Frequency Guidance Strength.

Table 17 varies the fixed frequency-guidance weight. The setting $\beta = 0.28$ achieves the best result across all five metrics, while both smaller and larger values perform less favorably. This non-monotonic response shows that increasing the filtered-gradient contribution alone does not reliably strengthen protection, supporting a balanced combination of filtered and raw gradients.