VEIL: VISUAL-STATE EARLY-ANCHOR STABILIZATION FOR IDENTITY LEAKAGE SUPPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image personalization can bind a subject's visual identity to an identifier using only a few reference images, enabling unauthorized synthesis from publicly shared face photos. User-side defenses perturb images before release to impair downstream identity recovery while preserving visual quality. However, existing methods typically optimize denoising or cross-attention signals without jointly controlling the diffusion states used for supervision and the conditioning positions and image regions targeted during optimization. In this paper, we propose VEIL (**V**isual-state **E**arly-anchor **I**dentity **L**eakage suppression), a user-side defense that stabilizes perturbation optimization using fixed ultra-early denoising anchors and posterior-mean latent encoding. It further targets conditioning positions and image regions that contribute most to the protection objective through attribution-guided selection, saliency-weighted updates, and soft shrinkage. Under DreamBooth on CelebA-HQ and VGGFace2, VEIL achieves the highest SSIM and PSNR among the compared defenses on both datasets, with an SSIM of 0.94 and PSNRs of 37.94 and 38.21 dB. Across the four primary dataset–prompt settings, VEIL achieves or ties for the highest FDFR, with values ranging from 0.73 to 0.85. Protection also transfers across the evaluated Stable Diffusion variants and personalization procedures without re-optimizing the released images. Code is available at https://anonymous.4open.science/r/VEAL_demo.

1 INTRODUCTION

Text-to-image diffusion models support high-fidelity image synthesis and few-shot subject personalization (Rombach et al., 2022). Methods such as DreamBooth (Ruiz et al., 2023), Textual Inversion (Gal et al., 2023), Custom Diffusion (Kumari et al., 2023), SVDiff (Han et al., 2023), and LoRA-based adaptation (Hu et al., 2022) can associate a subject's visual appearance with an identifier using only a few reference images. For face identities, publicly available photographs can therefore be used to personalize a generator that synthesizes the depicted individual in novel contexts without consent.

User-side protection addresses this risk by modifying images before release so that downstream personalization recovers less identity while the released images remain visually faithful, as illustrated in Figure 1. Existing defenses construct bounded perturbations by optimizing diffusion training or personalization objectives (Le et al., 2023; Xue et al., 2024; Wang et al., 2024). Attention-based defenses augment the diffusion training objective with an additional attention-suppression loss to disrupt subject-condition association (Xu et al., 2024; Liu et al., 2024b; Li et al., 2025), while fidelity-oriented approaches improve the visual quality of protected images (Wu et al., 2025; Mi et al., 2026). These directions improve different aspects of protection, but strong post-personalization suppression and high released-image fidelity remain difficult to achieve simultaneously. This trade-

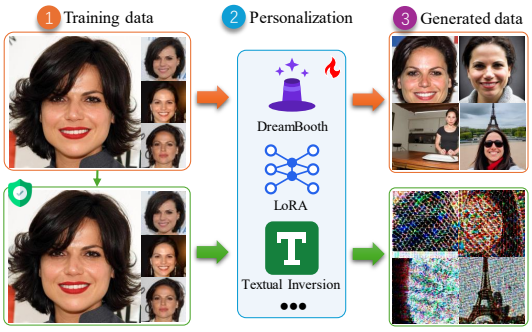


Figure 1: User-side protection with VEIL. Personalization recovers the identity from clean images but not from protected ones.

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.

off is shaped not only by the objective being optimized, but also by how the resulting gradient is generated and applied during protection optimization. Three choices are particularly consequential: which diffusion states provide supervision, which conditioning positions are suppressed, and which image regions receive perturbation updates. Temporally, random timestep sampling changes the supervised noise regime across iterations, while adaptive selection may also vary the selected timesteps as the protected image evolves (Wang et al., 2024). Because different timesteps induce different noise levels, latent structures, and denoising signals, successive updates may be computed from inconsistent visual states (Appendix E, Figure 9). Along the conditioning dimension, existing attention-based defenses generally suppress broad attention maps or target a predefined identifier position (Liu et al., 2024b; Li et al., 2025). Broad suppression ignores differences in how conditioning positions contribute to the protection objective, whereas fixed-position targeting assumes that a predefined position consistently captures the relevant interaction. In image space, conventional sign-based updates discard relative gradient magnitudes and spread perturbations across broad regions, including pixels with limited contribution to the current objective. Under a fixed perturbation budget, these choices can devote part of the available update to temporally inconsistent or low-contribution components, intensifying the trade-off between protection and fidelity.

To address the above limitations, we propose VEIL (**V**isual-state **E**arly-anchor **I**dentity **L**eakage suppression), a user-side defense for high-fidelity face identity protection. VEIL first stabilizes denoising supervision using a compact, fixed set of ultra-early denoising anchors together with posterior-mean latent encoding. Throughout, *ultra-early* refers to the smallest diffusion timesteps, that is, the low-noise end of the schedule. The fixed anchors retain substantial source-image structure and provide consistent low-noise states across iterations, while averaging their gradients avoids dependence on a single timestep. Posterior-mean encoding further removes the variability introduced by VAE posterior sampling. Using this stabilized denoising objective, VEIL estimates the first-order contribution of each conditioning position and selects the highest-contribution positions for cross-attention suppression. The selected positions are recomputed as the protected image evolves, allowing the attention objective to follow the current protection state rather than a predefined identifier position. Saliency weighting and soft shrinkage convert this gradient into a spatially selective perturbation by attenuating low-magnitude updates and removing small accumulated components. The surrogate denoiser is periodically adapted to the evolving protected images so that subsequent perturbation updates are computed against its response to the current samples.

We evaluate VEIL on CelebA-HQ (Karras et al., 2018) and VGGFace2 (Cao et al., 2018), where it attains the highest SSIM (0.94) and PSNR (37.94 and 38.21 dB) among the compared defenses and achieves or ties for the highest FDFR (0.73 to 0.85) in all four primary dataset–prompt settings. The same protected images remain effective across the evaluated Stable Diffusion variants and personalization procedures without re-optimization. Our contributions are summarized as follows:

- We identify temporal consistency, conditioning-position selectivity, and spatial update selectivity as three factors that shape the protection-fidelity trade-off under a bounded perturbation budget.
- We propose VEIL, which addresses the three factors through fixed ultra-early anchors with posterior-mean encoding, attribution-guided conditioning-position selection, and saliency-weighted updates with soft shrinkage.
- Experiments across datasets, personalization procedures, model variants, identifier bindings, and image transformations show that VEIL combines high released-image fidelity with effective identity suppression and transfers without re-optimization.

2 Related Work

2.1 Personalized Text-to-Image Diffusion

Personalized text-to-image diffusion adapts a pretrained generative model to a target subject from a small set of reference images. Textual Inversion learns a subject-specific token embedding while keeping the generative model frozen (Gal et al., 2023). DreamBooth instead fine-tunes the denoising network to associate a rare identifier with the target subject (Ruiz et al., 2023). Custom Diffusion (Kumari et al., 2023) and SVDiff (Han et al., 2023) restrict adaptation to attention-related or compact parameter subspaces, whereas LoRA-based personalization uses low-rank parameter updates (Hu et al., 2022). Although these methods adapt different parameter subsets, they all enable a

subject's visual identity to be reproduced from only a few reference images. When such images are collected and used without consent, the same capability can enable unauthorized identity synthesis across novel contexts, motivating user-side protection before release.

2.2 Protection Against Personalization

User-side protection modifies images before publication to reduce their utility for unauthorized downstream personalization (Le et al., 2023), whereas model-side defenses such as IDDM (Dai et al., 2026) modify the personalized model to reduce identity linkability in its outputs. For instance, Glaze (Shan et al., 2023) cloaks artistic styles against unauthorized imitation. AdvDM (Liang et al., 2023) maximizes the diffusion training loss. Mist (Liang & Wu, 2023) combines semantic and textural objectives. PhotoGuard (Salman et al., 2023) disrupts generative image editing through encoder- or diffusion-level representations. These methods establish the general setting of pre-release generative protection but are not designed specifically to prevent identity recovery through few-shot face personalization.

Face anti-personalization defenses can be organized by the training signal they intervene on and the fidelity constraints imposed on released images. Denoising-oriented methods include Anti-DreamBooth (Le et al., 2023) that perturbs the denoising supervision used by DreamBooth-style fine-tuning, SDS-based protection (Xue et al., 2024) that adapts score-distillation objectives, and SimAC (Wang et al., 2024) that further introduces adaptive timestep selection to emphasize influential diffusion states. Attention-based defenses instead intervene in subject-condition association: Anti-Diffusion (Li et al., 2025) suppresses broad cross-attention responses, DisDiff (Liu et al., 2024b) targets a predefined identifier position such as `sks`, and CAAT (Xu et al., 2024) disrupts attention-mediated concept association. Fidelity-oriented methods, including MYOPIA (Wu et al., 2025) and VCPro (Mi et al., 2026), improve the visual quality of protected images by reducing perceptible distortion, as does InMark (Liu et al., 2024a) by concentrating protection on influential pixels. Other defenses address attacker-side variability: MetaCloak (Liu et al., 2024c) uses meta-learning, PAP (Wan et al., 2024) removes the dependence on the inference prompt, and AntiPure (Yang et al., 2025) targets purification-resistant perturbations.

These defenses improve individual components of protection, but none controls the three choices jointly: supervision is drawn from timesteps that change across iterations, conditioning positions are either all suppressed or fixed in advance, and the spatial update is typically spread uniformly under a sign-based rule. Each choice can spend part of the budget on components that are temporally inconsistent or weakly related to the objective. VEIL addresses the three choices jointly through fixed ultra-early anchors with posterior-mean encoding, attribution-guided conditioning-position selection, and saliency-weighted shrinkage updates.

3 Preliminaries and Problem Formulation

3.1 Latent Diffusion and Personalization

We consider a pretrained latent diffusion model with a VAE encoder $\mathcal{E}_\phi$ and a conditional denoising network f_{θ_0}. Let $\mathcal{X} = [-1, 1]^{H\times W\times C}$ denote the normalized image space. For $x \in \mathcal{X}$, the encoder parameterizes a latent posterior $q_\phi(z_0 \mid x)$, and the forward diffusion process constructs

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon, \quad z_0 \sim q_\phi(\cdot \mid x), \quad \varepsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $t \in \{0, \dots, T-1\}$ and $\bar{\alpha}_t$ is determined by the noise schedule. Under text condition c, the denoising network is trained with

$$\mathcal{L}_{\text{diff}}(\theta) = \mathbb{E}_{x, z_0, t, \varepsilon, c}\left[\|f_\theta(z_t, t, c) - y_t\|_2^2\right], \tag{2}$$

where y_t is the target determined by the model prediction parameterization.

Let $\mathcal{D} = \{x_i\}_{i=1}^n$ denote a reference set depicting a target identity, and let c_{id} contain a subject identifier. A personalization algorithm $\mathcal{A}$ produces

$$\theta_{\mathcal{D}} = \mathcal{A}\big(\theta_0, \mathcal{D}, c_{\text{id}}\big), \tag{3}$$

where $\theta_{\mathcal{D}}$ may include fully fine-tuned weights, parameter-efficient adapters, or learned token embeddings. Given an inference prompt p, the personalized generator samples $\hat{x} \sim G_{\theta_{\mathcal{D}}}(\cdot \mid p)$. We use $S_{\mathrm{id}}(\hat{x}, \mathcal{D})$ to denote an identity-recovery score, with larger values indicating stronger recovery of the identity represented by $\mathcal{D}$.

3.2 Threat Model

Attacker. The attacker collects publicly released face images and uses a personalization algorithm, as defined in Eq. (3), to adapt a text-to-image diffusion model. The attacker controls the personalization procedure, training configuration, subject identifier, and inference prompts. We consider an oblivious attacker who applies an existing personalization procedure to the released images without tailoring it to the protection mechanism.

Defender. The defender can modify only the images before release and has no access to the attacker's subsequent personalization process or adapted model. The defender has white-box access only to a surrogate model, which need not match the attacker's model or personalization setup.

3.3 Problem Formulation

The feasible protected reference sets are

$$\mathcal{B}_\epsilon(\mathcal{D}) = \left\{ \tilde{\mathcal{D}} = \{\tilde{x}_i\}_{i=1}^n \;\middle|\; \tilde{x}_i = \Pi_{\mathcal{X}}(x_i + \delta_i),\ \|\delta_i\|_\infty \le \epsilon,\ \forall i \right\}, \tag{4}$$

where $\Pi_{\mathcal{X}}$ projects onto the valid image space. The attacker personalizes on $\tilde{\mathcal{D}} \in \mathcal{B}_\epsilon(\mathcal{D})$:

$$\theta_{\tilde{\mathcal{D}}} = \mathcal{A}\big(\theta_0, \tilde{\mathcal{D}}, c_{\mathrm{id}}\big). \tag{5}$$

For a fixed attacker configuration, let $\mathcal{P}$ denote the distribution over inference prompts. The ideal protection objective is

$$\tilde{\mathcal{D}}^* = \arg\min_{\tilde{\mathcal{D}} \in \mathcal{B}_\epsilon(\mathcal{D})} \mathbb{E}_{\substack{p \sim \mathcal{P}, \\ \hat{x} \sim G_{\theta_{\tilde{\mathcal{D}}}}(\cdot|p)}} \left[S_{\mathrm{id}}\big(\hat{x}, \mathcal{D}\big) \right]. \tag{6}$$

Direct optimization of Eq. (6) requires differentiating through the attacker's personalization and generation processes. We therefore optimize a pre-personalization proxy based on the surrogate denoising objective, which is directly coupled to the training signal used during downstream personalization. To make this proxy reliable under a bounded budget, VEIL stabilizes its temporal support and selectively applies the gradient across conditioning positions and image regions.

4 Method

4.1 Overview

As illustrated in Figure 2, VEIL organizes protection optimization into three coupled stages. Fixed ultra-early anchors and posterior-mean latent encoding first provide consistent denoising supervision across perturbation iterations (②). This stabilized objective supplies the attribution signal used to identify the conditioning positions that contribute most strongly to denoising disruption, and the selected positions are further suppressed through the cross-attention loss (③). The denoising and attention objectives are then combined across anchors to form the joint protection gradient. Finally, saliency weighting and soft shrinkage convert this gradient into a spatially selective perturbation that prioritizes high-contribution regions while removing small accumulated components (④).

4.2 Temporal Stabilization

Fixed ultra-early anchors. We use *anchor* to denote a diffusion timestep at which the denoising objective is evaluated during perturbation optimization. The noise level, latent structure, and denoising signal vary across diffusion timesteps, making the perturbation gradient timestep-dependent. Changing the supervised timestep across iterations therefore changes the diffusion state from which

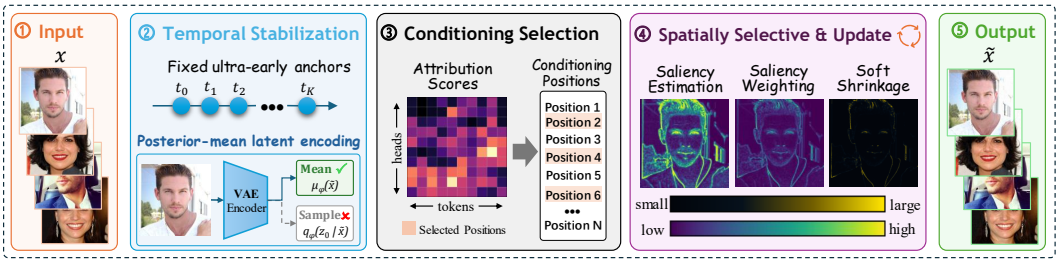


Figure 2: Overview of VEIL. The framework comprises temporal stabilization (②), attribution-guided conditioning-position selection (③), and a spatially selective perturbation update (④).

each update is computed. Let $\mathcal{T} \subset \{0, \ldots, T-1\}$ denote the fixed set of ultra-early denoising anchors, taken from the smallest timesteps of the schedule. At these low-noise states, the noisy latent retains substantial input-specific structure. They are also the states at which identity is resolved: denoising recovers identity-irrelevant content at high noise levels and restores identity-relevant detail at low noise levels (Lin et al., 2025). Reusing the same anchor set across iterations maintains consistent temporal support, while averaging across anchors reduces dependence on any single timestep.

Posterior-mean encoding. Sampling from the VAE posterior introduces gradient variability in addition to the diffusion noise. During perturbation optimization, VEIL uses the posterior mean:

$$\tilde{z}_0 = \mu_\phi(\tilde{x}), \tag{7}$$

where $\mu_\phi(\tilde{x})$ is the mean of $q_\phi(z_0 \mid \tilde{x})$. For each $t \in \mathcal{T}$, we sample a fresh $\varepsilon \sim \mathcal{N}(0, I)$ and form

$$\tilde{z}_t = \sqrt{\bar{\alpha}_t}\, \tilde{z}_0 + \sqrt{1 - \bar{\alpha}_t}\, \varepsilon. \tag{8}$$

The fixed anchors and posterior-mean encoding remove distinct sources of variation in the protection gradient. The anchors keep the supervised timesteps unchanged across iterations, whereas posterior-mean encoding fixes the latent representation before diffusion noise is added. Their combination makes successive updates more comparable while retaining multi-timestep supervision.

Denoising disruption. Let f_{θ_s} denote the surrogate denoiser, and let c_{sur} denote the empty-prompt embedding used as a fixed surrogate condition. The denoising-disruption loss at anchor t is

$$\mathcal{L}_{\text{denoise}}(t) = \mathbb{E}_\varepsilon \left[\left\| f_{\theta_s}\left(\tilde{z}_t, t, c_{\text{sur}}\right) - y_t \right\|_2^2 \right], \tag{9}$$

where y_t is the target determined by the diffusion prediction parameterization. At each perturbation iteration, the expectation is estimated using one fresh noise draw per anchor. VEIL maximizes $\mathcal{L}_{\text{denoise}}(t)$ with respect to the protected image, thereby increasing the surrogate denoising error induced by that image.

4.3 Conditioning-Position Selection

Although c_{sur} is obtained from the empty prompt, its encoded sequence contains N conditioning positions and produces nonuniform cross-attention responses. These positions contribute differently to the denoising-disruption objective. VEIL therefore scores each position by the absolute value of its first-order contribution and selects the highest-scoring positions for cross-attention suppression.

Let $A^{(\ell,h)}(t) \in \mathbb{R}^{U \times N}$ denote a cross-attention map at anchor t, where ℓ and h index the attention layer and head, respectively, and U is the number of spatial positions at the selected attention resolution. Let $\mathcal{J}$ denote the layer–head pairs used for attribution. For conditioning position i, we define

$$s_i(t) = \frac{1}{|\mathcal{J}|U} \sum_{(\ell,h)\in\mathcal{J}} \sum_{u=1}^{U} \left| A_{u,i}^{(\ell,h)}(t) \frac{\partial \mathcal{L}_{\text{denoise}}(t)}{\partial A_{u,i}^{(\ell,h)}(t)} \right|. \tag{10}$$

This gradient-times-activation score estimates the local first-order contribution of each conditioning position to the denoising-disruption objective in absolute magnitude, adapting the component-attribution formulation of CAD (Nguyen et al., 2025) from model parameters to conditioning positions. Using the denoising objective for attribution keeps conditioning-position selection aligned with the current protection signal rather than with a predefined identifier position.

Let $k = \max(1, \lfloor \rho N \rfloor)$, where $\rho \in (0, 1]$ is the selection ratio. The corresponding binary mask is

$$M_{\mathrm{cond},i}(t) = \mathbb{I}\left[i \in \mathrm{Top}_k\left(\{s_j(t)\}_{j=1}^{N}\right)\right]. \tag{11}$$

The attribution scores and mask are recomputed from the current protected image and detached before evaluating the perturbation gradient, so no gradient is propagated through the top-k selection.

We aggregate the cross-attention maps over $\mathcal{J}$ and bound their values as

$$\bar{A}(t) = \mathrm{clip}\left(\frac{\kappa}{|\mathcal{J}|}\sum_{(\ell,h)\in\mathcal{J}} A^{(\ell,h)}(t), 0, 1\right), \tag{12}$$

where $\kappa > 0$ is a fixed scaling factor. The clip bounds the contribution of each response so that a few strong activations cannot dominate the loss. The attribution-guided attention loss is

$$\mathcal{L}_{\mathrm{attn}}(t) = -\frac{\left\|\bar{A}(t) \odot M_{\mathrm{cond}}(t)\right\|_F^2}{\max(U \left\|M_{\mathrm{cond}}(t)\right\|_1, 1)}. \tag{13}$$

Here, $M_{\mathrm{cond}}(t)$ is broadcast over spatial positions. Because $\mathcal{L}_{\mathrm{attn}}(t) \le 0$, maximizing it drives the cross-attention energy at the selected conditioning positions toward zero.

4.4 Joint Protection Objective

The complete protection objective combines denoising disruption and attribution-guided attention suppression across the fixed anchor set:

$$\mathcal{L}_{\mathrm{VEIL}} = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\left[\mathcal{L}_{\mathrm{denoise}}(t) + \lambda_{\mathrm{attn}}\mathcal{L}_{\mathrm{attn}}(t)\right], \tag{14}$$

where $\lambda_{\mathrm{attn}} \ge 0$ controls the attention-suppression term.

4.5 Spatially Selective Update

A conventional sign update assigns the same step magnitude to every nonzero gradient component and therefore discards their relative magnitudes. Under a bounded perturbation budget, this may allocate updates to pixels that contribute little to the joint protection objective. VEIL instead derives a spatial weight from the joint gradient and applies soft shrinkage to the accumulated perturbation.

Let $\delta^{(r)} = \tilde{x}^{(r)} - x$ denote the perturbation at iteration r, with $\delta^{(0)} = 0$. The gradient of the joint objective and its channel-averaged magnitude are

$$g^{(r)} = \nabla_\delta \mathcal{L}_{\mathrm{VEIL}}\left(\delta^{(r)}\right), \quad m^{(r)} = \mathrm{mean}_{\mathrm{ch}}\left(\left|g^{(r)}\right|\right). \tag{15}$$

We convert $m^{(r)}$ into a bounded spatial weight:

$$w^{(r)} = \left[\mathrm{clip}\left(\frac{m^{(r)}}{\mathrm{Quantile}\left(m^{(r)}, q_{\mathrm{sal}}\right) + 10^{-12}}, 0, 1\right)\right]^{\gamma}, \tag{16}$$

where $q_{\mathrm{sal}} \in (0, 1)$ is the reference quantile and $\gamma > 0$ controls the concentration of the spatial weights. The weight map is broadcast over image channels.

The element-wise soft-shrinkage operator is

$$\mathrm{Shrink}_\tau(a) = \mathrm{sign}(a) \odot \max(|a| - \tau, 0). \tag{17}$$

The perturbation update is

$$\begin{aligned} \bar{\delta}^{(r+1)} &= \Pi_{[-\epsilon,\epsilon]}\left(\mathrm{Shrink}_\tau\left(\delta^{(r)} + \alpha w^{(r)} \odot \mathrm{sign}\left(g^{(r)}\right)\right)\right), \\ \tilde{x}^{(r+1)} &= \Pi_{\mathcal{X}}\left(x + \bar{\delta}^{(r+1)}\right), \quad \delta^{(r+1)} = \tilde{x}^{(r+1)} - x, \end{aligned} \tag{18}$$

where α is the ascent step size and $\tau = \tau_s \epsilon$ is the shrinkage threshold. Saliency weighting scales the sign update according to the relative spatial gradient magnitude. Soft shrinkage sets sufficiently small accumulated components to zero and contracts the remaining components without imposing a hard spatial mask. Together, these operations concentrate the bounded perturbation on spatial components that contribute more strongly to the joint protection objective.

Alternating optimization. For each reference image, the surrogate is initialized from the pre-trained model as $\theta_s = \theta_0$. After an initial block of R perturbation iterations, we perform E alternating rounds of surrogate adaptation and perturbation optimization. In each round, the surrogate is updated for S steps on the current protected image by minimizing Eq. (2) under c_{sur}, then fixed for the next R perturbation iterations. Algorithm 1 in Appendix A provides the complete procedure.

5 Evaluation

5.1 Setup

Datasets. We evaluate VEIL on CelebA-HQ (Karras et al., 2018) and VGGFace2 (Cao et al., 2018) using the identity splits and preprocessing protocol of Anti-DreamBooth (Le et al., 2023). Each dataset contains 50 identities, with four 512×512 reference images per identity used for protection and downstream personalization. The main DreamBooth (DB) evaluations use all 50 identities from each dataset, while the remaining analyses use the first 10 identities from each evaluated dataset.

Baselines. We compare VEIL with Anti-DreamBooth (Anti-DB) (Le et al., 2023), score-distillation-based protection (SDS-) (Xue et al., 2024), VCPro (Mi et al., 2026), MYOPIA (Wu et al., 2025), Anti-Diffusion (Anti-Diff) (Li et al., 2025), CAAT (Xu et al., 2024), DisDiff (Liu et al., 2024b), and SimAC (Wang et al., 2024). All methods use their official implementations; additive pixel-space methods use an ℓ_∞ budget of $\epsilon = 5 \times 10^{-2}$, while MYOPIA keeps its EOT-based default of $\epsilon = 5 \times 10^{-1}$.

Evaluation metrics. Released-image fidelity is measured by SSIM (↑) (Wang et al., 2004), PSNR (↑), and LPIPS (↓) (Zhang et al., 2018). For personalization outputs, FDFR (↑) measures the fraction of generated images in which RetinaFace (Deng et al., 2020) detects no face, ISM (↓) measures ArcFace (Deng et al., 2019) similarity to the clean reference identity, SER-FIQ (Terhörst et al., 2020) (↓) measures the recognition utility of the generated face, and CLIP-F (↓) follows VCPro (Mi et al., 2026) and scores face quality with CLIP (Radford et al., 2021). ISM and SER-FIQ are averaged over the generations in which a face is detected. Detailed metric definitions and aggregation rules are provided in Appendix B.2.

Implementation details. Protected images are generated using Stable Diffusion 2.1-base (SD v2.1) (Rombach et al., 2022) with an empty-prompt surrogate and fixed anchors $\mathcal{T} = \{0, 25, 50, 75, 100, 125\}$. Protection runs $R = 16$ iterations, followed by $E = 6$ alternating rounds of surrogate adaptation and protection optimization. We set $\rho = 0.6$, $\lambda_{\text{attn}} = 1.75$, and random seed 42. For downstream evaluation, SD v2.1 is personalized with DreamBooth for 1,000 steps using *"a photo of sks person"*, learning rate 5×10^{-7}, and batch size 2. Each model generates 32 images per identity under P1–P6. P1 is held out from personalization, P2 matches the training prompt, and P3–P6 introduce contextual variations. Table 1 reports P1 and P2, while Appendix C.1 reports P3–P6. In all tables, **bold** and underlined values mark the best and second-best results among the compared methods. Protection hyperparameters, personalization and generation protocols, and baseline, robustness, and transfer settings are provided in Table 4, Appendix B.4, and B.5, respectively.

5.2 Main Results

Quantitative results. In Table 1, ORG denotes personalization on unprotected images and serves as the identity-recovery reference. VCPro and MYOPIA achieve low LPIPS, but their FDFR is at most 0.02 in all four settings. SDS- incurs substantially greater released-image distortion while retaining near-zero FDFR. Anti-DB, Anti-Diff, CAAT, DisDiff, and SimAC achieve stronger suppression, but their released-image fidelity is consistently lower than that of VEIL. VEIL obtains the highest SSIM (0.94) and PSNR (37.94/38.21 dB) on both datasets. It also achieves the lowest LPIPS on CelebA-HQ (0.12) and the second-lowest LPIPS on VGGFace2 (0.14). For post-personalization suppression, VEIL achieves or ties for the highest FDFR in all four settings, reaching 0.85 and 0.83 on CelebA-HQ, and 0.84 and 0.73 on VGGFace2. It ranks first or second for SER-FIQ and CLIP-F in every setting. For ISM, VEIL ranks first or second in three settings, while CAAT is lower on CelebA-HQ under P2 (0.30 versus 0.34). VEIL is the only evaluated defense that attains both the highest SSIM and PSNR and the highest or tied-highest FDFR.

Table 1: Released-image fidelity and post-personalization suppression under DreamBooth (P1, P2).

Dataset	Method	Protected-image fidelity			Prompt **P1**				Prompt **P2**			
		SSIM ↑	PSNR ↑	LPIPS ↓	SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓	SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓
CelebA-HQ	ORG	1.00	80.00	0.00	0.73	0.02	0.47	0.34	0.75	0.00	0.71	0.56
	Anti-DB	0.90	34.34	0.20	0.40	0.45	0.24	0.26	0.31	0.40	0.35	0.39
	SDS-	0.74	29.81	0.30	0.77	0.00	0.52	0.35	0.76	0.00	0.62	0.49
	VCPro	0.92	37.40	0.13	0.73	0.02	0.42	0.22	0.72	0.01	0.59	0.36
	MYOPIA	0.91	33.83	0.13	0.77	0.02	0.49	0.37	0.70	0.00	0.51	0.29
	Anti-Diff	0.93	35.70	0.18	0.45	0.36	0.27	0.29	0.19	0.55	0.32	0.36
	CAAT	0.91	34.46	0.20	0.32	0.47	0.26	0.26	0.16	0.61	**0.30**	0.32
	DisDiff	0.92	35.04	0.15	0.22	0.79	0.21	0.18	0.20	0.82	0.34	0.34
	SimAC	0.92	35.03	0.15	0.22	0.79	**0.18**	**0.17**	0.24	0.74	0.34	0.34
	VEIL	**0.94**	**37.94**	**0.12**	**0.16**	**0.85**	0.20	0.18	**0.15**	**0.83**	0.34	**0.26**
VGGFace2	ORG	1.00	80.00	0.00	0.72	0.01	0.49	0.32	0.75	0.01	0.67	0.48
	Anti-DB	0.90	34.52	0.21	0.40	0.43	0.27	0.14	0.35	0.32	0.35	0.31
	SDS-	0.74	29.87	0.31	0.77	0.00	0.54	0.33	0.75	0.00	0.62	0.43
	VCPro	0.90	36.40	0.16	0.73	0.02	0.44	0.12	0.71	0.01	0.59	0.30
	MYOPIA	0.92	34.35	**0.11**	0.77	0.02	0.51	0.34	0.72	0.00	0.55	0.33
	Anti-Diff	0.92	36.01	0.20	0.43	0.34	0.30	0.18	0.24	0.46	0.34	0.27
	CAAT	0.90	34.65	0.21	0.32	0.51	0.26	0.11	0.24	0.49	0.32	0.25
	DisDiff	0.91	35.23	0.17	0.18	0.83	0.22	0.11	0.29	0.64	0.37	0.26
	SimAC	0.91	35.20	0.17	**0.17**	**0.84**	**0.19**	**0.09**	0.26	0.67	0.38	0.26
	VEIL	**0.94**	**38.21**	0.14	0.18	**0.84**	0.22	0.10	**0.22**	**0.73**	**0.30**	**0.16**

Table 2: Robustness on CelebA-HQ under preprocessing and purification, averaged over P1 and P2.

Method	JPEG-Q70			GBlur-K5			DiffPure			GrIDPure		
	FDFR ↑	ISM ↓	CLIP-F ↓	FDFR ↑	ISM ↓	CLIP-F ↓	FDFR ↑	ISM ↓	CLIP-F ↓	FDFR ↑	ISM ↓	CLIP-F ↓
ORG	0.02	0.57	0.42	0.04	0.57	0.37	0.02	0.53	0.45	0.03	0.52	0.44
Anti-DB	0.50	0.29	0.31	0.14	**0.46**	0.30	0.02	0.54	0.44	0.03	0.52	0.41
SDS-	0.00	0.58	0.44	0.06	0.51	0.28	0.03	0.54	0.43	0.03	0.51	0.38
VCPro	0.01	0.49	0.26	0.02	0.53	0.31	0.04	0.53	0.42	0.04	0.53	0.40
MYOPIA	0.03	0.50	0.34	0.04	0.49	**0.26**	0.04	0.55	0.44	**0.05**	0.54	**0.36**
Anti-Diff	0.51	0.32	0.31	0.16	0.48	0.33	0.03	0.54	0.42	0.04	0.53	0.40
CAAT	0.62	0.30	0.29	0.07	0.50	0.38	**0.05**	0.50	0.43	0.04	**0.48**	0.40
DisDiff	0.88	**0.21**	0.22	0.07	0.52	**0.26**	0.04	0.53	0.43	0.04	0.54	0.40
SimAC	0.82	0.29	0.27	0.06	0.54	0.31	0.03	0.54	0.44	0.04	0.54	0.38
VEIL	**0.91**	0.24	**0.18**	**0.17**	**0.46**	0.28	0.04	**0.45**	**0.41**	0.03	0.50	**0.36**

Qualitative results. Appendix Figures 5 and 6 present qualitative comparisons of the released images and personalized generations, respectively. The protected images remain visually close to their clean references, while models trained on them generate faces with weaker coherence and identity resemblance across portrait and contextual prompts (P1–P6).

5.3 Robustness

We evaluate robustness on CelebA-HQ under JPEG compression, Gaussian blur, DiffPure (Nie et al., 2022), and GrIDPure (Zhao et al., 2024) (Table 2). Across all four transformations, VEIL ranks first or second on ISM and CLIP-F. The transformations differ mainly in how much of the FDFR gain they remove. JPEG compression leaves most of it intact, and VEIL retains the highest FDFR (0.91) and the lowest CLIP-F (0.18). Gaussian blur lowers FDFR to at most 0.17 for every method, while VEIL keeps the highest FDFR and ties for the lowest ISM. Diffusion-based purification removes most of the FDFR gain for all evaluated methods. Relative to ORG, however, VEIL still lowers ISM under DiffPure (0.45 versus 0.53), whereas the other methods remain between 0.50 and 0.55.

5.4 Transferability

All transfer evaluations reuse the same protected images without re-optimization, thereby isolating protection transfer across attacker-side configurations.

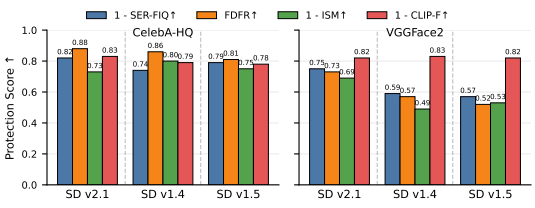


Figure 3: Transfer from SD v2.1 to SD v1.4 and SD v1.5 under P2.

Across diffusion backbones. We generate protected images using the SD v2.1 surrogate and directly reuse them to personalize SD v1.4 and SD v1.5. Figure 3 reports results under P2. On CelebA-HQ, FDFR changes only slightly from the

SD v2.1 source (0.88) to SD v1.4 (0.86) and SD v1.5 (0.81). Prompt-level comparisons are reported in Table 6 in Appendix D.1.

Across personalization procedures. We reuse the same protected images under DreamBooth-LoRA (LoRA), Textual Inversion (TI), and SVDiff. Figure 4(a) reports the results on CelebA-HQ under P2. VEIL achieves FDFR values of 0.94, 0.62, and 0.86 under DreamBooth-LoRA, Textual Inversion, and SVDiff, respectively. Transfer is weakest under Textual Inversion, which learns only a token embedding and leaves the denoiser unchanged, whereas DreamBooth-LoRA and SVDiff adapt the denoiser against which the protection is optimized. Full prompt-level comparisons are reported in Table 7 in Appendix D.2.

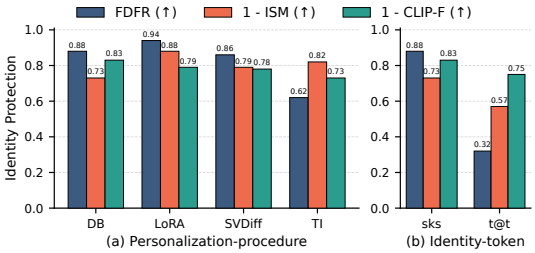


Figure 4: Transfer across personalization procedures (a) and identifier bindings (b) under P2.

Across identifier bindings. We replace `sks` with `t@t` in both the DreamBooth training and inference prompts while keeping the protected images unchanged. Figure 4(b) reports the results on CelebA-HQ under P2. Without re-optimization, VEIL achieves an FDFR of 0.32, an ISM of 0.43, and a CLIP-F score of 0.25 under the alternative identifier. Full P1 and P2 results are provided in Table 8 in Appendix D.3.

5.5 Ablation Studies

Core components. Table 3 reports the component ablation averaged over both datasets and the two primary evaluation prompts. Adding the attention-suppression loss and conditioning-position selection progressively improves FDFR, CLIP-F, and LPIPS. Replacing posterior-mean encoding with posterior sampling degrades both FDFR and CLIP-F. Removing the spatially selective update yields the highest FDFR but substantially worsens LPIPS and CLIP-F, because sign updates then spread the perturbation over all pixels. The complete model achieves the lowest CLIP-F while retaining the second-best LPIPS and FDFR among the evaluated variants.

Table 3: Module ablation of VEIL. *SSP*: spatially selective perturbation update; VAE_μ: posterior-mean latent encoding; *Attn.*: attention-suppression loss; *Cond.*: attribution-guided conditioning-position selection.

SSP	VAE_μ	*Attn.*	*Cond.*	LPIPS↓	FDFR↑	CLIP-F↓
✓	✓	✗	✗	0.1436	0.7563	0.1548
✓	✓	✓	✗	0.1340	0.8038	<u>0.1382</u>
✗	✓	✓	✓	0.1628	**0.8623**	0.1416
✓	✗	✓	✓	**0.1332**	0.7695	0.1566
✓	✓	✓	✓	<u>0.1338</u>	<u>0.8330</u>	**0.1291**

Number of ultra-early anchors. Figure 7 in Appendix E.1 evaluates the effect of the number of anchors. Protection improves up to the six-anchor configuration, whereas additional anchors provide no clear gain and increase LPIPS. We therefore use six anchors by default. Moving the anchors to intermediate or late timesteps lowers P2 FDFR on CelebA-HQ from 0.88 to 0.48 and 0.08, and random placement also underperforms (Table 9).

Appendix E further shows that surrogate adaptation surpasses a static surrogate at equal iterations (Table 10) and that $\lambda_{\text{attn}} = 1.75$ balances fidelity and suppression (Figure 10).

6 Conclusion

To mitigate unauthorized identity synthesis from public face images, we proposed VEIL, a user-side defense suppressing identity recovery under diffusion personalization while preserving released-image fidelity. VEIL integrates temporally stabilized denoising supervision, attribution-guided conditioning-position selection, and spatially selective perturbation refinement. Across CelebA-HQ and VGGFace2, VEIL attains the highest SSIM and PSNR and achieves or ties for the highest FDFR in all four primary dataset–prompt settings, while transferring across Stable Diffusion variants and personalization procedures without re-optimization. **Future work** will improve transferability across diverse personalization configurations, strengthen robustness to purification-based preprocessing, and reduce optimization cost via efficient multi-timestep gradient aggregation.

AI USE STATEMENT

Large language models were used only for language editing, such as grammar correction and stylistic polishing. They were not used for research ideation, experimental design, data analysis, or generating any scientific content. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work follows the ICLR Code of Ethics. VEIL is a user-side defense that lets individuals protect their publicly shared face photos against unauthorized text-to-image personalization. All experiments use the publicly available CelebA-HQ and VGGFace2 benchmarks under the identity splits released by Anti-DreamBooth (Le et al., 2023), and no additional human-subject data were collected. Generated images are used solely for evaluation and are not distributed.

REPRODUCIBILITY STATEMENT

The method is described in Section 4, with the full procedure in Algorithm 1 (Appendix A). The experimental setup, including datasets, baselines, and metrics, is given in Section 5.1, and the complete hyperparameters, metric definitions, and generation, baseline, robustness, and transfer protocols are provided in Appendix B. All datasets are public, and all experiments use a fixed random seed of 42. An anonymized reference implementation, covering protected-image generation, the downstream DreamBooth pipeline, and sample data, is available at `https://anonymous.4open.science/r/VEAL_demo`.

REFERENCES

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A Dataset for Recognising Faces Across Pose and Age. In *FG*, pp. 67–74, 2018.

Linyan Dai, Xinwei Zhang, Haoyang Li, Qingqing Ye, and Haibo Hu. IDDM: Identity-decoupled personalized diffusion models with a tunable privacy-utility trade-off. *arXiv preprint arXiv:2604.00903*, 2026.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, pp. 4690–4699, 2019.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *CVPR*, pp. 5203–5212, 2020.

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 *ICLR*, 2023.

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. SVDiff: Compact parameter space for diffusion fine-tuning. In *ICCV*, 2023.

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 *ICLR*, 2022.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, pp. 1931–1941, 2023.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N. Tran, and Anh Tuan Tran. Anti-DreamBooth: Protecting Users from Personalized Text-to-Image Synthesis. In *ICCV*, pp. 2116–2127, 2023.

Zheng Li, Liangbin Xie, Jiantao Zhou, Xintao Wang, Haiwei Wu, and Jinyu Tian. Anti-Diffusion: Preventing abuse of modifications of diffusion-based models. In *AAAI*, pp. 10582–10590, 2025.

Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *arXiv*, 2023.

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 *ICML*, pp. 20763–20786, 2023.

Xiao Lin, Yuge Huang, Jianqing Xu, Yuxi Mi, Shuigeng Zhou, and Shouhong Ding. UIFace: Unleashing inherent model capabilities to enhance intra-class diversity in synthetic face recognition. In *ICLR*, 2025.

Hanwen Liu, Zhicheng Sun, and Yadong Mu. Countering personalized text-to-image generation with influence watermarks. In *CVPR*, pp. 12257–12267, 2024a.

Yisu Liu, Jinyang An, Wanqian Zhang, Dayan Wu, Jingzi Gu, Zheng Lin, and Weiping Wang. Disrupting diffusion: Token-level attention erasure attack against diffusion-based customization. In *MM*, pp. 3587–3596, 2024b.

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 *CVPR*, pp. 24219–24228, 2024c.

Xiaoyue Mi, Fan Tang, You Wu, Juan Cao, Peng Li, and Yang Liu. Visual-friendly concept protection via selective adversarial perturbations. In *AAAI*, pp. 39033–39041, 2026.

Quang H Nguyen, Hoang Phan, and Khoa D Doan. Unveiling concept attribution in diffusion models. In *NeurIPS*, 2025.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *ICML*, pp. 16805–16827, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 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 *CVPR*, 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 *ICML*, 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 *USENIX Security*, pp. 2187–2204, 2023.

Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *CVPR*, pp. 5651–5660, 2020.

Cong Wan, Yuhang He, Xiang Song, and Yihong Gong. Prompt-agnostic adversarial perturbation for customized diffusion models. In *NeurIPS*, 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 *CVPR*, pp. 12047–12056, 2024.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, pp. 600–612, 2004.

Zhihao Wu, Yushi Cheng, Tianyang Sun, Xiaoyu Ji, and Wenyuan Xu. MYOPIA: protecting face privacy from malicious personalized text-to-image synthesis via unlearnable examples. In *AAAI*, pp. 905–913, 2025.

Jingyao Xu, Yuetong Lu, Yandong Li, Siyang Lu, Dongdong Wang, and Xiang Wei. Perturbing attention gives you more bang for the buck: Subtle imaging perturbations that efficiently fool customized diffusion models. In *CVPR*, pp. 24534–24543, 2024.

Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion based mimicry through score distillation. In *ICLR*, 2024.

Wenkui Yang, Jie Cao, Junxian Duan, and Ran He. Towards robust defense against customization via protective perturbation resistant to diffusion-based purification. In *ICCV*, pp. 19290–19300, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by stable diffusion? In *CVPR*, pp. 24398–24407, 2024.

APPENDIX OVERVIEW

This appendix provides the algorithmic, experimental, and diagnostic details supporting the main paper:

- Appendix A presents the complete alternating optimization procedure of VEIL, including fixed-anchor protection optimization, attribution-guided conditioning-position selection, the spatially selective update, and surrogate adaptation.
- Appendix B details the data splits, aggregation rules, metric definitions, protection configuration, downstream personalization settings, baseline configurations, and robustness and transfer protocols.
- Appendix C reports results on the contextual prompts P3–P6 and qualitative comparisons of released images and personalized generations.
- Appendix D reports prompt-level transfer results across Stable Diffusion variants, personalization procedures, and identifier bindings.
- Appendix E presents additional ablation and diagnostic analyses of anchor count and placement, timestep selection, optimization stability, surrogate adaptation, and sensitivity to the attention weight and conditioning ratio.

A ALGORITHM

The regression target in Eq. (9) follows the prediction type of the surrogate scheduler. SD v2.1-base uses ε-prediction, so the target is the sampled Gaussian noise and carries no gradient with respect to the perturbation.

Algorithm 1 expands the alternating procedure summarized in Section 4. An initial protection block is followed by E rounds of surrogate adaptation and protection optimization. Within each protection iteration, every anchor in $\mathcal{T}$ is evaluated with an independent diffusion-noise draw, and the averaged gradient is transformed by the spatially selective update.

B EXPERIMENTAL SETUP

B.1 DATASETS AND AGGREGATION

We use the preprocessed identity splits released by Anti-DreamBooth (Le et al., 2023), comprising 50 identities from CelebA-HQ (Karras et al., 2018) and 50 identities from VGGFace2 (Cao et al., 2018). For each identity, four images from Set_B are resized to 512×512 and used for protection and downstream personalization. The DreamBooth comparison uses all 50 identities from each dataset, whereas the robustness, transfer, and ablation analyses use the first 10 identities from each evaluated dataset.

All metrics are first averaged over the images associated with each identity under a fixed prompt and experimental setting, and are then macro-averaged across identities. Each identity therefore contributes equally to the reported dataset-level result. Prompt-level tables report each prompt separately, whereas results described as averaged over multiple prompts assign equal weight to the corresponding prompt-level identity averages.

B.2 METRIC DEFINITIONS

The evaluation separates fidelity of the released protected images from the outcomes produced after downstream personalization.

B.2.1 RELEASED-IMAGE FIDELITY

For each clean-protected pair $(x_{i,j}, \tilde{x}_{i,j})$, we compute SSIM (Wang et al., 2004), PSNR, and LPIPS (Zhang et al., 2018). SSIM and PSNR measure structural and pixel-level preservation, respectively, whereas LPIPS measures perceptual dissimilarity. Higher SSIM and PSNR and lower LPIPS

Algorithm 1 Complete protection procedure of VEIL

Require: Reference set $\mathcal{D} = \{x_i\}_{i=1}^{n}$; pretrained denoiser f_{θ_0}; anchor set $\mathcal{T}$; budget ϵ; step size α; adaptation rounds E; surrogate steps per round S; protection iterations per block R; hyperparameters $\lambda_{\text{attn}}, \rho, \kappa, \tau_s, q_{\text{sal}}, \gamma$.
Ensure: Protected set $\tilde{\mathcal{D}}$.

$\tilde{\mathcal{D}} \leftarrow \varnothing$.
for each $x_i \in \mathcal{D}$ **do**
  Copy the pretrained surrogate: $\theta_s \leftarrow \theta_0$; set c_{sur} to the empty-prompt embedding.
  Initialize $\delta_i \leftarrow 0$ and $\tilde{x}_i \leftarrow x_i$.
  for $e = 0$ to E **do**
    if $e > 0$ **then**
      for $s = 1$ to S **do**
        Sample a VAE posterior latent, a timestep uniformly from the full diffusion trajectory, and Gaussian noise.
        Update θ_s by minimizing Eq. (2) on the current $\tilde{x}_i$ under c_{sur}.
      end for
    end if
    for $r = 1$ to R **do**
      Initialize $g_i \leftarrow 0$.
      for each $t \in \mathcal{T}$ **do**
        Sample $\varepsilon \sim \mathcal{N}(0, I)$, encode $\tilde{z}_0 = \mu_\phi(\tilde{x}_i)$, and construct $\tilde{z}_t$ using Eq. (8).
        Compute $\mathcal{L}_{\text{denoise}}(t)$ and retain the cross-attention maps indexed by $\mathcal{J}$ for attribution.
        Compute $M_{\text{cond}}(t)$ using Eqs. (10)–(11) and $\mathcal{L}_{\text{attn}}(t)$ using Eq. (13).
        $g_i \leftarrow g_i + \nabla_{\delta_i} [\mathcal{L}_{\text{denoise}}(t) + \lambda_{\text{attn}} \mathcal{L}_{\text{attn}}(t)]$.
      end for
      $g_i \leftarrow g_i / |\mathcal{T}|$.
      Compute the current spatial weight w_i from g_i using Eqs. (15)–(16), and update δ_i and $\tilde{x}_i$ using Eq. (18).
    end for
  end for
  $\tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \cup \{\tilde{x}_i\}$.
end for
return $\tilde{\mathcal{D}}$.

indicate better preservation. Pairwise scores are averaged within each identity before dataset-level macro-averaging.

B.2.2 Post-Personalization Outcomes

Let $\mathcal{G}_{i,p}$ denote the set of images generated for identity i under prompt p, and let $d(G) \in \{0, 1\}$ indicate whether RetinaFace (Deng et al., 2020) detects a face in G.

Face Detection Failure Rate. We define

$$\text{FDFR}_{i,p} = 1 - \frac{1}{|\mathcal{G}_{i,p}|} \sum_{G \in \mathcal{G}_{i,p}} d(G). \tag{19}$$

Higher FDFR indicates that fewer generated images contain a detectable face. Because FDFR measures generation validity rather than identity similarity, it is interpreted jointly with ISM and the face-quality diagnostics below.

Identity Similarity Metric. Let $e(\cdot)$ denote the ArcFace (Deng et al., 2019) embedding. We construct an identity prototype from the clean references $\mathcal{R}_i$:

$$\bar{e}_i = \frac{1}{|\mathcal{R}_i|} \sum_{x \in \mathcal{R}_i} \frac{e(x)}{\|e(x)\|_2}. \tag{20}$$

Let $\mathcal{G}_{i,p}^{+} = \{G \in \mathcal{G}_{i,p} \mid d(G) = 1\}$ denote the generated images in which a face is detected. The identity-prompt score is

$$\mathrm{ISM}_{i,p} = \begin{cases} \frac{1}{|\mathcal{G}_{i,p}^{+}|} \sum_{G \in \mathcal{G}_{i,p}^{+}} \cos(e(G), \bar{e}_i), & |\mathcal{G}_{i,p}^{+}| > 0, \\ 0, & |\mathcal{G}_{i,p}^{+}| = 0. \end{cases} \tag{21}$$

The average is taken over the detected generations only, and the zero value applies when no face is detected in any generation for that identity and prompt. Generations without a detectable face are accounted for separately by FDFR.

SER-FIQ. Let $q_{\mathrm{face}}(G)$ denote the SER-FIQ (Terhörst et al., 2020) score computed from the detected face crop. SER-FIQ follows the same aggregation convention as ISM:

$$\text{SER-FIQ}_{i,p} = \begin{cases} \frac{1}{|\mathcal{G}_{i,p}^{+}|} \sum_{G \in \mathcal{G}_{i,p}^{+}} q_{\mathrm{face}}(G), & |\mathcal{G}_{i,p}^{+}| > 0, \\ 0, & |\mathcal{G}_{i,p}^{+}| = 0. \end{cases} \tag{22}$$

As with ISM, the average is taken over the detected generations only. Higher SER-FIQ indicates greater recognition utility, whereas lower values indicate weaker usable face evidence.

CLIP-F. Following VCPro (Mi et al., 2026), we compute CLIP-F using CLIP ViT-B/32 (Radford et al., 2021) and the prompts `good face` and `bad face`. For each generated image G,

$$\text{CLIP-F}(G) = P(\texttt{good face} \mid G) - P(\texttt{bad face} \mid G). \tag{23}$$

By construction, $\text{CLIP-F}(G) \in [-1, 1]$. Scores are averaged over all generated images within each identity and prompt. Lower values indicate weaker face-like visual evidence. CLIP-F measures face quality rather than identity similarity.

B.3 Protection Configuration

Table 4 lists the complete configuration used to generate protected images. Images are processed independently, and each image uses a fresh surrogate initialized from SD v2.1. The perturbation budget ϵ and the step size α are both expressed in the $[-1, 1]$ image space used by the diffusion pipeline, so $\epsilon = 5 \times 10^{-2}$ corresponds to 2.5×10^{-2} in $[0, 1]$ units. The conditioning ratio ρ and the attention weight λ_{attn} were selected on the first 10 identities of each dataset, which are also the identities used for the robustness, transfer, and ablation analyses; the main DreamBooth comparison evaluates all 50 identities per dataset. During protection optimization, we use the VAE posterior mean. During surrogate adaptation, we sample from the VAE posterior and draw timesteps uniformly from the full diffusion trajectory.

For the anchor-count ablation, we use the ordered candidate sequence $\{0, 25, 50, 75, 100, 125, 150, 175\}$ and retain the corresponding prefix for each reported anchor count.

B.4 Downstream Personalization and Generation

The main attacker trains DreamBooth with SD v2.1-base using the protected images and the instance prompt *"a photo of sks person"*. The U-Net and text encoder are optimized for 1,000 steps with AdamW, learning rate 5×10^{-7}, a constant schedule without warmup, batch size 2, and gradient accumulation 1. Prior preservation uses the class prompt *"a photo of person"*, 200 class images, and weight 1.0. Training uses 512×512 center-cropped images and bf16 precision.

For both quantitative and qualitative evaluation, each personalized model generates 32 images per identity for each of the following six prompts:

- **P1:** *"a dslr portrait of sks person"*;
- **P2:** *"a photo of sks person"*;
- **P3:** *"a close-up photo of sks person, high details"*;
- **P4:** *"a photo of sks person in front of eiffel tower"*;

Table 4: Configuration used for VEIL in the main experiments.

Component	Configuration
Surrogate backbone	Stable Diffusion 2.1-base
Surrogate condition	Empty-prompt embedding
Image space	$[-1, 1]$, resolution 512×512
Perturbation budget	$\epsilon = 5 \times 10^{-2}$
PGD step size	$\alpha = 4 \times 10^{-3}$
Anchor set	$\{0, 25, 50, 75, 100, 125\}$
Protection iterations	$R = 16$ per block
Adaptation rounds	$E = 6$
Surrogate steps	$S = 20$ per round
Surrogate learning rate	5×10^{-7}
Attention blocks	Down- and up-sampling blocks
Attention resolution	16×16
Conditioning ratio	$\rho = 0.6$
Attention weight	$\lambda_{\text{attn}} = 1.75$
Attention scaling	$\kappa = 100$
Saliency quantile	$q_{\text{sal}} = 0.95$
Weight exponent	$\gamma = 1$
Shrinkage ratio	$\tau_s = 5 \times 10^{-4}$
Random seed	42

- **P5:** *"a photo of sks person looking at the mirror"*;
- **P6:** *"a photo of sks person standing in front of the table holding a knife"*.

P1 is held out from personalization, P2 matches the instance training prompt, and P3–P6 evaluate identity recovery across contextual variations. All prompts use 100 denoising steps, a classifier-free guidance scale of 7.5, a resolution of 512×512, and shared generation seeds across methods.

B.5 Baseline, Robustness, and Transfer Protocols

Baselines. We use the official implementations and recommended configurations of Anti-DreamBooth (Anti-DB), SDS-, VCPro, MYOPIA, Anti-Diffusion (Anti-Diff), CAAT, DisDiff, and SimAC. Additive pixel-space perturbation methods are evaluated with $\epsilon = 5 \times 10^{-2}$ in the normalized image space. MYOPIA retains its official EOT-based configuration and recommended budget $\epsilon = 5 \times 10^{-1}$ because reducing it to 5×10^{-2} removes its intended operating behavior.

Preprocessing and purification. Every transformation is applied to the released protected images before downstream DreamBooth training. The attacker then retrains DreamBooth from the corresponding pretrained checkpoint. JPEG uses quality factor 70, and Gaussian blur uses a 5×5 kernel. DiffPure (Nie et al., 2022) uses the `google/ddpm-celebahq-256` DDPM purifier with 50 noising-denoising steps. GrIDPure (Zhao et al., 2024) follows its official implementation with the guided-diffusion ImageNet model, 10 purification steps, 20 iterations, blending coefficient 0.1, and 256×256 grid crops with stride 128.

Transfer across Stable Diffusion variants. Protected images generated with the SD v2.1 surrogate are reused without re-optimization to train DreamBooth models initialized from SD v1.4 and SD v1.5. The identities, prompts, training steps, and evaluation protocol are held fixed. ORG is trained separately on clean images for each target backbone.

Cross-personalization transfer. The same protected images optimized with the SD v2.1 surrogate are reused without re-optimization under DreamBooth-LoRA, Textual Inversion, and SVDiff. The identities and evaluation prompts remain fixed, so only the downstream personalization procedure changes. DreamBooth-LoRA trains rank-16 adapters on the U-Net attention projections for 2,000 steps with a learning rate of 1×10^{-4}, while keeping the text encoder frozen. Textual Inversion learns a single `<sks>` embedding initialized from `human` for 3,000 steps, using a base learning rate of 5×10^{-4} with learning-rate scaling enabled. SVDiff optimizes only the spectral-shift parameters of the U-Net and text encoder for 500 steps, using learning rates of 1×10^{-3} and 1×10^{-6} for matrix-valued and one-dimensional shifts, respectively. DreamBooth-LoRA and SVDiff use prior

preservation with 200 class images. For evaluation, each procedure generates 32 images per prompt with a guidance scale of 7.5; DreamBooth-LoRA and SVDiff use 100 denoising steps, whereas Textual Inversion uses 50.

Identifier-binding sensitivity. The protected images remain unchanged, while every occurrence of `sks` in the DreamBooth training and inference prompts is replaced with `t@t`. All remaining settings are held fixed. This protocol evaluates whether the same protected images transfer to an alternative identifier binding.

C Additional Main Results

C.1 Contextual DreamBooth Prompts

Table 5 evaluates P3–P6, which introduce close-up, landmark, reflection, and object-interaction contexts. Across the eight dataset–prompt settings, VEIL achieves the highest FDFR in all cases and the best or second-best SER-FIQ and CLIP-F in every case. For ISM, it is best or tied for best in six settings, second-best on VGGFace2 under P3, and slightly behind SimAC and DisDiff on CelebA-HQ under P5. The results confirm that the protection observed under P1 and P2 extends to more diverse generation contexts.

C.2 Qualitative Results

Figure 5 compares clean images, protected images, and amplified residual maps, illustrating the spatial distribution of the perturbations produced by each method. Figure 6 presents DreamBooth outputs generated after training on the corresponding protected images.

Table 5: Post-personalization suppression under four additional DreamBooth prompts (**P3**–**P6**). The exact prompt templates are defined in Appendix B.4. **Bold** and underlined values denote the best and second-best results, respectively.

Dataset	Method	Prompt **P3**				Prompt **P4**			
		SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓	SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓
CelebA-HQ	ORG	0.66	0.19	0.47	0.38	0.54	0.22	0.16	0.09
	Anti-DB	0.05	0.96	0.11	0.18	0.28	0.45	0.07	0.00
	SDS-	0.75	0.14	0.55	0.40	0.47	0.38	0.09	-0.02
	VCPro	0.62	0.29	0.36	0.29	0.52	0.24	0.13	0.10
	MYOPIA	0.76	0.17	0.52	0.35	0.41	0.52	0.11	-0.03
	Anti-Diffusion	0.03	0.94	0.08	0.20	0.23	0.46	0.07	-0.04
	CAAT	0.04	0.94	0.12	0.18	0.19	0.47	0.07	-0.02
	DisDiff	0.02	0.98	0.02	0.06	0.14	0.55	0.06	-0.10
	SimAC	0.01	0.99	0.03	0.07	0.15	0.56	0.06	-0.08
	VEIL	**0.00**	**1.00**	**0.01**	**0.03**	**0.07**	**0.68**	**0.05**	**-0.11**
VGGFace2	ORG	0.63	0.13	0.48	0.34	0.53	0.26	0.22	0.04
	Anti-DB	0.09	0.88	0.19	0.17	0.28	0.43	0.11	-0.03
	SDS-	0.73	0.08	0.57	0.39	0.51	0.23	0.19	-0.06
	VCPro	0.60	0.16	0.41	0.24	0.55	0.18	0.21	0.05
	MYOPIA	0.71	0.20	0.51	0.36	0.45	0.38	0.19	-0.05
	Anti-Diffusion	0.12	0.78	0.25	0.18	0.29	0.34	0.12	-0.08
	CAAT	0.07	0.85	0.17	0.15	0.23	0.40	0.10	-0.06
	DisDiff	0.07	0.92	0.12	0.07	0.13	0.51	0.09	-0.14
	SimAC	**0.02**	0.96	**0.08**	0.07	0.14	0.56	**0.08**	-0.12
	VEIL	0.03	**0.97**	0.11	**0.03**	**0.08**	**0.68**	**0.08**	**-0.17**
Dataset	Method	Prompt **P5**				Prompt **P6**			
		SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓	SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓
CelebA-HQ	ORG	0.58	0.03	0.44	0.34	0.63	0.42	0.22	-0.12
	Anti-DB	0.27	0.46	0.21	0.09	0.27	0.51	0.07	-0.15
	SDS-	0.58	0.08	0.36	0.31	0.37	0.52	0.09	-0.13
	VCPro	0.57	0.05	0.41	0.25	0.62	0.51	0.18	**-0.16**
	MYOPIA	0.62	0.06	0.40	0.29	0.40	0.48	0.12	-0.07
	Anti-Diffusion	0.28	0.40	0.25	0.04	0.22	0.64	0.08	-0.14
	CAAT	0.24	0.46	0.23	0.05	0.19	0.62	0.08	-0.15
	DisDiff	0.10	0.85	0.16	0.04	0.06	0.82	0.05	**-0.16**
	SimAC	0.12	0.81	**0.15**	0.04	0.07	0.82	0.06	-0.15
	VEIL	**0.08**	**0.87**	0.17	**-0.01**	**0.02**	**0.93**	**0.03**	-0.15
VGGFace2	ORG	0.58	0.02	0.42	0.30	0.64	0.32	0.28	-0.07
	Anti-DB	0.27	0.36	0.24	0.05	0.31	0.40	0.12	-0.17
	SDS-	0.58	0.07	0.40	0.26	0.52	0.23	0.20	-0.15
	VCPro	0.58	0.03	0.42	0.24	0.63	0.34	0.25	-0.18
	MYOPIA	0.59	0.04	0.40	0.25	0.51	0.21	0.20	-0.12
	Anti-Diffusion	0.29	0.30	0.27	0.03	0.29	0.43	0.13	-0.17
	CAAT	0.23	0.42	0.24	0.02	0.24	0.47	0.11	-0.17
	DisDiff	0.10	0.83	0.20	0.03	0.10	0.73	0.09	-0.18
	SimAC	0.10	0.80	0.19	0.03	0.09	0.74	**0.08**	-0.18
	VEIL	**0.05**	**0.86**	**0.16**	**-0.03**	**0.05**	**0.80**	**0.08**	**-0.19**

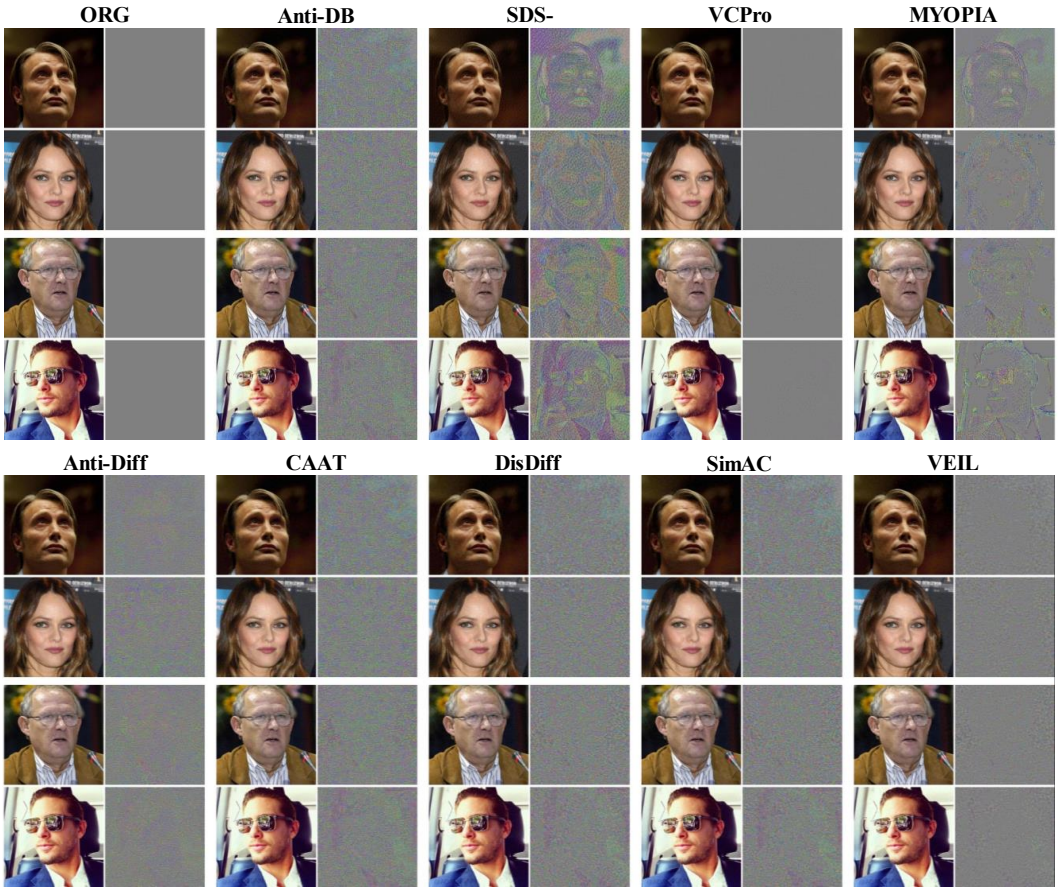
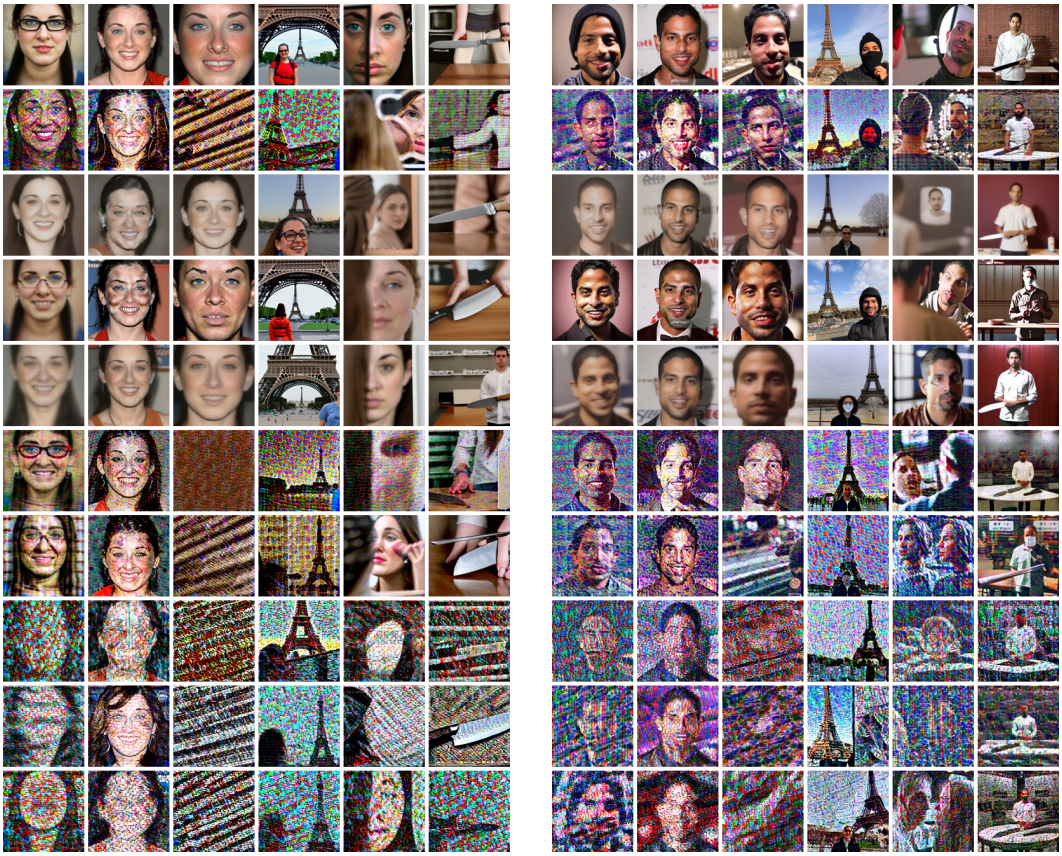


Figure 5: Released protected images and residual perturbations, with ORG and the nine defenses arranged in two rows of five. Within each method, the left column shows the protected image, or the clean image for ORG, and the right column shows the residual; the top two rows are CelebA-HQ identities and the bottom two rows are VGGFace2 identities. Residuals are the absolute clean–protected difference magnified five times and clipped to the display range, and show only spatial structure rather than perturbation magnitude.



(a) CelebA-HQ (b) VGGFace2

Figure 6: Post-personalization generations from DreamBooth models trained on protected images. Rows, from top to bottom, correspond to ORG, Anti-DB, SDS-, VCPro, MYOPIA, Anti-Diff, CAAT, DisDiff, SimAC, and VEIL. Columns, from left to right, correspond to P1: *"a dslr portrait of sks person"*, P2: *"a photo of sks person"*, P3: *"a close-up photo of sks person, high details"*, P4: *"a photo of sks person in front of eiffel tower"*, P5: *"a photo of sks person looking at the mirror"*, and P6: *"a photo of sks person standing in front of the table holding a knife"*.

D Additional Transferability Results

D.1 Transfer across Stable Diffusion Variants

Table 6 reports prompt-level transfer across Stable Diffusion variants on CelebA-HQ. The protected images are generated once with SD v2.1 and are then used without re-optimization to train DreamBooth models initialized from SD v1.4 and SD v1.5. Compared with the corresponding ORG results, VEIL increases FDFR and reduces ISM under both variants, with larger changes under P2.

Table 6: Prompt-level transfer across Stable Diffusion variants on CelebA-HQ. ORG denotes personalization of the corresponding variant using clean images. **Bold** and underlined entries denote the best and second-best defense results within each model-prompt block, respectively.

Model	Method	**P1**: `a dslr portrait of sks person`				**P2**: `a photo of sks person`			
		SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓	SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓
SD1.4	ORG	0.69	0.01	0.47	0.46	0.77	0.01	0.75	0.61
	Anti-DB	0.60	0.10	0.26	0.38	0.13	0.59	0.18	0.28
	SDS-	0.73	0.02	0.45	0.44	0.70	0.00	0.56	0.38
	VCPro	0.68	0.02	0.36	0.37	0.74	0.00	0.61	0.25
	MYOPIA	0.72	0.00	0.50	0.48	0.70	0.00	0.64	0.32
	Anti-Diff	**0.54**	0.09	0.24	0.36	0.12	0.69	0.20	0.24
	CAAT	0.57	0.08	**0.22**	**0.33**	0.16	0.63	0.22	0.25
	DisDiff	0.59	0.14	0.30	0.43	0.11	**0.97**	**0.06**	0.21
	SimAC	0.55	**0.18**	0.27	0.37	**0.06**	0.95	0.11	**0.20**
	VEIL	0.60	0.11	0.30	0.45	0.26	0.86	0.20	0.21
SD1.5	ORG	0.69	0.00	0.46	0.48	0.75	0.00	0.71	0.58
	Anti-DB	0.58	0.14	**0.20**	**0.34**	0.21	0.51	0.25	0.28
	SDS-	0.73	0.00	0.46	0.42	0.62	0.00	0.44	0.24
	VCPro	0.71	0.02	0.39	0.44	0.73	0.01	0.60	0.31
	MYOPIA	0.71	0.00	0.56	0.47	0.69	0.01	0.58	0.22
	Anti-Diff	**0.56**	0.16	0.28	0.37	0.06	0.74	0.21	**0.20**
	CAAT	0.63	0.16	0.21	0.35	0.10	0.73	0.16	0.21
	DisDiff	0.57	**0.28**	0.26	0.35	**0.01**	**1.00**	**0.00**	0.21
	SimAC	0.63	0.08	0.29	0.42	0.05	0.99	0.06	0.24
	VEIL	0.58	0.21	0.40	0.46	0.21	0.81	0.25	0.22

D.2 Transfer across Personalization Procedures

Table 7 reports prompt-level transfer on CelebA-HQ. Under DreamBooth-LoRA, all methods provide limited separation on the held-out P1 prompt, for which ORG itself exhibits relatively low identity recovery. Under P2, VEIL obtains the lowest SER-FIQ and the second-highest FDFR, while DisDiff and SimAC achieve lower ISM. Under Textual Inversion, VEIL is strongest on SER-FIQ, FDFR, and ISM for P1, whereas Anti-DB, Anti-Diff, and DisDiff are stronger on several P2 metrics. Under SVDiff, VEIL obtains the best SER-FIQ and FDFR for both prompts, the best ISM on P1, and the best CLIP-F on P2. Transfer strength therefore depends on both the adaptation mechanism and the evaluation prompt.

D.3 Transfer across Identifier Bindings

Table 8 reports prompt-level results after replacing `sks` with `t@t` in both the DreamBooth training and inference prompts. On P1, VEIL obtains the lowest SER-FIQ and ISM and the highest FDFR. On P2, it obtains the lowest CLIP-F, whereas SimAC and DisDiff achieve stronger FDFR and SER-FIQ.

Table 7: Prompt-level transfer across downstream personalization procedures on CelebA-HQ. **Bold** and underlined entries denote the best and second-best defense results within each procedure–prompt block, respectively.

Procedure	Method	P1: `a dslr portrait of sks person`				P2: `a photo of sks person`			
		SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓	SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓
DreamBooth-LoRA	ORG	0.79	0.00	0.26	0.03	0.74	0.01	0.66	0.34
	Anti-DB	**0.68**	**0.03**	**0.19**	**-0.09**	0.48	0.13	0.36	0.18
	SDS-	0.80	0.00	0.28	0.20	0.72	0.00	0.57	0.46
	VCPro	0.77	0.02	0.25	0.03	0.72	0.00	0.57	0.31
	MYOPIA	0.80	0.01	0.29	0.09	0.62	0.18	0.49	0.30
	Anti-Diff	0.75	0.01	0.22	-0.03	0.44	0.11	0.36	**0.10**
	CAAT	0.72	0.01	**0.19**	-0.02	0.21	0.60	0.24	0.11
	DisDiff	0.74	0.02	0.28	-0.01	0.06	**0.97**	**0.07**	0.20
	SimAC	0.75	0.02	0.26	-0.01	0.09	0.91	0.11	0.20
	VEIL	0.73	0.01	0.25	0.01	**0.03**	0.94	0.12	0.21
TI	ORG	0.72	0.00	0.60	0.50	0.72	0.00	0.61	0.51
	Anti-DB	0.51	0.59	0.12	**0.18**	0.47	0.60	**0.15**	0.24
	SDS-	0.71	0.17	0.36	0.40	0.71	0.29	0.36	0.34
	VCPro	0.68	0.01	0.41	0.30	0.68	0.02	0.43	0.26
	MYOPIA	0.69	0.08	0.46	0.32	0.65	0.13	0.47	0.32
	Anti-Diff	0.49	0.55	0.17	0.20	**0.41**	**0.70**	0.17	**0.16**
	CAAT	0.52	0.45	0.14	0.24	0.57	0.49	0.19	0.23
	DisDiff	0.42	0.65	0.11	0.22	0.44	0.64	**0.15**	0.24
	SimAC	0.42	0.64	0.16	0.23	0.53	0.57	0.19	0.28
	VEIL	**0.28**	**0.77**	**0.10**	0.23	0.43	0.62	0.18	0.27
SVDiff	ORG	0.69	0.02	0.37	0.27	0.73	0.02	0.70	0.55
	Anti-DB	0.35	0.49	0.15	0.12	0.28	0.62	0.24	0.38
	SDS-	0.72	0.02	0.27	0.19	0.76	0.00	0.57	0.48
	VCPro	0.65	0.02	0.30	0.27	0.72	0.02	0.58	0.43
	MYOPIA	0.69	0.05	0.31	0.15	0.74	0.01	0.65	0.45
	Anti-Diff	0.46	0.29	0.20	0.09	0.19	0.68	0.23	0.34
	CAAT	0.36	0.48	0.15	0.10	0.20	0.69	**0.20**	0.41
	DisDiff	0.34	0.51	0.15	**-0.04**	0.36	0.52	0.30	0.40
	SimAC	0.38	0.50	0.14	-0.02	0.46	0.47	0.38	0.39
	VEIL	**0.09**	**0.87**	**0.08**	-0.02	**0.14**	**0.86**	0.21	**0.22**

Table 8: Prompt-level identifier-binding sensitivity on CelebA-HQ with `sks` replaced by `t@t`. **Bold** and underlined entries denote the best and second-best defense results for each prompt, respectively.

Method	**P1**: `a dslr portrait of t@t person`				**P2**: `a photo of t@t person`			
	SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓	SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓
Anti-DB	0.56	0.19	0.23	0.23	0.47	0.18	0.43	0.34
SDS-	0.76	0.00	0.35	0.24	0.78	0.00	0.68	0.50
VCPro	0.74	0.00	0.40	0.25	0.76	0.01	0.64	0.33
MYOPIA	0.76	0.00	0.37	0.21	0.74	0.00	0.65	0.41
Anti-Diff	0.64	0.06	0.27	0.14	0.38	0.28	0.41	0.29
CAAT	0.49	0.17	0.23	0.16	**0.33**	0.38	**0.33**	0.30
DisDiff	0.49	0.40	0.23	**0.05**	0.41	0.56	0.42	0.28
SimAC	0.59	0.22	0.21	0.10	0.37	**0.61**	0.40	0.33
VEIL	**0.44**	**0.44**	**0.18**	0.07	0.47	0.32	0.43	**0.25**

E Additional Ablation Studies

E.1 Number of Ultra-Early Anchors

Figure 7 reports the anchor-count ablation discussed in Section 5.5. Each configuration retains the corresponding prefix of the ordered candidate sequence given in Appendix B.3.

E.2 Anchor Placement

Table 9 varies the placement of the six anchors while keeping all other settings fixed. Fixed ultra-early anchors give the highest FDFR and the lowest ISM under both prompts. Moving the fixed anchors to intermediate or late timesteps sharply weakens suppression, and late anchors provide

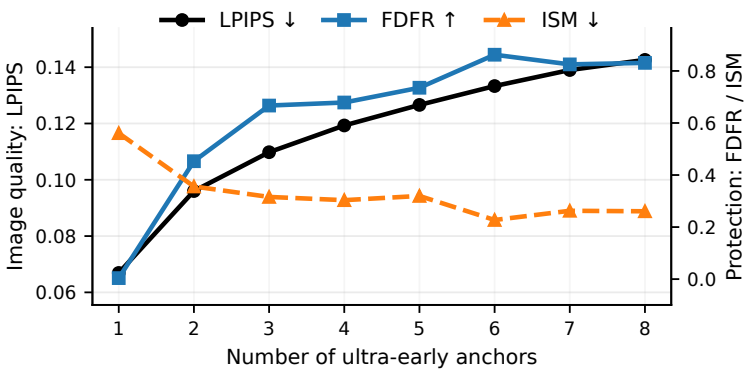


Figure 7: Effect of the number of ultra-early anchors, averaged over both datasets and P1–P2. The left axis reports LPIPS of the protected images; the right axis reports post-personalization FDFR and ISM.

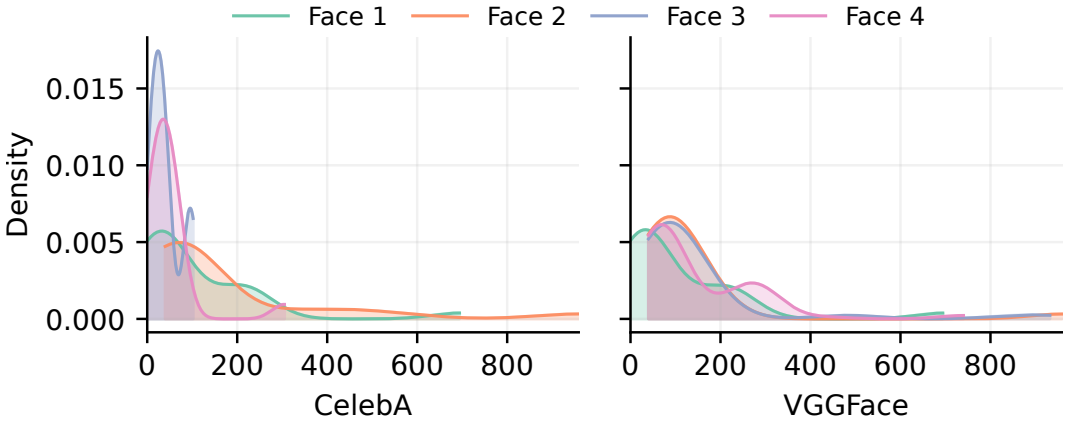


Figure 8: Empirical distribution of timesteps selected by SimAC on CelebA-HQ and VGGFace2.

almost no protection (FDFR of 0.04 and 0.08). All variants reach the same maximum per-pixel perturbation and a comparable mean absolute perturbation of 39% to 54% of ϵ, with the mid-range anchors using the most. Within the same early region, fixed anchors improve FDFR over randomly sampled ones by 0.12 and 0.14 at comparable fidelity. Random sampling over the full schedule attains a lower SER-FIQ under P2 but increases LPIPS from 0.13 to 0.17.

Table 9: Effect of anchor placement on CelebA-HQ. All variants use six anchors; random variants resample the anchors from the listed range at every iteration.

Type	Anchors	Protected-image fidelity			Prompt **P1**				Prompt **P2**			
		SSIM ↑	PSNR ↑	LPIPS ↓	SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓	SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓
Fixed	$\{0, 25, \ldots, 125\}$ **(VEIL)**	**0.94**	37.98	**0.13**	**0.07**	**0.95**	**0.09**	**0.15**	0.18	**0.88**	**0.27**	**0.17**
	$\{400, 425, \ldots, 525\}$	0.91	36.02	0.20	0.35	0.54	0.21	0.33	0.28	0.48	0.38	0.33
	$\{800, 825, \ldots, 925\}$	0.93	37.81	0.15	0.62	0.04	0.33	0.17	0.61	0.08	0.53	0.26
	$\{0, 200, 400, 600, 800, 999\}$	**0.94**	**38.25**	0.14	0.35	0.57	0.25	0.27	0.23	0.63	0.41	0.26
Random	$[0, 125]$	**0.94**	38.20	**0.13**	0.20	0.83	0.20	**0.15**	0.25	0.74	0.38	0.22
	$[0, 999]$	0.93	37.13	0.17	0.11	0.83	0.16	0.16	**0.13**	0.81	0.33	0.27

E.3 Distribution of Adaptively Selected Timesteps

Figure 8 reports the empirical timestep distribution selected by SimAC on CelebA-HQ and VGGFace2. The selected timesteps concentrate at the low-noise end of the schedule rather than being distributed uniformly. This observation motivates restricting the candidate region to early timesteps, while the anchor-count, anchor-placement, and optimization-stability analyses evaluate the specific fixed-anchor configuration.

E.4 Optimization Stability of Temporal Anchors

Figure 9 compares denoising-disruption trajectories for a representative CelebA-HQ image. Attention suppression and conditioning-position selection are disabled to isolate the temporal construction of the denoising objective. Random single-timestep optimization exhibits the largest fluctuations, early-region sampling reduces the fluctuations, multi-timestep accumulation produces a smoother trajectory, and the fixed ultra-early anchors yield the smoothest trajectory in this diagnostic example. Each curve is independently min-max normalized, so the comparison concerns within-run trajectory smoothness rather than absolute loss magnitude.

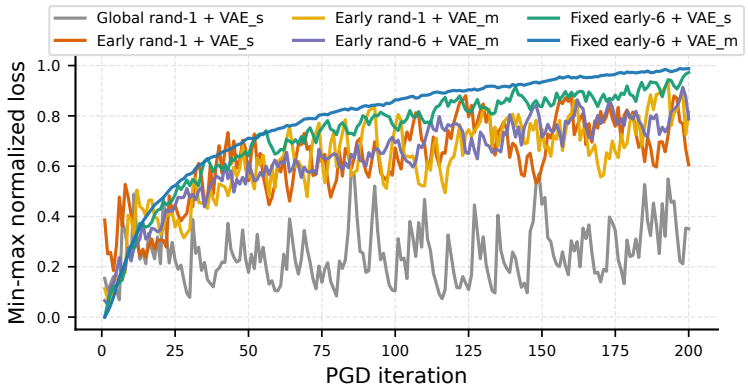


Figure 9: Denoising-objective stability on one representative CelebA-HQ image. Each trajectory is independently min-max normalized.

E.5 Surrogate Adaptation

Table 10 compares the alternating procedure with a static surrogate that is never updated, keeping the same 112 perturbation iterations and all other settings. Surrogate adaptation raises FDFR in all four settings, by 0.04 to 0.21, and lowers SER-FIQ, ISM, and CLIP-F in three of the four settings each. The static surrogate yields a 0.24 dB higher PSNR, whereas surrogate adaptation yields a lower LPIPS.

Table 10: Effect of surrogate adaptation. Bold marks the better variant within each dataset. Both variants run 112 perturbation iterations; the static variant never updates the surrogate.

Dataset	Surrogate	Protected-image fidelity			Prompt **P1**				Prompt **P2**			
		SSIM ↑	PSNR ↑	LPIPS ↓	SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓	SER-FIQ ↓	FDFR ↑	ISM ↓	CLIP-F ↓
CelebA-HQ	**Adaptive (VEIL)**	**0.94**	37.98	**0.13**	**0.07**	**0.95**	**0.09**	**0.15**	**0.18**	**0.88**	**0.27**	**0.17**
	Static	**0.94**	**38.22**	0.14	0.21	0.78	0.25	0.19	0.25	0.67	0.35	0.31
VGGFace2	**Adaptive (VEIL)**	**0.94**	38.00	**0.13**	**0.12**	**0.77**	0.29	**0.03**	0.25	**0.73**	**0.31**	0.18
	Static	**0.94**	**38.24**	0.14	0.23	0.73	**0.27**	0.09	**0.23**	0.55	0.43	**0.14**

E.6 Sensitivity to the Attention Weight

Figure 10 evaluates the sensitivity to λ_{attn} with all remaining settings fixed. The selected value $\lambda_{\text{attn}} = 1.75$ reduces both ISM and CLIP-F while keeping LPIPS close to the lower-weight configurations.

E.7 Sensitivity to the Conditioning Ratio

Figure 11 varies the conditioning ratio ρ, which sets the fraction of conditioning positions retained for attention suppression, with all remaining settings fixed. FDFR stays between 0.82 and 0.86 and LPIPS between 0.133 and 0.134 across the evaluated range, so protection is not sensitive to the exact value of ρ. We use $\rho = 0.6$ by default.

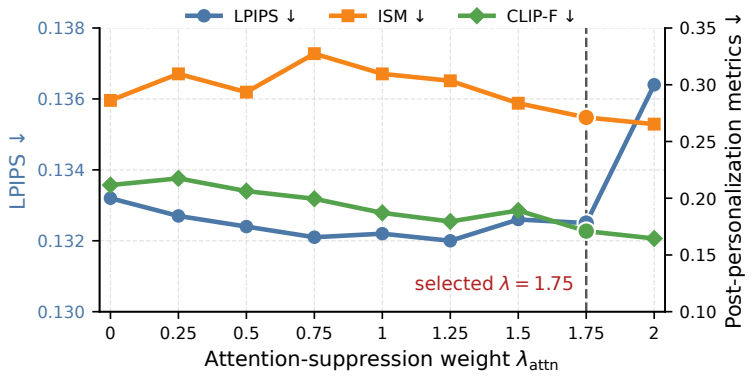


Figure 10: Sensitivity to the attention-suppression weight λ_{attn}. The left axis reports LPIPS of the protected images; the right axis reports post-personalization ISM and CLIP-F.

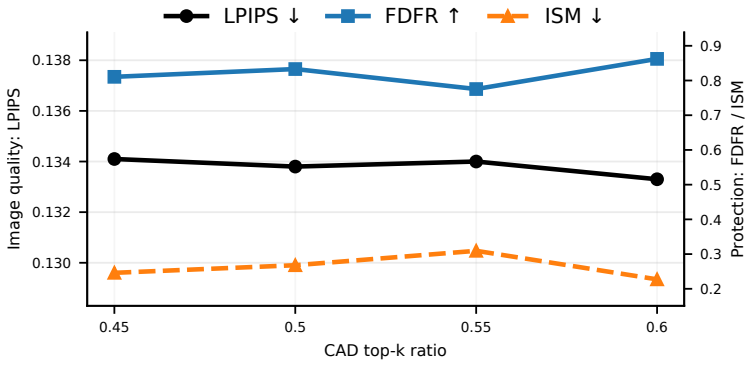


Figure 11: Sensitivity to the conditioning ratio ρ. The left axis reports LPIPS of the protected images; the right axis reports post-personalization FDFR and ISM.