Manifold-Aligned Adversarial Perturbation for Anti-Customization under Diffusion-based Purification

Ruotian Liu[1] **Zhen Chen**[1] **Zhiguo Yang**[1] **Xiangyu Yin**[2] **Peipei Xu**[3*] **Wenjie Ruan**[1*]
[1] USTC [2] Chalmers University of Technology [3] University of Dundee

Abstract

Anti-customization techniques aim to protect visual content from unauthorized diffusion model personalization by adding imperceptible perturbations to training images. However, such protection can be significantly weakened by diffusion-based purification, which may remove adversarial perturbations before customization. We study representative anti-customization methods under diffusion purification and observe that existing methods exhibit similar off-manifold perturbation structures that are largely suppressed by purification. We then provide a principled geometric analysis that characterizes purification as a local contraction toward the data manifold, primarily suppressing perturbation components along normal directions. Motivated by this insight, we propose Manifold-Aligned Customization protection (**MAC**), a practical purification-aware framework that optimizes downstream anti-customization objectives through a differentiable purifier and uses a structured perturbation parameterization to bias updates toward purification-retentive, manifold-aligned directions. Across experiments on CelebA-HQ, VGGFace2-HQ, and WikiArt under diffusion-based purification, MAC substantially improves protection effectiveness across most prompts and domains. Finally, frequency analysis and ImageNet and MS-COCO retention studies show that spectral signatures cannot predict purification outcomes, providing additional evidence for the proposed manifold-alignment perspective beyond a single data domain. Our code is available at `https://github.com/Trusted-LLM/MAC`.

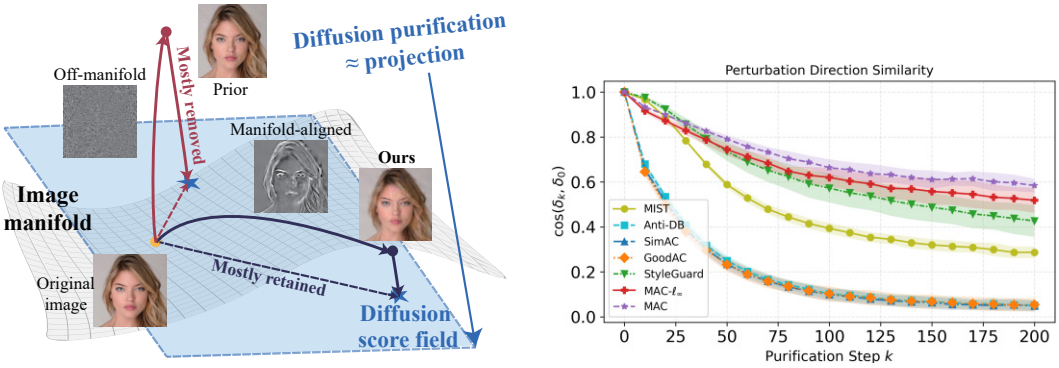
1 Introduction

Customization methods such as DreamBooth Hu et al. [2022], Ruiz et al. [2023], Kumari et al. [2023] adapt large diffusion models to a small set of training images to capture a target visual concept. However, they also raise significant concerns when visual content is scraped, redistributed, or exploited for unauthorized personalization without consent Carlini et al. [2023]. Conversely, anti-customization methods aim to protect visual content by adding imperceptible perturbations so that downstream customization either fails or produces outputs with degraded concept fidelity Liang and Wu [2023], Onikubo and Matsui [2024], Van Le et al. [2023], Wan et al. [2024], Wang et al. [2024a], Xu et al. [2025], Yang et al. [2025]. A practical complication arises from the fact that diffusion-based purifiers (e.g., DiffPure-style preprocessing) can be applied to training images before customization, potentially attenuating or removing protection perturbations Nie et al. [2022], Song et al. [2024], Bai et al. [2024], Pei et al. [2025]. In realistic threat models, an adversary can freely apply sufficiently strong diffusion purification to remove protection perturbations. Our experiments show that even recent state-of-the-art anti-customization methods, including StyleGuard Li et al. [2025], that are explicitly designed with purification as a threat model, are significantly weakened under strong purification. This creates the following key research questions:

What kinds of perturbations remain effective after diffusion purification, and how should anti-customization adversarial perturbations be designed under this threat?

*Corresponding author

Preprint.



(a) Geometric intuition of diffusion purification.

(b) Perturbation dynamics under diffusion purification.

Figure 1: **Geometric intuition and perturbation dynamics.** (a) Diffusion purification acts as a projection-like contraction toward the image manifold, removing off-manifold perturbations while retaining manifold-aligned ones. (b) We track perturbation direction similarity $\cos(\delta_k, \delta_0)$ under increasing purification strength, where $\delta_k = x_{\text{pur}}^{(k)} - x_{\text{orig}}$ and $\delta_0 = x_{\text{adv}} - x_{\text{orig}}$ denotes the initial perturbation. MAC shows higher similarity, indicating stronger purification-preserved alignment.

A common intuition attributes purification robustness to frequency characteristics Han et al. [2021], Chen et al. [2022], Onikubo and Matsui [2024], Pei et al. [2025], Yang et al. [2025]. Although frequency correlates with perceptual smoothness, it does not capture the geometric mechanism that governs purification: alignment with normal versus tangent directions of the data manifold. We therefore treat frequency statistics as descriptive rather than causal and design our analysis to separate spectral appearance from manifold alignment.

This work provides a *principled geometric analysis* of diffusion purification and uses it to guide purification-aware anti-customization. We model natural images as concentrating near a low-dimensional manifold in a high-dimensional space Tenenbaum et al. [2000], Belkin and Niyogi [2003]. **Figure** 1(a) illustrates the key intuition: score-driven reverse dynamics act as a projection-like contraction toward the image manifold, preferentially suppressing perturbation components in normal, off-manifold directions while largely preserving tangent, manifold-aligned components. **Figure** 1(b) provides empirical support by tracking perturbation direction similarity under increasing purification strength. Existing methods rapidly lose directional consistency after purification, whereas **MAC** maintains substantially higher similarity, suggesting that its perturbations are better aligned with directions preserved by diffusion purification.

Based on these insights, we instantiate a practical *purification-aware* anti-customization framework, termed **MAC**. MAC optimizes downstream anti-customization objectives on purified images and backpropagates through a differentiable purifier by unrolling a finite number of reverse denoising steps Kassis et al. [2024], Wang et al. [2024b]. To bias optimization toward manifold-aligned perturbations in practice, we adopt a structured, locally coherent perturbation parameterization via a differentiable content-guided filter. This design is compatible with existing strong anti-customization objectives and can be inserted as a drop-in component in standard protection pipelines. Beyond facial protection benchmarks, we further evaluate DreamBooth protection on WikiArt Saleh and Elgammal [2015] and conduct perturbation-retention analysis on ImageNet Deng et al. [2009] and MS-COCO Lin et al. [2014]. These results provide complementary evidence for the broad applicability of the proposed manifold-alignment perspective across visual domains and purification settings.
In summary, the contribution of this paper is as follows:

1. We present a geometric analysis of diffusion-based purification, characterizing it as a local contraction toward the image manifold that selectively suppresses off-manifold perturbations while preserving components aligned with purification-retentive manifold directions.
2. We propose **MAC**, a purification-aware anti-customization framework that operationalizes manifold alignment through structured perturbation parameterization and end-to-end differentiation through the purifier.
3. We demonstrate that MAC improves protection effectiveness across facial datasets, prompts, visual domains, and purification settings. Complementary results support our explanation:

frequency-domain analysis shows that spectral signatures alone cannot explain purification robustness, while WikiArt DreamBooth experiments and ImageNet/MS-COCO retention studies provide evidence for the proposed purification-retention mechanism.

2 Related Work

2.1 Customization of Latent Diffusion Models

Customization techniques such as DreamBooth and its variants enable adapting large text-to-image diffusion models to a specific subject using only a few reference images Hu et al. [2022], Ruiz et al. [2023]. Subsequent work has extended this line of research to multi-concept customization and more parameter-efficient adaptations Gal et al. [2022], Kumari et al. [2023]. These methods primarily focus on improving subject fidelity and generalization under the assumption that the training images are clean. In contrast, our work considers a threat model where the training data are intentionally perturbed and further processed by diffusion-based purification before customization.

2.2 Anti-Customization with Protective Perturbations

To mitigate misuse of customization, prior work on anti-customization adds imperceptible perturbations to user images to disrupt downstream fine-tuning or degrade identity consistency Liang and Wu [2023], Van Le et al. [2023], Wan et al. [2024], Wang et al. [2024a], Xu et al. [2025]. Most approaches construct perturbations via iterative PGD-based optimization Madry et al. [2017] against the customization objective. AntiPure Yang et al. [2025] explicitly targets diffusion purification by emphasizing patch-wise high-frequency perturbations, while StyleGuard Li et al. [2025] perturbs latent-space style statistics and incorporates purifier-aware objectives, largely treating purification as an external transformation. However, prior works do not **explicitly optimize perturbations through purification dynamics**. Instead, our work focuses on the *dynamics of diffusion-based purification*, analyzing purification from a geometric perspective and designing perturbations aligned with directions preserved by the purification process, as evidenced by consistent post-purification behavior across purification strengths.

2.3 Diffusion-based Adversarial Purification

Diffusion models have been widely studied as a defense mechanism for adversarial purification, where noisy inputs are iteratively denoised toward the data distribution Nie et al. [2022], Bai et al. [2024], Song et al. [2024]. Several works analyze the robustness and limitations of these defenses, and propose attacks that differentiate through the purification process Kang et al. [2023], Wang et al. [2024b], Kassis et al. [2024]. Recent studies further associate purification behavior with frequency characteristics of perturbations Wang et al. [2020], Pei et al. [2025]. Complementary to prior work that relies primarily on *empirical or spectral explanations*, we adopt a *geometric perspective* that interprets diffusion purification as a contraction toward the data manifold, and leverage this view to derive actionable design principles for purification-aware anti-customization.

3 Geometric Analysis of Diffusion-based Purification

In this section, we present a geometric theoretical analysis of diffusion-based purification to understand its behavior from the perspective of data manifolds and score-based dynamics. Our analysis is grounded in the manifold hypothesis Roweis and Saul [2000], Tenenbaum et al. [2000], Belkin and Niyogi [2003], Li and He [2025], where natural images are assumed to concentrate around a low-dimensional embedded manifold in a high-dimensional ambient space.

We first characterize diffusion-based purification as a score-driven dynamical system Song et al. [2020] that attracts samples toward the data manifold by contracting normal directions (**Section** 3.1). Based on this characterization, we analyze how different perturbation components behave under purification and show that off-manifold perturbations are effectively suppressed, while manifold-aligned perturbations can be preserved (**Section** 3.2). Finally, we revisit frequency-based explanations and demonstrate that frequency characteristics are not the fundamental factor governing purification behavior, but rather a secondary consequence of manifold alignment (**Section** 3.3).

3.1 Diffusion-based purification as manifold projection

Following the SDE formulation of score-based diffusion models Song et al. [2020], we consider the forward-time stochastic process

$$dx(t) = f(t)\,x(t)\,dt + g(t)\,dw(t), \qquad x(0) \sim p_{\mathrm{data}}, \tag{1}$$

with marginal density p_t for $x(t)$, the corresponding reverse-time SDE (integrated from $t = T$ to 0) is

$$dx(t) = \Big[f(t)\,x(t) - g(t)^2 \nabla_x \log p_t(x(t))\Big]\,dt + g(t)\,d\bar{w}(t).$$

We denote the true score by $s^\star(x,t) := \nabla_x \log p_t(x)$. A trained diffusion model provides a score estimator $s_\theta(x,t) \approx s^\star(x,t)$.

Purification is implemented by discretizing the reverse dynamics with $s^\star$ replaced by s_θ. We isolate the score-driven update as the dominant manifold-attracting component and absorb other sampler-dependent drifts into the perturbation terms below Nie et al. [2022].

$$x_{k-1} = x_k + \eta_k\, s_\theta(x_k, t_k) + \xi_k, \tag{2}$$

where $\eta_k > 0$ and ξ_k denotes injected sampling noise.

Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact C^2 embedded manifold with positive reach, and assume a manifold-smoothing model for the marginal:

$$p_t(x) \;\approx\; \int_{\mathcal{M}} p_{\mathcal{M}}(y)\,\mathcal{N}\big(x; y, \sigma(t)^2 I\big)\,\mathrm{dVol}_{\mathcal{M}}(y). \tag{3}$$

For x in the tubular neighborhood where the metric projection is unique, define $\pi_{\mathcal{M}}(x) := \arg\min_{y\in\mathcal{M}} \|x-y\|_2$ and $n(x) := x - \pi_{\mathcal{M}}(x)$ so that $\|n(x)\|_2 = \mathrm{dist}(x,\mathcal{M})$.

To reveal the geometric nature of purification in Eqn. (2), the key is to understand how the score field behaves near $\mathcal{M}$. The next proposition shows that its leading term acts as a *normal force* pulling samples toward the manifold.

Proposition 3.1 (Normal-dominant Score Field)**.** *Under Eqn.* (3)*, for $\sigma(t)$ sufficiently small and x in a tubular neighborhood of radius strictly smaller than* $\mathrm{reach}(\mathcal{M})$*,*

$$s^\star(x,t) \;=\; -\frac{1}{\sigma(t)^2}\,n(x) \;+\; \tau_t(x), \tag{4}$$

where $\tau_t(x)$ is a remainder term whose normal *component is curvature-controlled. Let $\Pi_{\mathrm{nor}}(x)$ denote the orthogonal projector onto the normal space $\mathcal{N}_{\pi_{\mathcal{M}}(x)}\mathcal{M}$. Then there exists a constant $C > 0$ (depending on curvature bounds of $\mathcal{M}$ and regularity of $p_{\mathcal{M}}$, and uniform over t in any fixed range where Eqn.* (3) *holds) such that*

$$\big\|\Pi_{\mathrm{nor}}(x)\,\tau_t(x)\big\| \;\le\; C\,\|n(x)\|, \tag{5}$$

for all x in the tubular neighborhood.

The proof is deferred to **Appendix** A.1.

Substituting Eqn. (4) into Eqn. (2) yields the one-step decomposition

$$x_{k-1} = x_k - \alpha_k n(x_k) + \eta_k \tau_{t_k}(x_k) + \xi_k + \eta_k e_k, \quad \alpha_k = \eta_k/\sigma(t_k)^2. \tag{6}$$

where $e_k := s_\theta(x_k,t_k) - s^\star(x_k,t_k)$ denotes the score estimation error. To expose the leading-order geometric behavior, we drop the perturbation terms $\eta_k\,\tau_{t_k}(x_k)$, ξ_k, and $\eta_k\,e_k$ and obtain the idealized projection-like update

$$x_{k-1} \;\approx\; x_k - \alpha_k\, n(x_k). \tag{7}$$

Thus Eqn. (7) implements a relaxed projection step toward $\mathcal{M}$ along $n(x_k)$.

Proposition 3.2 (Local Contraction toward $\mathcal{M}$)**.** *Assume x_k stays in the tubular neighborhood where $\pi_{\mathcal{M}}$ is unique. If $\alpha_k \in (0,1)$ and the update follows Eqn.* (7)*, then*

$$\mathrm{dist}(x_{k-1},\mathcal{M}) \;\le\; (1-\alpha_k)\,\mathrm{dist}(x_k,\mathcal{M}).$$

The proof is deferred to **Appendix** A.2. (A one-line intuition is that $\mathrm{dist}(x_{k-1},\mathcal{M}) \le \|x_{k-1} - \pi_{\mathcal{M}}(x_k)\| = (1-\alpha_k)\|n(x_k)\|$.)

Sections 3.1–3.3 characterize the leading-order behavior of purification in a tubular neighborhood of $\mathcal{M}$ by focusing on the dominant normal drift. Lower-order score remainders τ_t, score estimation errors e_k Kang et al. [2023], sampler-dependent drifts, and sampling noise ξ_k are treated as perturbations and absorbed into error terms (**Appendix** A.2).

3.2 Off-manifold vs. Manifold-Aligned Perturbations

Let $x \in \mathcal{M}$ be a clean sample on the data manifold, and consider a perturbed input $x_a = x + \delta$ that lies within the tubular neighborhood where the metric projection $\pi_{\mathcal{M}}$ is unique. At x, the ambient space admits the orthogonal decomposition Do Carmo and Flaherty Francis [1992]

$$\delta = \delta_{\tan} + \delta_{\mathrm{nor}}, \quad \delta_{\tan} \in \mathcal{T}_x\mathcal{M}, \quad \delta_{\mathrm{nor}} \in \mathcal{N}_x\mathcal{M}, \tag{8}$$

where $\mathcal{T}_x\mathcal{M}$ and $\mathcal{N}_x\mathcal{M}$ denote the tangent and normal spaces of $\mathcal{M}$ at x, respectively. We refer to $\delta_{\tan}$ as the *manifold-aligned component* and δ_{nor} as the *off-manifold component* of the perturbation.

Building on the idealized purification dynamics discussed above, applying Eqn. (7) to the perturbed input x_a yields

$$x_{a,k-1} \;\approx\; x_{a,k} - \alpha_k\, n(x_{a,k}), \tag{9}$$

where $n(x) = x - \pi_{\mathcal{M}}(x)$ denotes the normal displacement defined in **Section** 3.1.

To compare how perturbations evolve under purification, we consider the (idealized) purified clean trajectory $\{x_k\}$ and the purified perturbed trajectory $\{x_{a,k}\}$ induced by Eqn. (7). We define the trajectory-wise perturbation

$$\delta^{(k)} := x_{a,k} - x_k.$$

At each step, we decompose $\delta^{(k)}$ using the tangent/normal spaces at the clean point x_k:

$$\delta^{(k)} = \delta^{(k)}_{\tan} + \delta^{(k)}_{\mathrm{nor}}, \qquad \delta^{(k)}_{\tan} \in \mathcal{T}_{x_k}\mathcal{M}, \quad \delta^{(k)}_{\mathrm{nor}} \in \mathcal{N}_{x_k}\mathcal{M}.$$

For sufficiently small perturbations such that the first-order expansion of the projection map applies,

$$\pi_{\mathcal{M}}(x+\delta) = x + \delta_{\tan} + O(\|\delta\|^2), \qquad n(x+\delta) = \delta_{\mathrm{nor}} + O(\|\delta\|^2).$$

Theorem 3.3 (Selective Suppression with Controlled Tangent Drift)**.** *Under the local assumptions of Proposition 3.2 and the one-step decomposition Eqn.* (6)*, one purification step admits the component-wise form*

$$\delta^{(k-1)}_{\mathrm{nor}} = (1-\alpha_k)\delta^{(k)}_{\mathrm{nor}} + \Delta^{(k)}_{\mathrm{nor}}, \qquad \delta^{(k-1)}_{\tan} = \delta^{(k)}_{\tan} + \Delta^{(k)}_{\tan}.$$

where $\Delta^{(k)}_{\mathrm{nor}}$ and $\Delta^{(k)}_{\tan}$ collect the effects of the score remainder τ_{t_k}, score error e_k, sampler-dependent drifts, sampling noise ξ_k, and curvature-induced variation of tangent/normal spaces across steps. Moreover, the dominant term in the normal update is the contraction $(1-\alpha_k)\delta^{(k)}_{\mathrm{nor}}$, while $\Delta^{(k)}_{\tan}$ captures the (typically smaller) drift in the tangent component.

The proof is deferred to **Appendix** A.3.

A further implication is that manifold-aligned perturbations tend to be more visually inconspicuous, as they correspond to locally coherent variations along the data manifold rather than introducing off-manifold artifacts. We first provide a perturbation morphology analysis in the **Appendix** C.1, contrasting off-manifold and manifold-aligned perturbations, **which reveals clear structural differences between the two**. We then report clean–residual similarity analyses in **Appendix** C.2 to provide additional quantitative evidence.

3.3 Why Frequency Is Not the Root Cause

Let $x \in \mathcal{M}$ and $x_a = x + \delta$ lie in the tubular neighborhood. By Proposition 3.3, purification contracts the off-manifold (normal) component of δ to leading order, while the tangent component is affected only through perturbation terms. In particular, after K idealized reverse steps, the residual off-manifold magnitude satisfies

$$\|\delta^{(K)}_{\mathrm{nor}}\| \;\lesssim\; \Big(\prod_{k=1}^{K}(1-\alpha_k)\Big)\,\|\delta_{\mathrm{nor}}\| \;+\; O(\|\delta\|^2), \tag{10}$$

where $\alpha_k = \eta_k/\sigma(t_k)^2$ and $O(\|\delta\|^2)$ collects curvature-controlled terms. Eqn. (10) indicates that purification behavior is governed by manifold alignment (normal vs. tangent), rather than being determined by frequency descriptors alone.

To formalize this point, we introduce a generic frequency decomposition. Let $U \in \mathbb{R}^{d\times d}$ be an orthonormal Fourier (or any fixed orthonormal spectral) basis and let $\{\Omega_b\}_{b=1}^{B}$ be a partition of indices. Define the band projectors $P_{\Omega_b} := U^\top \mathrm{diag}(\mathbf{1}_{\Omega_b})U$ and the band-energy profile

$$\Phi(\delta) := \big(\|P_{\Omega_1}\delta\|, \ldots, \|P_{\Omega_B}\delta\|\big) \in \mathbb{R}^B_+. \tag{11}$$

Proposition 3.4 (Insufficiency of Band-Energy Profiles)**.** *Let $\Pi_{\mathrm{nor}}(x)$ denote the orthogonal projector onto $\mathcal{N}_{\pi_{\mathcal{M}}(x)}\mathcal{M}$. Assume there exists a band b such that the restriction of $\Pi_{\mathrm{nor}}(x)$ to the band subspace $S_b := \mathrm{Im}(P_{\Omega_b})$ is neither identically zero nor the identity on S_b. Then there exist perturbations δ_1, δ_2 such that*

$$\Phi(\delta_1) = \Phi(\delta_2) \quad \textit{but} \quad \|\Pi_{\mathrm{nor}}(x)\delta_1\| \neq \|\Pi_{\mathrm{nor}}(x)\delta_2\|.$$

Consequently, frequency profiles $\Phi(\delta)$ alone cannot predict purification outcomes, which are governed (to leading order) by the off-manifold component $\Pi_{\mathrm{nor}}(x)\delta$.

The proof is deferred to **Appendix** A.4.

Proposition 3.5 (Non-necessity of Band-Energy Profiles)**.** *For any achievable profile $v \in \mathbb{R}_+^B$ and any $\varepsilon > 0$, there exists a perturbation δ such that $\Phi(\delta) = v$ while $\|\Pi_{\mathrm{nor}}(x)\delta\| \leq \varepsilon$. In particular, a perturbation can concentrate energy in specific frequency bands while remaining manifold-aligned in the local geometric sense.*

The proof is deferred to **Appendix** A.5.

Propositions 3.4–3.5 show that frequency profiles of the form Eqn. (11) are neither sufficient nor necessary to predict purification outcomes, and therefore cannot serve as a causal explanation by themselves. We provide empirical evidence supporting these implications in **Section** C.8.

4 Manifold-Aligned Adversarial Attack

This section instantiates the geometric insights in **Section** 3.1–3.2 into a practical *purification-aware* adversarial attack pipeline for anti-customization. Our goal is not to introduce a new attack primitive, but to demonstrate that perturbations exhibiting the *manifold-aligned behavior* characterized in **Section** 3 can be constructed in practice and remain effective under diffusion-based purification.

4.1 Attack Objective under Purification

Let $x \in \mathbb{R}^d$ be an input image and $\mathcal{P}_{t^*}$ denote the diffusion-based purification operator (forward noising to t^* followed by reverse denoising). We optimize the perturbation by maximizing a downstream objective evaluated on the purified output, $\max_{\delta \in \mathcal{C}} \mathcal{L}_{\mathrm{task}}(\mathcal{P}_{t^*}(x+\delta))$. where $\mathcal{C}$ encodes a perceptual or norm-based constraint. In our implementation, $\mathcal{L}_{\mathrm{task}}$ follows enhanced attack losses proposed in prior anti-customization methods (e.g., the losses used in GoodAC Xu et al. [2025] and SimAC Wang et al. [2024a], as well as the timestep selection strategy adopted by SimAC). Our emphasis, however, is not on proposing new objectives, but on analyzing how the resulting perturbations interact with diffusion-based purification.

4.2 Manifold-alignment Constraint

Directly computing the tangent space $\mathcal{T}_x\mathcal{M}$ of the unknown image manifold is intractable in high dimensions. Motivated by **Section** 3.2, we impose an explicit *manifold-alignment bias* by restricting perturbations to structured, locally coherent variations.

Concretely, we parameterize adversarial examples as $x_{\mathrm{adv}}(\vartheta) := \mathcal{H}_\vartheta(x)$, where $\mathcal{H}_\vartheta$ is a differentiable filtering operator that produces content-guided perturbations. This choice is inspired by the low-frequency/structured optimization component used in Kassis et al. [2024], but we interpret it through the manifold lens: under appropriate configurations, the resulting perturbations empirically exhibit stronger manifold alignment and thus are less affected by purification.

To further encourage manifold alignment, we regularize the filtering operator by a local consistency penalty, $\mathcal{R}_{\mathrm{filt}}(\vartheta; x) := \mathrm{Dist}(\mathcal{H}_\vartheta(x), x)$. where $\mathrm{Dist}(\cdot,\cdot)$ measures patch-level deviation between the filtered output and the guidance image. We keep the notation abstract in the main text and provide the exact instantiation in the **Appendix** B.5, including the specific design of $\mathcal{H}_\vartheta$, patch extraction, robust statistics, and norm choices.

Combining the downstream objective with the manifold-alignment bias yields

$$\max_{\vartheta} \left\{ \mathcal{L}_{\mathrm{task}}\big(\mathcal{P}_{t^*}(x_{\mathrm{adv}}(\vartheta))\big) - \lambda \mathcal{R}_{\mathrm{filt}}(\vartheta; x) - \gamma \mathcal{R}_{\mathrm{budget}}(\vartheta; x) \right\}. \tag{12}$$

where $\mathcal{R}_{\mathrm{budget}}$ enforces a perturbation budget (e.g., ℓ_p or perceptual distance), and $\lambda, \gamma > 0$ balance the trade-off.

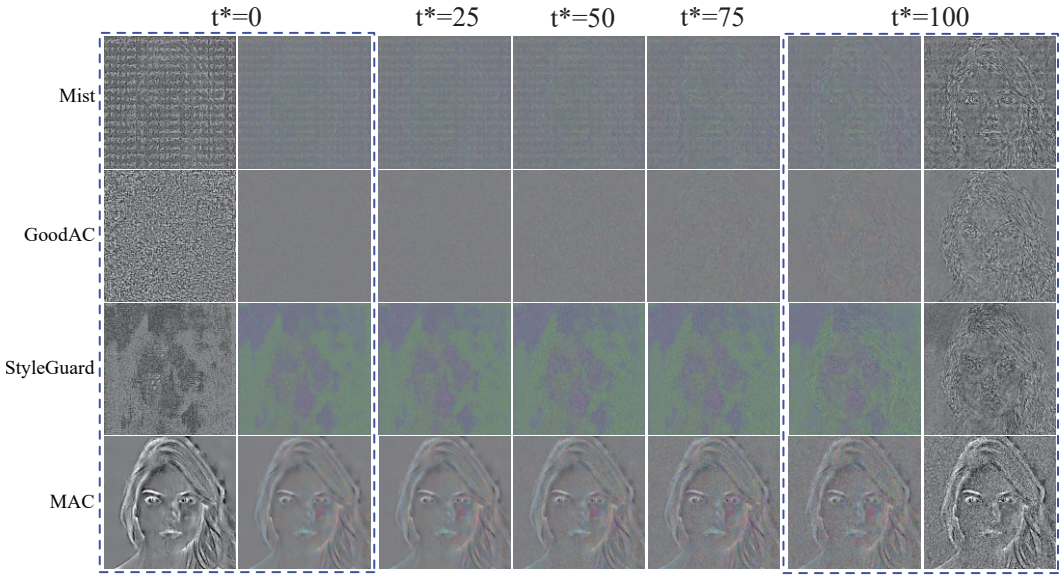


Figure 2: Evolution of perturbations under increasing diffusion purification. At each purification step t^*, the perturbation is visualized as $\mathcal{P}_{t^*}(x_{\text{adv}}) - x_{\text{original}}$. Fine stripe-like structures produced by existing methods are progressively suppressed, while MAC maintains a stable morphology with mainly noise accumulation. For clarity, enhanced visualizations of the perturbations at $t^* = 0$ and $t^* = 100$ are additionally provided.

4.3 Purification-Aware Optimization

Optimizing Eqn. (12) requires gradients through the diffusion purifier. We adopt *purification-aware differentiation* techniques introduced in DiffBreak Kassis et al. [2024] to backpropagate through the diffusion-based purifier, treating $\mathcal{P}_{t^*}$ as a differentiable module by unrolling a finite number of reverse denoising steps. The update takes the generic form

$$\vartheta \leftarrow \vartheta + \eta \nabla_{\vartheta} \big[\mathcal{L}_{\text{task}}\big(\mathcal{P}_{t^*}(x_{\text{adv}}(\vartheta))\big) - \lambda \mathcal{R}_{\text{filt}}(\vartheta; x) - \gamma \mathcal{R}_{\text{budget}}(\vartheta; x) \big]. \tag{13}$$

where η is the step size.

All implementation details and hyperparameters are reported in **Appendix** B.3 and **Appendix** B.2; the algorithmic pseudocode is provided in **Appendix** B.1.

5 Experiments

5.1 Setup

Datasets. We evaluate on CelebA-HQ Karras et al. [2017] and VGGFace2-HQ Cao et al. [2018] following prior anti-customization methods. We further include WikiArt for artistic-style customization and ImageNet and MS-COCO for natural-image perturbation-retention analysis, examining whether the proposed purification-retention mechanism extends beyond facial images.
Models. All quantitative results are on Stable Diffusion v2.1 (SD-v2.1) Rombach et al. [2022a] under the standard protocol in Van Le et al. [2023] and Li et al. [2025]. We additionally test black-box transfer to SD-v1.5 Rombach et al. [2022b] in **Appendix** C.3.
Baselines. We use DreamBooth Ruiz et al. [2023] for customization. We compare against Mist Liang and Wu [2023], Anti-DreamBooth (Anti-DB) Van Le et al. [2023], SimAC Wang et al. [2024a], GoodAC Xu et al. [2025], and StyleGuard Li et al. [2025], using official code and recommended hyperparameters. All methods share the same purifier.
Evaluation Metrics. For facial personalization on CelebA-HQ and VGGFace2-HQ, we report FDFR using RetinaFace Deng et al. [2020], ISM Van Le et al. [2023], SER-FIQ Terhorst et al. [2020], and BRISQUE Mittal et al. [2012]. For non-facial evaluations, we use task-appropriate metrics: FID and Precision for WikiArt Saleh and Elgammal [2015] artistic-style customization. Detailed metric definitions and interpretations are provided in **Appendix** B.4.

Table 1: Results on CelebA-HQ. FDFR is scaled by $\times 100$. **Bold** denotes the best value, underlined denotes the second-best distinct value. For MAC, subscripts indicate the relative change with respect to GoodAC†. $\uparrow/\downarrow$ indicate increase/decrease, respectively.

	"a photo of sks person"				"a dslr portrait of sks person"			
Method	ISM↓	FDFR↑	BRISQUE↑	SER-FIQ↓	ISM↓	FDFR↑	BRISQUE↑	SER-FIQ↓
Mist	0.65	2.21	20.36	0.75	0.46	18.20	7.30	0.75
Anti-DB	0.70	2.80	17.19	0.77	0.47	17.20	6.85	0.76
SimAC	0.69	2.71	17.84	0.77	0.45	17.71	7.10	0.74
GoodAC†	0.67	2.75	17.62	0.76	0.44	18.00	7.45	0.72
StyleGuard	0.70	2.27	19.32	0.78	0.46	20.98	**13.63**	0.76
MAC-ℓ_∞	0.56	6.26	23.60	0.72	0.36	33.11	7.54	0.67
MAC (w/o pure)	0.56	5.78	15.53	0.73	0.38	30.67	11.09	0.63
MAC	**0.48**$_{\downarrow 28.4\%}$	**7.91**$_{\uparrow 187.6\%}$	**26.04**$_{\uparrow 47.8\%}$	**0.67**$_{\downarrow 11.8\%}$	**0.32**$_{\downarrow 27.3\%}$	**36.86**$_{\uparrow 104.8\%}$	9.72$_{\uparrow 30.5\%}$	**0.60**$_{\downarrow 16.7\%}$
	"a photo of sks person looking at the mirror"				"a photo of sks person in front of eiffel tower"			
Method	ISM↓	FDFR↑	BRISQUE↑	SER-FIQ↓	ISM↓	FDFR↑	BRISQUE↑	SER-FIQ↓
Mist	0.41	7.95	16.40	0.55	**0.11**	29.90	**20.40**	0.34
Anti-DB	0.42	6.70	15.60	0.58	0.14	**33.80**	18.90	0.35
SimAC	0.42	7.29	16.14	0.56	0.13	31.04	19.41	0.33
GoodAC†	0.40	7.85	16.90	0.53	0.12	30.10	20.10	**0.32**
StyleGuard	0.46	9.22	16.14	0.58	0.17	29.22	19.91	0.38
MAC-ℓ_∞	0.37	11.78	17.41	0.54	0.12	26.00	18.99	0.34
MAC (w/o pure)	0.40	11.32	16.42	0.54	0.12	23.78	17.79	0.37
MAC	**0.34**$_{\downarrow 15.0\%}$	**13.72**$_{\uparrow 74.8\%}$	**17.68**$_{\uparrow 4.6\%}$	**0.49**$_{\downarrow 7.5\%}$	0.12$_{0.0\%}$	24.17$_{\downarrow 19.7\%}$	17.48$_{\downarrow 13.0\%}$	0.38$_{\uparrow 18.8\%}$

Purification Settings. We use DiffPure-style Nie et al. [2022] preprocessing: forward noising to a high-noise timestep and DDPM reverse sampling back toward $t{=}0$. Unless stated otherwise, we add noise at $t_0 \in [950, 999]$ and run 50 reverse denoising steps (a fixed schedule shared by all methods) without EOT, which is sufficient to reach a stronger purification regime while keeping computational cost tractable. A sensitivity study over reverse timesteps is reported in **Appendix** C.4. Black-Box transfer to additional defense and updated purification settings are evaluated in **Appendix** C.7.

Perturbation Budgets. We consider two threat models. (i) *Fair* ℓ_∞*:* $\|x_{\text{adv}} - x\|_\infty \leq 16/255$ (default for main tables; Anti-DB/SimAC/GoodAC/StyleGuard and MAC-ℓ_∞). (ii) *LPIPS:* Zhang et al. [2018] $\text{LPIPS}(x, x_{\text{adv}}) \leq 0.25$ for MAC (beyond ℓ_∞ while preserving perceptual similarity). Mist follows its official setting and does not enforce an explicit ℓ_∞ or LPIPS bound.

5.2 Experimental Results

Main results on CelebA-HQ. Table 1 summarizes the main results on CelebA-HQ under DiffPure-style purification. Overall, MAC achieves stronger protection on most prompts, with clear improvements in identity-related metrics such as ISM and FDFR. Importantly, MAC-ℓ_∞ also remains competitive under the same ℓ_∞ constraint as prior methods, indicating that the improvement is not simply caused by a larger perturbation budget. In addition, MAC (w/o pure) consistently underperforms full MAC, suggesting that explicitly optimizing through the purifier is beneficial under diffusion-based purification. Additional qualitative visualizations are provided in **Appendix** D.

The *"a photo of sks person in front of eiffel tower"* prompt is less diagnostic. In this setting, DreamBooth customization itself often fails to produce stable facial content across methods, making face-centric metrics less sensitive and less aligned with visual protection quality. We therefore treat this prompt as a stress case and avoid over-interpreting small margins on identity metrics.

Perturbation morphology under purification. Figure 2 visualizes how perturbations evolve as diffusion purification becomes stronger. For baseline methods such as Mist, GoodAC, and StyleGuard, the initial stripe-like or noisy structures are progressively suppressed. **At stronger purification levels, semantic facial structures re-emerge in the perturbation visualization, accompanied by the removal of most stripe-like or noisy structures.** This behavior is consistent with our manifold-alignment hypothesis: score-driven contraction preferentially attenuates normal, off-manifold perturbation components, effectively pulling the purified image back toward the natural image manifold. By contrast, MAC maintains a more stable morphology across purification steps, suggesting that its perturbations better aligned with directions preserved by diffusion purification. **Appendix** C.5 reports perturbation-retention analyses on VGGFace2, WikiArt, ImageNet and MS-COCO, showing that the same trend appears across facial, artistic, and natural-image domains.

Table 2: Ablation study of optional components under the prompt "a photo of sks person" on CelebA-HQ for **MAC**. TS: Time-step selection; EL: Enhanced attack loss from prior work.

TS	EL	Filter	DiffGrad	ISM↓	FDFR↑	BRISQUE↑	SER-FIQ↓
✗	✗	✗	✗	0.70	2.80	17.47	0.81
✓	✓	✗	✗	0.67	2.75	17.69	0.77
✗	✗	✓	✓	0.62	4.23	21.57	0.71
✗	✓	✓	✓	0.60	4.93	23.01	0.73
✓	✗	✓	✓	<u>0.54</u>	<u>5.50</u>	**25.43**	<u>0.68</u>
✓	✓	✗	✓	0.65	5.12	22.33	0.69
✓	✓	✓	✗	0.64	5.27	22.13	<u>0.68</u>
✓	✓	✓	✓	**0.48**	**8.11**	<u>24.65</u>	**0.66**

Table 3: Evaluation of text-to-image protection methods under different transformations for DreamBooth on the WikiArt dataset. Metrics reported are FID↑ (higher is better) and Precision↓ (lower is better). **Bold** denotes the best value, and <u>underlined</u> values denote the second-best distinct value.

Method	No Prep.		Crop+Resize		Gauss. Noise		DiffPure		Noise Up.	
	FID	Prec.	FID	Prec.	FID	Prec.	FID	Prec.	FID	Prec.
No Protect	96.80	0.46	116.12	0.55	98.82	0.35	95.45	0.47	94.86	0.50
Mist	158.25	**0.00**	150.84	0.02	160.91	**0.00**	97.60	0.25	95.96	0.48
Anti-DB	133.48	0.02	97.03	0.18	131.66	0.03	98.18	0.23	88.69	0.57
SimAC	168.72	**0.00**	158.90	<u>0.01</u>	155.80	<u>0.01</u>	99.50	0.23	96.01	0.44
GoodAC	**186.98**	**0.00**	**179.15**	**0.00**	<u>175.93</u>	**0.00**	102.10	0.21	93.00	0.55
StyleGuard	<u>179.51</u>	**0.00**	158.40	0.02	**187.73**	**0.00**	127.80	0.09	113.72	0.15
MAC-ℓ_∞	164.28	<u>0.01</u>	167.22	**0.00**	166.80	**0.00**	<u>186.46</u>	<u>0.04</u>	<u>180.79</u>	<u>0.01</u>
MAC	176.34	**0.00**	<u>172.73</u>	0.03	173.48	**0.00**	**223.30**	**0.00**	**212.98**	**0.00**

Ablation study. Table 2 studies the contribution of different components in MAC. The full model achieves the strongest overall protection, confirming that time-step selection, enhanced attack loss, the content-guided filter, and DiffGrad provide complementary benefits. In particular, removing the filter leads to a substantial degradation in downstream protection, showing that the explicit manifold-alignment mechanism is crucial for MAC. Without this structured perturbation parameterization, the optimized perturbations are less biased toward purification-retentive directions and are therefore more easily suppressed by the purifier. A more detailed ablation analysis is provided in **Appendix** C.6.

Evaluation beyond facial images. Table 3 reports DreamBooth protection results on the WikiArt dataset under multiple transformations. Following an evaluation setup adapted from StyleGuard, we evaluate protection by comparing generations from DreamBooth models trained on clean versus protected WikiArt images, using 150 generated images for each class in our evaluation. This experiment demonstrates that MAC remains effective in the artistic-image domain under diffusion-based purification. Together with the ImageNet perturbation-retention analysis in **Appendix** C.5, these results provide complementary evidence that the proposed purification-retention mechanism is not limited to facial images. **Complete frequency-domain experience are reported in Appendix C.8.**

6 Conclusion

We study adversarial perturbations under diffusion-based purification and analyze **why some adversarial perturbations remain effective after purification**. Our geometric analysis suggests that robustness is closely related to *manifold alignment*: perturbation components preserved along the purification trajectory are more likely to survive, while off-manifold components tend to be suppressed. Motivated by this view, we propose a purification-aware attack pipeline that encourages structured and locally coherent perturbations as a practical proxy for manifold-aligned optimization. Experiments across datasets, prompts, diffusion backbones, and purification settings show that MAC consistently preserves protection effectiveness after purification. Further analyses of purification dynamics, ablations, and frequency behavior indicate that robustness is not determined by a fixed spectral signature alone, but by the preservation of effective perturbation components during purification. Overall, our findings establish purification retention as a key principle for robust anti-customization.

Acknowledgments and Disclosure of Funding

Funding. This research received no external funding.

Competing Interests. The authors declare no competing interests.

References

Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593, 2014.

Mingyuan Bai, Wei Huang, Tenghui Li, Andong Wang, Junbin Gao, Cesar F Caiafa, and Qibin Zhao. Diffusion models demand contrastive guidance for adversarial purification to advance. In *Forty-first international conference on machine learning*, 2024.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003. ISSN 0899-7667.

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

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. ISBN 1538623358.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pages 5253–5270, 2023.

Yiting Chen, Qibing Ren, and Junchi Yan. Rethinking and improving robustness of convolutional neural networks: a shapley value-based approach in frequency domain. *Advances in neural information processing systems*, 35:324–337, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020.

Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian geometry*, volume 2. Springer, 1992.

Lawrence C Evans. *Partial differential equations*, volume 19. American mathematical society, 2022.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3): 418–491, 1959.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Sicong Han, Chenhao Lin, Chao Shen, and Qian Wang. Rethinking adversarial examples exploiting frequency-based analysis. In *International Conference on Information and Communications Security*, pages 73–89. Springer, 2021.

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

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

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. *ICLR*, 1(2):3, 2022.

Mintong Kang, Dawn Song, and Bo Li. Diffattack: Evasion attacks against diffusion-based adversarial purification. *Advances in Neural Information Processing Systems*, 36:73919–73942, 2023.

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

Andre Kassis, Urs Hengartner, and Yaoliang Yu. Diffbreak: Is diffusion-based purification robust? *arXiv preprint arXiv:2411.16598*, 2024.

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*, pages 1931–1941, 2023.

John M Lee. Smooth manifolds. In *Introduction to smooth manifolds*, pages 1–29. Springer, 2003.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Yanjie Li, Wenxuan Zhang, Xinqi Lyu, Yihao Liu, and Bin Xiao. Styleguard: Preventing text-to-image-model-based style mimicry attacks by style perturbations. *arXiv preprint arXiv:2505.18766*, 2025.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

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.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. ISSN 1057-7149.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

Takuto Onikubo and Yusuke Matsui. High-frequency anti-dreambooth: Robust defense against personalized image synthesis. *arXiv preprint arXiv:2409.08167*, 2024.

Gaozheng Pei, Ke Ma, Yingfei Sun, Qianqian Xu, and Qingming Huang. Diffusion-based adversarial purification from the perspective of the frequency domain. *arXiv preprint arXiv:2505.01267*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022a.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022b.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000. ISSN 1095-9203.

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*, pages 22500–22510, 2023.

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

Kaiyu Song, Hanjiang Lai, Yan Pan, and Jian Yin. Mimicdiffusion: Purifying adversarial perturbation via mimicking clean diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24665–24674, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Elias M Stein and Rami Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press, 2009.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000. ISSN 1095-9203.

Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5651–5660, 2020.

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

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*, pages 12047–12056, 2024a.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8684–8694, 2020.

Kaibo Wang, Xiaowen Fu, Yuxuan Han, and Yang Xiang. Diffhammer: Rethinking the robustness of diffusion-based adversarial purification. *Advances in Neural Information Processing Systems*, 37: 89535–89562, 2024b.

Yihan Wang, Yiwei Lu, Xiao-Shan Gao, Gautam Kamath, and Yaoliang Yu. Bridgepure: Limited protection leakage can break black-box data protection. *Advances in Neural Information Processing Systems*, 38:28175–28209, 2026.

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

Haotian Xue and Yongxin Chen. Pixel is a barrier: Diffusion models are more adversarially robust than we think. *arXiv preprint arXiv:2404.13320*, 2024.

Wenkui Yang, Jie Cao, Junxian Duan, and Ran He. Towards robust defense against customization via protective perturbation resistant to diffusion-based purification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24398–24407, 2024.

Appendix Roadmap

This appendix provides additional theoretical details, implementation details, experimental analyses, and qualitative visualizations.

Appendix Section	Page
A. Proofs	14
A.1 Proof of Proposition 3.1	14
A.2 Proof of Proposition 3.2	16
A.3 Proof of Theorem 3.3	17
A.4 Proof of Proposition 3.4	18
A.5 Proof of Proposition 3.5	19
B. Implementation Details	20
B.1 Algorithm Pseudocode	20
B.2 MAC-ℓ_∞ Variant	20
B.3 Hyperparameters	21
B.4 Evaluation Metrics	21
B.5 Manifold-alignment Filter	22
C. Additional Experimental Results	24
C.1 Perturbation Morphology Analysis	24
C.2 Input-Level Similarity Analysis	25
C.3 Additional Results on Transferability and VGGFace2-HQ	25
C.4 Sensitivity to Purifier Reverse Timesteps	27
C.5 Perturbation Retention across Different Datasets	28
C.6 Full Ablation Study with All Evaluation Metrics	30
C.7 Black-Box Transfer to Additional Purification Defenses	31
C.8 Frequency-domain Analysis	32
D. Visualization	35
D.1 CelebA-HQ Results on MAC	35
D.2 VGGFace2-HQ Results on MAC	36
D.3 CelebA-HQ Results on Baseline Methods	37
D.4 WikiArt Results on MAC	39
E. Limitations and Broader Impacts	40

A Proof

A.1 Proof of Proposition 3.1

Proof of Proposition 3.1. Fix t and denote $\sigma := \sigma(t)$. Let x lie in the tubular neighborhood $\{x : \operatorname{dist}(x, \mathcal{M}) < \operatorname{reach}(\mathcal{M})\}$ so that the metric projection $y := \pi_{\mathcal{M}}(x)$ is unique Federer [1959], Lee [2003]. Write

$$n := n(x) = x - y \in \mathcal{N}_y\mathcal{M}, \qquad \|n\| = \operatorname{dist}(x, \mathcal{M}).$$

Let $m = \dim(\mathcal{M})$. Since $\mathcal{M}$ is C^2, in a neighborhood of y we may parameterize $\mathcal{M}$ as a C^2 graph over the tangent space $T_y\mathcal{M}$ Lee [2003]: for u in a neighborhood of $0 \in \mathbb{R}^m$,

$$\varphi(u) = y + Uu + h(u),$$

where $U \in \mathbb{R}^{d\times m}$ has orthonormal columns spanning $T_y\mathcal{M}$, $h(u) \in \mathcal{N}_y\mathcal{M}$, and (by C^2 regularity / bounded curvature) there exists $C_h > 0$ such that

$$\|h(u)\| \le C_h\|u\|^2, \qquad \|\nabla h(u)\| \le C_h\|u\|. \tag{14}$$

Let $J(u)$ denote the associated volume element, so $J(u) = 1 + O(\|u\|^2)$ and $\nabla J(u) = O(\|u\|)$ on this chart. Define $a(u) := p_{\mathcal{M}}(\varphi(u))J(u)$; by compactness and regularity, a is C^1 and bounded above/below on the chart.

Using $x = y + n$ and $U^\top n = 0$ (because $n \in \mathcal{N}_y\mathcal{M}$), we expand the squared distance with a chain of equalities:

$$\begin{aligned}
\|x - \varphi(u)\|^2 &= \|y + n - (y + Uu + h(u))\|^2 \\
&= \|n - Uu - h(u)\|^2 \\
&= \|n\|^2 + \|u\|^2 + \|h(u)\|^2 - 2\langle n, h(u)\rangle + 2\langle Uu, h(u)\rangle \\
&= \|n\|^2 + \|u\|^2 + R(u; n),
\end{aligned} \tag{15}$$

where the remainder satisfies the chain bound (from (14)):

$$\begin{aligned}
|R(u;n)| &\le \|h(u)\|^2 + 2|\langle n, h(u)\rangle| + 2|\langle Uu, h(u)\rangle| \\
&\le C_h^2\|u\|^4 + 2\|n\|\, C_h\|u\|^2 + 2\|u\|\, C_h\|u\|^2 \\
&\le C_R\big(\|n\|\,\|u\|^2 + \|u\|^3 + \|u\|^4\big)
\end{aligned} \tag{16}$$

for some constant C_R depending only on curvature bounds and chart regularity.

By the manifold-smoothing model (3), which is standard in score-based generative modeling Alain and Bengio [2014], Song et al. [2020], restricting to a sufficiently small chart (the complement contributes an exponentially small tail under compactness and Gaussian decay), we obtain

$$\begin{aligned}
p_t(x) &\propto \int_{\mathbb{R}^m} a(u) \exp\Big(-\frac{\|x - \varphi(u)\|^2}{2\sigma^2}\Big)\, du \\
&\overset{(15)}{=} \exp\Big(-\frac{\|n\|^2}{2\sigma^2}\Big) \int_{\mathbb{R}^m} a(u) \exp\Big(-\frac{\|u\|^2}{2\sigma^2}\Big) \exp\Big(-\frac{R(u;n)}{2\sigma^2}\Big)\, du.
\end{aligned} \tag{17}$$

Define

$$Z(n) := \int_{\mathbb{R}^m} a(u) \exp\Big(-\frac{\|u\|^2}{2\sigma^2}\Big) \exp\Big(-\frac{R(u;n)}{2\sigma^2}\Big)\, du, \tag{18}$$

so that (17) implies, up to an x-independent constant,

$$\log p_t(x) = -\frac{\|n\|^2}{2\sigma^2} + \log Z(n) + \operatorname{const}(t). \tag{19}$$

Taking gradients, and noting that the *normal* component of ∇_x coincides with ∇_n in tubular coordinates, we have the chain identity

$$\begin{aligned}
\Pi_{\mathrm{nor}}(x)\, s^\star(x,t) &= \Pi_{\mathrm{nor}}(x)\nabla_x \log p_t(x) \\
&= \nabla_n \log p_t(x) \\
&= \nabla_n\Big(-\frac{\|n\|^2}{2\sigma^2} + \log Z(n)\Big) \\
&= -\frac{1}{\sigma^2} n + \nabla_n \log Z(n).
\end{aligned} \tag{20}$$

Thus the decomposition (4) holds with

$$\tau_t(x) := s^\star(x,t) + \frac{1}{\sigma^2} n(x), \qquad \Pi_{\mathrm{nor}}(x)\tau_t(x) = \nabla_n \log Z(n(x)).$$

It remains to prove $\|\nabla_n \log Z(n)\| \le C\|n\|$ for n corresponding to x in the tubular neighborhood, uniformly over t in any fixed range where (3) holds.

Differentiate $Z(n)$ under the integral sign (justified by dominated convergence Evans [2022], Stein and Shakarchi [2009]; dominated convergence follows from (16), boundedness of a, and Gaussian decay). From (18),

$$\begin{aligned}\nabla_n Z(n) &= \int a(u) e^{-\frac{\|u\|^2}{2\sigma^2}} \nabla_n \exp\Big(-\frac{R(u;n)}{2\sigma^2}\Big)\, du \\ &= -\frac{1}{2\sigma^2}\int a(u) e^{-\frac{\|u\|^2}{2\sigma^2}} e^{-\frac{R(u;n)}{2\sigma^2}} \nabla_n R(u;n)\, du. \end{aligned} \tag{21}$$

From (15), the dependence on n enters only through $-2\langle n, h(u)\rangle$, hence

$$\nabla_n R(u;n) = -2h(u). \tag{22}$$

Substituting (22) into (21) yields the clean ratio form

$$\begin{aligned}\nabla_n \log Z(n) &= \frac{\nabla_n Z(n)}{Z(n)} \\ &= \frac{1}{\sigma^2}\cdot \frac{\int a(u) e^{-\frac{\|u\|^2}{2\sigma^2}} e^{-\frac{R(u;n)}{2\sigma^2}}\, h(u)\, du}{\int a(u) e^{-\frac{\|u\|^2}{2\sigma^2}} e^{-\frac{R(u;n)}{2\sigma^2}}\, du}. \end{aligned} \tag{23}$$

Now observe that the "tilt" induced by $R(u;n)$ is weak on the Gaussian scale $\|u\| = O(\sigma)$. Indeed, by (16), for x in the tubular neighborhood and σ sufficiently small,

$$\begin{aligned} -\frac{R(u;n)}{2\sigma^2} &\le \frac{C_R}{2\sigma^2}\big(\|n\|\,\|u\|^2 + \|u\|^3 + \|u\|^4\big) \\ &\le C_1\|n\|\frac{\|u\|^2}{\sigma^2} + C_2\frac{\|u\|^3}{\sigma^2} \end{aligned} \tag{24}$$

for constants C_1, C_2 (absorbing $\|u\|^4/\sigma^2$ into $\|u\|^3/\sigma^2$ on a small neighborhood). Using also $\|h(u)\| \le C_h\|u\|^2$, we bound the numerator of (23) by

$$\begin{aligned}\Big\| \int a(u) e^{-\frac{\|u\|^2}{2\sigma^2}} e^{-\frac{R(u;n)}{2\sigma^2}} h(u)\, du\Big\| &\le \int |a(u)| e^{-\frac{\|u\|^2}{2\sigma^2}} e^{-\frac{R(u;n)}{2\sigma^2}} \|h(u)\|\, du \\ &\le C\int e^{-\frac{\|u\|^2}{2\sigma^2}} \exp\Big(C_1\|n\|\frac{\|u\|^2}{\sigma^2}\Big)\exp\Big(C_2\frac{\|u\|^3}{\sigma^2}\Big)\,\|u\|^2\, du. \end{aligned} \tag{25}$$

Apply the change of variables $u = \sigma z$ to make the σ-scaling explicit:

$$\begin{aligned}\text{RHS of (25)} &= C\sigma^{m+2}\int e^{-\frac{\|z\|^2}{2}} \exp\big(C_1\|n\|\|z\|^2\big)\exp\big(C_2\sigma\|z\|^3\big)\,\|z\|^2\, dz \\ &\le C'\sigma^{m+2}\int e^{-\frac{\|z\|^2}{4}} \exp\big(C_1\|n\|\|z\|^2\big)\,\|z\|^2\, dz \\ &\le C''\sigma^{m+2}\,\|n\|, \end{aligned} \tag{26}$$

where the second inequality uses $\exp(C_2\sigma\|z\|^3) \le \exp(\|z\|^2/4)$ for sufficiently small σ (absorbing it into the Gaussian tail), and the last inequality uses that the integrand is even in z and the first non-vanishing dependence on n in the normalized ratio is linear in $\|n\|$ (equivalently, $\nabla_n \log Z(0) = 0$ and $\nabla_n \log Z(n)$ is Lipschitz near 0 Fefferman et al. [2016]; this Lipschitzness follows from the same dominated-convergence control applied to $\nabla_n^2 Z$).

Similarly, the denominator in (23) admits a uniform lower bound of order σ^m:

$$\begin{aligned} Z(n) &= \int a(u) e^{-\frac{\|u\|^2}{2\sigma^2}} e^{-\frac{R(u;n)}{2\sigma^2}}\, du \\ &\geq c \int e^{-\frac{\|u\|^2}{2\sigma^2}} \exp\Big(- C_3\|n\| \frac{\|u\|^2}{\sigma^2} - C_4 \frac{\|u\|^3}{\sigma^2} \Big)\, du \\ &= c\sigma^m \int e^{-\frac{\|z\|^2}{2}} \exp\big(- C_3\|n\|\|z\|^2 - C_4\sigma\|z\|^3 \big)\, dz \\ &\geq c'\sigma^m, \end{aligned} \tag{27}$$

for σ small enough and $\|n\|$ bounded within the tubular neighborhood.

Combining (23), (26), and (27) gives the desired normal bound as a chain:

$$\begin{aligned} \|\nabla_n \log Z(n)\| &\leq \frac{1}{\sigma^2} \cdot \frac{C''\sigma^{m+2}\|n\|}{c'\sigma^m} \\ &\leq C\|n\|. \end{aligned} \tag{28}$$

Together with (20), this proves (4) and (5). □

A.2 Proof of Proposition 3.2

Lemma A.1 (Leading-order normal drift with absorbed perturbations)**.** *Assume x_k stays in a tubular neighborhood where $\pi_{\mathcal{M}}$ is unique Lee [2003], and the one-step purification update can be written as*

$$x_{k-1} = x_k - \alpha_k n(x_k) \;+\; \varepsilon_k, \qquad \alpha_k = \frac{\eta_k}{\sigma(t_k)^2}, \tag{29}$$

where the aggregated perturbation ε_k collects all non-leading terms, e.g.

$$\varepsilon_k := \eta_k\, \tau_{t_k}(x_k) \;+\; \eta_k\, e_k \;+\; \xi_k \;+\; d_k,$$

with d_k denoting sampler-dependent drifts (if any) not included in the dominant normal term. Then the distance to $\mathcal{M}$ satisfies the perturbed contraction bound

$$\operatorname{dist}(x_{k-1}, \mathcal{M}) \;\leq\; (1-\alpha_k)\operatorname{dist}(x_k, \mathcal{M}) \;+\; \|\varepsilon_k\|. \tag{30}$$

If in addition the normal component of τ_{t_k} is curvature-controlled as in Proposition 3.1, i.e. $\|\Pi_{\mathrm{nor}}(x_k)\tau_{t_k}(x_k)\| \leq C\|n(x_k)\|$, then

$$\operatorname{dist}(x_{k-1}, \mathcal{M}) \;\leq\; \big(1-\alpha_k+\eta_k C\big)\operatorname{dist}(x_k, \mathcal{M}) \;+\; \|\xi_k\|+\eta_k\|e_k\|+\|d_k\|+\eta_k\|\Pi_{\tan}(x_k)\tau_{t_k}(x_k)\|. \tag{31}$$

Proof. Let $y_k := \pi_{\mathcal{M}}(x_k) \in \mathcal{M}$ and $n_k := n(x_k) = x_k - y_k$, so that $\|n_k\| = \operatorname{dist}(x_k, \mathcal{M})$ Federer [1959]. From (29),

$$\begin{aligned} \operatorname{dist}(x_{k-1}, \mathcal{M}) &= \min_{y\in\mathcal{M}} \|x_{k-1} - y\| \\ &\leq \|x_{k-1} - y_k\| \\ &= \|x_k - \alpha_k n_k + \varepsilon_k - y_k\| \\ &= \|(x_k - y_k) - \alpha_k n_k + \varepsilon_k\| \\ &= \|(1-\alpha_k)n_k + \varepsilon_k\| \\ &\leq (1-\alpha_k)\|n_k\| + \|\varepsilon_k\| \\ &= (1-\alpha_k)\operatorname{dist}(x_k, \mathcal{M}) \;+\; \|\varepsilon_k\|, \end{aligned} \tag{32}$$

which proves (30).

For the refined bound, decompose $\tau_{t_k}(x_k)$ into normal and tangent parts: $\tau_{t_k} = \Pi_{\mathrm{nor}}\tau_{t_k} + \Pi_{\tan}\tau_{t_k}$, and expand $\varepsilon_k = \eta_k\tau_{t_k}(x_k) + \eta_k e_k + \xi_k + d_k$ to get

$$\begin{aligned} \|\varepsilon_k\| &\leq \eta_k\|\Pi_{\mathrm{nor}}(x_k)\tau_{t_k}(x_k)\| + \eta_k\|\Pi_{\tan}(x_k)\tau_{t_k}(x_k)\| + \eta_k\|e_k\| + \|\xi_k\| + \|d_k\| \\ &\leq \eta_k C\|n_k\| + \eta_k\|\Pi_{\tan}(x_k)\tau_{t_k}(x_k)\| + \eta_k\|e_k\| + \|\xi_k\| + \|d_k\|, \end{aligned} \tag{33}$$

where the second line uses Proposition 3.1. Substituting (33) into (30) yields (31). □

A.3 Proof of Theorem 3.3

Proof of Theorem 3.3. Write the one-step purified updates in the decomposed form (6) for both the clean and perturbed trajectories:

$$\begin{aligned} x_{k-1} &= x_k - \alpha_k n(x_k) + r_k, \\ x_{a,k-1} &= x_{a,k} - \alpha_k n(x_{a,k}) + r_{a,k}, \end{aligned} \tag{34}$$

where we collect all non-leading terms into

$$r_k := \eta_k \tau_{t_k}(x_k) + \xi_k + \eta_k e_k + d_k, \qquad r_{a,k} := \eta_k \tau_{t_k}(x_{a,k}) + \xi_{a,k} + \eta_k e_{a,k} + d_{a,k},$$

and $d_k, d_{a,k}$ denote sampler-dependent drifts (if any) absorbed into perturbations. Subtracting the two recursions gives the exact perturbation recursion

$$\begin{aligned} \delta^{(k-1)} &:= x_{a,k-1} - x_{k-1} \\ &= \big(x_{a,k} - x_k\big) \;-\; \alpha_k\big(n(x_{a,k}) - n(x_k)\big) \;+\; \big(r_{a,k} - r_k\big) \\ &= \delta^{(k)} - \alpha_k\big(n(x_k + \delta^{(k)}) - n(x_k)\big) + \Delta_r^{(k)}, \end{aligned} \tag{35}$$

where $\Delta_r^{(k)} := r_{a,k} - r_k$ aggregates the effects of τ_{t_k}, score error, sampling noise, and sampler-dependent drifts.

Let $\Pi_{\tan}(x_k)$ and $\Pi_{\mathrm{nor}}(x_k)$ be the orthogonal projectors onto $\mathcal{T}_{x_k}\mathcal{M}$ and $\mathcal{N}_{x_k}\mathcal{M}$, respectively, and recall the decomposition

$$\delta^{(k)} = \delta_{\tan}^{(k)} + \delta_{\mathrm{nor}}^{(k)}, \qquad \delta_{\tan}^{(k)} := \Pi_{\tan}(x_k)\delta^{(k)}, \quad \delta_{\mathrm{nor}}^{(k)} := \Pi_{\mathrm{nor}}(x_k)\delta^{(k)}.$$

Under the stated local assumptions (tubular neighborhood, sufficiently small perturbations), the projection map $\pi_{\mathcal{M}}$ admits the first-order expansion at x_k Federer [1959], Fefferman et al. [2016]:

$$\pi_{\mathcal{M}}(x_k + \delta) = \pi_{\mathcal{M}}(x_k) + \Pi_{\tan}(x_k)\delta + E_\pi^{(k)}(\delta), \qquad \|E_\pi^{(k)}(\delta)\| \le C_\pi \|\delta\|^2, \tag{36}$$

where C_π depends on curvature bounds (uniform on the tubular neighborhood). Since $n(z) = z - \pi_{\mathcal{M}}(z)$, we obtain the corresponding expansion for $n(\cdot)$:

$$\begin{aligned} n(x_k + \delta) - n(x_k) &= \big(x_k + \delta - \pi_{\mathcal{M}}(x_k + \delta)\big) - \big(x_k - \pi_{\mathcal{M}}(x_k)\big) \\ &= \delta - \big(\pi_{\mathcal{M}}(x_k + \delta) - \pi_{\mathcal{M}}(x_k)\big) \\ &= \delta - \Pi_{\tan}(x_k)\delta - E_\pi^{(k)}(\delta) \\ &= \Pi_{\mathrm{nor}}(x_k)\delta + E_n^{(k)}(\delta), \end{aligned} \tag{37}$$

with $E_n^{(k)}(\delta) := -E_\pi^{(k)}(\delta)$ and hence $\|E_n^{(k)}(\delta)\| \le C_\pi\|\delta\|^2$. Substituting (37) into (35) yields

$$\begin{aligned} \delta^{(k-1)} &= \delta^{(k)} - \alpha_k\Big(\Pi_{\mathrm{nor}}(x_k)\delta^{(k)} + E_n^{(k)}(\delta^{(k)})\Big) + \Delta_r^{(k)} \\ &= \Big(I - \alpha_k \Pi_{\mathrm{nor}}(x_k)\Big)\delta^{(k)} + \underbrace{\Big(\Delta_r^{(k)} - \alpha_k E_n^{(k)}(\delta^{(k)})\Big)}_{=:\widetilde{\Delta}^{(k)}} \\ &= \delta_{\tan}^{(k)} + (1-\alpha_k)\delta_{\mathrm{nor}}^{(k)} + \widetilde{\Delta}^{(k)}. \end{aligned} \tag{38}$$

To obtain component-wise recursions, project (38) onto the tangent/normal spaces. Note that the decomposition at time $(k-1)$ is taken with respect to the clean point x_{k-1}, hence we explicitly account for the variation of projectors:

$$\delta_{\tan}^{(k-1)} = \Pi_{\tan}(x_{k-1})\delta^{(k-1)}, \qquad \delta_{\mathrm{nor}}^{(k-1)} = \Pi_{\mathrm{nor}}(x_{k-1})\delta^{(k-1)}.$$

Using (38), we derive the normal component as a chain:

$$\begin{aligned} \delta_{\mathrm{nor}}^{(k-1)} &= \Pi_{\mathrm{nor}}(x_{k-1})\delta^{(k-1)} \\ &= \Pi_{\mathrm{nor}}(x_{k-1})\Big(\delta_{\tan}^{(k)} + (1-\alpha_k)\delta_{\mathrm{nor}}^{(k)} + \widetilde{\Delta}^{(k)}\Big) \\ &= (1-\alpha_k)\Pi_{\mathrm{nor}}(x_{k-1})\delta_{\mathrm{nor}}^{(k)} \;+\; \Pi_{\mathrm{nor}}(x_{k-1})\delta_{\tan}^{(k)} \;+\; \Pi_{\mathrm{nor}}(x_{k-1})\widetilde{\Delta}^{(k)} \\ &= (1-\alpha_k)\delta_{\mathrm{nor}}^{(k)} \;+\; \underbrace{(1-\alpha_k)\big(\Pi_{\mathrm{nor}}(x_{k-1}) - \Pi_{\mathrm{nor}}(x_k)\big)\delta_{\mathrm{nor}}^{(k)} + \Pi_{\mathrm{nor}}(x_{k-1})\delta_{\tan}^{(k)} + \Pi_{\mathrm{nor}}(x_{k-1})\widetilde{\Delta}^{(k)}}_{=:\,\Delta_{\mathrm{nor}}^{(k)}}, \end{aligned} \tag{39}$$

where we used $\Pi_{\mathrm{nor}}(x_k)\delta_{\mathrm{nor}}^{(k)} = \delta_{\mathrm{nor}}^{(k)}$ and inserted $\Pi_{\mathrm{nor}}(x_k)$ to isolate the dominant contraction term. Similarly, projecting onto the tangent space yields

$$\begin{aligned}\delta_{\mathrm{tan}}^{(k-1)} &= \Pi_{\mathrm{tan}}(x_{k-1})\delta^{(k-1)} \\ &= \Pi_{\mathrm{tan}}(x_{k-1})\Big(\delta_{\mathrm{tan}}^{(k)} + (1-\alpha_k)\delta_{\mathrm{nor}}^{(k)} + \widetilde{\Delta}^{(k)}\Big) \\ &= \delta_{\mathrm{tan}}^{(k)} + \underbrace{\big(\Pi_{\mathrm{tan}}(x_{k-1}) - \Pi_{\mathrm{tan}}(x_k)\big)\delta_{\mathrm{tan}}^{(k)} + (1-\alpha_k)\Pi_{\mathrm{tan}}(x_{k-1})\delta_{\mathrm{nor}}^{(k)} + \Pi_{\mathrm{tan}}(x_{k-1})\widetilde{\Delta}^{(k)}}_{=:\,\Delta_{\mathrm{tan}}^{(k)}},\end{aligned} \tag{40}$$

where we used $\Pi_{\mathrm{tan}}(x_k)\delta_{\mathrm{tan}}^{(k)} = \delta_{\mathrm{tan}}^{(k)}$.

The two identities (39)–(40) give the claimed component-wise form

$$\delta_{\mathrm{nor}}^{(k-1)} = (1-\alpha_k)\delta_{\mathrm{nor}}^{(k)} + \Delta_{\mathrm{nor}}^{(k)}, \qquad \delta_{\mathrm{tan}}^{(k-1)} = \delta_{\mathrm{tan}}^{(k)} + \Delta_{\mathrm{tan}}^{(k)}.$$

By construction, $\Delta_{\mathrm{nor}}^{(k)}$ and $\Delta_{\mathrm{tan}}^{(k)}$ collect exactly the effects of (i) the score remainder through $\widetilde{\Delta}^{(k)}$ (which contains $\Delta_r^{(k)}$ and the quadratic projection linearization error $E_n^{(k)}$), (ii) the score estimation errors and sampling noise inside $\Delta_r^{(k)}$, (iii) sampler-dependent drifts $d_k, d_{a,k}$, and (iv) the curvature-induced variation of tangent/normal spaces across steps via the projector differences $\Pi_{\mathrm{tan}}(x_{k-1}) - \Pi_{\mathrm{tan}}(x_k)$ and $\Pi_{\mathrm{nor}}(x_{k-1}) - \Pi_{\mathrm{nor}}(x_k)$. Finally, the dominant term in the normal update is explicitly $(1-\alpha_k)\delta_{\mathrm{nor}}^{(k)}$, establishing selective suppression of the off-manifold component, while the tangent component has no multiplicative contraction term at leading order and hence evolves primarily through the drift term $\Delta_{\mathrm{tan}}^{(k)}$. □

A.4 Proof of Proposition 3.4

Proof of Proposition 3.4. Fix x and abbreviate $\Pi := \Pi_{\mathrm{nor}}(x)$. Let b be a band satisfying the assumption that the restriction $\Pi|_{S_b}$ is neither identically zero nor the identity on $S_b := \mathrm{Im}(P_{\Omega_b})$ Horn and Johnson [2012]. Equivalently, there exist vectors $u, v \in S_b$ such that

$$\Pi u \neq 0, \qquad \Pi v \neq v. \tag{41}$$

Since S_b is a linear subspace and Π is a linear map, the set $\{\Pi z : \ z \in S_b,\ \|z\| = 1\}$ is not a singleton; in particular the Rayleigh quotient $z \mapsto \|\Pi z\|^2$ is not constant on the unit sphere in S_b. Hence there exist unit vectors $z_1, z_2 \in S_b$ such that

$$\|z_1\| = \|z_2\| = 1, \qquad \|\Pi z_1\| \neq \|\Pi z_2\|. \tag{42}$$

Now define perturbations

$$\delta_1 := z_1, \qquad \delta_2 := z_2.$$

Because $z_1, z_2 \in S_b = \mathrm{Im}(P_{\Omega_b})$, they are supported entirely in band b, i.e. $P_{\Omega_b}\delta_i = \delta_i$ and $P_{\Omega_{b'}}\delta_i = 0$ for all $b' \neq b$. Therefore their band-energy profiles satisfy the chain equality

$$\begin{aligned}\Phi(\delta_1) &= \big(\|P_{\Omega_1}\delta_1\|, \ldots, \|P_{\Omega_B}\delta_1\|\big) \\ &= \big(0, \ldots, 0, \underbrace{\|\delta_1\|}_{=1}, 0, \ldots, 0\big) \\ &= \big(0, \ldots, 0, \underbrace{\|\delta_2\|}_{=1}, 0, \ldots, 0\big) \\ &= \big(\|P_{\Omega_1}\delta_2\|, \ldots, \|P_{\Omega_B}\delta_2\|\big) = \Phi(\delta_2).\end{aligned} \tag{43}$$

On the other hand, by (42),

$$\|\Pi\delta_1\| = \|\Pi z_1\| \neq \|\Pi z_2\| = \|\Pi\delta_2\|.$$

This constructs δ_1, δ_2 with identical band-energy profiles but different off-manifold magnitudes, proving the proposition. □

A.5 Proof of Proposition 3.5

Proof of Proposition 3.5. Fix x and abbreviate $\Pi := \Pi_{\mathrm{nor}}(x)$. Let $v = (v_1, \dots, v_B) \in \mathbb{R}_+^B$ be an achievable band-energy profile and let $\varepsilon > 0$. By achievability (under the fixed pair $(\Pi, \{P_{\Omega_b}\})$), for each band b there exists a unit vector $u_b \in S_b := \mathrm{Im}(P_{\Omega_b})$ such that

$$\|u_b\| = 1, \qquad \|\Pi u_b\| \le \frac{\varepsilon}{\sqrt{B}\,\max\{1, \|v\|_2\}}. \tag{44}$$

Define

$$\delta := \sum_{b=1}^{B} v_b\, u_b. \tag{45}$$

We first verify $\Phi(\delta) = v$ using the orthogonality of the band projectors. Since $\{\Omega_b\}$ is a partition and U is orthonormal, the subspaces S_b are mutually orthogonal and P_{Ω_b} is the orthogonal projector onto S_b. Therefore, for each b,

$$\begin{aligned} P_{\Omega_b}\delta &= P_{\Omega_b} \sum_{j=1}^{B} v_j u_j \\ &= \sum_{j=1}^{B} v_j P_{\Omega_b} u_j \\ &= \sum_{j=1}^{B} v_j\, \mathbf{1}_{\{j=b\}}\, u_j \\ &= v_b u_b, \end{aligned} \tag{46}$$

and hence

$$\|P_{\Omega_b}\delta\| = \|v_b u_b\| = v_b. \tag{47}$$

Combining (47) over $b = 1, \dots, B$ yields $\Phi(\delta) = v$.

It remains to bound the off-manifold magnitude $\|\Pi\delta\|$. By linearity of Π and the triangle/Cauchy–Schwarz inequalities,

$$\begin{aligned} \|\Pi\delta\| &= \Big\|\Pi \sum_{b=1}^{B} v_b u_b\Big\| \\ &\le \sum_{b=1}^{B} v_b \|\Pi u_b\| \\ &\le \Big(\sum_{b=1}^{B} v_b^2\Big)^{1/2} \Big(\sum_{b=1}^{B} \|\Pi u_b\|^2\Big)^{1/2} \\ &\le \|v\|_2 \cdot \Big(B \cdot \frac{\varepsilon^2}{B\,\max\{1, \|v\|_2\}^2}\Big)^{1/2} \\ &\le \varepsilon, \end{aligned} \tag{48}$$

where the penultimate inequality uses (44). Thus we constructed δ with $\Phi(\delta) = v$ and $\|\Pi_{\mathrm{nor}}(x)\delta\| \le \varepsilon$, which proves the proposition. □

B Implementation Details

B.1 Algorithm Pseudocode

Algorithm 1 MAC: Surrogate-guided purification-aware optimization with time-step selection (TS)

Require: Training set $\mathcal{D}_{\text{train}}$; images to protect $\mathcal{D}_{\text{adv}}$ with corresponding originals O; (optional) class set $\mathcal{D}_{\text{class}}$; pretrained diffusion model (f_θ, e_ϕ) and scheduler $\mathcal{S}$; VAE (omitted for brevity); differentiable purifier $\mathcal{P}_{t^*}$, prompt p; outer iterations T, surrogate update steps T_f, inner steps T_a; filter $\mathcal{H}_\vartheta(\cdot; O)$, modifier m; timestep candidate set $\mathcal{T}$ (TS); loss weights $\lambda_{\text{filt}}, \lambda_{\text{dist}}, \lambda_{\text{reg}}$ and β, γ; budget set $\mathcal{C}$.

Ensure: Protected images X.

Initialize $X \leftarrow \mathcal{D}_{\text{adv}}$; set $O \leftarrow X$ (fixed copy); initialize (ϑ, m) and optimizer over (ϑ, m)
for $i = 1$ to T **do** ▷ Outer iterations
  Surrogate update.
    $(f_{\theta'}, e_{\phi'}) \leftarrow \text{Clone}(f_\theta, e_\phi)$
    $(f_{\theta'}, e_{\phi'}) \leftarrow \text{Train}(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{class}}, p; T_f)$ ▷ with prior preservation if enabled
  for $k = 1$ to T_a **do** ▷ Inner optimization
    (1) Filter parameterization (manifold-alignment bias).
      $X_b \leftarrow \tanh(\text{arctanh}(X + m))$
      $\tilde{X} \leftarrow \mathcal{H}_\vartheta(X_b; O)$
    (2) Differentiable purification (gradient flows through $\mathcal{P}_{t^*}$).
      $X_p \leftarrow \mathcal{P}_{t^*}(\tilde{X})$
    (3) Time-step selection (TS) and task loss.
      Sample $t \sim \mathcal{T}$; compute base denoising loss $\ell_{\text{mse}}(X_p, t)$ using $(f_{\theta'}, e_{\phi'}, \mathcal{S})$
      Compute auxiliary losses: $\ell_{\text{mix1}}, \ell_{\text{mix3}}, \ell_{\text{canny}}, \ell_{\text{tgt}}$
      $\ell_{\text{task}} \leftarrow \ell_{\text{mse}} + \text{stopgrad}(\ell_{\text{mix1}}) + \text{stopgrad}(\ell_{\text{mix3}}) + \beta\,\text{stopgrad}(\ell_{\text{canny}}) + \gamma\,\text{stopgrad}(\ell_{\text{tgt}})$
    (4) Regularizers and update.
      $\mathcal{L}_{\text{filt}} \leftarrow \lambda_{\text{filt}} \cdot \text{FilterLoss}(\tilde{X}; O)$
      $\mathcal{L}_{\text{dist}} \leftarrow \lambda_{\text{dist}} \cdot \text{DistLoss}(\tilde{X}, O)$
      $\mathcal{L}_{\text{reg}} \leftarrow \lambda_{\text{reg}} \cdot \text{RegLoss}(\tilde{X}, O)$
      $\mathcal{L} \leftarrow \mathcal{L}_{\text{filt}} + \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{reg}} - \ell_{\text{task}}$
      Take one optimizer step on (ϑ, m) using $\nabla_{\vartheta,m}\mathcal{L}$
      $X \leftarrow \Pi_{\mathcal{C}}(\tilde{X})$
  end for
  Main model update.
    $(f_\theta, e_\phi) \leftarrow \text{Train}(X, \mathcal{D}_{\text{class}}, p; T_f)$
end for
return X

In our implementation, the surrogate update follows the training procedure of Anti-DreamBooth Van Le et al. [2023]. The auxiliary loss terms ℓ_{mix1} and ℓ_{mix3}, as well as the time-step selection strategy, are adopted from SimAC Wang et al. [2024a]. The edge-based regularization ℓ_{canny} and the target feature loss ℓ_{tgt} follow GoodAC Xu et al. [2025]. We implement differentiable purification by unrolling the reverse diffusion process and adopt a learnable, content-guided filtering parameterization following DiffBreak Kassis et al. [2024]. These components are integrated into a unified purification-aware optimization framework, while our contribution lies in the manifold-aligned interpretation and the structured perturbation parameterization that improves robustness under purification.

B.2 MAC-ℓ_∞ Variant

In addition to the default MAC setting, we implement a constrained variant, denoted as MAC-ℓ_∞, to enable fair comparison under a strict ℓ_∞ threat model. Specifically, we first generate adversarial images using the standard MAC optimization procedure, and then project the resulting perturbations onto an ℓ_∞ ball with radius $16/255$ around the original image. Formally, given the MAC output

x_{MAC} and the clean image x, we compute

$$x_{\text{MAC-}\ell_\infty} = \Pi_{\|\cdot\|_\infty \le 16/255}(x_{\mathrm{MAC}} - x) + x,$$

followed by clipping to the valid pixel range.

This post-hoc projection does not modify the optimization procedure of MAC, but enforces a strict ℓ_∞ constraint on the final adversarial images. We use this variant only for evaluation under the fair ℓ_∞ protocol.

B.3 Hyperparameters

Unless otherwise specified, all hyperparameters are fixed across methods. We use $T{=}50$ outer iterations, with $T_f{=}3$ surrogate (and main-model) training steps and $T_a{=}6$ inner-loop updates per iteration. Surrogate and main models are optimized with AdamW (learning rate $5{\times}10^{-7}$, $\beta_1{=}0.9$, $\beta_2{=}0.999$, weight decay 10^{-2}, $\epsilon{=}10^{-8}$), while the learnable filter parameters are optimized with Adam, using learning rates 0.05 for the kernels and 0.008 for the modifier. Gradient clipping with maximum norm 1.0 is applied throughout. Diffusion-based purification follows a DiffPure-style configuration using the DDPM-CelebAHQ-256 model. Both during optimization and evaluation, purification is implemented by forward noising to a timestep sampled from $[950, 999]$ and unrolling the reverse denoising process back to $t{=}0$, with mixed-precision training for efficiency. The same purifier configuration is used for all methods during evaluation. For time-step selection (TS), a candidate set $\mathcal{T}_i$ is constructed offline for each image, and timesteps are sampled uniformly from $\mathcal{T}_i$ during optimization; when $|\mathcal{T}_i| \le 100$, the set is fixed thereafter. Under the ℓ_∞ protocol, we enforce the constraint $\|x_{\mathrm{adv}} - x\|_\infty \le 16/255$, for perceptual-budget experiments, we enforce $\mathrm{LPIPS}(x, x_{\mathrm{adv}}) \le 0.25$. We set `eot_iters=1` (i.e., no EOT), and all experiments use fixed random seeds for reproducibility. Unless otherwise specified, experiments are conducted on 30 individuals.For **Tables** 1, 7, and 6, we report averages over 50 individuals. All experiments are conducted on NVIDIA H200 GPUs, requiring approximately 80GB of GPU memory at runtime. The optimization time for each protected image is approximately 30 minutes.

B.4 Evaluation Metrics

We use task-specific evaluation metrics for facial personalization and artistic-style customization, as the two settings capture different notions of protection. For facial personalization on CelebA-HQ and VGGFace2-HQ, we follow the evaluation protocol of Anti-DreamBooth Van Le et al. [2023]. For artistic-style customization on WikiArt, we follow the evaluation protocol adapted from StyleGuard Li et al. [2025]. In both cases, stronger protection means that the customized model fails to preserve the target identity or artistic style after being trained on protected images.

Facial personalization metrics. For facial personalization, we report four complementary metrics that measure identity preservation, face detectability, image quality, and face recognizability. Since protection effectiveness under purification is multi-faceted, we do not rely on any single metric.

ISM↓. ISM measures identity similarity between generated images and the reference identity. Lower ISM indicates stronger identity disruption and therefore stronger protection.

FDFR↑. FDFR measures the face detection failure rate using RetinaFace Deng et al. [2020]. Higher FDFR indicates that generated outputs fail to preserve detectable facial content, directly reflecting stronger disruption of downstream facial personalization. In practice, we compute FDFR over all generated images, scale it by $\times 100$, and include detection failures in the denominator.

BRISQUE↑. BRISQUE is a no-reference image-quality metric. In our protection setting, higher BRISQUE indicates stronger degradation of the naturalness or visual quality of generated outputs.

SER-FIQ↓. SER-FIQ measures face image quality and recognizability. Lower SER-FIQ indicates that the generated face is less recognizable or less useful for identity preservation.

Together, these metrics capture different failure modes of facial personalization: identity similarity, face detectability, perceptual naturalness, and face recognizability. We therefore interpret protection effectiveness based on consistent trends across metrics, prompts, and datasets, rather than on a single scalar score.

Artistic-style customization metrics. For WikiArt, we evaluate style protection by comparing generations from DreamBooth models trained on clean artworks and protected artworks. Specifically, we first fine-tune a DreamBooth model on clean artworks and generate a set of images using the style-customization prompt. We then fine-tune another DreamBooth model on protected artworks and generate another set of images using the same prompt. FID and Precision are computed between these two generated sets.

FID↑. FID measures the distributional distance between clean-trained and protected-trained generations in the Inception feature space. In this setting, higher FID indicates that protected generations deviate more from clean generations, suggesting stronger disruption of style mimicry.

Precision↓. Precision measures the fraction of protected-model generations that remain within the feature-space manifold of clean-model generations. Lower Precision indicates less overlap with clean generations and therefore stronger protection against style imitation.

Unlike standard image-generation evaluation, here FID and Precision are computed between clean-trained and protected-trained generations, rather than between generated images and real images. Their interpretation is therefore task-specific: higher FID and lower Precision indicate stronger disruption of the target artistic style.

B.5 Manifold-alignment Filter

This appendix specifies the concrete instantiation of the differentiable content-guided filtering operator $\mathcal{H}_\vartheta$ and the associated structural regularizer $\mathcal{R}_{\text{filt}}$ used in **Section** 4.2. Our implementation follows the differentiable filtering formulation introduced in DiffBreak Kassis et al. [2024], with adaptations tailored to the anti-customization setting and the optimization of structured perturbations.

Overview. Given an input image $u \in \mathbb{R}^{C\times H\times W}$ and a guidance image $x \in \mathbb{R}^{C\times H\times W}$, the operator $\mathcal{H}_\vartheta(u; x)$ produces a locally coherent output by aggregating values from a spatial neighborhood of u using (i) learnable spatial aggregation weights and (ii) content-guided similarity weights computed from x. The operator is applied sequentially over a fixed multi-scale set of kernels. In all stages, the guidance image is fixed to the original clean image x, while the input to each stage is the output of the previous one.

Patch extraction and receptive field. For each kernel $k_b = (k_h^{(b)}, k_w^{(b)})$, we extract overlapping $k_h^{(b)} \times k_w^{(b)}$ patches from both u and x with unit stride, after reflection padding to avoid boundary artifacts. The filter produces an output at every spatial location; the effective receptive field is therefore determined solely by the kernel size. Let $P_i^{(b)}(\cdot) \in \mathbb{R}^{C\times k_h^{(b)}\times k_w^{(b)}}$ denote the patch extracted at location i, and let $P_{i,j}^{(b)}(\cdot) \in \mathbb{R}^C$ denote the value at offset index j within that patch.

Learnable spatial weights. For each kernel k_b, we maintain a set of learnable spatial parameters $\tilde{w}^{(b)}$ in the log domain. To reduce redundancy, only a canonical quadrant of the kernel is parameterized; the full kernel is reconstructed via reflection padding. The parameters are shared across all spatial locations and normalized using a softmax over patch offsets:

$$w_j^{(b)} = \frac{\exp\big(\tilde{w}_j^{(b)}\big)}{\sum_{j'} \exp\big(\tilde{w}_{j'}^{(b)}\big)}, \qquad \sum_j w_j^{(b)} = 1. \tag{49}$$

This construction ensures non-negativity, unit-sum normalization, and numerical stability, constraining the purely spatial aggregation to be a convex local average.

Content-guided similarity weighting. To incorporate structural information from the guidance image x, the spatial weights are modulated by a content similarity term. For each location i and patch offset j, we compute an ℓ_p content distance between the guidance pixel at the center and its neighbor:

$$d_{i,j}^{(b)} = \big\|x_{i+j} - x_i\big\|_p^2, \tag{50}$$

where $p = 1$ is used in all experiments. The squared distance is converted into a similarity weight via a Gaussian kernel

$$\kappa(d) = \exp\left(-\frac{d}{2\sigma_{\text{color}}^2}\right), \tag{51}$$

with $\sigma_{\text{color}} = 0.1$. The final aggregation weights are obtained by multiplicative modulation followed by renormalization:

$$\hat{w}_{i,j}^{(b)} = \frac{w_j^{(b)}\,\kappa\big(d_{i,j}^{(b)}\big)}{\sum_{j'} w_{j'}^{(b)}\,\kappa\big(d_{i,j'}^{(b)}\big) + \varepsilon}, \tag{52}$$

where $\varepsilon = 10^{-8}$ prevents numerical degeneracy.

Filtering operation. With the normalized weights, the filtered output for kernel k_b is

$$\mathcal{H}_\vartheta^{(b)}(u; x)_i = \sum_j \hat{w}_{i,j}^{(b)}\, P_{i,j}^{(b)}(u). \tag{53}$$

Applying the operator sequentially across all kernels yields the final output

$$\mathcal{H}_\vartheta(u; x) = \mathcal{H}_\vartheta^{(B)}(\cdot\,; x) \circ \cdots \circ \mathcal{H}_\vartheta^{(1)}(\cdot\,; x)\,(u), \tag{54}$$

where each stage takes the previous stage output as input while reusing the same fixed guidance image x.

Structural consistency regularization. To encourage local coherence while remaining robust to outliers, we introduce a structural regularizer based on patch-wise median statistics. For each kernel k_b, we compute a median map from the guidance image by taking the channel-wise median over spatial offsets within each local patch and broadcasting it back to the image grid. Denoting this median map by $m^{(b)}(x)$, the regularizer is defined as

$$\mathcal{R}_{\text{filt}}(\vartheta; x) = \sum_{b=1}^{B} \frac{1}{CHW} \sum_i \sum_j \big\| x_{i+j} - m_i^{(b)}(x) \big\|_p^2, \tag{55}$$

with $p = 2$. This term penalizes locally incoherent deviations while allowing structure-preserving variations aligned with the guidance image. The regularizer is weighted by $\lambda_{\text{filt}} = 0.015$ in the overall optimization objective.

Hyperparameters. All filter-related hyperparameters are fixed across experiments. We use a kernel set

$$\mathcal{K} = \{(21, 5), (5, 5), (17, 33), (7, 7), (17, 17), (5, 5), (3, 3)\}, \qquad B = 7,$$

applied sequentially in the listed order. Reflection padding and unit stride are used throughout, with box size $(1, 1)$ corresponding to per-pixel output. The content similarity bandwidth is set to $\sigma_{\text{color}} = 0.1$, the renormalization constant to $\varepsilon = 10^{-8}$, and all spatial weights $\tilde{w}^{(b)}$ are initialized to zero, yielding uniform kernels at initialization.

Discussion. Overall, the differentiable filter $\mathcal{H}_\vartheta$ provides a structured parameterization that biases optimization toward locally coherent, content-aligned variations. Under a local linear approximation of the data manifold, this acts as a practical proxy for promoting manifold-aligned directions, complementing the theoretical analysis in **Section** 3.1.

C Additional Experimental Results

C.1 Perturbation Morphology Analysis

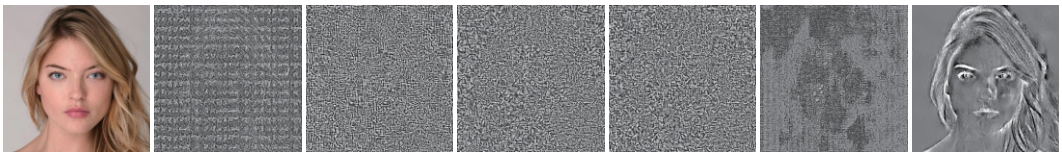


Figure 3: Perturbation morphology ($x_{\mathrm{adv}} - x_{\mathrm{original}}$) produced by different protection methods, visualized with enhanced contrast to highlight spatial structure. PGD-based baselines show fine stripe-like patterns, while StyleGuard introduces additional coarse semantic structures.

This section provides a qualitative analysis of perturbation morphology to complement the quantitative results in the main text. We visualize pixel-space perturbations $x_{\mathrm{adv}} - x_{\mathrm{original}}$ under different visualization schemes to examine their spatial organization and evolution under diffusion-based purification. We focus on visually salient properties such as locality, anisotropy, and coherent stroke-like or semantically related structures. These properties help distinguish unstructured sampling noise from perturbations with organized components that may be relevant to purification robustness.

Figure 3 compares perturbations produced by different protection methods using a visualization scheme that emphasizes spatial structure. Anti-DreamBooth, SimAC, and GoodAC exhibit similar perturbation morphologies dominated by fine *stripe-like structures*. This similarity is consistent with their shared PGD-based optimization backbone, although their objectives differ. Since SimAC extends Anti-DreamBooth and GoodAC further augments SimAC with additional objectives, we use GoodAC as a representative baseline for this method family in the following discussion.

We also include StyleGuard as a comparison method. StyleGuard is designed to improve robustness against diffusion-based purification. Its perturbations retain some fine stripe-like components, but also introduce coarser semantically related facial structures. This observation is consistent with our manifold-alignment perspective: perturbations with more image-structured components are more likely to contain directions that remain effective after purification, whereas purely fine-grained or off-manifold components are more easily suppressed.

Figure 4 presents a direct comparison using raw residual visualizations, covering Gaussian noise, Mist, GoodAC, MAC, and StyleGuard. Unlike the isotropic randomness of Gaussian noise, GoodAC and StyleGuard produce irregular, thin stripe-like structures. Mist yields wave-like patterns while retaining coarse semantic outlines of the subject. StyleGuard also shows noticeable color deviations, which correspond to the color distortions observed in images generated by DreamBooth models trained on StyleGuard-protected data. In contrast, MAC produces perturbations with more spatially coherent structures, consistent with its design goal of encouraging locally organized perturbations that better survive purification.

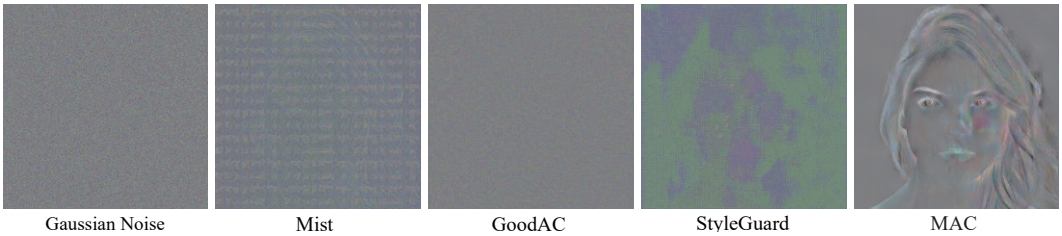


Figure 4: Residual visualization of perturbations. GoodAC produces irregular stripe-like structures distinct from Gaussian noise; Mist yields wave-like patterns with coarse semantic outlines; StyleGuard exhibits both stripe-like structures and noticeable color deviations; MAC exhibits more spatially coherent perturbations.

C.2 Input-Level Similarity Analysis

Table 4: Clean–perturbed perceptual similarity on CelebA-HQ. CLIP and DINO report cosine similarity (↑), and LPIPS reports perceptual distance (↓). Values are reported as mean ± std.

Method	CLIP↑	DINO↑	LPIPS↓
Mist	0.46 ± 0.0123	0.06 ± 0.0045	0.84 ± 0.0312
Anti-DB	0.47 ± 0.0108	0.09 ± 0.0061	0.80 ± 0.0284
SimAC	0.47 ± 0.0097	0.09 ± 0.0058	0.80 ± 0.0279
GoodAC	0.48 ± 0.0102	0.09 ± 0.0060	0.80 ± 0.0267
StyleGuard	0.52 ± 0.0102	0.11 ± 0.0060	0.79 ± 0.0267
MAC-ℓ_∞	$\mathbf{0.64 \pm 0.0089}$	$\mathbf{0.50 \pm 0.0127}$	$\mathbf{0.62 \pm 0.0214}$
MAC	$\underline{0.63 \pm 0.0075}$	$\underline{0.46 \pm 0.0119}$	$\underline{0.64 \pm 0.0196}$

We analyze two complementary input-level properties on CelebA-HQ. First, we measure the similarity between the clean image x and the protected image x_{adv}, as reported in **Table** 5. This evaluates whether a protection method preserves the perceptual and semantic content of the original image before DreamBooth fine-tuning. Second, we measure the similarity between the clean image x and the added residual $\delta = x_{\mathrm{adv}} - x$, as reported in **Table** 4. This evaluates whether the perturbation itself contains image-aligned semantic or structural components, rather than being purely noise-like.

In **Table** 5, MAC-ℓ_∞ achieves the best LPIPS score and the highest DINO similarity, while maintaining competitive CLIP similarity. This indicates that the bounded MAC variant preserves the visual structure of the clean image and introduces substantially less perceptible changes. The full MAC variant is less constrained and therefore introduces stronger changes that are more beneficial for downstream protection after purification. Nevertheless, MAC still obtains the second-best LPIPS score, showing that its perturbations remain perceptually close to the clean image.

Table 5: Perceptual similarity metrics between the clean image x and the adversarially perturbed image x_{adv} on CelebA-HQ. CLIP and DINO report cosine similarity, and LPIPS measures perceptual distance.

Metric	**Mist**	**Anti-DB**	**SimAC**	**GoodAC**	**StyleGuard**	**MAC-ℓ_∞**	**MAC**
CLIP↑	0.909	0.904	<u>0.916</u>	**0.926**	0.912	0.895	0.801
DINO↑	0.830	0.907	0.908	<u>0.909</u>	0.905	**0.914**	0.814
LPIPS↓	0.516	0.416	0.425	0.413	0.458	**0.143**	<u>0.182</u>

Table 4 provides a complementary clean–residual analysis. Compared with prior baselines, MAC and MAC-ℓ_∞ show stronger similarity between the clean image and the added residual under CLIP and DINO, together with lower LPIPS distance. This suggests that the learned residuals are less like arbitrary high-frequency noise and more compatible with image-structured variations. These results support our geometric interpretation: perturbations that are more aligned with the local image manifold tend to be more perceptually coherent and more likely to survive diffusion purification.

C.3 Additional Results on Transferability and VGGFace2-HQ

We provide additional experiments to further evaluate the robustness and generalizability of the proposed method. Specifically, we consider two complementary settings: black-box transfer from Stable Diffusion v2.1 to Stable Diffusion v1.5 on CelebA-HQ, and additional evaluation on VGGFace2-HQ under Stable Diffusion v2.1.

Black-box transfer to Stable Diffusion v1.5. We first evaluate the black-box transferability of the proposed method by applying perturbations generated on Stable Diffusion v2.1 to Stable Diffusion v1.5 on CelebA-HQ, without any re-optimization. This setting tests whether the learned perturbations remain effective when the personalization model differs from the one used during perturbation generation.

Overall, the results on Stable Diffusion v1.5 closely mirror those on Stable Diffusion v2.1. MAC maintains strong protection after transfer and continues to outperform prior baselines in most settings. We observe a slight degradation in FDFR compared to the results on v2.1, while other perceptual and identity-related metrics remain comparable or are marginally improved. A plausible explanation is that Stable Diffusion v1.5 has lower baseline generation fidelity than v2.1, which may obscure face detection failures while also reducing the sensitivity of perceptual and identity-based metrics. This

Table 6: Results on CelebA-HQ under black-box transfer to Stable Diffusion v1.5. FDFR is scaled by $\times 100$. **Bold** denotes the best value, underlined denotes the second-best distinct value. For MAC, subscripts indicate the relative change with respect to GoodAC†. $\uparrow/\downarrow$ indicate increase/decrease, respectively.

	"a photo of sks person"				"a dslr portrait of sks person"			
Method	ISM$\downarrow$	FDFR$\uparrow$	BRISQUE$\uparrow$	SER-FIQ$\downarrow$	ISM$\downarrow$	FDFR$\uparrow$	BRISQUE$\uparrow$	SER-FIQ$\downarrow$
Mist	0.62	1.45	20.30	0.70	0.54	2.65	14.82	0.66
Anti-DB	0.71	1.20	22.15	0.69	0.48	3.35	12.69	0.71
SimAC	0.69	1.32	19.44	0.73	0.52	2.85	15.92	0.63
GoodAC†	0.63	1.95	21.80	0.65	0.46	3.48	12.95	0.68
StyleGuard	0.62	0.98	25.83	0.75	0.48	1.67	**19.99**	0.72
MAC-ℓ_∞	0.53	5.35	27.99	0.60	0.46	2.85	14.62	0.57
MAC (w/o pure)	0.58	4.80	25.11	0.67	0.42	3.20	12.54	**0.52**
MAC	**0.48**$_{\downarrow 23.8\%}$	**6.35**$_{\uparrow 225.6\%}$	**29.80**$_{\uparrow 36.7\%}$	**0.56**$_{\downarrow 13.8\%}$	**0.36**$_{\downarrow 21.7\%}$	**3.51**$_{\uparrow 0.9\%}$	16.98$_{\uparrow 31.1\%}$	0.65$_{\downarrow 4.4\%}$
	"a photo of sks person looking at the mirror"				"a photo of sks person in front of eiffel tower"			
Method	ISM$\downarrow$	FDFR$\uparrow$	BRISQUE$\uparrow$	SER-FIQ$\downarrow$	ISM$\downarrow$	FDFR$\uparrow$	BRISQUE$\uparrow$	SER-FIQ$\downarrow$
Mist	0.50	6.65	23.14	0.58	0.16	52.80	**16.97**	0.18
Anti-DB	0.54	5.85	22.20	0.63	0.11	54.90	14.35	0.12
SimAC	0.47	6.10	23.48	0.55	0.14	56.00	13.91	0.16
GoodAC†	0.51	6.85	22.84	0.59	**0.09**	57.60	15.84	**0.10**
StyleGuard	0.48	4.12	25.76	0.67	0.19	55.49	16.68	0.20
MAC-ℓ_∞	0.44	9.45	22.76	0.44	0.17	54.20	16.42	0.14
MAC (w/o pure)	0.41	8.97	21.39	**0.40**	0.15	51.69	13.63	0.12
MAC	**0.36**$_{\downarrow 29.4\%}$	**11.82**$_{\uparrow 72.7\%}$	**26.95**$_{\uparrow 18.0\%}$	0.48$_{\downarrow 18.6\%}$	0.12	**58.73**$_{\uparrow 2.0\%}$	11.43	0.11

Table 7: Results on VGGFace2-HQ. FDFR is scaled by $\times 100$. **Bold** denotes the best value, underlined denotes the second-best distinct value. For MAC, subscripts indicate the relative change with respect to GoodAC†. $\uparrow/\downarrow$ indicate increase/decrease, respectively.

	"a photo of sks person"				"a dslr portrait of sks person"			
Method	ISM$\downarrow$	FDFR$\uparrow$	BRISQUE$\uparrow$	SER-FIQ$\downarrow$	ISM$\downarrow$	FDFR$\uparrow$	BRISQUE$\uparrow$	SER-FIQ$\downarrow$
Mist	0.72	2.60	17.10	0.69	0.57	14.90	7.20	0.70
Anti-DB	0.71	2.67	17.38	0.68	0.56	15.33	7.50	0.69
SimAC	0.69	2.69	17.72	0.73	0.49	16.28	8.33	0.72
GoodAC†	0.68	3.10	18.20	0.70	0.47	17.10	8.80	0.69
StyleGuard	0.71	2.42	19.66	0.73	0.48	19.75	**14.18**	0.74
MAC-ℓ_∞	0.58	6.06	22.34	**0.66**	0.48	27.27	7.78	0.66
MAC (w/o pure)	0.60	5.85	20.90	0.67	0.50	26.40	7.30	0.67
MAC	**0.54**$_{\downarrow 20.6\%}$	**6.93**$_{\uparrow 123.6\%}$	**25.02**$_{\uparrow 37.5\%}$	**0.66**$_{\downarrow 5.7\%}$	**0.41**$_{\downarrow 12.8\%}$	**32.13**$_{\uparrow 87.9\%}$	9.84$_{\uparrow 11.8\%}$	**0.64**$_{\downarrow 7.2\%}$
	"a photo of sks person looking at the mirror"				"a photo of sks person in front of eiffel tower"			
Method	ISM$\downarrow$	FDFR$\uparrow$	BRISQUE$\uparrow$	SER-FIQ$\downarrow$	ISM$\downarrow$	FDFR$\uparrow$	BRISQUE$\uparrow$	SER-FIQ$\downarrow$
Mist	0.51	5.50	18.10	0.54	0.22	26.90	21.10	0.39
Anti-DB	0.50	5.67	18.29	0.53	0.21	27.33	**21.43**	0.38
SimAC	0.44	6.79	18.13	0.54	0.15	30.26	20.09	0.34
GoodAC†	0.42	7.30	18.70	0.52	**0.14**	**31.00**	20.50	**0.33**
StyleGuard	0.46	7.90	17.33	0.56	0.19	29.63	20.18	0.38
MAC-ℓ_∞	0.43	4.85	**18.84**	0.53	0.22	19.09	18.68	0.43
MAC (w/o pure)	0.44	4.70	18.40	0.54	0.23	18.40	18.20	0.44
MAC	**0.38**$_{\downarrow 9.5\%}$	**10.67**$_{\uparrow 46.2\%}$	18.16$_{\downarrow 2.9\%}$	**0.51**$_{\downarrow 1.9\%}$	0.15$_{\uparrow 7.1\%}$	23.07$_{\downarrow 25.6\%}$	18.02$_{\downarrow 12.1\%}$	0.39$_{\uparrow 18.2\%}$

suggests that downstream evaluation metrics are affected not only by the perturbations but also by the intrinsic generation quality of the target model.

At the prompt level, transfer performance varies across prompts. For "a dslr portrait of sks person", we observe limited additional improvement over strong baseline methods under Stable Diffusion v1.5: although the perturbations introduce visible degradation in the generated images, their effect is often insufficient to consistently cause face detection failure. For prompts such as "a photo of sks person in front of eiffel tower", DreamBooth itself may fail to generate stable facial content, making face-centric metrics less reliable and limiting the interpretability of transfer gains.

Additional results on VGGFace2-HQ. We also report additional results on VGGFace2-HQ under Stable Diffusion v2.1 in **Table 7**. MAC remains competitive and improves upon prior baselines in most settings, although the overall gains are weaker than those on CelebA-HQ. We attribute this difference to dataset characteristics. CelebA-HQ images typically contain cleaner, more centered, and more consistent facial details, leading to more stable identity signals and more sensitive downstream metrics. In contrast, VGGFace2-HQ contains greater variations in pose, lighting, expression, and image quality, which reduce baseline identifiability and leave less headroom for protection improvements under purification.

Overall, these additional results show that MAC generalizes beyond the main CelebA-HQ evaluation. The transferred results on Stable Diffusion v1.5 remain broadly consistent with the main observations on Stable Diffusion v2.1, and the VGGFace2-HQ results further support the effectiveness of MAC on a more diverse face dataset.

C.4 Sensitivity to Purifier Reverse Timesteps

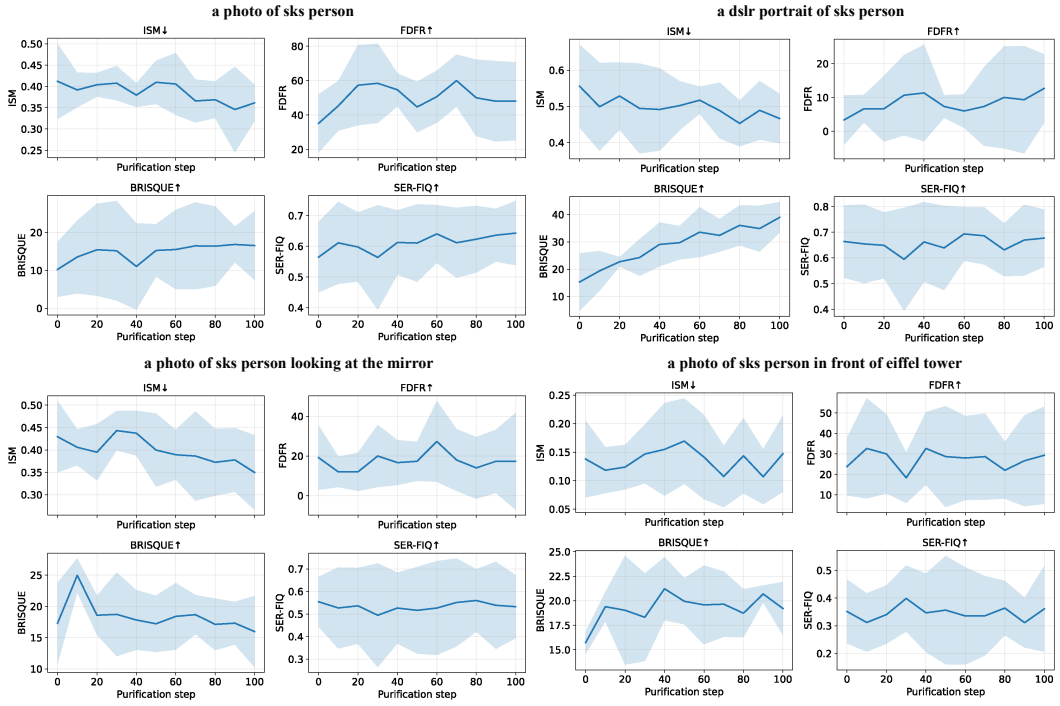


Figure 5: Sensitivity of protection performance to purifier reverse timesteps. Each curve reports the mean and standard deviation of evaluation metrics over multiple runs. Results are shown for four prompts, with primary attention given to prompts that produce stable facial content under DreamBooth. Performance remains stable or improves mildly as purification strength increases, indicating strong robustness of MAC to purifier configuration.

In this section, we describe the evaluation protocol used to analyze the sensitivity of protection performance to the strength of diffusion-based purification. Let $\mathcal{P}_k$ denote a diffusion-based purifier instantiated with k reverse denoising steps. For each protected image x_{adv}, we generate a purified image $\mathcal{P}_k(x_{\text{adv}})$ for $k \in \{0, 10, 20, \ldots, 100\}$, and perform perturbation visualization and DreamBooth fine-tuning with respect to the original image x. The resulting models are evaluated using standard identity and image-quality metrics, including ISM, FDFR, BRISQUE, and SER-FIQ, while all other training settings, prompts, and evaluation protocols are kept identical across different values of k, ensuring that observed differences are attributable solely to the choice of the purifier reverse timestep.

Figure 5 reports the sensitivity of protection performance to the number of reverse diffusion steps used in the purifier. We primarily focus on the prompts *"a photo of sks person"* and *"a dslr portrait of sks person"*, as these prompts consistently produce stable facial content under DreamBooth fine-tuning. The remaining prompts exhibit substantial instability in face generation due to the

customization process itself, and are therefore included only for *completeness* but not used for drawing *core conclusions*.

A notable observation is that **increasing the strength of purification does not lead to a monotonic degradation of protection performance**. Instead, metrics such as FDFR, ISM, and SER-FIQ remain stable and in several cases even exhibit mild improvements as the reverse timestep increases. This behavior can be attributed to two complementary effects that arise at different purification regimes.

When the number of reverse timesteps is relatively small, the purifier operates close to the configuration used during training, where it is explicitly differentiable and participates in end-to-end optimization. In this regime, gradients propagated through the purifier can amplify optimization bias, which may negatively affect downstream DreamBooth training and generation quality. This effect is also reflected in the main results, where **MAC w/o purification** consistently underperforms the full **MAC** configuration (**Tables** 1 and 7), As a result, performance around the training-time purification strength (e.g., $\sim$50 steps) can appear worse than that under both weaker and stronger purification settings.

As the number of reverse steps further increases, strong diffusion-based purification inevitably introduces a substantial amount of stochastic noise, a phenomenon that has also been observed in prior work Nie et al. [2022] on diffusion-based adversarial purification. In this regime, generation quality becomes increasingly dominated by the injected diffusion noise rather than by the residual adversarial perturbation. Consequently, the relative influence of the protection perturbation is partially masked, leading to flattened or even slightly improved metric trends.

Overall, these results indicate that **MAC exhibits strong robustness to the choice of purifier reverse timesteps.** Even when evaluated under purification strengths that significantly exceed those used during training, protection effectiveness does not collapse. This empirical robustness is consistent with our earlier qualitative and quantitative analyses, which show that MAC perturbations preserve their morphology and structural alignment under diffusion-based purification.

C.5 Additional Evaluations

We compare GoodAC and MAC-ℓ_∞-direct (abbreviated as MAC-direct) under matched budgets. Both methods are optimized from scratch, with projection after every update to enforce the ℓ_∞ constraint. Table 8 shows that MAC-direct achieves lower ISM and LPIPS at all three budgets. FDFR also improves at $8/255$ and $16/255$, while both methods provide negligible FDFR at $4/255$.

Table 8: Comparison under matched ℓ_∞ budgets.

Budget	Method	ISM ↓	FDFR ↑	LPIPS ↓
$4/255$	GoodAC	0.79	0.00%	0.269
	MAC-direct	0.72	0.01%	0.047
$8/255$	GoodAC	0.74	1.82%	0.357
	MAC-direct	0.67	2.80%	0.082
$16/255$	GoodAC	0.67	2.75%	0.413
	MAC-direct	0.59	5.87%	0.129

We additionally evaluate IMPRESS Cao et al. [2023] using its released implementation and a ten-pair configuration of BridgePure Wang et al. [2026], denoted BridgePure-10. The latter trains a separate removal model for each protection method using ten paired clean–protected images. Comparisons use identical identities, protected-image counts, personalization protocols, prompts, and metrics. Table 9 shows mixed, numerically close results under IMPRESS. BridgePure-10 reduces both methods' FDFR to near zero, revealing a failure boundary under paired-data, protection-specific removal.

Figure 6 evaluates perturbation retention under diffusion purification across four datasets: VGGFace2, WikiArt, ImageNet, and MS-COCO. For each method, we measure the cosine similarity between the perturbation direction after k purification steps and the original perturbation direction, i.e., $\cos(\delta_k, \delta_0)$. A higher value indicates that the perturbation direction is better preserved through the purification process.

Table 9: Protection effectiveness under additional removal attacks, with DiffPure as a reference.

	ISM ↓		FDFR ↑	
Removal method	GoodAC	MAC-direct	GoodAC	MAC-direct
DiffPure	0.67	0.59	2.75%	5.87%
IMPRESS	0.62	0.63	3.33%	3.87%
BridgePure-10	0.79	0.70	0.04%	0.09%

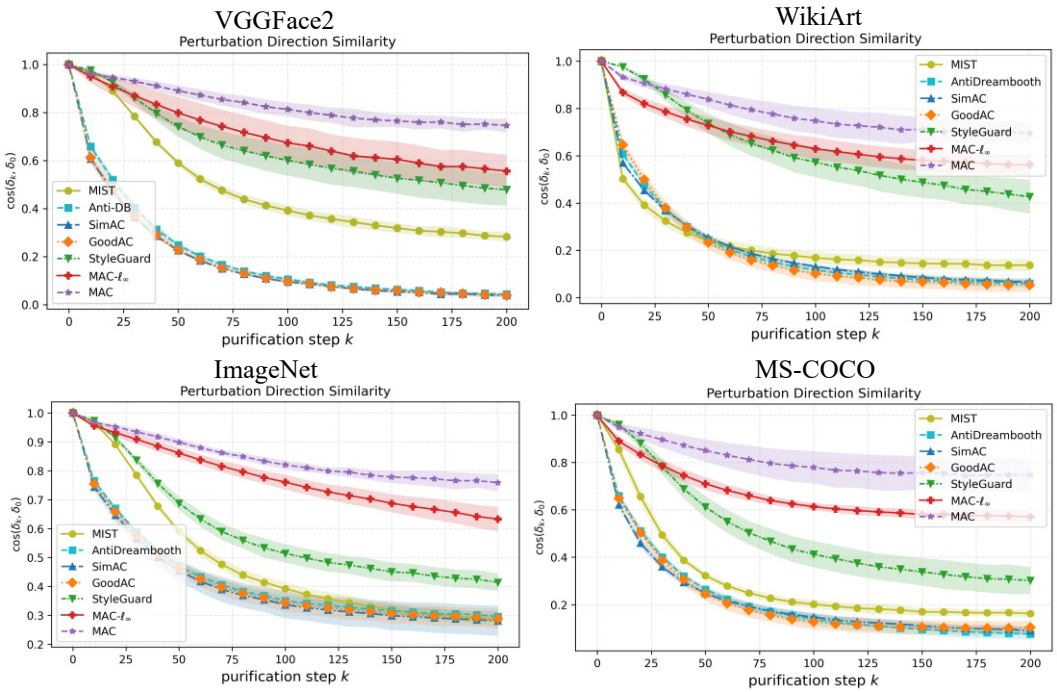


Figure 6: Perturbation direction retention under diffusion purification across different datasets. We report the cosine similarity between the purified perturbation direction δ_k and the original perturbation direction δ_0 over purification steps k on VGGFace2, WikiArt, ImageNet, and MS-COCO. Existing methods rapidly lose directional coherence under purification, whereas MAC consistently preserves substantially higher perturbation direction similarity across datasets. This indicates that the purification-retentive property of MAC is not specific to a particular dataset or visual domain.

Across all datasets, existing methods exhibit a rapid decline in perturbation direction similarity. This suggests that their effective perturbation components are largely suppressed or randomized by diffusion purification, causing the purified perturbations to deviate substantially from their original optimized directions. In contrast, MAC consistently maintains much higher direction similarity over the entire purification trajectory. This trend is especially clear at larger purification steps, where the gap between MAC and prior methods becomes more pronounced.

The results also show that the proposed purification-retentive behavior is not restricted to face datasets. On VGGFace2, MAC preserves stable perturbation directions under diverse facial appearances. On WikiArt, the same trend holds despite the shift from facial identity images to artistic style images. Similar behavior is further observed on ImageNet and MS-COCO, which contain more diverse object categories and natural scenes. These observations indicate that the geometric mechanism identified in the main paper is shared across different visual domains rather than being tied to a specific dataset.

Overall, these cross-dataset results provide additional evidence for the manifold-alignment perspective. Diffusion purification tends to suppress perturbation components that are not aligned with purification-retentive directions, while MAC biases the optimized perturbations toward directions that remain

stable under the purification dynamics. This explains why MAC maintains stronger downstream protection after purification and further supports the generality of the proposed mechanism.

C.6 Full Ablation Study with All Evaluation Metrics

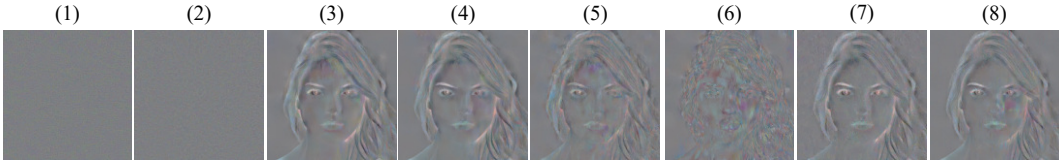


Figure 7: Perturbation visualizations corresponding to the ablation settings in **Table** 10. From left to right, the panels correspond to rows (1)–(8) in **Table** 10, visualizing $x_{\text{adv}} - x_{\text{original}}$ under each configuration.

To better understand how different components affect perturbation structure, **Figure** 7 provides a qualitative visualization of the corresponding perturbations and complements the quantitative results by providing a comprehensive ablation analysis of the optional components. Each panel visualizes $x_{\text{adv}} - x_{\text{original}}$ and is aligned with the rows in **Table** 10: from left to right, the eight panels correspond to configurations (1)–(8) in the table.

Role of the learned Filter. The most salient observation is that the learned *Filter* constitutes the **primary driver of purification-robust protection**. Configurations that include the Filter consistently outperform those without it across all metrics, most notably in FDFR and ISM. By introducing a structured and spatially coherent parameterization, the Filter explicitly biases optimization toward perturbations that align with the data manifold, rather than dispersing energy into off-manifold directions. As a result, the generated perturbations are both harder to suppress under diffusion-based purification and more effective during subsequent DreamBooth customization.

This effect is particularly evident when comparing configurations without any structural constraint (e.g., rows 1–2) to those that incorporate the Filter (rows 3–5 and 7–8), where a substantial performance gap emerges. These results confirm that **manifold-aligned structure, rather than loss design alone, is the key factor enabling robustness under purification**.

Effect of enhanced loss (EL). The enhanced loss (EL) plays a *secondary but non-negligible* role. When combined with the Filter, EL provides a consistent performance boost, as reflected by higher FDFR and improved perceptual metrics. However, EL alone is insufficient to yield strong protection: configurations that include EL but omit the Filter show only marginal gains over the baseline. This behavior indicates that, under purification, strengthening the objective cannot compensate for an unsuitable perturbation parameterization.

Notably, the configuration that excludes EL while retaining the Filter (configuration 5) achieves the second-best overall performance. This result highlights that structural constraints and differentiable purification dominate the effectiveness of the attack, while EL primarily acts as an amplifier rather than a fundamental enabler.

Time-step selection (TS) and DiffGrad. The contributions of time-step selection (TS) and DiffGrad are more configuration-dependent and less consistent than those of the Filter. In practice, the TS strategy tends to favor small time steps close to $t \approx 0$ Wang et al. [2024a], corresponding to the late-stage denoising phase of the reverse diffusion process. Our ablation results show that incorporating TS can still yield moderate performance gains, indicating that even in the presence of a purifier, effective adversarial optimization benefits from emphasizing late-stage, detail-oriented denoising steps.

This behavior can be explained by the reverse diffusion dynamics themselves, perturbations that are biased toward early denoising stages (i.e., large time steps) are more likely to lose effectiveness during the subsequent long reverse diffusion trajectory. Due to the strong stochasticity and repeated resampling in high-noise regimes, directional signals injected at large time steps tend to be progressively randomized and attenuated by the diffusion dynamics themselves. This inherent property of

Table 10: Ablation study of optional components under the prompt "a photo of sks person" on CelebA-HQ for **MAC**. TS: Time-step selection; EL: Enhanced attack loss from prior work.

TS	EL	Filter	DiffGrad	ISM↓	FDFR↑	BRISQUE↑	SER-FIQ↓
✗	✗	✗	✗	0.70	2.80	17.47	0.81
✓	✓	✗	✗	0.67	2.75	17.69	0.77
✗	✗	✓	✓	0.62	4.23	21.57	0.71
✗	✓	✓	✓	0.60	4.93	23.01	0.73
✓	✗	✓	✓	<u>0.54</u>	<u>5.50</u>	**25.43**	<u>0.68</u>
✓	✓	✗	✓	0.65	5.12	22.33	0.69
✓	✓	✓	✗	0.64	5.27	22.13	<u>0.68</u>
✓	✓	✓	✓	**0.48**	**8.11**	<u>24.65</u>	**0.66**

diffusion models explains why TS-centric configurations, while sometimes beneficial, generally fail to match the overall performance of Filter-dominant designs.

DiffGrad introduces additional variability. When enabled, perturbations tend to exhibit stronger color shifts and more textured patterns, which correlate with changes in perceptual metrics and are also reflected in downstream DreamBooth generations, where color artifacts become more pronounced. While such behavior can increase attack strength in some cases, it does not provide the same level of stability or robustness as the Filter-based structure.

Structural constraint versus implicit alignment via differentiable purification. An interesting intermediate case arises when no explicit Filter is used, but gradients are allowed to flow through a differentiable purifier (configuration 6). In this setting, **perturbations already exhibit a tendency to concentrate around facial structures**, indicating that the purifier itself induces a weak form of implicit manifold alignment. Nevertheless, these perturbations remain coarse and irregular, and their performance lags behind configurations that employ an explicit structural constraint. This comparison further emphasizes that while differentiable purification can encourage alignment, *explicitly modeling structure via the Filter is substantially more effective*.

Overall, the ablation results demonstrate that optimal protection performance arises from a Filter-centric design, optionally augmented by EL and, to a lesser extent, TS or DiffGrad. No single auxiliary component can replace the role of structured, manifold-aligned parameterization. These findings provide strong empirical support for our central claim that manifold alignment, rather than objective complexity or heuristic enhancements, is the dominant factor governing purification-robust anti-customization.

C.7 Black-Box Transfer to Additional Purification Defenses

Table 11: Evaluation of text-to-image protection methods under different purification settings for DreamBooth on CelebA-HQ under the prompt *"a photo of sks person"*. Metrics reported are FDFR↑ (higher is better) and ISM↓ (lower is better). **Bold** denotes the best value, and underlined values denote the second-best distinct value.

Method	No Prep.		DiffPure		Noise Up.		GrIDPure		PDMPure	
	FDFR↑	ISM↓	FDFR↑	ISM↓	FDFR↑	ISM↓	FDFR↑	ISM↓	FDFR↑	ISM↓
No Protect	0.00	0.85	0.00	0.77	0.00	0.79	0.00	0.82	0.00	0.82
GoodAC	**96.87**	**0.04**	2.75	0.67	1.63	0.77	1.58	0.71	1.17	0.76
StyleGuard	<u>86.48</u>	<u>0.06</u>	2.27	0.70	1.44	0.61	1.12	0.68	0.95	<u>0.75</u>
MAC-ℓ_∞	3.84	0.67	<u>6.26</u>	<u>0.56</u>	<u>6.87</u>	<u>0.57</u>	<u>4.32</u>	<u>0.52</u>	<u>3.86</u>	<u>0.65</u>
MAC	5.78	0.56	**7.91**	**0.48**	**7.52**	**0.50**	**6.27**	**0.44**	**5.60**	**0.56**

We further evaluate different protection methods against a broader set of purification-based defenses, including DiffPure Nie et al. [2022], noisy upscaling Hönig et al. [2024], GrIDPure Zhao et al. [2024], and PDMPure Xue and Chen [2024]. All experiments are conducted on CelebA-HQ under the DreamBooth prompt *"a photo of sks person"*. We report FDFR↑ and ISM↓, where higher

FDFR indicates stronger identity disruption and lower ISM indicates lower identity similarity in the generated images after personalization.

As shown in **Table 11**, existing methods such as GoodAC and StyleGuard remain highly effective when no preprocessing is applied, but their protection substantially degrades after purification-based preprocessing. In particular, their FDFR values drop sharply under DiffPure, noisy upscaling, GrID-Pure, and PDMPure, while their ISM values generally increase. This indicates that the perturbations produced by these methods are largely suppressed by purification, making the protected images less effective for disrupting subsequent DreamBooth fine-tuning.

In contrast, MAC consistently preserves stronger protection after purification. Across all considered purification defenses, MAC achieves the best FDFR and the lowest or competitive ISM. The bounded variant MAC-ℓ_∞ also outperforms existing baselines in terms of post-purification FDFR, although it is weaker than the full MAC variant. These results suggest that the improvement of MAC is not merely due to perturbation magnitude or an unconstrained perturbation space, **but is closely related to its purification-aware optimization design.**

Overall, these results further support our main claim that purification robustness is crucial for reliable anti-customization protection. By encouraging perturbations to align with purification-retentive directions, MAC produces adversarial changes that are more likely to survive defensive preprocessing and remain effective in downstream personalization.

C.8 Frequency-domain Analysis

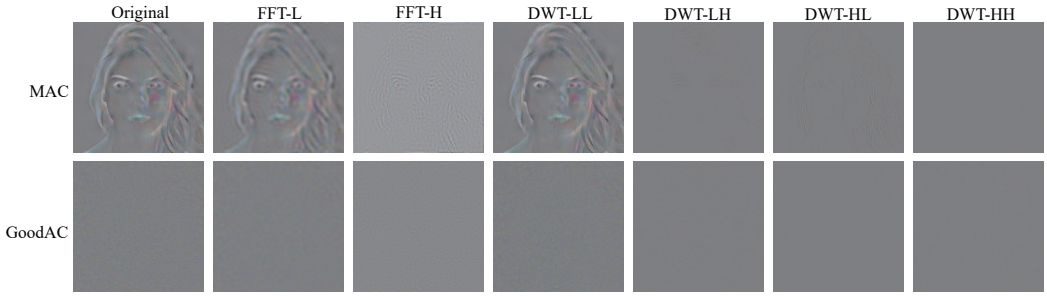


Figure 8: Visualization of component-wise perturbation transfer for MAC and GoodAC. Each column shows the perturbation component reconstructed by applying the same frequency or wavelet operator to the adversarial and clean images and taking their difference. DWT high-frequency subbands (LH/HL/HH) yield near-zero responses, while effective perturbations concentrate in FFT-H and DWT-LL for GoodAC, and in FFT-L and DWT-LL for MAC.

In this section, we conduct a component-wise analysis of protection effectiveness through frequency- and wavelet-domain decompositions. Prior work Chen et al. [2022] has shown that diffusion-based purification severely disrupts both the frequency content and phase structure of adversarial perturbations. Rather than treating spectral patterns as a stand-alone mechanism, we use these decompositions as *diagnostic probes* to test whether protection can be recovered by transferring individual transform components back to the clean image, and how such transferability relates to the structure of the underlying perturbation. This analysis complements **Section** 3.3 by testing whether protection effectiveness can be attributed to specific transform components in a transferable manner.

Given a protected image x_{adv} and its clean counterpart x, we construct component-wise variants by applying the same frequency or wavelet operator to both images and taking their difference. Specifically, for a component operator $\mathcal{T}_k$ (e.g., FFT-L/FFT-H or DWT-LL/LH/HL/HH), we define

$$x_{\mathrm{adv}}^{(k)} = x + \big(\mathcal{T}_k(x_{\mathrm{adv}}) - \mathcal{T}_k(x)\big). \tag{56}$$

This formulation is adopted because directly decomposing the raw perturbation $\delta = x_{\mathrm{adv}} - x$ is numerically unstable in practice, owing to phase sensitivity and boundary effects

Table 12: **High-frequency energy measured by FFT and DCT. Values are percentages (%). *Orig.* is a reference.**

Method	FFT HF	DCT HF
Orig.	3.63	7.83
Mist	4.99	19.95
Anti-DB	4.19	13.08
SimAC	4.08	13.42
GoodAC	4.02	13.44
StyleGuard	5.14	12.70
MAC-ℓ_∞	3.25	12.22
MAC	3.02	15.76

Table 13: **Component-wise transfer analysis for MAC under the prompt "a photo of sks person"**. Each row reports protection performance after transferring a single frequency or wavelet component back to the clean image. *Energy share* denotes the explained energy proportion (mean ± std, in %) of each component relative to the full perturbation. **FDFR is scaled by** $\times 100$**.**

Component	FDFR↑	ISM↓	BRISQUE↑	SER-FIQ↓	Energy share (%)
w/o δ	0.00	0.85	15.59	0.66	0.00
Full δ	**8.67**	**0.57**	**22.42**	**0.52**	**100.00**
FFT-L	3.33	0.51	31.41	0.46	60.48 ± 6.30
FFT-H	1.62	0.62	23.56	0.54	37.05 ± 9.86
DWT-LL	4.67	0.53	34.07	0.46	88.94 ± 4.80
DWT-LH (H)	0.02	0.70	15.12	0.55	2.49 ± 0.24
DWT-HL (V)	0.01	0.76	15.23	0.58	8.57 ± 3.13
DWT-HH (D)	0.01	0.85	14.45	0.56	1.27 ± 0.32

Table 14: **Component-wise transfer analysis for GoodAC under the prompt "a photo of sks person"**. Each row reports protection performance after transferring a single frequency or wavelet component back to the clean image. *Energy share* denotes the explained energy proportion (mean ± std, in %) of each component relative to the full perturbation. FDFR is scaled by $\times 100$.

Component	FDFR↑	ISM↓	BRISQUE↑	SER-FIQ↓	Energy share (%)
w/o δ	0.00	0.85	15.59	0.66	0.00
Full δ	**96.87**	**0.04**	**37.49**	**0.02**	**100.00**
FFT-L	0.00	0.84	15.89	0.58	11.74 ± 3.97
FFT-H	**93.49**	0.11	32.40	0.02	86.79 ± 4.30
DWT-LL	**85.23**	0.54	39.18	0.13	60.65 ± 3.06
DWT-LH (H)	0.01	0.81	15.42	0.67	14.16 ± 0.71
DWT-HL (V)	0.00	0.85	15.77	0.60	14.73 ± 0.66
DWT-HH (D)	0.01	0.82	15.36	0.65	8.01 ± 0.58

in frequency and wavelet transforms. Operating on full images and then taking differences yields more stable and interpretable component transfers.

To quantify the relative contribution of each component, we measure its *explained energy share*, which captures how much of the full perturbation energy is explained by the corresponding component direction. Let

$$\delta := x_{\text{adv}} - x, \qquad \delta^{(k)} := \mathcal{T}_k(x_{\text{adv}}) - \mathcal{T}_k(x).$$

Before computing energy statistics, we remove per-channel mean values from both δ and $\delta^{(k)}$ to eliminate trivial DC biases induced by global intensity shifts. The explained energy share of component k is then defined as

$$\text{EnergyShare}_k = 100 \cdot \frac{\left\|\text{proj}_{\delta^{(k)}}(\delta)\right\|_2^2}{\|\delta\|_2^2} = 100 \cdot \frac{\langle \delta, \delta^{(k)} \rangle^2}{\|\delta^{(k)}\|_2^2 \, \|\delta\|_2^2}, \tag{57}$$

where $\text{proj}_{\delta^{(k)}}(\delta)$ denotes the orthogonal projection of δ onto $\delta^{(k)}$. By construction, $\text{EnergyShare}_k \in [0, 100]$, so the reported values reflect *alignment-based* contribution rather than raw magnitude.

Tables 13 and 14 summarize the component-wise transfer results for **MAC** and **GoodAC**, respectively. Two complementary observations emerge. First, the *spectral allocation* differs markedly between the two methods. GoodAC exhibits an extreme FFT-high concentration, i.e., the perturbation is almost entirely aligned with the FFT-H component, consistent with the visually fine-grained, texture-like patterns commonly produced by gradient-based iterative optimization. In contrast, MAC shows a much *less skewed* distribution, with a **noticeably larger fraction of its perturbation explained by lower-frequency structure**. This is consistent with the design of MAC: the content-guided filter

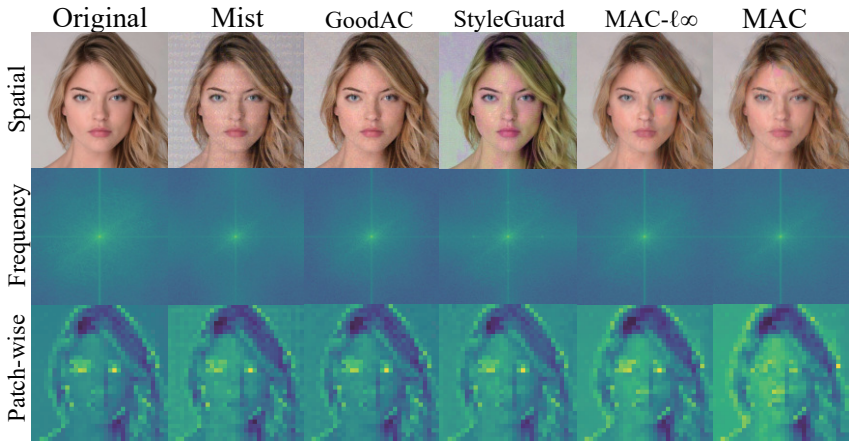


Figure 9: Spatial- and frequency-domain visualizations of images with different types of perturbations.

induces structured, locally coherent variations, which empirically bias updates toward smoother (often lower-frequency) components.

Figure 8 provides a qualitative visualization of the same component-wise transfer, and supports the energy trends in **Tables** 13 and 14. In particular, the high-frequency DWT subbands (LH/HL/HH) yield near-zero responses, appearing visually indistinguishable from uniform gray regions, consistent with their negligible transferability. For GoodAC, the perturbation manifests primarily as fine-grained, stripe-like textures concentrated in FFT-H, with a substantial portion also present in DWT-LL. For MAC, prominent structure is retained in DWT-LL and, to a lesser extent, in the Fourier components, indicating a more balanced allocation across bases. Overall, the effective transferred perturbations exhibit **structured and spatially coherent patterns rather than arbitrary spectral energy**, reinforcing that frequency statistics are surface-level correlates of the underlying manifold-consistent structure. Second, this difference is *not* resolved by a single "low-frequency vs. high-frequency" narrative because FFT and DWT capture different bases. Indeed, GoodAC also exhibits a large explained share in DWT-LL—often comparable to, though slightly smaller than, FFT-H—indicating that a substantial portion of its perturbation aligns with the low-pass wavelet subspace even when its global Fourier energy concentrates in high frequencies. This discrepancy highlights that transform-specific partitions (FFT vs. DWT) can attribute "where the energy lies" differently, even for the same underlying perturbation.

Most importantly, *transferability does not follow energy concentration*. For **MAC**, protection effectiveness is preserved primarily when either FFT-H or DWT-LL is transferred, while other components yield negligible protection. For **GoodAC**, protection transfer is dominated by FFT-H, with DWT-LL retaining only partial effectiveness, and the remaining DWT high-frequency subbands (LH/HL/HH) largely failing despite non-trivial explained shares in some cases. Therefore, neither "more low-frequency" nor "more high-frequency" alone predicts which component preserves protection.

The above behavior is more naturally explained by *geometric alignment* rather than spectral placement. Frequency decompositions partition perturbations by basis functions, whereas diffusion purification acts through score-driven dynamics that preferentially suppress *off-manifold (normal)* components while comparatively preserving *manifold-aligned (tangent)* components. This interpretation aligns with the selective suppression mechanism in Proposition 3.3 and the tangent/normal decomposition discussion in **Section** 3.2. A component can be low-frequency in one basis yet still contain substantial off-manifold structure, and conversely, a component can concentrate in high frequencies while remaining locally coherent and tangent-dominant. This explains why (i) FFT-H and DWT-LL can both carry effective protection under MAC, despite representing different "frequency" notions, and (ii) components with sizeable explained shares may still fail to transfer protection if their variations are not aligned with the data manifold.

Taken together, these results suggest that spectral statistics are best interpreted as *diagnostic descriptors* of perturbation appearance, while purification-robust protection is governed more directly by whether the perturbation preserves manifold-consistent structure.

D Visualization

D.1 CelebA-HQ Results on MAC



Figure 10: Qualitative results on CelebA-HQ. Left: original images, corresponding perturbations, and protected images. Right: images generated *after diffusion-based purification* under four prompts: *"a photo of sks person"*, *"a dslr portrait of sks person"*, *"a photo of sks person looking at the mirror"*, and *"a photo of sks person in front of eiffel tower"*.

D.2 VGGFace2-HQ Results on MAC



Figure 11: Qualitative results on VGGFace2-HQ. Left: original images, corresponding perturbations, and protected images. Right: images generated *after diffusion-based purification* under four prompts: *"a photo of sks person"*, *"a dslr portrait of sks person"*, *"a photo of sks person looking at the mirror"*, and *"a photo of sks person in front of eiffel tower"*.

D.3 CelebA-HQ Results on Baseline Methods

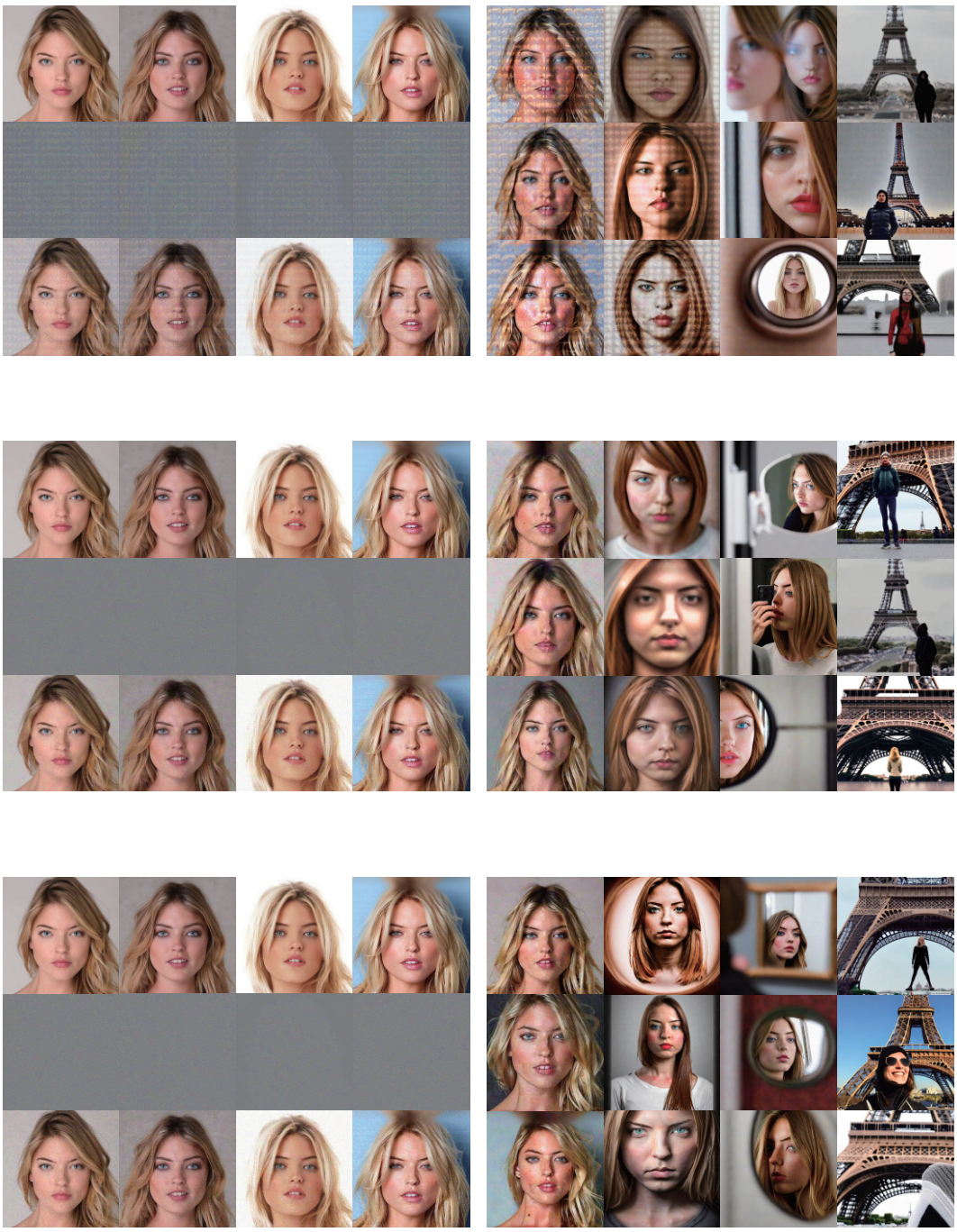


Figure 12: Qualitative results on CelebA-HQ. **From top to bottom: Mist, Anti-DreamBooth, and SimAC**. Left: original images, corresponding perturbations, and protected images. Right: images generated *after diffusion-based purification* under four prompts: *"a photo of sks person"*, *"a dslr portrait of sks person"*, *"a photo of sks person looking at the mirror"*, and *"a photo of sks person in front of eiffel tower"*.

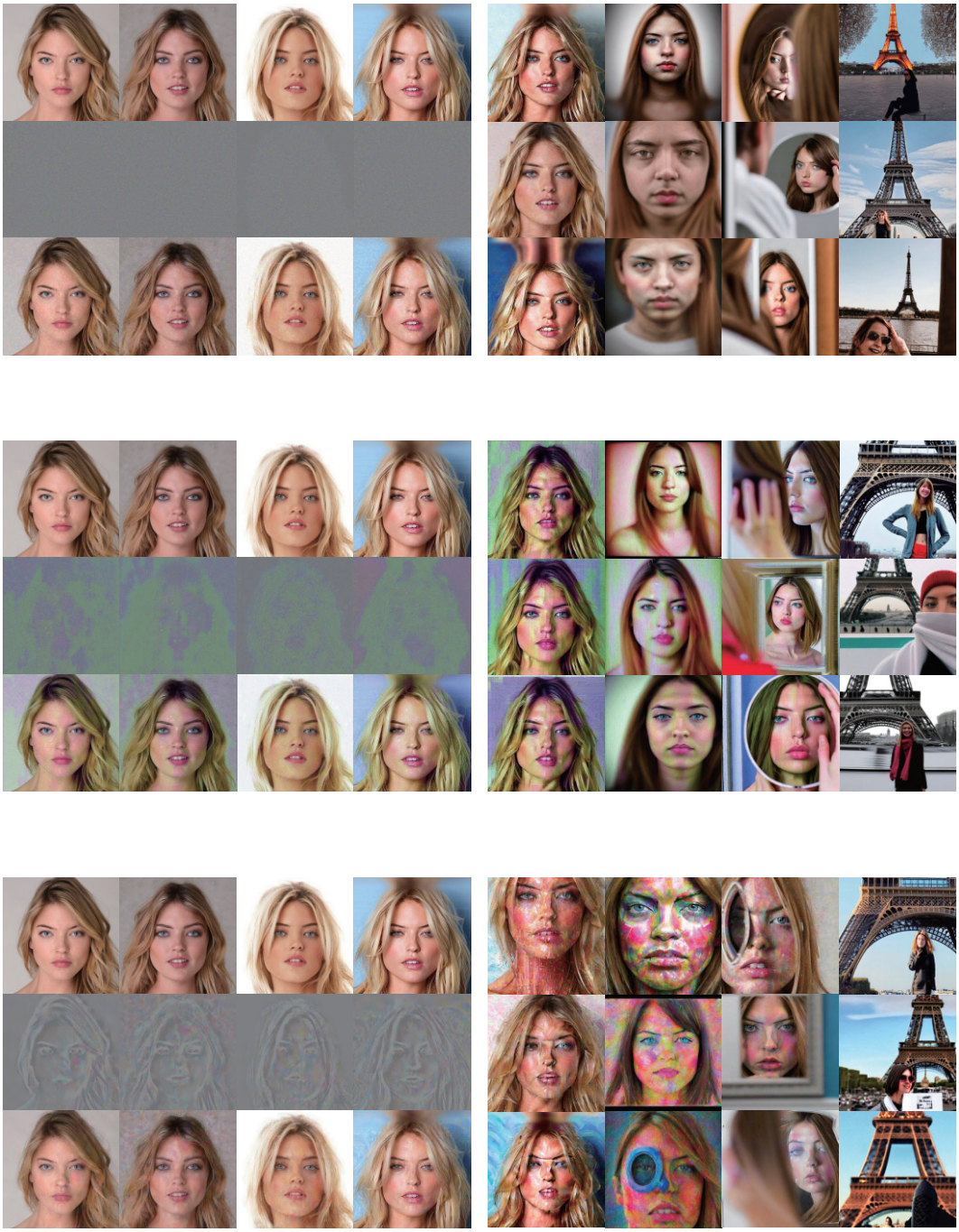


Figure 13: Qualitative results on CelebA-HQ. **From top to bottom: GoodAC, StyleGuard, and MAC-ℓ_∞**. Left: original images, corresponding perturbations, and protected images. Right: images generated *after diffusion-based purification* under four prompts: *"a photo of sks person"*, *"a dslr portrait of sks person"*, *"a photo of sks person looking at the mirror"*, and *"a photo of sks person in front of eiffel tower"*.

D.4 WikiArt Results on MAC

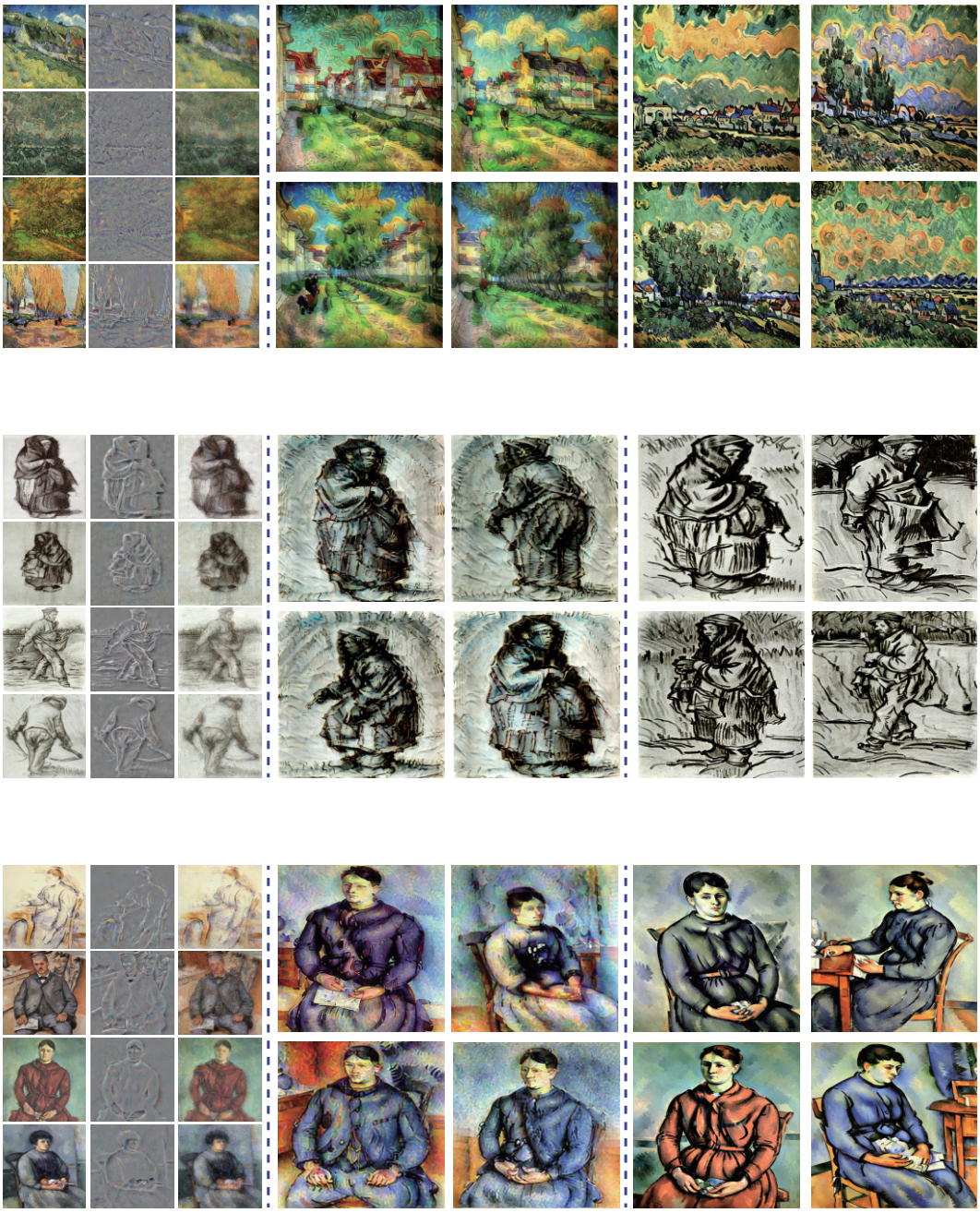


Figure 14: Qualitative results on WikiArt. The layout is divided into three blocks: **Left:** original artworks, corresponding perturbations, and protected images; **Middle:** images generated after diffusion-based purification under MAC protection; **Right:** images generated without any protection or perturbation. The prompt used for generation is *"a sks painting"*.

E Limitations and Broader Impacts

Limitations. The geometric analysis in this work is local: it characterizes purification dynamics in a neighborhood of the data manifold under standard smooth-manifold and score-approximation assumptions. These assumptions are useful for explaining the selective suppression effect, but they do not provide a global characterization of diffusion purification over the entire image space. Moreover, the proposed content-guided filter is a practical mechanism for encouraging locally coherent, manifold-aligned perturbations, rather than an exact estimator of tangent directions or an exact tangent-space projector. Therefore, the empirical effectiveness of MAC should be interpreted as evidence that the proposed structured parameterization promotes purification-retentive directions, rather than as a formal guarantee of exact manifold alignment.

Broader impacts. This paper presents work whose goal is to advance machine learning research on robust visual-content protection. Our contributions are primarily methodological and aimed at improving our understanding of how diffusion-based purification interacts with protective perturbations and model personalization. The most direct positive impact is to provide clearer theoretical and empirical tools for analyzing robustness and failure modes in generative modeling pipelines. In the context of anti-customization, this may help artists, photographers, and individuals better protect visual content from unauthorized personalization, style mimicry, or identity misuse.

At the same time, the techniques and insights developed here could be used in different ways depending on the surrounding context. They may support more effective privacy-preserving or misuse-resistant deployment of generative models, but they could also inform stronger attacks against existing protection mechanisms or disrupt benign personalization workflows. We do not introduce new data collection practices, release a new generative model, or deploy a real-world system, and our experiments are conducted on standard academic benchmarks under controlled settings. Therefore, we do not anticipate immediate harms from the experiments themselves, but we acknowledge the general dual-use nature of research in security and generative modeling. We encourage responsible use in consent-based content-protection settings.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes].

   Justification: The abstract and introduction state the paper's main contributions: a geometric analysis of diffusion-based purification, a purification-aware anti-customization method, and empirical evaluation across facial, artistic, and natural-image domains. The claims are supported by the theoretical analysis in Section 3 and the experimental results in Section 4 and Appendix C.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes].

   Justification: The paper discusses the main limitation that the geometric analysis is local and relies on standard smooth-manifold and score-approximation assumptions. It also clarifies that the proposed content-guided filter is a practical proxy for encouraging manifold-aligned perturbations rather than an exact tangent-space projector.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes].

   Justification: The theoretical results state the required local manifold, tubular-neighborhood, score-decomposition, and projection assumptions. Complete proofs of the propositions and theorem are provided in Appendix A.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes].

   Justification: The paper specifies the datasets, base models, baselines, purification settings, perturbation budgets, prompts, evaluation metrics, and implementation hyperparameters. Additional algorithmic details, pseudocode, and hyperparameters are provided in Appendix B.

   Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes].

   Justification: The paper uses publicly available datasets, models, and baseline implementations, and provides detailed implementation and hyperparameter descriptions. We do not include a public code release at submission time to preserve anonymity, but we plan to release anonymized or public code and reproduction instructions upon acceptance.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes].

   Justification: The main text describes the experimental setup, including datasets, models, baselines, prompts, metrics, purification settings, and perturbation budgets. Appendix B further reports the algorithm pseudocode, optimizer settings, training iterations, purification configuration, and filter hyperparameters.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes].

Justification: Our paper evaluated the statistical significance.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes].

   Justification: We provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes].

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes].

Justification: The paper discusses both potential positive societal impacts and negative societal impacts of the work performed.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A].

Justification: societal impacts of the work performed.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes].

Justification: The paper cites the datasets, models, and baseline methods used in the experiments, including CelebA-HQ, VGGFace2-HQ, WikiArt, ImageNet, MS-COCO, Stable Diffusion, DreamBooth, and prior anti-customization baselines. We use these assets for research evaluation and follow their intended academic-use settings.

Guidelines:

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

13. **New assets**

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

    Answer: [N/A].

    Justification: The paper does not introduce or release a new dataset, benchmark, pretrained model, or other standalone asset. The contribution is a method and analysis evaluated on existing public datasets and models.

    Guidelines:

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

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

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

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All evaluations are conducted on public image datasets and generated outputs from diffusion personalization pipelines.

    Guidelines:

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

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

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

Answer: [N/A].

Justification: The paper does not conduct human-subject studies, crowdsourcing, surveys, or user experiments. Therefore, IRB approval or an equivalent review is not applicable.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A].

Justification: The core method, theory, and experiments do not use LLMs as an important, original, or non-standard component. Any possible use of language tools for writing or formatting does not affect the methodology, experiments, or scientific conclusions.

Guidelines:

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