CAPURE: RETHINKING PURIFICATION-RESISTANT PERSONALIZATION DEFENSES VIA IDENTITY-AWARE DIFFUSION PURIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image diffusion models such as Stable Diffusion can learn a person's visual features from a few reference images, raising the risk of unauthorized personalized generation. Existing defenses against unauthorized personalization add protective perturbations to reference images. Recent purification-resistant protective perturbations further reduce the separability of perturbations from meaningful image content, making it difficult for existing purification methods based primarily on natural image priors to reliably neutralize their protective effects. We seek to bypass such protection to test whether existing purification paradigms overestimate its robustness. To this end, we analyze how protective perturbations disrupt downstream identity learning to identify a more direct target for restoring identity learnability. Experiments show that protective perturbations substantially suppress the cross-attention responses of generic person-related tokens within person regions. This observation motivates using these responses as a guidance signal for restoring identity learnability. Based on this finding, we propose CAPure, a purification method that directly restores identity learnability. It uses the cross-attention responses of person-related tokens to guide diffusion purification, recovering suppressed person-semantic responses and facilitating subsequent identity association, while constraining stepwise clean-image predictions to limit content distortion and identity drift. The method operates solely at the input purification stage and requires no modification to the downstream personalization pipeline. Experiments on CelebA-HQ and VGGFace2 show that CAPure more effectively restores identity learnability against various purification-resistant protection methods, revealing remaining vulnerabilities in existing purification-resistant protective perturbations.

1 INTRODUCTION

Text-to-image diffusion models such as Stable Diffusion (Rombach et al., 2022) can be personalized from only a few reference images to reproduce the visual identity of a specific person (Ruiz et al., 2023). This capability enables flexible subject-driven generation but raises concerns about unauthorized personalization of personal photographs. To mitigate this risk, existing defenses add visually imperceptible protective perturbations to reference images, disrupting the association between the personalized identifier and the target identity during fine-tuning (Van Le et al., 2023; Wang et al., 2024; Liu et al., 2024a). Before personalization, however, an attacker can apply input purification to remove protective perturbations and weaken their protective effects (Nie et al., 2022; Zhao et al., 2024; Zheng et al., 2026). In response, recent defenses improve robustness to input purification by adjusting the frequency characteristics or spatial distribution of protective perturbations, or by incorporating purification into perturbation optimization (Liu et al., 2024b; Onikubo & Matsui, 2024; Kim et al., 2025; Yang et al., 2025).

Unfortunately, robustness against existing purification methods can give a false sense of security. Representative approaches focus on removing perturbations while preserving image content. DiffPure, for example, was originally developed to defend image classifiers through diffusion-based adversarial purification (Nie et al., 2022). GrIDPure and UDAP address protective perturbations in Stable Diffusion through iterative denoising and a DDIM-based reconstruction objective, respec-

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

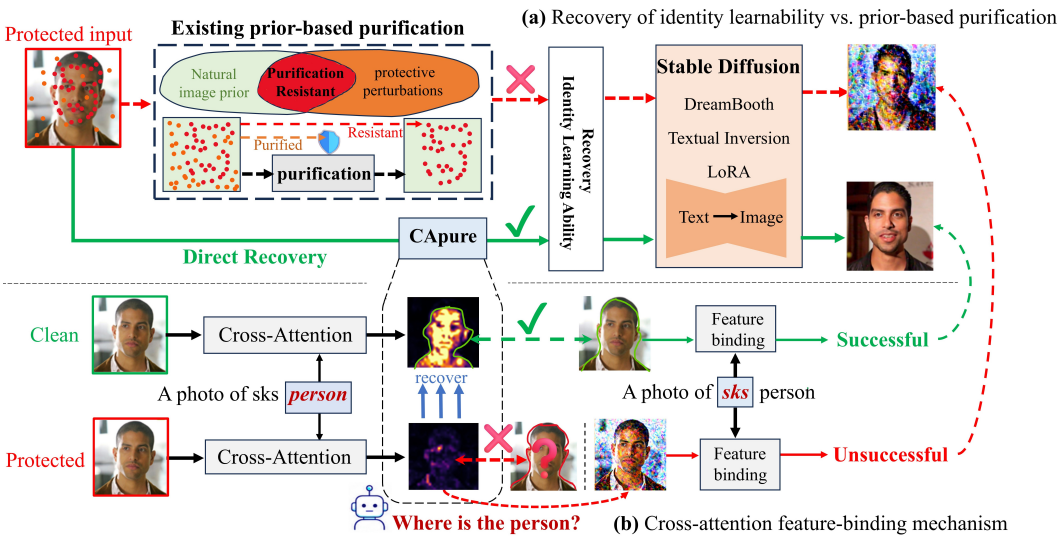

Figure 1: Overview of CAPure. (a) Purification-resistant perturbations can survive prior-based purification and hinder identity learning; CAPure targets the recovery of identity learnability. (b) These perturbations suppress person-token cross-attention responses and disrupt identifier–identity association. CAPure restores these responses to facilitate identity association during personalization.

tively (Zhao et al., 2024; Zheng et al., 2026). Although these methods can weaken protection, their purification mechanisms do not explicitly target the recovery of person-related semantic responses for downstream identity learning. This distinction matters because an attacker need not recover the original clean image: restoring the information needed for the personalization model to relearn the target identity suffices. Thus, failure to completely remove protective perturbations does not necessarily imply that the protected identity is fundamentally unrecoverable.

This raises the question: *what information should purification restore to make a protected identity learnable again?* Successful personalization links the identifier to the target person's appearance, but protective perturbations can redirect this association toward perturbation patterns (Liu et al., 2026). This failure calls for examining what the model responds to, beyond what the purified image looks like. We find that protective perturbations suppress person-token cross-attention responses, and the evaluated purification baselines do not fully recover them even when image appearance improves (Figs. 2 to 4). Strengthening these responses during purification makes faces more distinguishable from background artifacts in subsequent personalized generations. This improvement suggests that restoring person-related responses helps shift personalization from reproducing disruptive patterns toward learning person-related content, making response recovery a useful purification target for restoring identity learnability.

Motivated by this finding, we propose CAPure (Cross-Attention Pure), an identity-aware diffusion purification method that recovers suppressed person-related semantic responses to support identity learning. CAPure combines subject-related cross-attention guidance with stepwise reconstruction anchoring: the former guides denoising toward recovering person-related semantic responses, while the latter anchors the predicted clean image at each step to the protected input to preserve identity-specific appearance and structure. This stepwise design provides appearance feedback during denoising without a separate inversion–reconstruction trajectory for computing reconstruction loss. CAPure uses a frozen pretrained diffusion model, requires no matched clean counterpart for the protected input, and leaves downstream personalization unchanged.

Our contributions are: (i) We rethink purification-resistant personalization defenses from the perspective of identity learnability, identifying a mismatch between image restoration and the recovery of person-related semantic responses. (ii) We propose CAPure, which combines subject-related cross-attention guidance with stepwise reconstruction anchoring to improve identity learnability while limiting content distortion and identity drift. (iii) Experiments on CelebA-HQ and VGGFace2 show that CAPure achieves the highest identity similarity among the evaluated purifiers across all five defenses in the main evaluation. Qualitative comparisons further illustrate reduced residual textures and more coherent facial appearance in images generated by the personalized models.

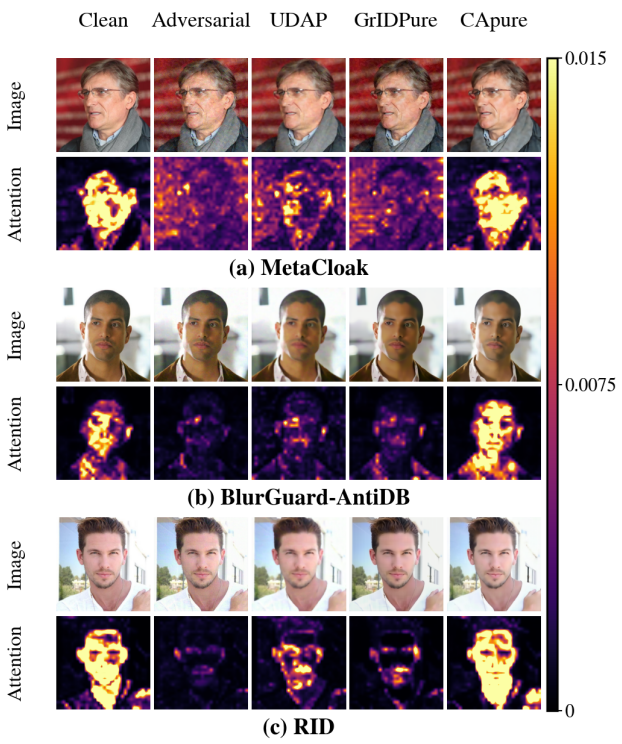

Figure 2: Reference images and their person-token cross-attention maps before and after purification.

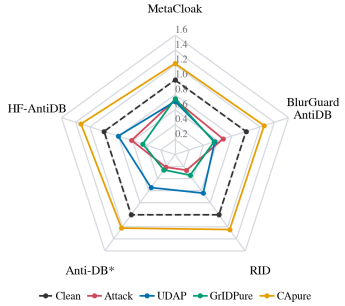

Figure 3: Mean person-token response, normalized by the clean reference.

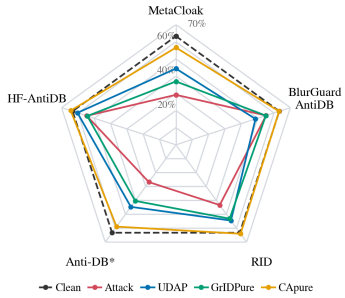

Figure 4: Mean fraction of person-token attention within the face.

2 Related Work

Protective Perturbations against Unauthorized Personalization. Anti-DreamBooth and subsequent methods, including SimAC, CAAT, DisDiff, and CoPSAM, optimize protective perturbations to disrupt identity learning (Van Le et al., 2023; Wang et al., 2024; Xu et al., 2024a; Liu et al., 2024a; Xu et al., 2024b). Recent defenses further seek to sustain protection after preprocessing and purification. MetaCloak and RID improve robustness through transformation-aware meta-optimization and diffusion-model ensembles, respectively (Liu et al., 2024b; Guo et al., 2024). HF-ADB concentrates stronger perturbations in high-frequency image regions, while BlurGuard adaptively adjusts the perturbation spectrum to resist purification (Onikubo & Matsui, 2024; Kim et al., 2025).

Input Purification against Protective Perturbations. Simple transformations such as Gaussian smoothing and JPEG compression can weaken protective perturbations but may degrade image quality (Van Le et al., 2023). DiffPure and GDMP use diffusion models for adversarial purification, originally targeting classifier robustness (Nie et al., 2022; Wang et al., 2022). For personalized diffusion models, GrIDPure preserves image structure through iterative, small-step diffusion over image grids (Zhao et al., 2024), while UDAP minimizes a DDIM-based reconstruction objective (Zheng et al., 2026). These methods emphasize perturbation removal and content preservation but can be limited by purification-resistant perturbations. DiffShortcut combines image restoration with contrastive decoupling to separate identity learning from perturbation patterns (Liu et al., 2026). However, its full framework requires changes to training and inference prompts, limiting its use with fixed downstream procedures. CAPure directly targets identity learnability during input purification, leaving downstream personalization and sampling unchanged.

Cross-Attention Analysis and Guidance in Diffusion Models. Cross-attention maps characterize the spatial correspondence between text tokens and image regions. Prompt-to-Prompt controls these maps for text-guided image editing (Hertz et al., 2023); DAAM aggregates them to analyze token-to-region correspondence (Tang et al., 2023); and Attend-and-Excite strengthens weak attention responses to subject tokens during generation to reduce subject omission (Chefer et al., 2023). Inspired by these works, we examine how protective perturbations affect the spatial responses to

person-related tokens and use these responses to guide input purification toward restoring identity learnability.

3 METHODOLOGY

CAPure aims to enable downstream personalization models to relearn the target person's identity from protected images.

3.1 MOTIVATION AND DESIGN RATIONALE

We first examine images generated after personalization on purified inputs. Under purification-resistant protection, existing purifiers can leave generated faces covered with textures resembling those in the background, as illustrated in Figure 6. Similar artifacts appear when our purifier uses reconstruction alone. Their presence in both facial and background regions suggests difficulty learning the person's appearance separately from these disruptive patterns.

We investigate this behavior through *person*-token cross-attention in frozen pretrained SD v1.5, using matched clean images, protected inputs, and purified counterparts. We use "a photo of sks person" as the text condition c. For these diagnostics, we extract attention maps at the fixed diffusion training timestep $t = 25$. These maps characterize how spatial image features respond to the person concept (Hertz et al., 2023; Tang et al., 2023; Chefer et al., 2023). Figure 2 shows that protective perturbations suppress responses within person regions, which existing purification does not consistently recover. Figures 3 and 4 quantify this effect using statistics averaged over 100 images per group: both the overall response and the fraction of attention within the face remain below the clean reference after UDAP or GrIDPure purification across the five evaluated defenses. The measurements are defined in Appendix A.

To guide purification using this suppression, we compare the overall person-token response with a clean-face reference. Under this text condition c, let $A_t(p; z_t)$ denote the head-averaged cross-attention weight of this token at spatial position p in the selected layer. We extract these weights when SD v1.5 predicts noise at timestep t.

For offline calibration, $\mathcal{X}_{\text{clean}}$ contains 1,164 clean face images from CelebA-HQ and VGGFace2. For each image x, we sample a VAE latent $z_0(x)$, apply forward diffusion to obtain $z_t(x)$, and extract attention with a single U-Net forward pass. We compute the spatially averaged response and its clean-face mean separately at each timestep:

$$m_t(z_t) = \frac{1}{|\Omega|} \sum_{p \in \Omega} A_t(p; z_t), \qquad \mu_t = \frac{1}{|\mathcal{X}_{\text{clean}}|} \sum_{x \in \mathcal{X}_{\text{clean}}} m_t\big(z_t(x)\big), \tag{1}$$

where Ω is the spatial grid of the attention map. Calibration retains only one scalar μ_t per timestep, shared across all protected inputs. It summarizes the average attention to *person* across clean faces, without requiring a matched clean counterpart of the protected image.

We next test whether person-token guidance during purification changes downstream generation. On VGGFace2 under RID and CelebA-HQ under Anti-DreamBooth, we compare diffusion purification without either guidance term against purification using only person-token attention guidance, denoted as mass-only. For each variant, we train DreamBooth on the purified images and generate images with the same prompt. In Figure 8, mass-only produces faces more clearly distinguishable from background artifacts, but their appearance can deviate substantially from the target individual. This intervention improves person depiction while exposing the need to preserve identity-specific details.

This identity drift motivates adding reconstruction for content preservation, following prior purification work (Zheng et al., 2026). We use the protected input as an appearance reference and constrain the clean-image prediction at each denoising step. This reference also contains protective perturbations, so reconstruction alone can retain patterns that disrupt personalization. Person-token guidance supplies an additional objective that penalizes the observed response deficit. In the combined variant, residual facial textures are reduced in the displayed examples, and identity similarity improves over either component alone in Table 4.

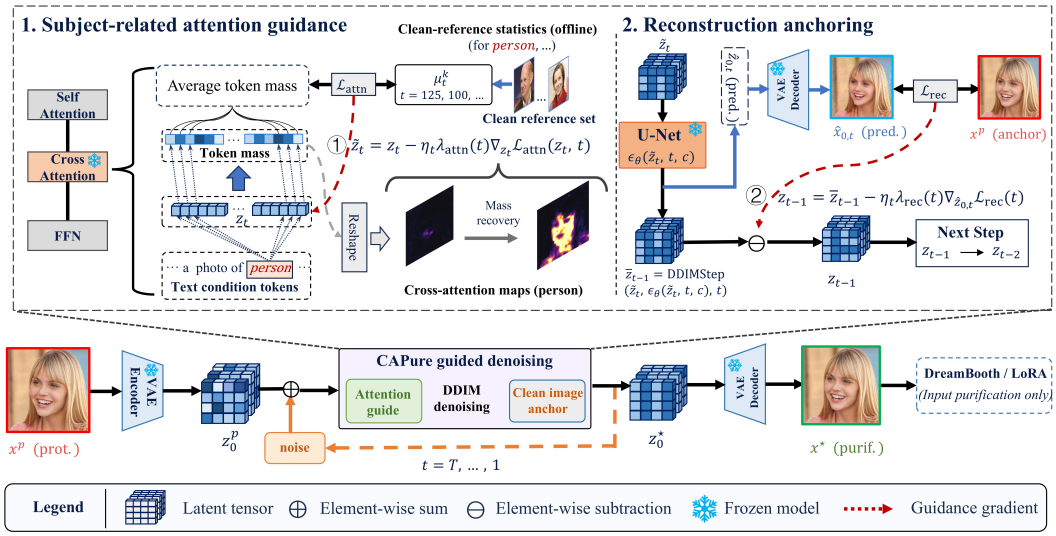

Figure 5: Overview of CAPure. Cross-attention guidance restores suppressed responses to person-related tokens, while stepwise reconstruction anchors predicted clean images to the protected input to limit appearance changes and identity drift.

As illustrated in Figure 5, given a protected image x^p, CAPure encodes it as $z_0^p = \mathcal{E}(x^p)$ and performs R rounds of noise addition and guided DDIM denoising using a frozen pretrained SD v1.5 model. The final latent is decoded into the purified image $x^\star = \mathcal{D}(z_0^\star)$. The process requires no matched clean counterpart of the protected input and leaves the downstream personalization procedure unchanged.

3.2 Person-Token Attention Guidance

With $m_t(z_t)$ and μ_t defined in Eq. (1), we strengthen attention only when its normalized shortfall from the clean-face average exceeds a tolerance τ, allowing response strength to vary across images:

$$\mathcal{L}_{\text{attn}}(z_t, t) = \text{SmoothL1}\left(\left[\frac{\mu_t - m_t(z_t)}{\mu_t + \varepsilon} - \tau\right]_+, 0\right), \tag{2}$$

where $[u]_+ = \max(u, 0)$ and $\varepsilon > 0$ ensures numerical stability. Normalization measures the shortfall relative to the clean-face attention level at that timestep. Once the normalized shortfall is at most τ, the loss is zero and no longer encourages an increase.

At each guidance step, we adjust the current latent to reduce this attention shortfall:

$$\widetilde{z}_t = z_t - \eta_t \lambda_{\text{attn}}(t) \nabla_{z_t} \mathcal{L}_{\text{attn}}(z_t, t), \tag{3}$$

where η_t is the step size and $\lambda_{\text{attn}}(t)$ controls the guidance strength. The denoiser then predicts noise from $\widetilde{z}_t$, incorporating the attention guidance into the subsequent DDIM transition.

3.3 Stepwise Reconstruction for Appearance Preservation

Increasing the average person-token response does not specify where the response should increase or ensure that the target person's appearance is preserved. Attention guidance may also increase person-token responses outside the face. We therefore constrain the clean-image prediction at each denoising step using the protected input x^p as an appearance reference. If an attention-guided update distorts the predicted face, the resulting discrepancy from the input contributes to the reconstruction error. This provides feedback on the visual effects of attention guidance without prescribing a target attention map or requiring a face mask. Given the attention-guided latent $\widetilde{z}_t$, we estimate the clean latent and decode it for comparison with the input:

$$\widehat{z}_{0,t} = \frac{\widetilde{z}_t - \sqrt{1 - \bar{\alpha}_t}\, \epsilon_\theta(\widetilde{z}_t, t, c)}{\sqrt{\bar{\alpha}_t}}, \qquad \widehat{x}_{0,t} = \mathcal{D}(\widehat{z}_{0,t}), \tag{4}$$

where ϵ_θ is the frozen noise predictor, $\bar{\alpha}_t$ is the cumulative signal coefficient, and $\mathcal{D}$ is the VAE decoder. We penalize the difference between this prediction and the protected input:

$$\mathcal{L}_{\text{rec}}(t) = \|\widehat{x}_{0,t} - x^p\|_2^2 . \tag{5}$$

Reusing the noise prediction in Eq. (4), CAPure performs the DDIM transition and then applies the reconstruction feedback:

$$\begin{aligned} \overline{z}_{t-1} &= \text{DDIMStep}\left(\widetilde{z}_t, \epsilon_\theta(\widetilde{z}_t, t, c), t\right), \\ z_{t-1} &= \overline{z}_{t-1} - \eta_t \lambda_{\text{rec}}(t) \nabla_{\widehat{z}_{0,t}} \mathcal{L}_{\text{rec}}(t), \end{aligned} \tag{6}$$

where $t-1$ denotes the next timestep in the reverse sampling schedule and $\lambda_{\text{rec}}(t)$ controls the reconstruction strength. We compute the reconstruction gradient through the decoder with respect to $\widehat{z}_{0,t}$ to correct $\overline{z}_{t-1}$, without backpropagating through the denoiser.

The reconstruction feedback uses the clean-image prediction already available within each denoising step. It requires no separate DDIM inversion–reconstruction trajectory as used to evaluate UDAP's reconstruction objective (Zheng et al., 2026).

4 Experiments

4.1 Experimental Setup

Datasets and personalization. We evaluate on CelebA-HQ (Karras et al., 2018) and VGGFace2 (Cao et al., 2018). The main evaluation includes 25 identities from each dataset, with four reference images per identity. We personalize Stable Diffusion v1.5 using DreamBooth on protected images or their purified counterparts, with models trained on clean images serving as references. Each personalized model generates 80 images with the prompt "a photo of sks person" for evaluation.

Protection methods and purification baselines. We evaluate Anti-DreamBooth as a conventional protection method and HF-ADB, BlurGuard, MetaCloak, and RID as defenses designed to resist purification. BlurGuard is applied in combination with Anti-DreamBooth. We compare CAPure with JPEG compression and three diffusion-based purifiers: DiffPure, GrIDPure, and UDAP. Protected images without purification serve as a control. Anti-DreamBooth uses a perturbation budget of $\epsilon = 0.05$ in the main evaluation; Section 4.5 examines a larger budget.

Evaluation metrics. We evaluate identity recovery and quality in the personalized models' generated images. ArcFace-based ISM (Deng et al., 2019) measures identity similarity to the target person, while BRISQUE (Mittal et al., 2012) and PIQE (Venkatanath et al., 2015) provide no-reference image-quality scores. Higher ISM and lower BRISQUE/PIQE are preferred. Since no-reference scores may not fully reflect visible artifacts, we interpret them alongside identity similarity and qualitative comparisons. We report scores for all methods, but restrict BRISQUE/PIQE rankings to diffusion-based purifiers.

Implementation details. Experiments use 512×512 images on a system with four NVIDIA GeForce RTX 3090 GPUs. In the main evaluation, DreamBooth fine-tunes both the U-Net and text encoder for 1,000 steps with a learning rate of 5×10^{-7} and 200 class images for prior preservation. CAPure performs 10 purification rounds, each with 20 denoising updates using a DDIM scheduler. The base coefficients for attention guidance and stepwise reconstruction are 0.075 and 0.35, respectively. Baselines use the default settings from their public implementations or configurations recommended in the original papers. Appendix A defines the attention measurements used in Section 3.1.

4.2 Main Results

Identity recovery across defenses. CAPure achieves the highest ISM across all ten dataset–defense combinations (Table 1). It improves identity recovery under conventional Anti-DreamBooth protection, with even larger gains over the strongest baseline under the four defenses explicitly designed to resist purification. CAPure also achieves the lowest PIQE among diffusion-based purifiers, reducing it by 13.3–36.2% relative to the best competing diffusion purifier in each setting. Figure 6 illustrates the corresponding improvement in facial appearance. JPEG leaves substantial visible corruption,

Table 1: Quantitative comparison of purification methods under different protection settings (SD v1.5 with DreamBooth).

Method	Anti-DreamBooth			BlurGuard +Anti-DreamBooth			HF-ADB			MetaCloak			RID		
	B↓	P↓	ISM↑	B↓	P↓	ISM↑	B↓	P↓	ISM↑	B↓	P↓	ISM↑	B↓	P↓	ISM↑
CelebA-HQ															
Clean reference (B / P / ISM): 15.88 / 31.86 / 0.71															
None	23.79	57.17	0.491	23.97	37.66	0.624	14.96	43.90	0.623	28.30	51.44	0.383	25.13	58.09	0.377
DiffPure	39.62	58.54	0.638	40.97	60.58	0.644	38.52	56.60	0.652	45.71	58.76	0.592	43.79	64.06	0.622
GrIDPure	29.23	58.71	0.671	30.00	57.66	0.670	42.66	66.74	0.650	31.69	67.49	0.513	35.50	59.11	0.656
JPEG	23.51	56.21	0.511	24.51	39.66	0.619	19.82	39.17	0.669	33.31	49.09	0.390	32.56	51.99	0.528
UDAP	45.23	67.62	0.623	51.82	65.73	0.654	42.81	67.36	0.617	40.53	68.36	0.539	42.60	67.81	0.641
CAPure	36.52	41.40	0.678	34.86	49.14	0.687	33.87	49.05	0.695	41.17	49.39	0.607	33.71	43.78	0.681
VGGFace2															
Clean reference (B / P / ISM): 15.87 / 28.92 / 0.66															
None	45.09	68.78	0.224	27.23	41.39	0.589	13.19	40.40	0.548	26.01	48.84	0.369	20.01	56.73	0.369
DiffPure	42.91	61.30	0.589	42.74	62.38	0.587	39.59	58.33	0.592	47.40	64.61	0.557	44.47	63.86	0.594
GrIDPure	35.28	62.91	0.626	37.68	64.45	0.610	49.05	66.44	0.611	34.30	69.32	0.476	36.16	62.80	0.600
JPEG	24.43	59.97	0.389	28.19	39.82	0.592	23.87	37.48	0.607	34.54	49.14	0.431	24.95	51.24	0.485
UDAP	46.90	66.80	0.610	49.35	68.17	0.596	39.16	56.36	0.615	42.22	68.23	0.529	44.67	65.52	0.608
CAPure	35.46	40.37	0.638	36.63	45.98	0.630	34.71	45.12	0.640	36.40	41.24	0.581	35.75	41.87	0.641

Note: B = BRISQUE; P = PIQE. Red/blue indicate the best/second-best results among diffusion-based purifiers for B/P and among all purifiers for ISM. Gray rows and clean references are excluded.

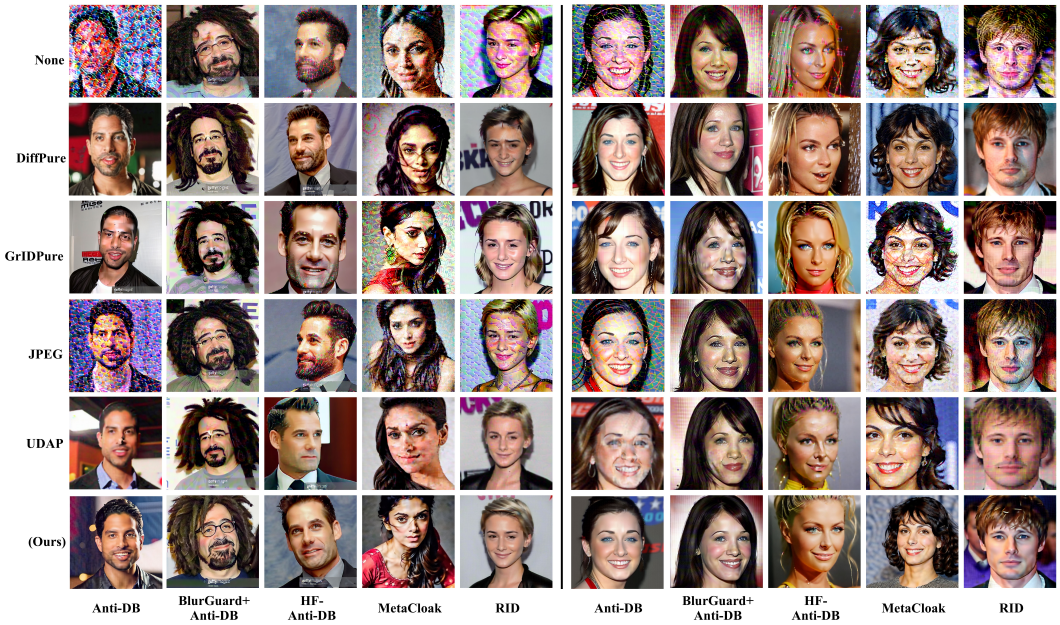

Figure 6: Images generated by DreamBooth models personalized on protected inputs or their purified counterparts under different protection settings.

while GrIDPure and UDAP still produce faces with residual textures or local distortions. DiffPure yields cleaner-looking faces in some examples, but its lower ISM indicates that visual cleanup alone does not ensure faithful identity recovery. Similarly, GrIDPure's lower BRISQUE in several settings does not translate into stronger identity recovery or consistently cleaner facial details. CAPure produces clearer facial features and more natural skin appearance, particularly in the MetaCloak examples where competing outputs remain visibly affected by colored textures. These results support more effective identity recovery with fewer artifacts in subsequent personalized generation.

Table 2: DreamBooth results on SD v2.1. Red/blue: best/second-best among purifiers.

Defense	Method	BRISQUE↓	PIQE↓	ISM↑
Anti-DreamBooth	None	38.63	71.34	0.259
	DiffPure	43.59	59.43	0.571
	GrIDPure	45.63	68.12	0.548
	UDAP	42.86	66.95	0.622
	CAPure	30.11	39.74	0.637
HF-ADB	None	38.19	51.67	0.536
	DiffPure	46.95	67.94	0.590
	GrIDPure	53.19	66.83	0.634
	UDAP	48.32	71.24	0.627
	CAPure	35.71	48.68	0.641

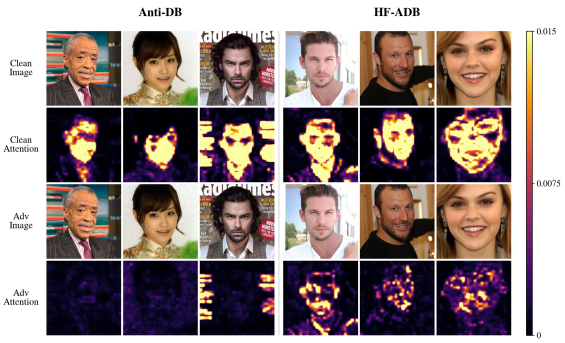

Figure 7: SD v1.5 person-token attention before and after protection targeting SD v2.1.

Table 3: ASPL perturbation-budget analysis on VGGFace2.

ϵ	Input	BRISQUE↓	PIQE↓	ISM↑
–	Clean	15.96	28.65	0.58
–	Clean + CAPure	32.93	43.49	0.55
0.01	Protected	27.68	35.21	0.52
0.01	Protected + CAPure	33.08	45.23	0.56
0.03	Protected	31.20	64.02	0.26
0.03	Protected + CAPure	35.59	42.28	0.55
0.15	Protected	56.73	75.87	–
0.15	Protected + CAPure	38.43	43.17	0.50

Table 4: Ablation on VGGFace2 under RID and CelebA-HQ under Anti-DreamBooth.

Setting	Variant	BRISQUE↓	PIQE↓	ISM↑
VGGFace2 RID	w/o both	41.94	38.56	0.338
	w/o $\mathcal{L}_{\text{rec}}$	37.65	41.98	0.356
	w/o $\mathcal{L}_{\text{attn}}$	41.47	44.50	0.627
	Full	35.75	41.87	0.641
CelebA-HQ Anti-DB	w/o both	43.45	39.19	0.242
	w/o $\mathcal{L}_{\text{rec}}$	31.80	45.67	0.277
	w/o $\mathcal{L}_{\text{attn}}$	43.88	44.50	0.648
	Full	36.52	41.40	0.678

4.3 Generalization Analysis

We evaluate transfer across diffusion backbones and sensitivity to protective perturbation strength.

Generalization to SD v2.1. We evaluate Anti-DreamBooth and HF-ADB on 15 identities per method and defense, using SD v2.1 for protection and DreamBooth while retaining CAPure's frozen SD v1.5 purifier. Despite this backbone mismatch, SD v1.5 still shows weaker person-token responses for protected examples than for matched clean images (Figure 7), indicating an observable guidance signal. CAPure achieves the highest ISM and lowest BRISQUE and PIQE under both defenses (Table 2), extending this advantage to the evaluated SD v2.1 settings.

Generalization across perturbation budgets. Table 3 evaluates CAPure on clean and ASPL-protected VGGFace2 images. Purifying clean inputs only slightly decreases ISM, suggesting that identity learnability is largely preserved without protection. Under ASPL protection, CAPure's ISM gradually declines across the tested budgets, reaching 0.50 even at $\epsilon = 0.15$, versus 0.55 on purified clean inputs. This small gap indicates substantial identity recovery even under strong perturbations.

4.4 Ablation Study

Table 4 and Figure 8 show that attention guidance alone makes faces more distinguishable from background artifacts but yields only a small ISM gain over the variant without either constraint. Stronger person-token responses thus cannot ensure individual identity preservation. Reconstruction alone recovers substantially more identity information, although visible facial textures remain in the displayed examples. Adding attention guidance further improves ISM in both settings and reduces these textures. These results highlight reconstruction's role in retaining individual appearance and the added benefit of person-related guidance for downstream identity learning.

4.5 Can Larger Perturbation Budgets Induce Purification Resistance?

Purification resistance may arise when removing protective perturbations without altering content needed for identity learning is difficult. This suggests that even a conventional defense could exhibit

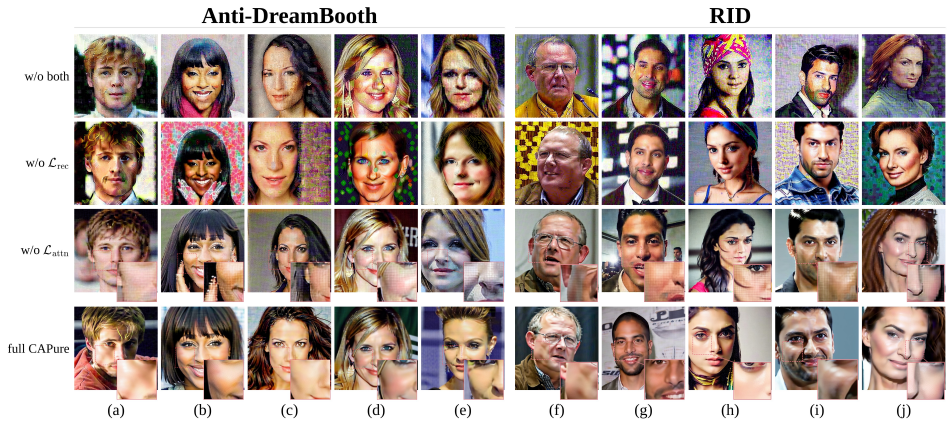

Figure 8: Generated-image ablation under Anti-DreamBooth and RID. Insets highlight residual facial textures.

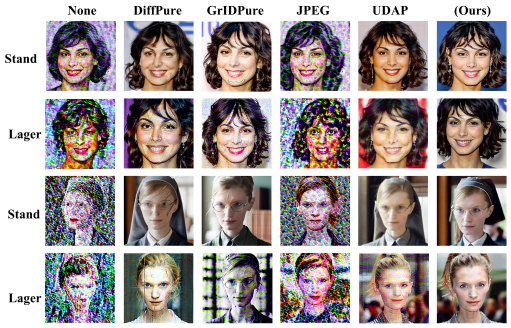

Figure 9: Generated images under Anti-DreamBooth protection with standard ($\epsilon = 0.05$) and larger ($\epsilon = 0.10$) perturbation budgets.

Table 5: Results under the larger-budget Anti-DreamBooth configuration ($\epsilon = 0.10$).

Dataset	Method	BRISQUE↓	PIQE↓	ISM↑
CelebA-HQ	None	43.25	70.54	0.153
	DiffPure	43.81	55.63	0.590
	GrIDPure	42.45	65.21	0.534
	JPEG	41.29	69.77	0.174
	UDAP	38.44	65.49	0.610
	CAPure	36.10	43.55	0.665
VGGFace2	None	64.22	74.30	0.077
	DiffPure	42.60	53.22	0.575
	GrIDPure	32.35	62.68	0.480
	JPEG	49.74	69.45	0.164
	UDAP	39.37	64.63	0.561
	CAPure	34.77	39.96	0.613

resistance to existing purifiers at a larger perturbation budget. We test this possibility by increasing the Anti-DreamBooth budget from $\epsilon = 0.05$ to $\epsilon = 0.10$. Table 5 reports the quantitative results, while Figure 9 compares generated images across the two budgets.

Increasing the perturbation budget of Anti-DreamBooth weakens existing purifiers, while CAPure better maintains identity recovery and produces more coherent faces. These results suggest that larger budgets can make a conventional defense resist existing purification methods, even without a dedicated purification-resistant design. CAPure's sustained advantage in this setting further supports guiding purification toward downstream identity learning by recovering person-related responses while preserving individual appearance.

5 Conclusion

We revisited purification-resistant personalization defenses through the lens of identity learnability. Our analysis shows that protective perturbations suppress person-token cross-attention responses, motivating CAPure's combination of attention guidance and stepwise reconstruction anchoring. This input-stage design restores person-related responses while limiting changes to individual appearance, without modifying downstream personalization. Across the evaluated defenses on CelebA-HQ and VGGFace2, CAPure achieves the highest identity similarity among the compared purifiers. Additional experiments support its effectiveness under backbone mismatch and increased perturbation budgets. These findings show that resistance to existing purification methods does not ensure protected identities remain unlearnable, motivating evaluation against purification objectives tailored to downstream identity recovery.

AI USE STATEMENT

We used generative AI tools to assist with literature searches, translation, drafting and revising text, and LaTeX formatting. These tools also provided feedback on methodology and experimental design. The authors developed the core method, made the final methodological and experimental decisions, conducted the experiments, and take full responsibility for the paper's content.

ETHICS STATEMENT

This work studies weaknesses in protective perturbations intended to prevent unauthorized personalization, with the aim of informing stronger defenses. CAPure could also facilitate unauthorized identity imitation or the generation of misleading personal images. Our experiments use CelebA-HQ and VGGFace2; results on these benchmarks do not establish effectiveness or fairness across all populations. The findings should not be interpreted as endorsing non-consensual personalization. Any application involving identifiable people should respect their consent, privacy, and control over their images, as well as applicable data-use terms, and should consider safeguards against impersonation and misuse.

REPRODUCIBILITY STATEMENT

Section 3 specifies CAPure's objectives, latent updates, and clean-reference calibration. Section 4.1 reports the datasets, personalization protocol, evaluation metrics, hardware, and main optimization settings. Appendix A details the attention extraction layer, aggregation procedure, diagnostic measurements, and calibration-set composition. The generalization experiments in Section 4.3 and component ablations in Section 4.4 further document the evaluation settings used to assess the method.

REFERENCES

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)*, 2018.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-Excite: Attention-based semantic guidance for text-to-image diffusion models, 2023.

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

Hanzhong Guo, Shen Nie, Chao Du, Tianyu Pang, Hao Sun, and Chongxuan Li. Real-time identity defenses against malicious personalization of diffusion models. *arXiv preprint arXiv:2412.09844*, 2024. URL https://arxiv.org/abs/2412.09844.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2208.01626.

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

Jinsu Kim, Yunhun Nam, Minseon Kim, Sangpil Kim, and Jongheon Jeong. BlurGuard: A simple approach for robustifying image protection against ai-powered editing. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0964.

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

Yixin Liu, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, and Lichao Sun. MetaCloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24219–24228, 2024b. doi: 10.1109/CVPR52733.2024.02286.

Yixin Liu, Ruoxi Chen, Xun Chen, and Lichao Sun. Rethinking and red-teaming protective perturbation in personalized diffusion models. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '26, pp. 937–948. Association for Computing Machinery, 2026. doi: 10.1145/3770854.3780303. URL https://doi.org/10.1145/3770854.3780303.

Anish Mittal, Anush Krishna Moorthy, and Alan C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16805–16827. PMLR, 2022. URL https://proceedings.mlr.press/v162/nie22a.html.

Takuto Onikubo and Yusuke Matsui. High-frequency Anti-DreamBooth: Robust defense against personalized image synthesis. In *ECCV 2024 Workshop on The Dark Side of Generative AIs and Beyond*, 2024. URL https://arxiv.org/abs/2409.08167.

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

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

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659, 2023. doi: 10.18653/v1/2023.acl-long.310.

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

N. Venkatanath, D. Praneeth, Bh. M. Chandrasekhar, S. S. Channappayya, and S. S. Medasani. Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications (NCC)*, pp. 1–6, 2015. doi: 10.1109/NCC.2015.7084843.

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

Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 2022. URL https://arxiv.org/abs/2205.14969.

Jingyao Xu, Yuetong Lu, Yandong Li, Siyang Lu, Dongdong Wang, and Xiang Wei. Perturbing attention gives you more bang for the buck: Subtle imaging perturbations that efficiently fool customized diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24534–24543, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xu_Perturbing_Attention_Gives_You_More_Bang_for_the_Buck_Subtle_CVPR_2024_paper.html.

Xide Xu, Muhammad Atif Butt, Sandesh Kamath, and Bogdan Raducanu. Privacy protection in personalized diffusion models via targeted cross-attention adversarial attack. In *NeurIPS 2024 Workshops: SafeGenAi*, 2024b. URL https://arxiv.org/abs/2411.16437.

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*, pp. 19290–19300, 2025.

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*, pp. 24398–24407, 2024. doi: 10.1109/CVPR52733.2024.02303.

Li Zheng, Liangbin Xie, Jiantao Zhou, and Yimin He. Universal adversarial purification with DDIM metric loss for Stable Diffusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(16):13414–13422, 2026. doi: 10.1609/aaai.v40i16.38345.

A ATTENTION-RESPONSE MEASUREMENTS

Figures 2, 3, and 4 use frozen pretrained SD v1.5 with the prompt "a photo of sks person." We extract cross-attention from the first attention module in the U-Net's second downsampling block, specifically `down_blocks.1.attentions.0.transformer_blocks.0.attn2`. For 512×512 inputs, this layer produces 32×32 attention maps. We select the *person* token and average over attention heads within this single layer. All measurements use the fixed diffusion training timestep $t = 25$; this index does not denote the 25th denoising update.

Figure 2 displays individual attention maps. For an image x, let $A_t^k(p; x)$ denote the head-averaged response at position p in the selected layer. For Figure 3, we first compute each image's spatially averaged response $m_t^k(x)$ as defined in Eq. (1), then average over 100 images in each group. We normalize each group mean by the corresponding clean-group mean.

To quantify the spatial allocation of person-token attention, we compute

$$S_t(x) = \frac{\sum_{p \in \Omega} A_t^{\text{person}}(p; x) M_x(p)}{\sum_{p \in \Omega} A_t^{\text{person}}(p; x)}, \tag{7}$$

where M_x is a soft face mask derived from the corresponding clean image and shared across matched inputs. Figure 4 reuses the same attention maps and reports the mean of the per-image face attention shares over 100 images in each group. The matched clean images and masks are used only for this diagnostic analysis. They are not required to purify a protected input; CAPure uses the offline reference statistics and the protected image itself as described in Section 3.

The offline calibration set $\mathcal{X}_{\text{clean}}$ contains 564 images from 47 CelebA-HQ identities and 600 images from 50 VGGFace2 identities, with 12 images per identity, for a total of 1,164 images from 97 identities. At each timestep t, attention is measured from directly forward-noised VAE latents, and μ_t is computed as the image mean in Eq. (1). These calibration latents are not taken from a denoising trajectory.