FACEDR-GUARD FOR ROBUST FACE-SWAP IDENTITY-REUSE DEFENSE IN SEMANTIC CODE SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Public face images can be exploited as reusable identity sources by face-swapping techniques. This identity-reuse threat differs from conventional face-recognition evasion: the key question is whether a protected image can still convey identity information to a downstream generative model. We introduce FaceDR-Guard, a proactive defense that operates on the semantic code of a frozen pretrained diffusion autoencoder. For each source image, FaceDR-Guard optimizes a bounded semantic-code offset through an end-to-end differentiable surrogate pipeline comprising image reconstruction, face swapping, and identity evaluation. The optimization suppresses the source identity retained in the swap output, while skin-aware multi-scale, chromatic-consistency, pixel-fidelity, and semantic-parsing objectives regularize visual distortion. We further employ short-chain optimization followed by long-chain rendering to reduce optimization cost and improve final reconstruction quality without updating any pretrained model weights. Under deployment shifts, experiments reveal a consistent advantage: FaceDR-Guard maintains strong identity-reuse suppression while conventional pixel-space attacks degrade, achieves robustness competitive with feature-space cloaking, and better preserves benign face-recognition utility. These findings highlight swap-output-aligned semantic reconstruction as a promising approach to robust identity-reuse defense.

1 INTRODUCTION

Recent advances in generative AI have made realistic face manipulation increasingly accessible. Face-swapping systems can transfer the identity of a source face to a target image while preserving the target's pose, expression, and background (Chen et al., 2020; Li et al., 2020; Gao et al., 2021). Consequently, a publicly available photo can be reused to synthesize manipulated faces without the owner's consent. We refer to this risk as the *identity-reuse* threat. Unlike the classical face-recognition (FR) threat, which only requires an image to evade an FR model, identity reuse persists even if the image evades FR: the image may still contain sufficient identity information for a face-swapping system to reproduce the source identity. Effective protection must therefore prevent published images from providing reusable identity information to the entire face-swapping pipeline.

Most proactive defenses, including Fawkes (Shan et al., 2020), LowKey (Cherepanova et al., 2021), PhotoGuard (Salman et al., 2023), CMUA-Watermark (Huang et al., 2022), and DiffProtect (Liu et al., 2023), optimize recognition or image-editing objectives directly on the protected image. More recent methods such as NullSwap (Wang et al., 2025), DeepProtect (Lee et al., 2026), and FaceShield (Jeong et al., 2025) target the face-swapping threat more explicitly. Nevertheless, two challenges remain. First, an objective defined on x' does not directly control the source identity retained in the swap output $F(x', t)$. Second, strong performance under one unmodified protocol does not guarantee persistent protection. In our unified evaluation, pixel-space attacks exceed a 95% defense success rate before post-processing, yet their performance deteriorates when the target face, image transformation, or swapping architecture changes. These observations motivate a defense whose objective matches the downstream identity-reuse criterion and whose adversarial signal survives deployment transformations without introducing conspicuous image distortion.

We address these problems with FaceDR-Guard, which performs the defensive modification in the semantic code of a frozen pretrained diffusion autoencoder. For each source image, we optimize a bounded semantic-code offset through a fully differentiable chain that includes reconstruction,

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.

face swapping, and identity evaluation. The loss directly minimizes the source identity retained in the swap output. Our carrier analysis justifies this design: pixel perturbations are attenuated by common image transformations, whereas decoder-weight optimization entangles reconstruction with identity suppression and incurs $2.3\times$ the chromatic drift of the semantic-code carrier. In contrast, a semantic-code offset alters the identity information available to the swapper while keeping the pretrained decoder as a fixed image prior. Skin-aware multi-scale, chromatic-consistency, and pixel losses constrain visible distortion. Short-chain optimization followed by long-chain rendering reduces optimization cost and improves final reconstruction without updating pretrained model weights. Our contributions are as follows:

- We formulate proactive face-swap protection as directly suppressing the source identity retained in the swap output and systematically compare pixel-space, decoder-weight, and semantic-code carriers under a unified criterion. This analysis exposes the dominant failure mode of each carrier and motivates the semantic code as the locus of the defensive change.
- We optimize a bounded offset against the identity cosine of the swap output while keeping all pretrained weights fixed. To jointly control visual distortion and computational cost, we combine skin-aware multi-scale, chromatic-consistency, and pixel-fidelity losses with a short-chain optimization stage followed by long-chain rendering.
- On CelebA-HQ, FaceDR-Guard reaches 99% defense success and maintains 99–100% under JPEG compression, resizing, and blurring, whereas Pixel-PGD and MIM decrease to 39–83% and 68–91%, respectively. Protection optimized against one surrogate target transfers to $K{=}8$ held-out targets (99.0% mean and 98.0% worst), an unseen swapper (100%), and two independent identity encoders (94–98%), demonstrating robust identity-reuse protection under deployment shifts.

2 Related Work

Proactive defenses. Face-recognition models compare identity embeddings (Taigman et al., 2014; Schroff et al., 2015; Deng et al., 2019; Meng et al., 2021; Chen et al., 2018), whereas face-swapping models extract identity from a source image and render it on a target face (Chen et al., 2020; Li et al., 2020; Gao et al., 2021). Most proactive defenses target recognition or a specific manipulation operation. Fawkes (Shan et al., 2020), LowKey (Cherepanova et al., 2021), and DiffProtect (Liu et al., 2023) alter the recognition embedding. CMUA-Watermark (Huang et al., 2022), initiative defense (Huang et al., 2021), anti-forgery noise (Wang et al., 2022), DF-RAP (Qu et al., 2024), and AMT-GAN (Hu et al., 2022) disrupt detection, attribute manipulation, or forgery generation, while PhotoGuard (Salman et al., 2023) targets diffusion-based editing. Recent face-swap defenses move closer to the downstream threat. NullSwap (Wang et al., 2025) uses image-space perturbations, DeepProtect (Lee et al., 2026) and SCOL (Lee et al., 2025) modify generative representations, and SGD-Guard (Back et al., 2026) and ID-Eraser (Luo et al., 2026) intervene in identity-bearing features. However, recognition- and operation-specific objectives do not directly suppress the source identity retained in the swap output. Moreover, prior evaluations differ in datasets, swappers, thresholds, and access assumptions, making their persistence and fidelity difficult to compare (Baek et al., 2026). To bridge this objective gap, we directly optimize the identity cosine of a differentiable swap output and re-evaluate all main baselines under a single unified protocol.

Generative-model-based face privacy protection. Generative representations have been used to reconstruct or anonymize faces while retaining visual quality (Liu et al., 2023; Le et al., 2025; Yang et al., 2025; Kung et al., 2024). DiffProtect (Liu et al., 2023) is the closest analogue to our use of a diffusion representation, but it optimizes the FR embedding of the protected image. Consequently, neither reconstruction nor anonymization alone determines whether a downstream swapper can reuse the source identity. Moreover, methods that modify generator parameters or internal representations entangle the defensive signal with image generation, which can increase visual drift and complicate deployment. In contrast, we place the signal in a bounded offset of the semantic code of a frozen diffusion autoencoder (Preechakul et al., 2022). The fixed decoder re-renders the protected image, while the offset is optimized against the swap-output identity loss. Pixel, decoder-weight, and semantic-code controls then quantify the trade-offs between identity suppression, persistence, and fidelity (§3.3).

Method	Venue	Reported DSR↑	Reported ID cos ↓	Source
WDP	–	36.2%	–	Lee et al., 2026 T. 2
CMUA-Watermark	AAAI'22	41.0%	0.921	Lee et al., 2026 T. 2
DF-RAP	TIFS'24	42.7%	0.468	Lee et al., 2026 T. 2
DeepProtect (Attribute-only)	CVPR'26	60.5%	–	Lee et al., 2026 T. 2
Fawkes	S&P'20	61.6%	–	Lee et al., 2026 T. 2
AMT-GAN	CVPR'22	68.8%	–	Lee et al., 2026 T. 2
CLIP2Protect	CVPR'24	79.0%	–	Lee et al., 2026 T. 2
DiffAM	CVPR'25	79.8%	–	Lee et al., 2026 T. 2
FaceShield†	ICCV'25	82.1%	–	Lee et al., 2026 T. 2
NullSwap‡	ICCV'25	–	0.217	Wang et al., 2025 T. 3
DeepProtect (Combined)‡	CVPR'26	94.8%	–	Lee et al., 2026 T. 2
SCOL‡	MM'25	95.1%	–	Back et al., 2026 T. 1
SGD-Guard‡	IJCAI'26	94.3%	–	Back et al., 2026 T. 1
ID-Eraser‡	arXiv'26	–	0.504§	Luo et al., 2026
FaceDR-Guard (ours, n=500)	–	**99%**	$\mathbf{-0.037}$	§4.2

Table 1: **Literature positioning on the identity-reuse surface.** Reported DSR is measured against SimSwap under each paper's protocol. Only the boldfaced row is evaluated in our setup, and values across rows are not directly comparable because the splits, swappers, K, and τ differ. NullSwap has no published code, while DeepProtect and SCOL require an unavailable VGGFace2-HQ feature bank. FaceShield uses a diffusion-centric threat model. SCOL/SGD-Guard results are quoted from SGD-Guard's Table 1. † denotes the diffusion-centric setting, ‡ a method not re-evaluated here, and § the protected-swap identity similarity averaged over five swappers as reported by ID-Eraser.

3 Methodology

3.1 Threat model and evaluation protocol

Let x be the owner's original image and x' the published protected image. An attacker who obtains x' can select a target face t and a swapper F to generate $F(x', t)$. Defense succeeds for x' when the generated face no longer resembles the owner's identity.

$$\mathrm{DSR}_{\mathrm{swap}}(x') \;=\; \mathbb{1}\Big[\;\cos\big(\mathrm{arc}(F(x',t)),\;\mathrm{arc}(x)\big)\;<\;\tau\;\Big], \qquad \tau = 0.30, \tag{1}$$

where $\mathrm{arc}(\cdot)$ is an ArcFace embedding and F is SimSwap. The threat model allows the attacker to choose an arbitrary target face t. Our *optimization* protocol uses a single fixed surrogate target t_{opt} (§3.4), and generalization to unseen targets is evaluated separately (§4.3). We also report the mean swap identity cosine $\cos_{\mathrm{def}} = \frac{1}{N}\sum_i \cos(\mathrm{arc}(F(x'_i,t)), \mathrm{arc}(x_i))$ and the standard fidelity metrics L_1, PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018) against the clean image x. All methods in one table are evaluated on *the same* stems, the *same* target face t, the same preprocessing, and the same evaluator, so numbers are directly comparable. The threshold $\tau = 0.30$ is fixed across our evaluations, and $\mathrm{DSR}_{\mathrm{swap}}$ reports the fraction of successful defenses.

3.2 Overview

Figure 1 provides an overview of the proposed pipeline. Given a source image x, the frozen DiffAE encoder maps it to a semantic code z_0. We optimize a bounded semantic-code offset δ and pass $z_0 + \delta$ with the frozen decoder to obtain the protected image x'. A face swapper F then combines x' with the surrogate target t_{opt}, and ArcFace measures the identity retained in the swap output. This identity signal is backpropagated to δ through the reconstruction and swapping chain. In parallel, the skin map from semantic parsing supplies the spatial weights for the multi-scale visual loss, and the pixel fidelity term constrains the overall image change. Throughout optimization, the encoder, decoder, swapper, and identity evaluator remain frozen; only the semantic-code offset is updated. We first analyze the choice of defensive carrier, then present semantic-code optimization and two-chain rendering, and finally define the loss functions.

3.3 Adversarial carrier

We consider three candidate carriers for the defensive modification: the pixel image, the decoder weights, and the semantic code. Our method adopts the semantic code as the carrier, optimizing

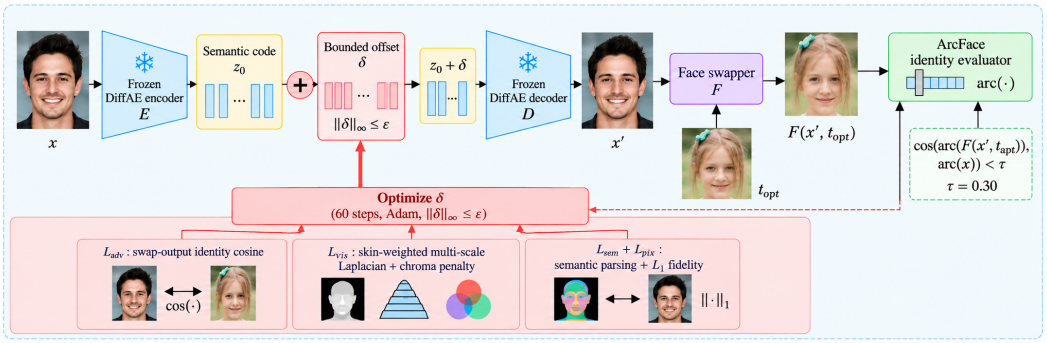

Figure 1: **Overview of FaceDR-Guard.** A bounded offset δ is optimized in semantic code space through a differentiable reconstruction and face-swapping chain. Face-aware fidelity losses preserve the source appearance, while the swap-output identity objective suppresses identity reuse. Short-chain optimization and long-chain rendering produce the protected image, which transfers to held-out target faces.

Table 2: Adversarial carriers and their dominant cost. Pixel-domain and latent pixel-attack methods in our protocol reach DSR_{swap} 96–100% at $L_1 \le 0.037$ but their protection degrades under post-processing (Table 6). The semantic-code carrier avoids this decay.

Carrier	Form	Identity effect	Dominant cost
Decoder weights	fine-tune G_θ	strong but unstable	reconstruction–identity conflict and color artifacts
Pixel	$x' = x + \delta_x$	indirect	high-frequency, visible, absorbed by pipeline
Semantic code	$z' = z + \delta$	direct	needs a fidelity protocol (§3.5)

a bounded offset and re-rendering the protected image through the unchanged pretrained decoder. Table 2 summarizes the three carriers and their dominant costs.

3.4 Semantic-code optimization and two-chain inference

For each source image x, we first encode it into a semantic code $z_0 = \text{Enc}(x)$ using the frozen pretrained DiffAE encoder. We then compute a texture state x_T via DDIM inversion with a short chain of $T_{\text{short}} = 4$. During optimization, a bounded semantic offset δ is added to z_0 and rendered as

$$x'_{T_{\text{short}}} = D\big(x_T;\ z_0 + \delta,\ T_{\text{short}}\big), \tag{2}$$

where $D(\cdot;\cdot,T)$ denotes the pretrained decoder run for T DDIM steps in the forward direction, with weights unchanged. The surrogate target t_{opt} is fixed during optimization, and the resulting objective is differentiated through the decoder, swapper, and ArcFace evaluator to update only δ. All pretrained networks remain frozen. We use Adam with learning rate 0.02 for 60 steps and project δ to $[-\varepsilon, \varepsilon]$ after every update. We set $\varepsilon = 0.05 \approx 0.18\,\sigma(z)$ to limit displacement in semantic code space. A larger budget of $\varepsilon = 0.4$ produces visible color speckle after long-chain rendering.

The short chain is used only for optimization because it requires 60×4 denoising steps. The final protected image is instead rendered with a long chain that restores texture, reducing the pure reconstruction L_1 from 0.057 at $T = 4$ to 0.011 at $T = 50$. Since the long chain can amplify the same semantic offset, the bound on δ is also necessary to control visual artifacts. We therefore render the final protected image as

$$x' = D\big(\text{DDIMInv}(x; z_0, T_{\text{long}});\ z_0 + \delta^*,\ T_{\text{long}}\big), \qquad T_{\text{long}} = 50. \tag{3}$$

This protocol reduces optimization cost while using the longer chain for the final reconstruction. Algorithm 1 gives the full per-image procedure.

3.5 LOSS FUNCTIONS

We optimize a bounded semantic-code offset with an identity loss defined on the swap output and two fidelity terms applied to the protected image. These objectives jointly suppress the source identity retained by the swapper while constraining visual distortion,

$$\delta^* = \arg\min_{\|\delta\|_\infty \le \varepsilon} \mathcal{L}_{\text{adv}}(\delta) + \lambda_{\text{vis}}\,\mathcal{L}_{\text{vis}} + \lambda_{\text{pix}}\,\mathcal{L}_{\text{pix}}. \tag{4}$$

where the adversarial term uses the fixed surrogate target t_{opt} and the complete differentiable swap chain

$$\mathcal{L}_{\text{adv}}(\delta) = \cos\Big(\text{arc}\big(F(x'_{T_{\text{short}}},\, t_{\text{opt}})\big),\ \text{arc}(x)\Big). \tag{5}$$

This term matches the evaluation criterion in Eq. (1). Optimization uses one target, while transfer to $K{=}8$ unseen targets is evaluated separately (§4.3).

Let $\Delta = x' - x$ be the pixel difference during optimization, $M_{\text{skin}} = \text{BiSeNet}(x)$ (Yu et al., 2018) the binary skin map, and $\mathbf{P} = 1 + sM_{\text{skin}}$ with $s = 2.5$. The visual term combines a skin-weighted Laplacian-pyramid penalty with a chroma penalty on the perturbation channels

$$\mathcal{L}_{\text{vis}} = \sum_l w_l \frac{\|\text{pyr}_l(\mathbf{P} \odot \Delta)\|_1}{N} + \gamma \frac{\|\Delta_R - \Delta_G\|_1 + \|\Delta_B - \Delta_G\|_1}{N}. \tag{6}$$

The pixel term is the mean L_1 magnitude of Δ and limits the overall change from the source image.

4 EXPERIMENTS AND ANALYSIS

4.1 EXPERIMENTAL SETTINGS

Dataset. We evaluate on the official 500-image CelebA-HQ benchmark at 256×256, adopted by the external baselines. Unless otherwise specified, each protected image is optimized against a single fixed surrogate target face t_{opt}. For held-out-target evaluation, we use $K{=}8$ target faces disjoint from t_{opt}. Cross-dataset transfer is measured on 200-image subsets of FFHQ and LADN with the same SimSwap surrogate, target, and configuration. For SCOL, robustness is evaluated on an $n{=}100$ subset because its released pipeline requires a feature gallery that is not available for a matched re-evaluation.

Comparison methods. In the main comparisons, we re-evaluate Pixel-PGD, MIM, DF-RAP, and AMT-GAN under a unified protocol with identical images, target face, preprocessing, and identity evaluator. Pixel-PGD and MIM perturb the protected image to evade an ensemble of FR embedders comprising IR-SE50, IR-152, and ArcFace; Pixel-PGD uses untargeted FR-embedding PGD, while MIM uses its momentum-based variant. For methods that cannot be re-evaluated under this protocol, we report their published numbers, which follow each paper's original setting. These methods include NullSwap, DeepProtect, SCOL, SGD-Guard, and ID-Eraser, subject to the stated availability constraints.

Evaluation metrics. For each protected image x'_i, we apply SimSwap with the surrogate target t_{opt} and measure the identity cosine of the swap output and the batch defense success rate

$$c_i = \cos\big(\text{arc}(F(x'_i, t_{\text{opt}})),\ \text{arc}(x_i)\big), \quad \text{DSR}_{\text{swap}} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbf{1}[c_i < \tau], \tag{7}$$

with $\tau = 0.30$. We report the mean swap identity cosine together with L_1, PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) against the original image. Robustness re-scores c_i after JPEG recompression, halved resolution, and Gaussian blur. For benign FR utility, we apply the same threshold directly to the protected image without swapping and report FR-DSR. Lower FR-DSR indicates higher retained recognition utility. All reported metrics use the same stems, target, preprocessing, and evaluator unless a subsection states otherwise.

Implementation details. DiffAE (ffhq256), SimSwap, ArcFace, and BiSeNet use off-the-shelf weights that remain fixed. We optimize δ with Adam at learning rate 0.02 for 60 steps, using $T_{\text{short}} = 4$, $T_{\text{long}} = 50$, $\varepsilon = 0.05$, $\lambda_{\text{vis}} = \lambda_{\text{pix}} = 1.0$, skin weight $s = 2.5$, and chroma weight $\gamma = 1.0$. Experiments use one NVIDIA GPU with about 16 GB of memory and require about one minute per image. Unless noted, results use $\tau = 0.30$ and evaluate the full 500-image CelebA-HQ batch.

Table 3: **Unified-protocol comparison on CelebA-HQ.** Results on 500 images with fixed target t_{opt}, common preprocessing, SimSwap, ArcFace, and τ=0.30. Fidelity metrics are computed against the clean image. Bold denotes the best value among defended methods. Clean is an unprotected reference.

Method	$\cos_{\text{def}} \downarrow$	$\text{DSR}_{\text{swap}} \uparrow$	$L_1 \downarrow$	PSNR↑	SSIM↑	LPIPS↓
Clean (no defense)	+0.804	0%	0.000	∞	1.000	0.000
Pixel-PGD	+0.030	96%	0.031	29.13	0.730	0.199
MIM	+0.007	98%	0.037	27.75	0.683	0.275
DF-RAP	+0.742	0%	**0.014**	**35.76**	**0.907**	**0.032**
AMT-GAN	+0.476	13.4%	0.063	22.44	0.781	0.075
FaceDR-Guard (Ours)	**−0.037**	**99%**	0.053	21.80	0.799	0.163

Table 4: **Benign face-recognition utility after protection.** FR-DSR on 500 protected images at τ=0.30.

	FR-DSR of protected image (↓, higher utility)		
Method	ArcFace	MobileFaceNet	MagFace
Clean (no defense)	0%	0%	0%
Pixel-PGD	100%	81%	88%
MIM	100%	88%	92%
FaceDR-Guard (Ours)	**61%**	**15%**	**80%**

4.2 Defense and Fidelity under One Protocol

Unified-protocol comparison. Identity-reuse defense and image fidelity are compared under a matched protocol. The quantitative results are reported in Table 3. All methods use the same 500 CelebA-HQ images, fixed target t_{opt}, preprocessing, SimSwap, and ArcFace evaluator at $\tau = 0.30$. The table shows that FaceDR-Guard reaches DSR_{swap} =99%, compared with 96% for Pixel-PGD and 98% for MIM. DF-RAP and AMT-GAN reach 0% and 13.4%, respectively, under the swap-output criterion. Among defended methods, DF-RAP has the best pixel-level fidelity, whereas FaceDR-Guard has the strongest swap-identity suppression.

Benign face-recognition utility. Direct face-recognition utility is measured on protected images. The results are reported in Table 4. We apply the recognition threshold directly to each protected image, without face swapping, using ArcFace, MobileFaceNet, and MagFace. Lower FR-DSR indicates higher retained recognition utility. The FR-DSR for FaceDR-Guard is 61%, 15%, and 80% under ArcFace, MobileFaceNet, and MagFace, respectively, versus 100%, 81–88%, and 88–92% for the FR-targeted baselines. Thus, the method suppresses identity reuse after swapping while retaining more direct-recognition utility than those baselines.

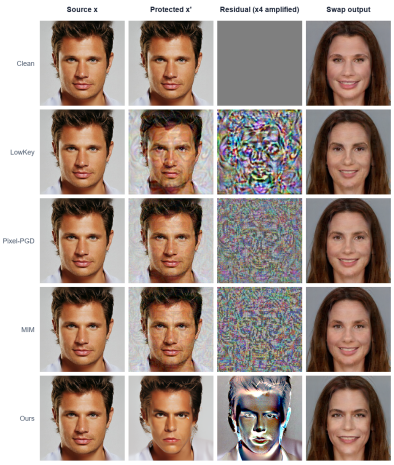

Figure 2: **Qualitative comparison on CelebA-HQ.** From left to right: source, protected image, amplified residual, and swap output.

Qualitative comparison. Figure 2 compares the source, protected image, amplified residual, and swap output of representative examples that follow the same protocol as Table 3. The residuals show visible high-frequency structure for the FR-targeted baselines, whereas the semantic-code reconstruction produces a smoother change. In the corresponding swap outputs, the source identity is visibly suppressed for FaceDR-Guard.

Table 5: **Held-out-target transfer.** Results on 500 images and $K=8$ target faces not used during optimization ($\tau=0.30$). Target spread is $\max_k \mathrm{DSR}_{\mathrm{swap}_k} - \min_k \mathrm{DSR}_{\mathrm{swap}_k}$.

Method	mean $\mathrm{DSR}_{\mathrm{swap}}$ ↑	worst-target $\mathrm{DSR}_{\mathrm{swap}}$ ↑	mean cos ↓	target spread↓
Clean (no defense)	0.0%	0.0%	+0.800	0%
Pixel-PGD	96.4%	95.4%	+0.035	2.6%
MIM	97.6%	96.8%	+0.011	2.2%
SCOL	57.2%	46.0%	+0.287	24.0%
FaceDR-Guard (Ours)	**99.0%**	**98.0%**	**−0.012**	**2.0%**

Table 6: **Detailed platform-robustness results.** $\mathrm{DSR}_{\mathrm{swap}}$ under six transformations. SCOL uses $n=100$ with a substituted LFW gallery.

Degradation	$\mathrm{DSR}_{\mathrm{swap}}$ ↑				
	Clean	Pixel-PGD	MIM	SCOL	FaceDR-Guard
none	0%	96%	98%	54%	**99%**
JPEG $q=75$	0%	83%	91%	55%	**99.4%**
JPEG $q=60$	0%	73%	88%	56%	**99.4%**
resize $0.5\times$	0%	53%	79%	55%	**99.2%**
blur $\sigma=1$	0%	55%	80%	56%	**99.2%**
blur $\sigma=2$	0%	39%	68%	58%	**99.8%**

4.3 Robustness to Deployment Conditions

A defense optimized against one surrogate target should remain effective under deployment variation. We test transfer to a different target face and robustness to platform post-processing with JPEG, resize, and blur.

Held-out target faces. Transfer across target identities is evaluated, with results in Table 5. We optimize each protected image against one surrogate target t_{opt} and then evaluate the same image against $K=8$ target faces that are disjoint from the optimization target. This tests whether protection transfers across target identities. Across the held-out targets, FaceDR-Guard reaches a mean DSR of 99.0% and a worst-target DSR of 98.0%, with mean swap cosine -0.012 and spread 2.0%. Pixel-PGD and MIM obtain mean DSRs of 96.4% and 97.6%, but their worst-target values fall to 95.4% and 96.8%, respectively.

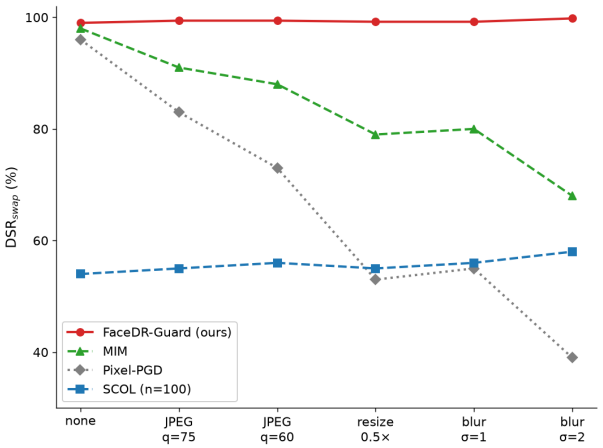

Figure 3: **Robustness to platform processing.** $\mathrm{DSR}_{\mathrm{swap}}$ under JPEG, resize, and blur on the 500-image batch.

Platform post-processing. Robustness to common transmission changes is evaluated. The trend is visualized in Figure 3 and detailed in Table 6. We re-score the same protected images after JPEG recompression, $0.5\times$ resizing, and two levels of Gaussian blur. The main batch contains 500 images. SCOL is reported on its available 100-image subset. Across the six conditions, FaceDR-Guard remains between 99.0% and 99.8%, whereas Pixel-PGD falls from 96% to 39% and MIM from 98% to 68% under the strongest blur. SCOL remains in the 54–58% range on its smaller subset. The contrast is consistent with semantic-code changes surviving image transformations while high-frequency pixel perturbations are attenuated.

Table 7: **Transfer across identity encoders.** Results on the same n=500 protected images with ArcFace, MagFace, and MobileFaceNet at τ=0.30.

	ArcFace		MagFace		MobileFaceNet	
Method	cos ↓	DSR_{swap} ↑	cos ↓	DSR_{swap} ↑	cos ↓	DSR_{swap} ↑
Clean (no defense)	+0.804	0%	+0.670	0%	+0.562	0%
Pixel-PGD	+0.030	96%	−0.031	**99%**	+0.051	98%
MIM	+0.007	98%	**−0.036**	**99%**	**+0.049**	**99%**
SCOL	+0.285	54%	+0.343	26%	+0.345	32%
FaceDR-Guard (Ours)	**−0.037**	**99%**	+0.053	98%	+0.143	94%

Table 8: **Transfer across face swappers.** SimSwap and InSwapper_{128} on n=200 images with the same identity evaluator and τ=0.30.

	SimSwap		InSwapper_128	
Method	cos ↓	DSR_{swap} ↑	cos ↓	DSR_{swap} ↑
Clean (no defense)	+0.805	0%	+0.633	0.5%
Pixel-PGD	+0.029	96%	+0.198	77%
MIM	+0.001	**99%**	+0.151	85%
FaceDR-Guard (Ours)	**−0.023**	98.5%	**+0.003**	**100%**

4.4 Cross-model transferability

Optimization uses one ArcFace embedder and one SimSwap surrogate. A real attacker can swap with a different architecture, and a defender may score with a different recognizer. We therefore evaluate transfer to independent identity encoders, an unseen swapping architecture, and held-out datasets.

Independent identity encoders. Transfer across identity encoders is evaluated. The results are reported in Table 7. We re-score the same protected images with MagFace and MobileFaceNet in addition to the ArcFace evaluator used during optimization. We keep τ=0.30 for all three encoders. Because their clean cosine distributions differ, this is a transferability check rather than a calibrated operating-point comparison. The defense obtains 99%, 98%, and 94% DSR under ArcFace, MagFace, and MobileFaceNet, respectively. Pixel-PGD and MIM reach 99% under MagFace, while MIM reaches 99% under MobileFaceNet. These results show transfer beyond the evaluator used for optimization but also motivate care when comparing fixed thresholds across encoders.

Unseen face swapper. Transfer to an unseen swapping architecture is evaluated in Table 8. We re-swap the protected images with InsightFace InSwapper_128, which is architecturally different from the SimSwap surrogate, while keeping the identity evaluator and threshold fixed. Under InSwapper_128, FaceDR-Guard reaches 100% DSR with cosine +0.003, while Pixel-PGD and MIM fall to 77% and 85% from 96% and 99% under SimSwap. Under the SimSwap columns, MIM has the highest DSR (99%), whereas FaceDR-Guard has the lowest cosine.

Cross-dataset transfer. Generalization beyond CelebA-HQ is evaluated on FFHQ and LADN. Results are reported in Table 9. We apply the same SimSwap surrogate, target, preprocessing, and configuration to 200-image subsets of FFHQ and LADN. The purpose is to test whether the protection transfers beyond CelebA-HQ without changing the evaluation rule. On FFHQ, FaceDR-Guard reaches 100%, 99.5%, and 100% DSR under ArcFace, MobileFaceNet, and MagFace. On LADN, the corresponding values are 97%, 90.5%, and 92.5%. The observed cross-dataset range is 90.5–100%. Clean ArcFace cosines of +0.772 (FFHQ) and +0.814 (LADN) indicate that the fixed threshold is conservative on these datasets.

4.5 Mechanism and ablation

Fidelity-term and optimization-step ablation. Table 10 isolates the fidelity terms and optimization budget on the same 30-image subset, with all other settings fixed. Lower $\cos_{\text{swap}}$ indicates stronger defense, whereas higher $\cos_{\text{FR}}$ and lower L_1 and chroma error indicate better fidelity. Removing the chroma penalty strengthens swap suppression (−0.073 versus −0.055) and slightly

Table 9: **Cross-dataset transfer.** CelebA-HQ, FFHQ, and LADN under the same SimSwap surrogate, target, and configuration (τ=0.30).

Dataset (n)	ArcFace (cos ↓ / DSR_{swap} ↑)	MobileFaceNet DSR_{swap} ↑	MagFace DSR_{swap} ↑
CelebA-HQ (500)	−0.037 / 99%	94%	98%
FFHQ (200)	−0.118 / 100%	99.5%	100%
LADN (200)	+0.021 / 97%	90.5%	92.5%

Table 10: **Fidelity-term and optimization-step ablation.** Each row removes one fidelity term or changes the optimization steps on the same 30-image subset. Bold marks column-wise extrema, not an overall ranking.

Variant	$\cos_{\text{swap}}$ ↓	$\cos_{\text{FR}}$ ↑	L_1 ↓	chroma↓
Full (ours)	−0.055	+0.019	0.056	**0.037**
w/o skin weighting	−0.060	+0.020	0.056	**0.037**
w/o chroma penalty	−0.073	+0.000	**0.055**	0.038
w/o pixel fidelity	−0.068	+0.004	0.058	0.039
w/o multi-scale term	−0.064	+0.007	0.058	0.040
steps = 30	−0.077	−0.001	0.056	**0.037**

reduces L_1, but lowers $\cos_{\text{FR}}$ to 0.000 and increases chroma error. Removing skin weighting has little effect on this subset, while removing pixel fidelity or the multi-scale term degrades both $\cos_{\text{FR}}$ and the distortion metrics. Reducing the budget to 30 steps gives the lowest $\cos_{\text{swap}}$ (−0.077), but also the lowest $\cos_{\text{FR}}$ (−0.001). Thus, the column-wise extrema reflect defense–fidelity trade-offs rather than a uniformly superior variant. The full objective provides the intended balance.

Fidelity–defense trade-off. We further sweep the code budget ε to measure the fidelity–defense trade-off; results are shown in Figure 4. The mean swap cosines are +0.235, −0.070, and −0.208 at ε=0.025, 0.05, and 0.10, with corresponding L_1 values of 0.035, 0.057, and 0.072.

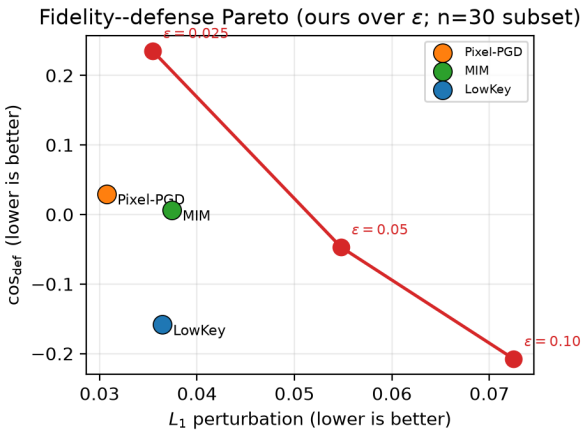

Figure 4: **Fidelity–defense trade-off.** Results for $\varepsilon \in \{0.025, 0.05, 0.10\}$ on a 30-image subset.

5 Conclusion

This work studies where defensive information should reside for face-swap identity-reuse protection. FaceDR-Guard places a bounded offset on the semantic code of a frozen diffusion autoencoder and optimizes it against the identity of the swap output. Under tested settings, the defense achieves 99% DSR on the 500-image CelebA-HQ benchmark following the unified protocol, with superior direct face-recognition utility preservation over FR-targeted baselines. It maintains 99.0–99.8% DSR against JPEG recompression, resizing and Gaussian blur, and generalizes to eight unseen target faces, an unseen InSwapper_{128}, and independent identity encoders. Cross-dataset tests on FFHQ and LADN further give 90.5–100% DSR under the same evaluation rule. These results support the central design choice of aligning optimization with the identity retained in the swap output and placing the perturbation in a semantic code rather than directly in pixels or model weights. Extending protection to broader generators and deployment conditions remains an important direction for future work.

AI USE DISCLOSURE

The authors used large language model (LLM) assistants to polish and restructure the manuscript, retrieve related work, and cross-check references and reported numbers. The authors conceived the threat model, adversarial-carrier analysis, method design, experiments, and interpretation of results. They also conducted all experiments. All LLM-generated text and suggestions were reviewed, verified, and edited by the authors, who take full responsibility for the content of this paper.

REFERENCES

Seung-hyeok Back, Do Hyun Ki, Juwan Kim, and Seok Bong Yoo. Robust, generalizable proactive face-swapping defense via semantic gradient divergence. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI-ECAI)*, 2026.

Joonhyuk Baek, Wonjune Seo, Jae-yun Kim, Saerom Park, and Hoki Kim. Proactive defense benchmark against deepfake generation. In *International Conference on Machine Learning (ICML)*, 2026.

Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *ACM MM*, 2020.

Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for improvement in face recognition. *arXiv:1804.07573*, 2018.

Valeria Cherepanova, Micah Goldblum, Harrison Foley, et al. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In *ICLR*, 2021.

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

Ge Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Infoswap: Information bottleneck theory guided face swapping. In *ECCV*, 2021.

Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *CVPR*, pp. 15014–15023, 2022.

Hao Huang, Yongtao Wang, Zhaoyu Chen, Yuze Zhang, Yuheng Li, Zhi Tang, Wei Chu, Jingdong Chen, Weisi Lin, and Kai-Kuang Ma. Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):989–997, 2022.

Qidong Huang, Jie Zhang, Wenbo Zhou, Weiming Zhang, and Nenghai Yu. Initiative defense against facial manipulation. In *AAAI*, 2021.

Jaehwan Jeong, Sumin In, Sieun Kim, Hannie Shin, Jongheon Jeong, Sang Ho Yoon, Jaewook Chung, and Sangpil Kim. Faceshield: Defending facial image against deepfake threats. In *ICCV*, 2025.

Han-Wei Kung, Tuomas Varanka, Sanjay Saha, and Nicu Sebe. Face anonymization made simple. *arXiv:2411.00762*, 2024.

Kevin Le et al. Diffprivate. *PoPETs*, 2025.

Eungi Lee, Jae Hyun Yoon, and Seok Bong Yoo. Scol: Style code orchestration in latent space for proactive face-swapping defense. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM)*, pp. 11472–11481, 2025.

Eungi Lee, Seung-hyeok Back, Hyung-Il Kim, and Seok Bong Yoo. Deepprotect: Proactive face-swapping defense using identity blending and attribute distortion. In *CVPR*, 2026.

Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. In *AAAI*, 2020.

Jiang Liu, Chun Pong Lau, Rama Chellappa, et al. Diffprotect: Generate adversarial examples with diffusion models for facial privacy protection. *Pattern Recognition*, 2023.

Junyan Luo, Peipeng Yu, Jianwei Fei, Shiya Zeng, Xiaoyu Zhou, Zhihua Xia, and Xiang Liu. Id-eraser: Proactive defense against face swapping via identity perturbation. *arXiv preprint arXiv:2604.21465*, 2026.

Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality. In *CVPR*, 2021.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 2022.

Zuomin Qu, Zuping Xi, Wei Lu, Xiangyang Luo, Qian Wang, and Bin Li. Df-rap: A robust adversarial perturbation for defending against deepfakes in real-world social network scenarios. *IEEE Transactions on Information Forensics and Security*, 2024.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-driven image editing. In *ICML*, 2023.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *USENIX Security*, 2020.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

Run Wang, Ziheng Huang, Zhikai Chen, Li Liu, Jing Chen, and Lina Wang. Anti-forgery: Towards a stealthy and robust deepfake disruption attack via adversarial perceptual-aware perturbations. In *IJCAI*, pp. 761–767, 2022.

Tianyi Wang, Shuaicheng Niu, Harry Cheng, Xiao Zhang, and Yinglong Wang. Nullswap: Proactive identity cloaking against deepfake face swapping. In *ICCV*, 2025.

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

Haoxin Yang, Yihong Lin, Xuemiao Xu, Yue Li, et al. Beyond inference intervention: Identity-decoupled diffusion for face anonymization. *arXiv:2510.24213*, 2025.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018.

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

APPENDIX

Semantic-code adversarial reconstruction. Algorithm 1 constructs one protected image while keeping the encoder, decoder, swapper, identity encoder, and skin parser fixed. It first encodes the source and computes a short-chain texture state. The only optimized variable is the semantic offset δ, which is initialized to zero and constrained elementwise to $[-\varepsilon, \varepsilon]$. At every iteration, the decoder renders a candidate protected image with $T{=}4$ denoising steps. The swapper combines this candidate with the fixed surrogate target, and the total objective combines the swap-output identity term with the visual and pixel fidelity terms defined in Section 3.5. Adam updates δ, after which projection enforces the bound. Once the 60-step optimization ends, the offset is detached from the computation graph and reused in a fresh $T{=}50$ reconstruction. This final long-chain pass produces the image released by the defense and does not update any pretrained parameter.

Algorithm 1 Semantic-code adversarial reconstruction (per image)

Require: source x, surrogate target t_{opt}, pretrained $\{\text{Enc}, D, F, \text{arc}, \text{BiSeNet}\}$ (weights fixed)

$z_0 \leftarrow \text{Enc}(x)$ ▷ $\mathbb{R}^{512}$ with fixed encoder weights
$x_T \leftarrow \text{DDIMInv}(x, z_0, 4)$ ▷ short-chain texture state
$\delta \leftarrow \mathbf{0}$ with $\delta.\mathbf{requires_grad_} = \mathbf{True}$
for step $= 1 \ldots 60$ **do**
  $x' \leftarrow D(x_T, z_0 + \delta, 4)$ ▷ decoder forward with gradients only for δ
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{adv}} + \lambda_{\text{vis}}\mathcal{L}_{\text{vis}} + \lambda_{\text{pix}}\mathcal{L}_{\text{pix}}$
  Update δ with Adam and set $\delta \leftarrow \text{clip}(\delta, -\varepsilon, \varepsilon)$
end for
$\delta^* \leftarrow \delta.\text{detach}()$
$x'_{\text{final}} \leftarrow D\big(\text{DDIMInv}(x, z_0, 50), z_0 + \delta^*, 50\big)$
return x'_{final}