

FaceArmor: A Universal Facial Image Protection Against Diffusion-Based Manipulations

Yiming Wang¹, Jiahao Chen¹, Qingming Li¹, Chunyi Zhou¹, Zhi Chen²,
Lingzhong Meng³, Jinbao Li^{4*}, and Shouling Ji^{1*}

¹ Zhejiang University, China

² University of Illinois Urbana-Champaign, USA

³ Chinese Academy of Sciences, China

⁴ Qilu University of Technology, China

ym_wang@zju.edu.cn, lijib@qlu.edu.cn, sjizju.edu.cn

Abstract. The rapid advancement of diffusion models has exacerbated privacy risks by enabling realistic manipulations of personal facial images. Existing proactive defenses primarily target model-specific internal modules under restrictive white-box assumptions, rendering their protective efficacy highly vulnerable to unseen architectures and manipulation paradigms. To bridge this gap, we propose **FaceArmor**, a universal facial image protection framework against both unknown generative architectures and real-world environmental transformations. Unlike prior model-dependent approaches, FaceArmor focuses on the intrinsic feature composition of facial images. It leverages an ensemble of surrogate extractors to decompose representations into a structural tier targeting latent reconstruction and fine textures, and a semantic tier obscuring prompt alignment and biometric identities. These complementary dimensions are universally exploited by mainstream diffusion-based manipulation methods. FaceArmor then employs an attention-based soft weighting scheme to generate effective adversarial perturbations across these heterogeneous features. By extracting intrinsic attention heatmaps directly from the surrogate models, this strategy smoothly decouples competing gradients and resolves spatial optimization conflicts. Extensive experiments demonstrate that FaceArmor not only preserves visual imperceptibility but also achieves exceptional zero-knowledge transferability, robustly thwarting diverse malicious manipulations across state-of-the-art architectures.

Keywords: Image Protection · Diffusion Models · Adversarial Defense

1 Introduction

The advent of diffusion models has significantly advanced customized image generation [1, 11, 33]. However, this newfound flexibility comes with a dark side: these tools that enable creativity can easily be repurposed for deception. Malicious

* Corresponding authors.

users can manipulate publicly available images, especially facial images, to fabricate hyper-realistic fake content that can mislead audiences, damage reputations, or even fuel disinformation campaigns [30, 39]. This threat landscape is diverse, defined by a range of manipulation paradigms, including *Image Inpainting* to regenerate masked regions [9], *Image Editing* for holistic modifications [14, 28], *Face Mimicry* for subject-specific fine-tuning [17, 34], and *Identity-Preserving Generation* guided by visual prompts [41, 45]. Therefore, how to protect facial photographs from manipulation by such a diverse ecosystem has become a critical research problem.

To mitigate this threat, several methods have sought to embed adversarial perturbations into images as an active defense strategy [35, 36, 38]. These perturbations are crafted to disrupt the generation pipeline of diffusion models, preventing them from editing faces correctly. However, existing methods are overwhelmingly characterized by a restrictive white-box assumption. They typically target specific internal components of a known model, such as the variational autoencoder (VAE) [7, 35] or the denoising U-Net [26, 43]. Since different manipulation scenarios inherently exploit distinct internal pathways of the generative pipeline, a defense overfitted to a single specific module naturally fails when exposed to an alternative attack paradigm. Table 1 empirically demonstrates this consequence, illustrating how current component-specific methods fail to maintain their protective efficacy across the full spectrum of manipulation scenarios.

To bridge this gap, we propose **FaceArmor**, a universal facial image protection framework. Moving away from the conventional paradigm of perturbing isolated internal modules [26, 35], we investigate the commonality of diverse face manipulation models. A core challenge in achieving universal transferability lies in the varying architectures across varying generative models. Optimizing adversarial perturbations against specific architectural components inevitably leads to representation overfitting. However, regardless of their distinct network designs, diffusion-based models share a common operational foundation: they extract and align both physical layout details and high-level conceptual semantics to generate plausible modifications. To exploit this vulnerability, FaceArmor bypasses model-dependent attack surfaces and instead maps the facial image into an intrinsic, architecture-invariant representation feature space.

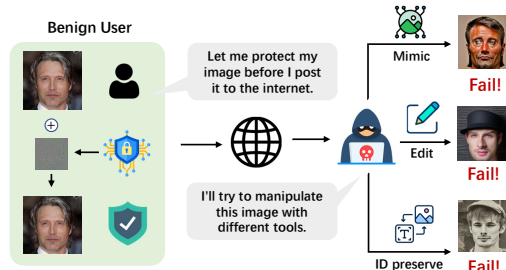


Fig. 1: The FaceArmor protection framework. (Left) A user applies an imperceptible perturbation before sharing images online. (Right) Malignous actors attempt to manipulate the protected image using various diffusion-based tools. FaceArmor renders these diverse, unknown attacks ineffective, causing them to fail.

Table 1: Effectiveness of existing image protection methods. ○ indicates that the method is not applicable or is ineffective in the corresponding scenario. ● indicates that the method is effective in the corresponding scenario.

	PhotoGuard [35]	EditShield [7]	Advdm [26]	Mist [25]	SDST [43]	FaceLock [40]	IDProtector [38]	FaceArmor
Image Inpainting	●	○	●	●	●	●	○	●
Image Editing	○	●	●	○	○	●	○	●
Face Mimicry	○	○	●	●	●	○	○	●
ID-Preserving Generation	○	○	○	○	○	○	●	●

This feature space is strategically designed to disrupt the generative process across two decoupled tiers. The first is the structural tier, which addresses the requirement that models must faithfully reconstruct physical layouts and localized details. By targeting foundational latent distributions and fine-grained reference features within this tier, FaceArmor degrades generative fidelity and induces severe textural collapse. The second is the semantic tier, which targets the models’ reliance on conceptual intent and identity consistency. Here, we decouple the cross-modal text-image bindings while simultaneously obscuring local biometric identities to neutralize prompt-driven edits and identity-preserving mimicry. By comprehensively attacking these intrinsic representation dimensions rather than isolated model components, FaceArmor achieves exceptional zero-knowledge transferability.

However, comprehensively optimizing adversarial perturbations across these diverse tiers induces spatially structured conflicts. Specifically, the gradients derived from these distinct representation dimensions target different semantic regions. For instance, the identity-concealing perturbations originating from the semantic tier are strictly localized to the facial area, whereas the texture-disrupting gradients derived from the structural tier predominantly affect the background and clothing. Naively aggregating these multi-objective gradients leads to severe interference, which not only degrades the overall protective efficacy but also provokes abrupt visual artifacts.

To resolve this spatial entanglement, we introduce an attention-based soft weighting scheme. FaceArmor extracts intrinsic attention heatmaps from the surrogate models during the forward pass. These internal heatmaps serve as adaptive soft weights to smoothly decouple the competing gradients. By integrating this spatial decoupling with PCGrad [46], our framework establishes a smooth transition gradient projection. This synergistic optimization guarantees a seamless gradient distribution, preventing unnatural visual artifacts and ensuring that perturbations across all representation tiers are robustly preserved without mutual interference.

We conduct extensive experiments across four distinct manipulation scenarios and four different diffusion models. The results validate that FaceArmor achieves stronger and more consistent protection than all prior methods, establishing its superior generalizability and robustness in real-world scenarios. In summary, our work makes the following contributions:

- We propose FaceArmor, a universal defense framework for facial image protection. By shifting the defense paradigm to an intrinsic representation feature space composed of structural and semantic tiers, our method effectively neutralizes diverse generative manipulations.
- We introduce an attention-guided spatial decoupling strategy to resolve the inherent gradient conflicts caused by multi-objective adversarial optimization. This approach leverages intrinsic attention heatmaps as adaptive soft weights and integrates a gradient projection mechanism to smoothly disentangle competing spatial updates.
- Extensive evaluations across four manipulation scenarios and four distinct generative architectures demonstrate the superior transferability of FaceArmor. Our framework maintains robust protection during zero-knowledge transfer to recent diffusion transformer models and exhibits strong resilience against common image transformations.

2 Related Work

2.1 Diffusion-Based Image Manipulation

Recent progress in diffusion models has underscored their profound efficacy in image manipulation [18, 21, 23]. These models possess the dual capabilities of performing direct edits on existing images [3, 28] and being fine-tuned to capture the nuanced features or styles of an image collection [29, 34]. Facial photography has emerged as a primary application area, where these techniques facilitate not just background alterations but also modifications to a subject’s expression and attire, extending to the complete synthesis of new identity-consistent portraits. The landscape of face manipulation is principally divided into four paradigms. Image Inpainting [9, 33] regenerates specified regions guided by text while masking preserved areas. Image Editing [13] applies holistic changes, exemplified by end-to-end models like InstructPix2Pix [3] and optimization-based methods like Ledits++ [2] that refine inversion and sampling processes. Face Mimicry achieves personalization by fine-tuning a model on a subject’s likeness—as seen in full model-tuning approaches like DreamBooth [34] or via lightweight, pluggable modules like LoRA [17]. Lastly, Identity-Preserving Generation leverages a source image as a visual prompt [24, 41, 45], typically by fusing identity features with textual guidance to direct the synthesis.

2.2 Defense against Malicious Manipulation

The powerful image manipulation capabilities of diffusion models, while facilitating customized image generation, also introduce significant security and privacy risks [47]. Malicious actors can exploit photos uploaded to social media to generate fabricated images, leading to infringements on personal privacy and reputation. To safeguard individuals from such violations, several image protection methods have been developed [44, 49]. Drawing inspiration from traditional

adversarial attacks [5, 12, 27], these methods add adversarial perturbations to an image to render subsequent manipulations by diffusion models ineffective. Prominent examples include PhotoGuard [35], EditShield [7], AdvDM [26], SDST [43], and more.

These approaches primarily target VAE [22] and the denoiser components of the diffusion model [16]. They employ adversarial attack techniques, such as Projected Gradient Descent (PGD) [27], to disrupt the generation process. This is typically achieved by either maximizing the distance between the latent representations of the protected and original images [7], or by minimizing the distance between the generated output and a benign, nonsensical target image [8]. Other methods adopt more specialized strategies, targeting components like attention maps [19], identity-specific information [38], or stylistic features [37] to achieve more focused protective effects in certain domains.

3 Preliminaries

To establish the technical foundation for our method, we first briefly review the core components of Latent Diffusion Models (LDMs) [16]. The synthesis of images in an LDM is achieved through a reverse denoising procedure operating in a low-dimensional latent space. This process involves three critical components:

1. **VAE Encoder ($\mathcal{E}$):** A pre-trained encoder, typically from a VAE, that maps an image x from the pixel domain to a compressed latent representation $z_0 = \mathcal{E}(x)$.
2. **Denoising U-Net (ϵ_θ):** This is the core generative network. During the reverse process, at each discrete timestep t , ϵ_θ is trained to predict the noise ϵ that was added to a latent z_t . This prediction is critically conditioned on both the timestep t and a conditioning embedding c , formulated as $\epsilon_\theta(z_t, t, c)$.
3. **VAE Decoder (D):** After the iterative denoising process yields a clean latent representation $\hat{z}_0$, the decoder maps this representation from the latent space back to the pixel domain to produce the final image $\hat{x}_0 = D(\hat{z}_0)$.

Prior defenses [25, 35] directly intervene in the functionality of the VAE encoder $\mathcal{E}$ or the denoising network ϵ_θ . However, such component-specific strategies result in poor generalizability to unseen architectures or manipulation paradigms that exploit alternative components.

4 Methodology

In this work, we propose FaceArmor, a universal and architecture-agnostic face protection framework. Given a clean facial image x , our objective is to craft an imperceptible adversarial perturbation δ such that the protected image $x' = x + \delta$ remains robust against diverse and potentially unknown manipulation paradigms. FaceArmor conceptualizes the protective mechanism as a representation feature space. Specifically, we optimize the perturbation across two fundamental representation tiers: structural tier and conceptual tier. Furthermore,

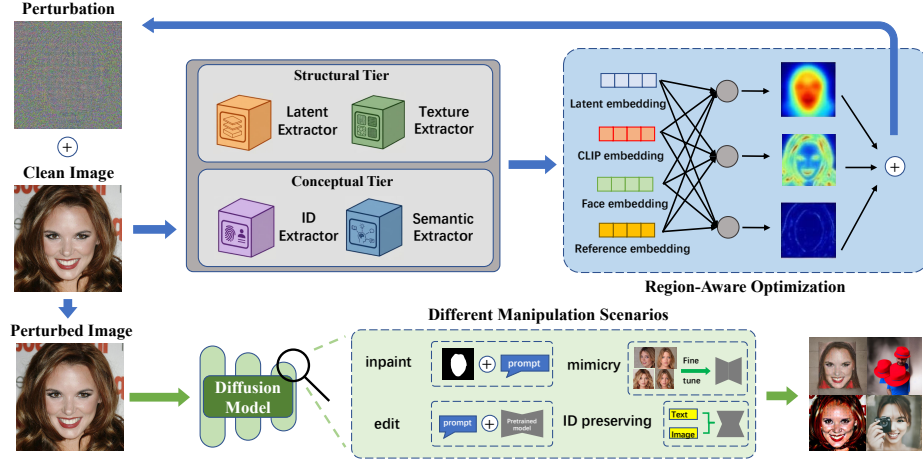


Fig. 2: The overall flow of FaceArmor. Blue arrows represent the optimization process for the adversarial perturbation. We optimize with the intrinsic features of the image. This optimization incorporates attention maps to yield the most effective adversarial perturbation. Green arrows represent the image generation process. When the adversarially perturbed image is used as input, malicious manipulations targeting the facial image are rendered ineffective across all tested image manipulation tasks.

to resolve the spatially structured conflicts inherent in this multi-objective optimization, we introduce a self-guided decoupling scheme. By leveraging intrinsic spatial heatmaps, this scheme fosters a gradient manifold to decouple competing objectives. The overall pipeline is illustrated in Figure 2.

4.1 Intrinsic Feature Space of Facial Images

Our methodology stems from the observation that diffusion-based manipulation networks, irrespective of their architectural designs or specific attack surfaces, universally rely on extracting and aligning an image’s structural and semantic characteristics. Instead of attacking model-specific internal modules, we propose to construct a semantically-decoupled defense manifold upon the image representation space. Specifically, we decouple the facial features into two fundamental tiers. The first tier encompasses structural regulators designed to disrupt foundational structures and fine-grained textures. The second tier consists of conceptual regulators aimed at confounding semantics and biometric identities. By comprehensively optimizing across these tiers, we ensure profound generalizability against diverse and unknown manipulation paradigms.

Structural Tier. This tier disrupts the network’s ability to faithfully reconstruct both the global macroscopic layout and the local textures of the input image.

- **Global Latent Representation.** The denoising process of latent diffusion models operates strictly within a compressed latent space, which governs

the structure of the generation. To establish a universal defense foundation, we optimize against an ensemble of VAE encoders $E = \{E_i\}_{i=1}^N$ sourced from two distinct generative architectures [31, 33]. We maximize the average Euclidean distance to corrupt the fundamental reconstruction pathway:

$$L_E(x, x + \delta) = \frac{1}{N} \sum_{i=1}^N \|E_i(x) - E_i(x + \delta)\|_2^2. \quad (1)$$

- **Fine-Grained Textural Representation.** To further degrade generative fidelity at the local structural level, we target the dense spatial features extracted by a pre-trained ReferenceNet R [6]. Perturbing these localized signals interferes with the network’s ability to maintain physical consistency, thereby inducing severe textural artifacts in the manipulated output:

$$L_R(x, x + \delta) = \|R(x) - R(x + \delta)\|_2^2. \quad (2)$$

Conceptual Tier. This tier focuses on breaking the conceptual alignment and identity consistency during the synthesis process.

- **Semantic Representation.** Prompt-guided manipulations heavily rely on the cross-modal alignment facilitated by vision-language models. To neutralize this vector, we decouple the correspondence between the textual description and the visual input using a pre-trained CLIP image encoder C . We introduce an untargeted objective to maximize the feature distance, effectively stripping the original semantic meaning:

$$L_C(x, x + \delta) = \|C(x) - C(x + \delta)\|_2^2. \quad (3)$$

Concurrently, a targeted objective actively steers the feature vector toward an unrelated target concept t_{target} and away from the original concept t_{orig} , enforcing a controlled semantic misalignment:

$$L_C^t = \frac{(C(x) - C(x + \delta)) \cdot (C(t_{orig}) - C(t_{target}))}{|C(x) - C(x + \delta)| \cdot |C(t_{orig}) - C(t_{target})|}. \quad (4)$$

- **Biometric Identity.** As the ultimate goal is to safeguard personal privacy, we explicitly target the biometric features extracted by ArcFace model F . We force the perturbed image to deviate substantially from the source identity. This ensures that any successful generation strictly yields an identity-incongruent portrait:

$$L_F(x, x + \delta) = \|F(x) - F(x + \delta)\|_2^2. \quad (5)$$

4.2 Self-Guided Spatial Decoupling

While optimizing perturbations across the aforementioned representation tiers yields a comprehensive defense, simultaneously aggregating these disparate objectives inherently induces spatially structured conflicts. For example, the gradient of the identity loss L_F is meaningful exclusively within the facial region,

whereas the gradient of the textural loss L_R is most critical in non-facial regions. Naively aggregating these multi-objective gradients causes severe optimization collisions, particularly along the semantic boundaries separating these regions. This spatial interference degrades the overall protective efficacy and leaves unnatural visual traces on the protected image.

To resolve this spatial entanglement, we introduce an attention-based soft weighting scheme. Instead of relying on external priors, FaceArmor directly extracts intrinsic continuous attention heatmaps generated by the surrogate extractors during the forward pass. These heatmaps serve as adaptive soft weights to smoothly decouple the competing objectives:

- **Facial Weight (W_{face}):** We extract the spatial attention map from the last convolutional layer of the ArcFace network while computing the identity embedding. This weight inherently highlights the core biometric regions essential for constraining L_F .
- **Semantic Weights (W_{fg} and W_{bg}):** We leverage the CLIP vision encoder to compute the Grad-CAM map conditioned on a target text prompt (e.g., “a person”). The normalized activation serves as the foreground weight W_{fg} for the targeted semantic loss, while the untargeted background loss is naturally guided by the complement weight, defined as $W_{bg} = \mathbf{1} - W_{fg}$.
- **Reference Weight (W_{ref}):** The ReferenceNet loss targets fine-grained textural and structural details. To prevent this perturbation from corrupting the facial identity, we invert the facial attention to focus strictly on non-facial regions (i.e., background and clothing), defined as $W_{ref} = \mathbf{1} - W_{face}$.
- **Global Weight (W_{latent}):** As the foundational latent embedding loss targets the global representation of the image, it is applied uniformly with $W_{latent} = \mathbf{1}$.

Smooth-Transition Gradient Projection. With the soft weights established, we spatially decouple the optimization by applying a Hadamard product between the normalized weights and their corresponding gradients. For each objective $j \in \{F, C, R, E\}$, the spatially-weighted gradient is computed as $\hat{g}_j = W_j \odot \nabla_{\delta} L_j(x, x + \delta)$.

Although this soft-masking mechanism smoothly distributes the update directions, overlapping transition regions along semantic boundaries still exhibit intersecting and conflicting gradients. To resolve these residual conflicts, we integrate the PCGrad algorithm to establish a synergistic smooth-transition projection. For the comprehensive set of weighted gradients $G = \{\hat{g}_F, \hat{g}_C, \hat{g}_R, \hat{g}_E\}$, if a conflict is detected (indicated by a negative inner product $\hat{g}_i \cdot \hat{g}_j < 0$), we project $\hat{g}_i$ onto the normal plane of $\hat{g}_j$. Critically, our continuous soft weights create an *overlapping gradient buffer zone* that prevents abrupt spatial cutoffs, fundamentally enhancing PCGrad’s efficacy and guaranteeing a seamless perturbation distribution without unnatural artifacts.

Finally, to guarantee that the perturbation remains robust against common image processing operations in real-world scenarios, we incorporate Expectation over Transformation (EOT). We define a transformation set T comprising Gaussian blur, JPEG compression, Gaussian noise, and random resizing. The overall

update gradient g_t^{update} at iteration t seamlessly aggregates these mechanisms:

$$g_t^{update} = \mathbb{E}_{t \sim T} [\text{PCGrad}(\{W_j \odot \nabla_{\delta} L_j(t(x), t(x + \delta_t))\}_{j \in \{F, C, R, E\}})] . \quad (6)$$

The adversarial perturbation is subsequently updated using the sign of this aggregated gradient and projected back into the L_{∞} -norm ball of radius η :

$$\delta_{t+1} = \Pi_{\eta}(\delta_t + \alpha \cdot \text{sign}(g_t^{update})) . \quad (7)$$

5 Experiments

5.1 Experimental Settings

Models. We evaluate FaceArmor across four prevalent manipulation paradigms: Image Inpainting, Image Editing, Face Mimicry, and Identity Preservation. To demonstrate architecture-agnostic robustness, our evaluation encompasses four generative models. We utilize the widely adopted U-Net-based models, Stable Diffusion (SD) v1.5 [33] and SD XL [31], as our primary testbeds. Specifically, we employ the base model for inpainting, InstructPix2Pix [3] for editing, Dream-Booth [34] for mimicry, and InstantID [41] for identity preservation. Furthermore, we incorporate the recent DiT-based architectures, SD 3 [10] and Flux [23], specifically for zero-shot transfer tests on the Image Inpainting task.

Baselines. We compare FaceArmor against seven defense methods: PhotoGuard [35], EditShield [7], AdvDM [26], Mist [25], SDST [43], FaceLock [40], and IDProtector [38]. Crucially, to simulate a realistic zero-knowledge transfer scenario, all baseline defenses are crafted targeting the widely used SD v1.5. The generated adversarial images are subsequently transferred directly to the architecturally similar SD XL, as well as the architecturally distinct SD 3 and Flux.

Datasets. Our evaluation utilizes a dataset of 800 images, comprising 100 distinct identities sampled from CelebA-HQ [20] and 100 from VGG-Face [4], with four images per identity. For each source image, we generate outputs using 10 different prompts, covering a range of modifications such as background replacement, facial feature alterations, accessory adjustments, and stylistic shifts.

Metrics. We assess performance using generation quality metrics (FID [15], LPIPS [48]) and source similarity metrics (SSIM [42], CLIP Score [32], FR Score [40]). Following established protocols [26], all methods use a perturbation budget of $\epsilon = 8/255$ with $N = 100$ iterations.

5.2 Evaluation Results

We first qualitatively compare FaceArmor with several representative baselines across four manipulation scenarios using SD v1.5. As illustrated in Figure 3, existing defenses offer limited or inconsistent protection. Methods such as Mist and SDST largely degrade the face through blurring, while others (e.g., FaceLock, IDProtector) either fail to block manipulations or introduce severe artifacts. In

Table 2: Quantitative evaluation across all four manipulation scenarios. Results are the mean performance on the combined CelebA-HQ and VGG-Face2 datasets. $\uparrow$ indicates higher is better, $\downarrow$ indicates lower is better. Best results are **bold**, second-best are underlined.

Method	Stable Diffusion 1.5					Stable Diffusion XL				
	FID $\uparrow$	LPIPS $\uparrow$	SSIM $\downarrow$	CLIP $\downarrow$	FR $\downarrow$	FID $\uparrow$	LPIPS $\uparrow$	SSIM $\downarrow$	CLIP $\downarrow$	FR $\downarrow$
(a) Image Inpainting										
PhotoGuard	117.41	0.3721	0.5754	0.8548	0.8784	106.39	0.3416	0.6013	0.8772	0.8960
EditShield	91.48	0.2384	0.6444	0.9388	0.9461	102.23	0.2049	0.6935	0.9458	0.9557
AdvDM	140.89	0.3733	<u>0.5439</u>	0.8689	0.8988	<u>127.30</u>	0.3483	<u>0.5691</u>	0.8857	0.9059
Mist	118.21	0.3663	0.5806	0.8524	0.8754	108.00	0.3442	0.6075	0.8792	0.8941
SDST	116.68	<u>0.3792</u>	0.5769	<u>0.8497</u>	0.8726	105.49	<u>0.3509</u>	0.5989	<u>0.8708</u>	0.8903
FaceLock	119.04	0.3512	0.5492	0.8866	<u>0.6752</u>	108.83	0.3274	0.5712	0.9045	<u>0.7109</u>
IDProtector	85.79	0.2606	0.6263	0.9377	0.9277	75.63	0.2302	0.6530	0.9429	0.9353
FaceArmor	<u>139.15</u>	0.4772	0.4385	0.8177	0.6162	129.30	0.4562	0.4617	0.8338	0.6355
(b) Image Editing										
PhotoGuard	93.88	0.4101	0.5159	0.7999	0.5705	81.58	0.3805	0.5460	0.8252	0.6011
EditShield	85.37	0.4277	0.4409	0.8401	0.6447	72.45	0.3925	0.4817	0.8580	0.6908
AdvDM	131.59	<u>0.5235</u>	<u>0.3748</u>	0.6187	0.4950	119.00	<u>0.4876</u>	<u>0.4002</u>	<u>0.6475</u>	0.5203
Mist	88.40	0.3973	0.5235	0.8204	0.7123	79.20	0.3700	0.5480	0.8375	0.7325
SDST	95.36	0.4408	0.4845	0.7609	0.5732	84.06	0.4107	0.5106	0.7876	0.6040
FaceLock	<u>132.80</u>	0.4932	0.3968	0.6635	0.3002	<u>120.90</u>	0.4620	0.4225	0.6855	0.3302
IDProtector	75.46	0.3924	0.4575	0.8599	0.5605	66.20	0.3625	0.4870	0.8765	0.5906
FaceArmor	163.43	0.5727	0.3121	<u>0.6247</u>	<u>0.3425</u>	150.10	0.5466	0.3355	0.6380	<u>0.3610</u>
(c) Face Mimicry										
PhotoGuard	127.77	0.5626	0.3205	0.7048	0.4668	117.63	0.5305	0.3512	0.7255	0.4978
EditShield	142.44	0.5986	0.2396	0.7160	0.3764	127.52	0.5601	0.2752	0.7347	0.4196
AdvDM	136.91	<u>0.6338</u>	0.2147	0.6751	<u>0.3716</u>	<u>126.44</u>	<u>0.6043</u>	<u>0.2650</u>	0.6938	<u>0.4061</u>
Mist	167.09	0.5933	0.2762	0.6721	0.3956	152.55	0.5605	0.3029	<u>0.6925</u>	0.4275
SDST	115.29	0.5706	0.3151	0.7140	0.4729	105.60	0.5429	0.3492	0.7317	0.5063
FaceLock	125.20	0.5828	0.2406	0.7194	0.3975	115.40	0.5561	0.2774	0.7397	0.4225
IDProtector	112.36	0.5946	0.3330	0.6996	0.4461	102.58	0.5648	0.3649	0.7132	0.4763
FaceArmor	<u>160.70</u>	0.6634	<u>0.2252</u>	<u>0.6745</u>	0.3657	<u>149.65</u>	0.6419	0.2484	0.6887	0.3857
(d) Identity Preservation Generation										
PhotoGuard	172.26	<u>0.5431</u>	0.3509	0.7605	0.8151	<u>158.77</u>	0.5153	0.3848	0.7729	0.8387
EditShield	167.51	0.4978	0.4053	0.7728	0.8356	153.63	0.4721	0.4396	0.7911	0.8480
AdvDM	162.19	0.5427	0.3521	<u>0.7517</u>	0.7940	148.65	0.5175	0.3821	<u>0.7759</u>	0.8119
Mist	157.39	0.5013	0.4063	0.7748	0.8136	144.80	0.4702	0.4680	0.7951	0.8294
SDST	<u>170.04</u>	0.5411	0.3547	0.7446	0.8208	156.46	<u>0.5233</u>	0.3780	0.7665	0.8363
FaceLock	159.19	0.4893	0.4147	0.7783	0.8083	145.55	0.4650	0.4403	0.7950	0.8200
IDProtector	161.05	0.5344	<u>0.3502</u>	0.7591	<u>0.6006</u>	147.62	0.5125	<u>0.3732</u>	0.7765	<u>0.6257</u>
FaceArmor	168.94	0.5533	0.3484	0.7647	0.5528	159.57	0.5336	0.3786	0.7834	0.5780

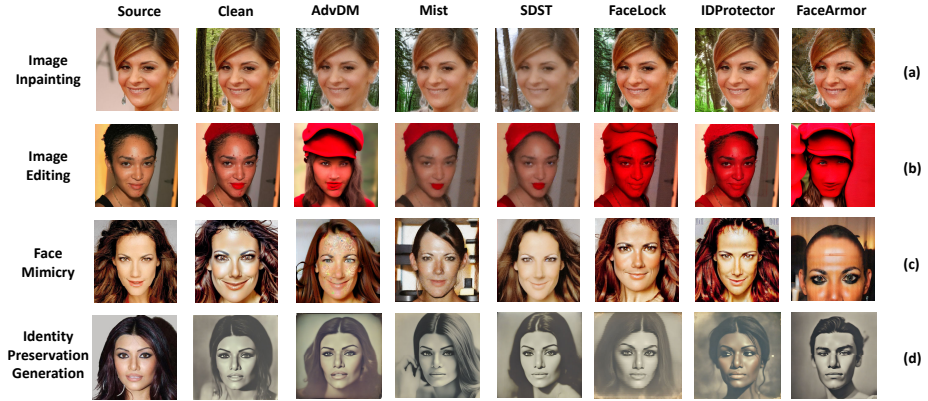


Fig. 3: Qualitative results of different defense methods. Four manipulation types are included. (a) Background replacement: “A person in the forest.” (b) Accessory adjustments: “Let the person wear a red hat.” (c) Facial feature alterations: “The face of a [V] person with a smile.” (d) Stylistic shifts: “Analog film photo of a person.”

contrast, FaceArmor consistently prevents successful edits across all scenarios while preserving visual plausibility. Additional qualitative results on other models are provided in the appendix.

We further conduct comprehensive quantitative evaluations, with results summarized in Table 2. FaceArmor consistently achieves the best or second-best performance across nearly all metrics and scenarios. Although certain baselines perform adequately on specific tasks (e.g., FaceLock on editing, IDProtector on identity preservation), their defensive efficacy drops precipitously when applied to other manipulation types. FaceArmor is the only framework that maintains robust, universal protection across the full spectrum of manipulations.

The results on SD XL additionally underscore FaceArmor’s significant advantage in cross-architecture transfer conditions. Despite SD XL sharing a partial architectural lineage with SD v1.5, baseline defenses exhibit weak transferability, resulting in diminished protection. FaceArmor, however, demonstrates a substantially larger performance margin. This compellingly validates our core hypothesis: systematically disrupting the intrinsic, semantically-decoupled feature space of an image yields far more reliable generalization than targeting model-specific internal representations.

5.3 Transferability to DiT-based Models

To demonstrate the effectiveness of FaceArmor against fundamentally different generative architectures, we evaluate its performance in a zero-knowledge transfer scenario. The adversarial images generated in our main experiments are directly transferred and evaluated on the Image Inpainting task using SD 3 and Flux, both based on DiT.

Table 3: Robustness evaluation against common transformations.

Method	No Transform		JPEG (QF=75)		Gaussian Noise ($\sigma = 0.05$)		Rotation ($\pm 10^\circ$)		Resizing (0.9x)	
	LPIPS $\uparrow$	SSIM $\downarrow$	LPIPS $\uparrow$	SSIM $\downarrow$	LPIPS $\uparrow$	SSIM $\downarrow$	LPIPS $\uparrow$	SSIM $\downarrow$	LPIPS $\uparrow$	SSIM $\downarrow$
EditShield	0.481	0.404	0.265	0.640	0.290	0.613	0.253	0.672	0.271	0.635
AdvDM	0.513	0.380	0.310	0.595	0.335	0.570	0.291	0.612	0.308	0.601
FaceLock	0.456	0.449	0.240	0.681	0.262	0.650	0.233	0.705	0.251	0.677
FaceArmor	0.537	0.325	0.431	0.449	0.425	0.458	0.402	0.480	0.435	0.445

Table 4: Quantitative evaluation of black-box transferability to the Stable Diffusion 3 model.

Method	FID $\uparrow$	LPIPS $\uparrow$	SSIM $\downarrow$	CLIP $\downarrow$	FR $\downarrow$
PhotoGuard	129.49	0.2562	0.7068	0.8762	0.9692
EditShield	117.50	0.2155	0.7667	0.8879	0.9885
Advdm	143.88	0.2830	0.7048	0.8480	0.9633
Mist	134.32	0.2561	0.6979	0.8729	0.9714
SDST	132.24	0.2550	0.7105	0.8724	0.9699
FaceLock	115.68	0.2424	0.7445	0.8794	0.9628
IDProtector	127.38	0.2466	0.7257	0.8785	0.9662
FaceArmor	156.44	0.3131	0.6839	0.7826	0.9479

Table 5: Quantitative evaluation of black-box transferability to the Flux model.

Method	FID $\uparrow$	LPIPS $\uparrow$	SSIM $\downarrow$	CLIP $\downarrow$	FR $\downarrow$
PhotoGuard	92.53	0.2234	0.7371	0.9141	0.9678
EditShield	83.22	0.1908	0.8011	0.9309	0.9879
Advdm	105.92	0.2824	0.7270	0.8946	0.9535
Mist	108.17	0.2350	0.7370	0.9148	0.9727
SDST	91.87	0.2280	0.7436	0.9170	0.9703
FaceLock	84.59	0.2196	0.7774	0.9214	0.9630
IDProtector	83.76	0.2276	0.7519	0.9216	0.9656
FaceArmor	115.23	0.2906	0.7218	0.7332	0.9464

The quantitative results presented in Table 4 and 5 reveal a critical vulnerability in existing defenses: while baseline methods exhibit competitive performance when the target model’s architecture is known, their protective efficacy collapses under cross-architecture transfer. In contrast, FaceArmor maintains robust protective performance, achieving the best results across all five metrics. This demonstrates that FaceArmor achieves profound architecture-agnostic generalization that prior methods lack.

5.4 Robustness to Image Transformations.

A practical defense must be robust to common image processing operations. We evaluate robustness by applying four transformations to the protected images before feeding them to the manipulation models: JPEG (Quality=75), Gaussian Noise ($\sigma = 0.05$), Rotation (random $\pm 10^\circ$), and Resizing ($0.9\times$ scale factor).

Table 3 reports the mean LPIPS and SSIM across all four manipulation tasks, baseline methods are brittle. A simple JPEG compression causes their protection to collapse, indicated by a sharp decrease in LPIPS and an increase in SSIM. In sharp contrast, FaceArmor’s LPIPS and SSIM scores remain stable across all transformations. This is a direct result of our EOT training strategy, which explicitly optimizes the perturbation to be invariant to these operations.

5.5 Ablation Study

Effectiveness of Feature-Level Objectives. To validate the effectiveness of four features, we conducted a component-wise ablation study, with aggregated results reported in Table 6. The $L_{\mathcal{E}}$ component provides a foundational

perturbation in the latent space. Integrating the semantic loss (L_C) disrupts high-level text-image correspondence, while the reference loss (L_R) degrades fine-grained textural fidelity. The biometric loss (L_F) specifically targets and obscures identity-centric features. The full FaceArmor model, which synergizes all four objectives, achieves the best comprehensive performance, confirming the necessity of our full-stack optimization strategy.

Effectiveness of Optimization

Strategies. To further investigate the effectiveness of our proposed mechanisms, we conducted a series of ablation studies. In these experiments, we configured three distinct defense settings: (1) Optimization using only the baseline loss, L_E . (2) Optimization using all losses from FaceArmor but without our region-aware optimization strategy. (3) The full FaceArmor model, using all losses

with the region-aware optimization strategy. We tested these settings across all scenarios and recorded the mean value for each metric, with the results presented in Figure 4. The results show that the performance of all three settings is broadly comparable on the FID metric. However, for all other metrics, the introduction of each new mechanism leads to improved performance over the preceding setting. This further demonstrates the effectiveness of the components proposed in our method.

Impact of Perturbation Budget and Iterations. We further investigated the impact of the perturbation budget η and the number of PGD iterations N on the protection effectiveness. The experimental results are presented in Figure 5. As the budget increases, the protective effectiveness of FaceArmor gradually improves. At a budget of 8/255, the method already achieves a strong protective effect. As this value is also commonly used in prior work, we selected a budget of 8/255 for our main experiments.

Similarly, increasing the number of iterations also enhances protection, with the most significant improvement occurring when the count rises from 10 to 50. When the number exceeds 100, the performance gains become marginal. Therefore, 100 iterations represents a suitable trade-off between computational cost and performance.

5.6 Computational Efficiency Analysis

To evaluate the computational overhead of FaceArmor, we measure the average optimization time per image, comparing FaceArmor with representative baseline defenses on a single NVIDIA A800 GPU.

As reported in Table 7, FaceArmor incurs a higher computational overhead compared to baseline defenses. This increased processing time is an inherent consequence of its comprehensive multi-objective optimization. While baseline meth-

Table 6: Ablation study on feature loss components. We report mean metrics across all four tasks.

Components					Metrics				
L_E	L_C	L_F	L_R		FID $\uparrow$	LPIPS $\uparrow$	SSIM $\downarrow$	CLIP $\downarrow$	FR $\downarrow$
✓					132.15	0.5215	0.3830	0.7604	0.7615
✓	✓				145.40	0.5329	0.3523	<u>0.6928</u>	0.7540
✓		✓			138.63	0.5086	0.3614	0.7533	0.4308
✓	✓	✓			162.01	<u>0.5510</u>	<u>0.3405</u>	0.7010	0.4796
✓	✓	✓	✓		<u>158.32</u>	0.5781	0.3205	0.6846	<u>0.4522</u>

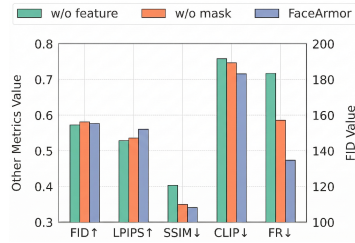


Fig. 4: Ablation study on the proposed mechanisms.

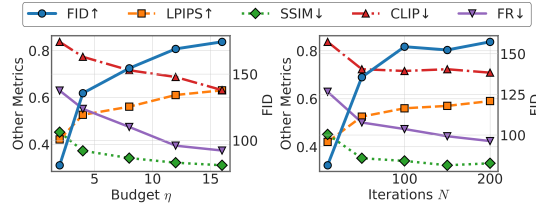


Fig. 5: Ablation study of budget and iterations.

ods achieve faster execution by overfitting to a single, model-specific attack surface, this narrow focus renders their protection highly brittle against unseen architectures and manipulation paradigms. In contrast, FaceArmor achieves superior performance across different manipulation paradigms by optimizing various representations. Traversing these representation tiers requires additional computation, yet it is the exact mechanism that guarantees our architecture-agnostic generalization.

Regarding real-world deployment, applying this protective perturbation is a one-time, offline preprocessing step executed before a user shares an image online. Given this application context, ultra-low latency or real-time processing is not a strict prerequisite. The additional computational cost represents an acceptable trade-off for achieving robust protection against a diverse ecosystem of malicious manipulations.

Table 7: Computational efficiency analysis with the average time per image.

Method	Avg. Time per Image (s) ↓
PhotoGuard	28.5
FaceLock	49.1
AdvDM	55.8
Mist	74.2
FaceArmor	90.2

6 Conclusion

In this paper, we propose FaceArmor, a protection method for facial images applicable to various diffusion-based face manipulation scenarios. By incorporating feature extraction models, we introduce perturbations at the feature level, preventing the perturbation from being over-specialized for a single task. Furthermore, by segmenting the facial image and applying region-aware optimization strategy, we mitigate the conflicts between various optimization objectives. Through extensive experiments, we demonstrate that FaceArmor remains effective across diverse face manipulation scenarios, and its comprehensive performance significantly surpasses previous methods.

Acknowledgements

This work was partly supported by the New Generation Artificial Intelligence-National Science and Technology Major Project under No. 2025ZD0123503, NSFC under No. U2441239 and U24A20336 and No. 62502433, the China Post-doctoral Science Foundation under No. 2024M762829, 2025M781523 and 2025M781522, Zhejiang Key Laboratory of Decision Intelligence under No. 2025E10006, Zhejiang Provincial Natural Science Foundation Exploration of China under No. LMS26F020003, State Key Laboratory of Cryptography and Digital Economy Security under No. KFYB2504, the Zhejiang Provincial Natural Science Foundation under No. LD24F020002, and the "Pioneer and Leading Goose" R&D Program of Zhejiang under No. 2025C02033 and 2025C01082.

References

1. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. *Computer Science*. **2**(3), 8 (2023)
2. Brack, M., Friedrich, F., Kornmeier, K., Tsaban, L., Schramowski, P., Kersting, K., Passos, A.: Ledits++: Limitless image editing using text-to-image models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 8861–8870 (2024)
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 18392–18402 (2023)
4. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. pp. 67–74. IEEE (2018)
5. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: *2017 IEEE Symposium on Security and Privacy (SP)*. pp. 39–57. IEEE (2017)
6. Chang, D., Shi, Y., Gao, Q., Xu, H., Fu, J., Song, G., Yan, Q., Zhu, Y., Yang, X., Soleymani, M.: Magicpose: realistic human poses and facial expressions retargeting with identity-aware diffusion. In: *Proceedings of the 41st International Conference on Machine Learning*. pp. 6263–6285 (2024)
7. Chen, R., Jin, H., Liu, Y., Chen, J., Wang, H., Sun, L.: Editshield: Protecting unauthorized image editing by instruction-guided diffusion models. In: *European Conference on Computer Vision*. pp. 126–142. Springer (2024)
8. Choi, J.S., Lee, K., Jeong, J., Xie, S., Shin, J., Lee, K.: Diffusionguard: A robust defense against malicious diffusion-based image editing. *arXiv preprint arXiv:2410.05694* (2024)
9. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. In: *The Eleventh International Conference on Learning Representations* (2023)
10. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: *Forty-first international conference on machine learning* (2024)

11. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
12. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
13. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
14. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. *ICLR* **1**(2), 3 (2022)
18. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Cao, L., Chen, S.: Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
19. Jeon, J., Kim, W.J., Ha, S., Son, S., Yoon, S.e.: Advpaint: Protecting images from inpainting manipulation via adversarial attention disruption. arXiv preprint arXiv:2503.10081 (2025)
20. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
21. Kavar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6007–6017 (2023)
22. Kingma, D.P., Welling, M., et al.: Auto-encoding variational bayes (2013)
23. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontekst: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
24. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked id embedding. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 8640–8650 (2024)
25. Liang, C., Wu, X.: Mist: Towards improved adversarial examples for diffusion models. arXiv preprint arXiv:2305.12683 (2023)
26. Liang, C., Wu, X., Hua, Y., Zhang, J., Xue, Y., Song, T., Xue, Z., Ma, R., Guan, H.: Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In: *Proceedings of the 40th International Conference on Machine Learning* (2024)
27. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
28. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)

29. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6038–6047 (2023)
30. News, B.: Inside the deepfake porn crisis engulfing korean schools. BBC News (2024)
31. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
34. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dream-booth: 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)
35. Salman, H., Khaddaj, A., Leclerc, G., Ilyas, A., Madry, A.: Raising the cost of malicious ai-powered image editing. arXiv preprint arXiv:2302.06588 (2023)
36. Shamshad, F., Naseer, M., Nandakumar, K.: Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20595–20605 (2023)
37. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 2187–2204 (2023)
38. Song, Y., Yang, P., Ci, H., Shou, M.Z.: Idprotector: An adversarial noise encoder to protect against id-preserving image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3019–3028 (2025)
39. Times, T.N.Y.: Taylor swift ai fake images controversy. The New York Times (2024)
40. Wang, H., Zhang, Y., Bai, R., Zhao, Y., Liu, S., Tu, Z.: Edit away and my face will not stay: Personal biometric defense against malicious generative editing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23806–23816 (2025)
41. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A., Li, H., Tang, X., Hu, Y.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
42. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
43. Xue, H., Liang, C., Wu, X., Chen, Y.: Toward effective protection against diffusion-based mimicry through score distillation. In: The Twelfth International Conference on Learning Representations, (2024)
44. Yang, W., Cao, J., Duan, J., He, R.: 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)

- 45. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
- 46. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. *Advances in neural information processing systems* **33**, 5824–5836 (2020)
- 47. Zhang, C., Hu, M., Li, W., Wang, L.: Adversarial attacks and defenses on text-to-image diffusion models: A survey. *Information Fusion* **114**, 102701 (2025)
- 48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 586–595 (2018)
- 49. Zhu, H., Pang, S., Lu, Z., Zhou, Y., Xue, M.: Gap-diff: protecting jpeg-compressed images from diffusion-based facial customization. In: *NDSS* (2025)