Edit Away and My Face Will not Stay: Personal Biometric Defense against Malicious Generative Editing

Hanhui Wang[1*], Yihua Zhang[2*], Ruizheng Bai[3], Yue Zhao[1], Sijia Liu[2†], Zhengzhong Tu[3†]

[1]University of Southern California, [2]Michigan State University, [3]Texas A&M University

* Equal Contribution, † Corresponding Authors

Paper Code

CVPR Nashville JUNE 11-15, 2025

Motivation

Can we design adversarial perturbations that cause edited images to lose their biometric information, making the edited image biometrically unrecognizable and thereby causing the edit to fail?

Contributions

- We present a novel perspective for protecting **personal images** from malicious editing by focusing on **making biometric features unrecognizable** after edits.
- We conduct critical analyses on quantitative evaluation metrics commonly used in image editing, exposing their vulnerabilities and the potential for manipulation to achieve deceptive results.
- We introduce **FaceLock**, which incorporates facial recognition models and feature embedding penalties to effectively protect against diffusion-based image editing.
- Extensive experiments demonstrate that **FaceLock** effectively alters human facial features against various editing prompts, achieving superior defense performance compared to baselines.

Editing Prompt: '*Let the person wear a police suit*'

Figure 1. Illustration of the two requirements of image editing: prompt fidelity and image integrity.

Related Works

[1] Salman Hadi et al. "Raising the Cost of Malicious AI-Powered Image Editing."

[2] Ruoxi Chen et al. "EditShield: Protecting Unauthorized Image Editing by Instruction-guided Diffusion Models." ECCV 2024.

What defines a successful image edit?

- **Prompt fidelity**: Edits should accurately reflect the instructions provided in the prompt. (Fig. 1 (b) and (d))
- **Image integrity**: Other elements in the image should remain intact after editing. (Fig. 1 (b) and (c))

FaceLock: Adversarial Biometrics Erasure

- FaceLock: Making edited images biometrically unrecognizable rather than blocking edits outright via perturbation optimization on facial disruption and feature embedding disparity.
- We formulate an optimization objective that jointly enforces biometric disruption and high-level feature disparity, defined as follows:

$$\boldsymbol{\delta} = \arg\max_{\|\boldsymbol{\delta}\|_\infty \le \epsilon} f_{\mathrm{FR}}(\mathcal{D}(\mathcal{E}(\mathbf{x}+\boldsymbol{\delta})),\mathbf{x}) + \lambda f_{\mathrm{FE}}(\mathcal{D}(\mathcal{E}(\mathbf{x}+\boldsymbol{\delta})),\mathbf{x}),$$

FR: the facial recognition loss, **FE**: feature embedding loss between the input images. $\mathcal{E}/\mathcal{D}$: Encoder/Decoder.

Pitfalls on Existing Evaluation Metrics

- CLIP-based scores overemphasize the presence of elements from the editing instructions, which often leads to prioritizing over-editing. (**Fig. 2**)
- SSIM and PSNR over-rely on differences between the edited image and the undefended source, potentially leading to a false sense of successful defense. (**Fig. 3**)
- We use LPIPS scores as a more robust alternative to SSIM and PSNR for evaluating high-level similarity between edited images.
- We propose to use the facial recognition (FR) similarity score to assess the preservation of biometric identity between the source and edited images.

Editing Prompt: '*Let the person's hair turn pink*'.

Figure 2. CLIP score (CLIP-S) of different editing results. The CLIP score provides a contradictory ranking (III > II > I) compared to the visual quality (I > II > III).

Editing Prompt: '*Let the person wear a hat*'.

Figure 3. SSIM and PSNR scores of different defenses. Defense II (d) receives better scores than Defense I (c) due to greater pixel differences from the unprotected edit (b), despite being less effective.

Experiment Results Highlights

Table 1. Quantitative evaluation on prompt fidelity (CLIP-S, PSNR, SSIM, LPIPS) and image integrity (CILP-I, FR). Arrows indicate whether a higher or lower value is preferred for a successful defense.

	Prompt Fidelity				Image Integrity	
Method	**CLIP-S ↓**	**PSNR ↓**	**SSIM ↓**	**LPIPS ↑**	**CLIP-I ↓**	**FR ↓**
No Defense	0.118±0.037	-	-	-	0.808±0.074	0.833±0.111
PhotoGuard Encoder attack	**0.108**±0.030	**15.44**±2.01	0.612±0.056	0.403±0.071	0.670±0.118	0.590±0.264
EditShield	0.110±0.026	17.74±2.20	0.593±0.072	0.382±0.071	0.677±0.096	0.641±0.231
Untargeted Encoder attack	0.116±0.023	16.74±2.27	**0.589**±0.084	0.371±0.094	0.653±0.090	0.563±0.236
CW L2 attack	0.115±0.031	19.64±2.46	0.701±0.060	0.247±0.062	0.733±0.089	0.725±0.173
VAE attack	0.114±0.034	19.40±1.70	0.715±0.039	0.251±0.060	0.786±0.061	0.846±0.097
FaceLock (ours)	0.114±0.024	17.11±2.36	**0.589**±0.079	**0.436**±0.065	**0.648**±0.089	**0.315**±0.109

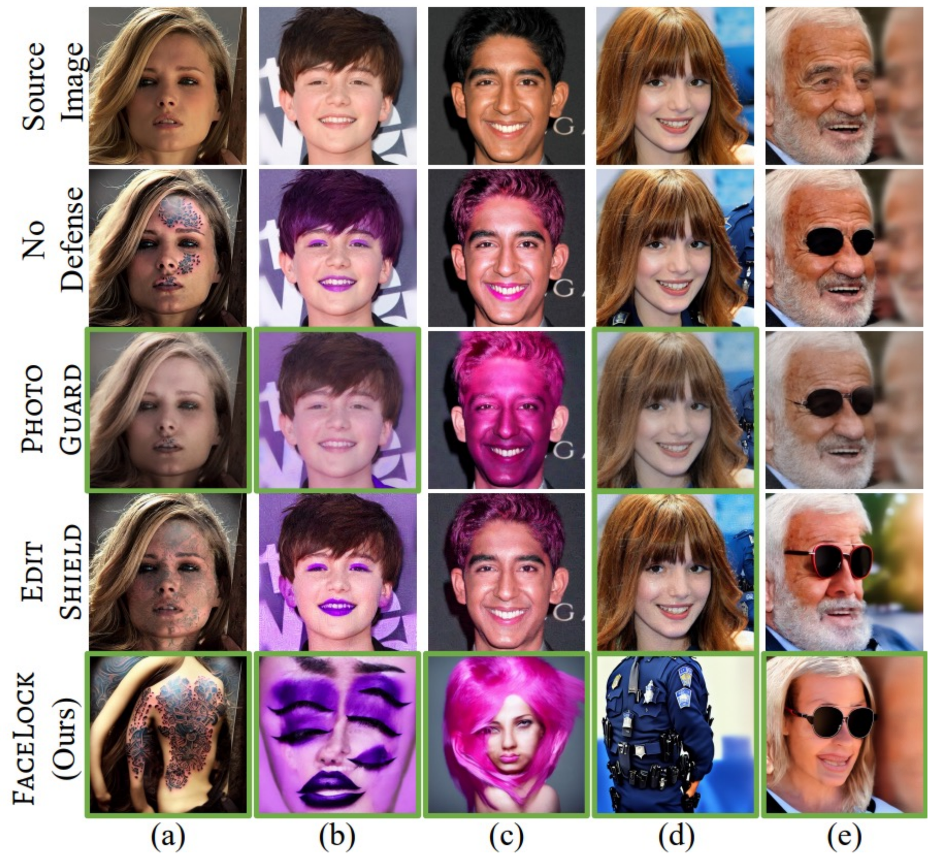

Figure 4. Qualitative results of defensive methods on various editing prompts.

Editing Prompt: '*Set the background in a library*'.

Figure 5. Qualitative results of edits on protected images after applying purification methods like color jitter and DiffPure. Compared to other methods, FaceLock more effectively prevents identity recovery after purification.

Contact: hanhuiwa@usc.edu, {zhan1908, liusiji5}@msu.edu, tzz@tamu.edu