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

Anonymous authors
Paper under double-blind review

Abstract

Recent advances in diffusion-based image editing have enabled highly realistic and accessible manipulation of facial images, raising serious concerns about biometric privacy and malicious misuse. FaceLock, introduced in Edit Away and My Face Will Not Stay: Personal Biometric Defense against Malicious Generative Editing, proposes an optimization-based defense that embeds subtle perturbations into images at publication time to induce identity distortion in downstream generative edits. The method claims prompt-agnostic effectiveness and strong performance across multiple editing scenarios, supported by open-source code. In this paper, we present a systematic reproducibility study of FaceLock that evaluates its technical, quantitative, and qualitative reproducibility. We assess whether the reported results can be obtained using the released codebase, analyze the correspondence between the paper's algorithmic description and its implementation, and document ambiguities that impact reproducibility. We further examine quantitative reproducibility by attempting to recover the reported performance trends and relative ranking against baselines. We, however, were not able to reproduce the originally reported performance trends, and our outputs were generally worse than those presented in the original paper. Beyond that, we expand the qualitative analysis to a broader set of image–prompt pairs and an additional, harder facial dataset to better test generalization behavior. While we obtained some successful outputs, only a small fraction of our qualitative results matched the consistently high quality reported by the authors. Finally, we introduce an extension to the FaceLock method that helps with robustness, and we critically examine the evaluation criteria used to measure defense effectiveness, highlighting limitations of prompt fidelity as a primary metric and arguing for a more explicit consideration of the trade-off between identity protection and preservation of the original image. We provide a link to our (currently anonymous and private) GitHub repository [1].

1 Introduction

In recent years, rapid progress in generative models for image editing, particularly diffusion-based methods, has enabled increasingly realistic transformations through natural language prompts, allowing facial attributes, expressions, and even semantic context to be altered with minimal effort and high visual fidelity (Brooks et al. (2022); Shuai et al. (2024)). While these advances unlock powerful creative applications, they also introduce significant privacy concerns when applied to images of real individuals. The ease with which facial images can be manipulated raises the risk of malicious misuse, including identity alteration, impersonation, non-consensual content creation, and misinformation (Korshunov & Marcel (2018); Tolosana et al. (2020)). Once an image is publicly released, its owner typically has little control over how it may be edited by third parties using powerful generative tools, creating an asymmetry between image ownership and editing capability that motivates the need for defensive mechanisms to protect biometric identity while naturally preserving the ability to share images online.

In response to these concerns, recent work has proposed biometric defense systems that modify images in subtle ways to disrupt downstream generative editing (Chen et al. (2023); Choi et al. (2024); Salman

[1] https://anonymous.4open.science/r/revisiting_facelock-F4D7

et al. (2023)). By embedding carefully designed perturbations at the time of image publication, some approaches aim to maintain visual fidelity of images while preventing unauthorized transformations, while others target complete image distortion, always without relying on platform-level enforcement. Within this paradigm, the paper "Edit Away and My Face Will Not Stay: Personal Biometric Defense against Malicious Generative Editing" (Wang et al. (2024)) introduces *FaceLock*, an optimization-based defense pipeline designed to operate under these principles. The work is particularly notable because, in contrast to much of the prior literature, it frames defense in terms of inducing biometric distortion in the edited output rather than making the protected image difficult to edit or completely deforming it. The authors report strong quantitative and qualitative performance across multiple metrics and editing scenarios, provide an open-source implementation, and evaluate their method on Celeba-HQ (Karras et al. (2017)), a publicly available dataset . Taken together, the methodological novelty, strong empirical claims, and open-source accessibility makes FaceLock a compelling candidate for reproducibility analysis.

Therefore, in this study, we conduct a systematic reproducibility assessment of FaceLock across multiple dimensions. We examine whether the reported results can be obtained using the released codebase, assess the clarity and completeness of the experimental setup, and analyze the correspondence between the paper's algorithmic descriptions and their implementation. We further evaluate quantitative reproducibility by assessing whether the main performance trends and relative ranking of FaceLock against proposed baselines can be recovered. Beyond these considerations, we extend the qualitative analysis by expanding the set of visual examples to better characterize FaceLock's behavior under image-prompt pairs and by incorporating an additional facial dataset. Finally, we introduce an extension to the FaceLock method that helps with robustness over editing seeds, and we critically examine the evaluation criteria used to assess defense effectiveness, arguing for evaluations that more explicitly capture the trade-off between biometric protection and preservation of the original image.

2 Scope of Reproducibility

The goal of our reproducibility study is to systematically evaluate the claims presented in Wang et al. (2024) with respect to their reproducibility across technical, quantitative, and qualitative dimensions. We also propose a new evaluation pipeline that we argue is more fair and effective for assessing biometric defense in generative editing settings. More explicitly, we investigate the following points:

- **Technical reproducibility:** determining whether the results reported by the authors can be obtained by executing the released code and following the documented experimental setup. This includes verifying the clarity and completeness of installation instructions, dataset preparation procedures, and configuration of hyperparameters. We also assess the fidelity of the algorithmic descriptions against the actual implementation, noting ambiguities that may hinder reproducibility and discrepancies that raise the question of whether or not the implemented FaceLock method is what is described in the paper.

- **Quantitative reproducibility:** attempting to recover the performance trends reported in the original paper. We focus on whether FaceLock's claimed overall quality and superiority over baseline defenses persists under reproduction. Our reproduction experiments are conducted on a representative subset of the full evaluation data described by the authors. As a result, our findings emphasize ranking replication rather than exact numeric agreement with published metrics.

- **Qualitative reproducibility and generalization analysis:** expanding the set of visual examples to include a broader range of source image and editing prompt pairs, and testing FaceLock across a new facial dataset that were not part of the original evaluation, namely FaceForensics++ (Rössler et al. (2019)). This allows us to assess how stable the defense behavior is when image content differs from the CelebA-HQ (Karras et al. (2017)) distribution used by the authors.

- **Critique of evaluation criteria:** examining the suitability of existing metrics for indicating and measuring defense effectiveness. We also propose additional measures that help capture the trade-off between biometric protection and preservation of the original image in publication time.

- **Method extension:** Building on the observations from our reproducibility study, particularly the identified shortcomings, we propose an extension to FaceLock that improves robustness to the randomness inherent in diffusion-based editing.

3 Methodology

3.1 Problem Setup

The authors of Wang et al. (2024), from now on referred to as **the Authors**, consider the problem of protecting an individual's biometric identity against malicious generative image editing. Given a facial image released publicly by its owner, the threat model assumes that an adversary may apply arbitrary diffusion-based editing methods with unconstrained natural language prompts. Different from most previous work in the field, the goal of the defense is not to prevent editing itself or completely distort the image, but to ensure that whatever the edited output looks like, it no longer preserves the subject's biometric identity. Importantly, the defense must operate prior to image publication, without access to the adversary's prompts or editing model. Following that, there's an unmeasured expectation that the protected image should remain visually natural and as close as possible to the original image so as not to affect sharing it instead of the original one.

Under this formulation, the defense produces a protected image by adding an optimization-computed subtle perturbation to the original input. The Authors define success in terms of low prompt fidelity and identity distortion in the edited results, which we'll further discuss in later sections. Specifically, after a protected image is edited, the resulting output should not follow what the prompt required and be dissimilar to the original identity in facial recognition embedding space. This framing shifts the objective from edit obstruction to biometric disruption, although they still account for low prompt fidelity, and emphasizes robustness across a wide range of prompts rather than protection against a predefined set of transformations.

3.2 Diffusion Model

All generative editing experiments in this work rely on pretrained diffusion-based image editing models. These models are used both to test a defense pipeline implementation via editing and within some defense pipelines themselves. The Authors use InstructPix2Pix (Brooks et al. (2022)), a model that operates using latent diffusion and is publicly available through the Hugging Face Diffusers library (von Platen et al. (2022)). The released model checkpoint [2] is fully pretrained and is used as-is in both the Authors' and our experiments.

3.3 Defense Models

We evaluate three image-level biometric defense methods: FaceLock, which is the primary focus of our reproducibility study, and two main defenses the Authors used as baselines, PhotoGuard (Choi et al. (2024)) and EditShield (Chen et al. (2023)). All three methods compute image-specific perturbations prior to the image editing process, but they differ substantially in their optimization objectives and their reliance on the underlying generative model.

3.3.1 PhotoGuard

PhotoGuard is a diffusion-based image protection method that generates adversarial perturbations via gradient-based optimization in the latent space of a diffusion model. Unlike approaches relying on implicit guidance, it explicitly differentiates through the generative pipeline by optimizing the input through the VAE encoder (Kingma & Welling (2022)) of Stable Diffusion, back-propagating gradients through the encoding process to manipulate the latent representation. Concretely, PhotoGuard applies projected gradient descent to update the input image by minimizing the distance between its latent representation and a predefined target latent, while constraining the perturbation using an ℓ_∞ budget and spatial masking to

[2] https://huggingface.co/timbrooks/instruct-pix2pix

restrict changes to selected regions and keep pixel values valid. Because the perturbations are tightly coupled to diffusion internals, the method requires white-box access to the VAE encoder.

3.3.2 EditShield

EditShield is an image-level defense against instruction-guided diffusion-based editing that disrupts the image conditioning signal by perturbing the latent representation used to guide the diffusion process. Unlike FaceLock, it does not aim to distort identity, but instead interferes with editing itself by optimizing perturbations so that the VAE latent of the protected image deviates from that of the original image, reducing the effectiveness of downstream edits. Formally, EditShield maximizes a latent inconsistency loss between the VAE-encoded latents of the original and protected images, while preserving visual fidelity through an ℓ_2 image-space regularizer. To improve robustness against common transformations, it further applies Expectation over Transformation (Athalye et al. (2018)), optimizing the latent inconsistency over a distribution of editing conditions while keeping the generation pipeline fixed.

3.3.3 FaceLock

FaceLock is an image-level biometric protection method designed to disrupt identity preservation in diffusion-based image editing models. Given an input image, FaceLock computes a bounded adversarial perturbation by optimizing directly in pixel space under an ℓ_∞ constraint. Rather than attempting to prevent edits from occurring or completely messing them up, the method aims to ensure that any edited output no longer preserves the original subject's biometric identity. The optimization objective combines two terms that balance identity distortion against visual fidelity. Specifically, it includes a face recognition loss that maximizes identity discrepancy between the protected and original images using a pretrained face recognition model, and a perceptual similarity loss based on LPIPS (Zhang et al. (2018)) to preserve visual fidelity.

FaceLock relies on pretrained face recognition and alignment models to define its identity-based optimization objective. Identity embeddings are computed using the ViT-KPRPE architecture (Kim et al. (2024)), trained with the AdaFace loss (Kim et al. (2023)) on the WebFace4M dataset [3] (Zhu et al. (2021)) available on Hugging Face). Prior to identity comparison, facial regions are extracted and aligned using a DFA-MobileNet face alignment model [4] (Kim et al. (2024)). These models are used exclusively during the defense optimization process to guide perturbation generation and are not involved in the downstream image editing stage.

3.3.4 FaceLock Extension: EoT FaceLock

We introduce EoT FaceLock, an extension of FaceLock designed to improve robustness to the stochasticity of diffusion-based editing. While the name is inspired by Expectation over Transformation from Athalye et al., our approach does not apply explicit transformations to the input image. Instead, it targets a failure mode observed in our qualitative study: high sensitivity to the random seed used during editing. This issue is particularly evident in Figure 8, where we attempt to reproduce the qualitative examples from Figure 6 of Wang et al. (2024) and find that FaceLock can succeed or fail depending on seed choice.

The core idea is to optimize the adversarial perturbation under an expectation over diffusion randomness, rather than for a single deterministic editing trajectory. Concretely, we simulate stochastic diffusion behavior during defense optimization by injecting scheduler-consistent noise into the VAE latent of the adversarial image at randomly sampled timesteps from the later denoising stages. We also replace the deterministic use of the latent mean with a (reparameterized) latent sample, introducing additional stochasticity. The defense loss is then computed as the average face-recognition loss over multiple such noisy latent samples, and gradients are taken through this expectation. This encourages the protected image to remain identity-disruptive under different diffusion noise realizations, improving robustness across seeds.

[3] https://huggingface.co/minchul/cvlface_adaface_vit_base_kprpe_webface4m

[4] https://huggingface.co/minchul/cvlface_DFA_mobilenet

Figure 1: EoT FaceLock (Expectation over Diffusion Randomness).

Inputs: original image x; adversarial image at ith iteration x^i; VAE encoder E, decoder D; scheduler noise function of diffusion editing pipeline $\mathcal{N}(\cdot, t)$; face recognition model f_{FR}; face recognition loss $\mathcal{L}_{\mathrm{FR}}$; EoT samples K; late diffusion timestep range $\mathcal{T}$.
Output: protected image $\tilde{x}$.

Procedure:

```
1. Set $\ell \leftarrow 0$
2. For $k = 1, \dots, K$:
    (a) Sample timestep $t \sim \mathcal{T}$
    (b) Encode latent distribution $(\mu, \sigma) \leftarrow E(x^i)$
    (c) Sample (reparameterized) latent $z \leftarrow \mu + \sigma \odot \eta$
    (d) Add scheduler-consistent noise $\hat{z} \leftarrow \mathcal{N}(z, t)$
    (e) Decode $\hat{x}^i \leftarrow D(\hat{z})$
    (f) Accumulate $\ell \leftarrow \ell + \mathcal{L}_{\mathrm{FR}}\big(f_{\mathrm{FR}}(\hat{x}^i), f_{\mathrm{FR}}(x)\big)$
3. Average $\ell \leftarrow \ell / K$
```

3.4 Datasets and Processing

The Authors primarily evaluate FaceLock on CelebA-HQ (Karras et al. (2017)), providing a download link and reporting experiments on 2000 images, though the paper does not specify the split used. Since the CelebA-HQ validation set contains exactly 2000 images (compared to 28,000 training images), we infer that evaluation was likely performed on the validation split.

A key reproducibility issue is the mismatch between the dataset resolution (1024 × 1024) and the released FaceLock implementation, which is configured for 512 × 512 inputs. Running the pipeline at full resolution did not fit in the 40GB A100 GPU memory available to us, and while memory can be reduced by removing unnecessary gradient computations, this is undocumented and requires manual code changes. Even with these fixes, the pipeline remains unstable at 1024 × 1024 due to numerical overflows, and the codebase contains no resizing logic despite pointing users to a higher-resolution dataset. The expected input resolution is only mentioned once in Wang et al. (2024) in Table A2 of the appendix, forcing users to infer the correct setup and implement their own resizing procedure.

We ran experiments on 1000 randomly selected images from the presumed validation split. Although this might reduce our ability to match the exact absolute performances reported, such a large fraction of the data should already provide consistent results and enable comparative evaluation across baselines, the primary focus of this work. To run the pipeline, we additionally apply explicit resizing using bicubic interpolation.

3.4.1 Additional Dataset: FaceForensics++

To assess the generalizability of FaceLock beyond the CelebA-HQ and the additional Flickr-Faces-HQ (Karras et al. (2019)) mentioned in the appendix of Wang et al. (2024), we extend our evaluation to FaceForensics++ (Rössler et al. (2019)), a large-scale benchmark for facial manipulation detection that contains real-world and manipulated online videos under varying compression levels. Compared to CelebA-HQ and Flickr-Faces-HQ, which consist of high-quality, well-centered face images, FaceForensics++ is more challenging due to lower visual fidelity, motion blur, compression artifacts, and higher variability in pose, illumination, and background, with faces not always centered and sometimes multiple faces per frame. We experiment on the 1000 benchmark frames, matching our CelebA-HQ evaluation size. To align with the released FaceLock implementation, we preprocessed each image by detecting faces and cropping around the largest detected face (when multiple faces were present), then resized all images to 512 × 512 using bicubic interpolation.

3.5 Experimental Setup

Having discussed the datasets and associated preprocessing, we now describe the experimental setup used in our reproduction study and highlight inconsistencies encountered in the released implementation. We first edited the original images using the 25 prompts provided by the Authors, spanning three categories - facial feature modifications, accessory adjustments, and background alterations - to cover a broad range of instruction-guided editing scenarios. We then defended the original images with PhotoGuard, EditShield and FaceLock, edited the defended images using the same prompts, and evaluated the resulting outputs. To reduce stochastic effects, we followed the Authors' recommended procedure of keeping the editing process consistent between original and defended images, including the same seed. For all editing pipelines, we followed the hyperparameter configurations reported in the appendix of the original paper (Table 1).

Table 1: Editing pipeline configuration used in our experiments, following the appendix hyperparameter settings from the original paper.

Editing Model	Inference Steps	Image Guidance Scale	Text Guidance Scale	Random Seed
Instruct-Pix2Pix	50	1.5	7.5	42

For the defense pipelines, in turn, we used the hyperparameters reported by the Authors for FaceLock (Table 2), but left the default defense parameters as they were for the baselines since the they never mention these parameter in Wang et al. (2024).

Table 2: FaceLock defense configuration used in our experiments, following the appendix hyperparameter settings from the original paper.

Perturbation Budget (ϵ)	Step Size (α)	Optimization Steps (N)	Weighting Parameter (λ)
0.02	0.003	100	0.2

While we executed the released implementation as provided to avoid introducing additional deviations, we observed multiple discrepancies between the paper and the code that may affect reproducibility. First, Algorithm 1 in the appendix states that the perturbation should be initialized from a normal distribution $\mathcal{N}(0, \mathbf{I})$. In contrast, the released code initializes the perturbation by sampling from a uniform distribution $\mathbf{U}(-\epsilon, \epsilon)$. This difference in initialization is not documented and may influence optimization behavior.

More critically, the optimization objective implemented in the code does not seem to correspond to Equation (5) of the paper, which defines the core FaceLock method. In the paper, f_{FE} is weighted by λ, whereas in the implementation this term (`loss_lpips`) is not multiplied by λ at all. Instead, the weighting parameter is applied to a regularization term (`loss_encoder`) that is not described anywhere in the paper. Moreover, the feature embedding loss is only activated after the first 25% of optimization steps, and the identity distortion loss f_{FR} (`loss_cvl` in the code) is only enabled after 35% of the iterations. In summary, the loss computation that would theoretically lead to the perturbation δ is rather implemented as:

$$\text{loss} = -\text{loss}_{\text{cvl}} \cdot \mathbf{1}_{i \geq 0.35N} + 0.2 \cdot \text{loss}_{\text{encoder}} + \text{loss}_{\text{lpips}} \cdot \mathbf{1}_{i > 0.25N}$$

We ratify that this staged optimization strategy is neither described nor motivated in the paper, despite Equation (5) being presented as the central methodological contribution, and it resembles the EditShield optimization objective with a few tweaks.

As for EoT FaceLock, we aimed to eliminate several of the arbitrary design choices we discussed. While we found the regularizer empirically important and therefore retained it, we removed the heuristic conditions controlling when `loss_cvl` and `loss_lpips` begin to take effect, and introduced an explicit weighting factor ρ for the f_{FE} term. Since this variant is explicitly designed for robustness, we also observed that the loss converges in substantially fewer iterations. All differences are summarized in Table 3.

Finally, the feature embedding disparity is computed using the LPIPS model in the implementation, even though LPIPS is introduced in the paper solely as an evaluation metric. Since FaceLock achieves strong LPIPS scores in the reported experiments, we believe that this design choice is important to disclose: the performance on LPIPS is not an emergent property of the method, but rather a direct consequence of

Table 3: EoT FaceLock defense configuration changes from vanilla FaceLock.

Loss Objective	Optimization Steps (N)	Regularizer Weight (λ)	Image Feat. Ext. Weight (ρ)
$-\text{loss}_{\text{cvl}} + \lambda \cdot \text{loss}_{\text{encoder}} + \rho \cdot \text{loss}_{\text{lpips}}$	30	0.2	0.8

optimizing for it. While optimizing for a given metric is not inherently problematic, the lack of transparency may lead readers to misinterpret the results.

Now to evaluate our experimental setup, we initially follow the methodology outlined in the original paper, assessing two main factors: prompt fidelity, which quantifies the disruption caused by editing through CLIP-S (Radford et al. (2021)), PSNR, SSIM (Wang et al. (2004)) and LPIPS scores, and image integrity, which measures how closely the edited image resembles the original through CLIP-I and FR scores. Once reproducibility is done, we repurposed the prompt fidelity metrics to assess image integrity between the original and defended images, an additional evaluation pipeline we better discuss in the next section.

4 Results

We now present the results of our evaluation from both quantitative and qualitative perspectives. The quantitative analysis provides a numerical comparison of FaceLock's (Wang et al. (2024)) performance across multiple metrics, focusing on its relative behavior compared to baseline methods and highlighting cases where we were unable to reproduce the performance trends reported by the Authors. The qualitative analysis offers complementary insights into the model's behavior, capturing nuanced failure modes that are not always reflected in aggregate metrics and that, in some cases, differ from the interpretation presented in the original paper.

4.1 Quantitative Results

As shown in Table 4, we were unable to reproduce several key results reported in the original paper. Importantly, this discrepancy does not primarily concern the absolute metric values, which are expected to vary given that our experiments were conducted on a subset of the data. Instead, our concerns center on the reproducibility of the reported performance trends and, in particular, the relative ranking between methods.

In our experiments, EditShield (Chen et al. (2023)) generally achieves better scores than the other methods. By contrast, both FaceLock and the version of PhotoGuard (Choi et al. (2024)) used in our experiments, consistent with the authors' original setup, performed a bit worse, with FaceLock consistently underperforming PhotoGuard in our evaluation. As a result, we are unable to confirm the Authors' claim that FaceLock provides state-of-the-art defense performance.

Table 4: Quantitative evaluation on prompt fidelity (CLIP-S, PSNR, SSIM, LPIPS) and image integrity (CLIP-I, FR). Arrows (↑ or ↓) indicate whether a higher or lower value is preferred for a successful defense. All results are averaged over the 25 prompts for editing.

	Prompt Fidelity				Image Integrity	
Method	CLIP-S ↓	PSNR ↓	SSIM ↓	LPIPS ↑	CLIP-I ↓	FR ↓
PhotoGuard	0.1123 ± 0.0321	17.78 ± 1.74	0.6464 ± 0.0455	0.3847 ± 0.0431	0.7528 ± 0.0994	0.7044 ± 0.2098
EditShield	0.1073 ± 0.0277	18.31 ± 2.34	**0.5655 ± 0.0740**	**0.4626 ± 0.0498**	0.7162 ± 0.1040	0.6042 ± 0.2336
FaceLock	0.1133 ± 0.0373	24.83 ± 2.20	0.7755 ± 0.0403	0.2573 ± 0.0274	0.8455 ± 0.0753	0.8207 ± 0.1167
Simple FaceLock	**0.0997 ± 0.0193**	**15.79 ± 1.59**	0.6271 ± 0.0839	0.3840 ± 0.0840	**0.5705 ± 0.0623**	**0.2477 ± 0.1205**
EoT FaceLock	0.1133 ± 0.0379	25.47 ± 2.12	0.7913 ± 0.0366	0.2378 ± 0.0247	0.8504 ± 0.0734	0.7807 ± 0.1018

PhotoGuard and EditShield achieve better prompt fidelity scores than FaceLock, but we view this difference as expected and not particularly informative. The former are designed to either suppress or distort editing, whereas FaceLock prioritizes removing biometric identity regardless of prompt adherence. This highlights a limitation in the evaluation protocol: prompt fidelity is an ill-suited metric for comparing defenses with such fundamentally different objectives. Despite this, the Authors report unexpectedly strong prompt fidelity for

FaceLock, achieving the best SSIM (Wang et al. (2004)) and LPIPS (Zhang et al. (2018)) and staying close to baselines on other metrics, which we found surprising and were unable to reproduce. In our experiments, FaceLock behaved more consistently with its stated identity-focused design, reinforcing our view that prompt fidelity gives a weak perspective for comparing the defenses.

Turning to image integrity, unlike prompt fidelity metrics, which we argued are poorly aligned with FaceLock's objective, these metrics are more relevant for evaluating FaceLock, and we therefore view their inclusion as a valuable step. However, FaceLock's performance does not follow the trends reported in the original paper and instead yields the worst results among baselines. Notably, FaceLock produces image integrity scores that are even worse than PhotoGuard, despite PhotoGuard being primarily designed for edit prevention and typically incurring low integrity only when it fails. This suggests that the proposed pipeline is unreliable and makes it difficult to determine when FaceLock meaningfully influences editing, a point we further substantiate in our qualitative analysis. Consequently, we were unable to confirm the Authors' claim that FaceLock achieves state-of-the-art performance even under their own proposed image integrity evaluation pipeline.

At the same time, the proposed metrics still create comparability issues across different defense goals. Low facial recognition (FR) similarity is meaningful for FaceLock (and also relevant for EditShield), but inappropriate for edit-prevention methods that preserve identity by design like PhotoGuard. Lastly, we find the introduced CLIP-I (Radford et al. (2021)) score problematic as well. Even within FaceLock, CLIP-I can vary widely depending on whether identity removal occurs via heavy visual distortion or more subtle facial changes, yet both outcomes may represent equally valid defenses and CLIP-I will only support one.

Following this analysis, we find that EoT FaceLock achieves largely the same results as vanilla FaceLock across most metrics, with the exception of FR, where it yields a modest improvement. This is expected, since EoT was designed to increase robustness to diffusion randomness (editing seeds). Notably, the improvement appears specifically on the metric we argue is most meaningful for evaluating FaceLock's objective. However, while EoT improves FR relative to FaceLock, its performance still remains far from the EditShield baseline, indicating that further improvements to the FaceLock-like methods are necessary.

Now more broadly, these difficult comparisons highlight a misalignment between current evaluation practices and the diversity of defense objectives: edit-prevention methods should be evaluated primarily using prompt fidelity, while disruption defenses should instead be assessed via facial recognition measures. That said, both categories require an explicit analysis of the trade-off between defense effectiveness and corruption of the protected image in order to properly contextualize their capabilities. Motivated by these observations, we propose an additional evaluation framework that directly captures this trade-off.

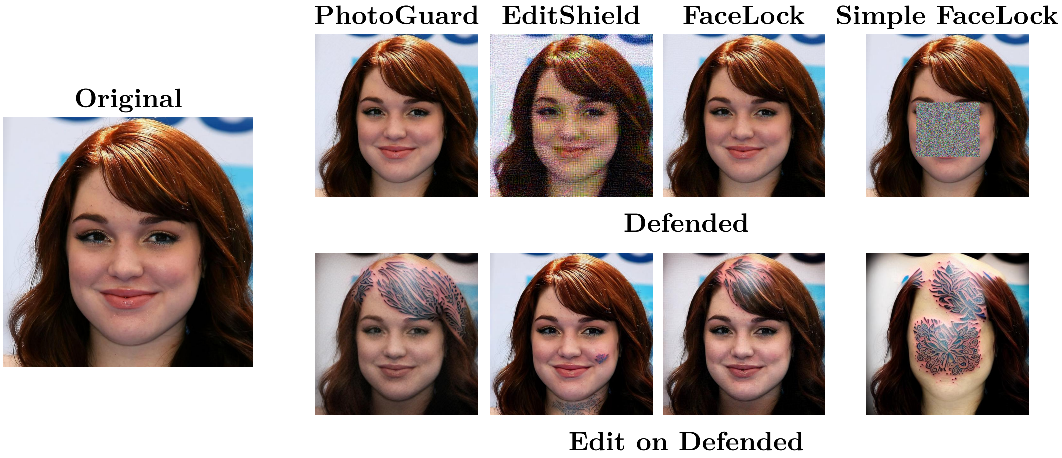
4.1.1 Additional Evaluation Pipeline

We propose evaluating image integrity **prior to editing** by comparing the original image to its protected counterpart using PSNR, SSIM, LPIPS, and CLIP-I. The first three capture low-level visual similarity and are particularly informative when defenses apply subtle perturbations, allowing us to quantify how much a protection method interferes with the original image. The latter, in contrast, measures high-level semantic similarity and is most relevant to identify defenses that introduce substantial distortions that may destroy image content. In such cases, downstream edits probably become meaningless and biometric information may be removed the cost of no longer being able to publish the intended image.

To illustrate this failure mode, we introduce a deliberately flawed baseline, Simple FaceLock, which uses FaceLock's face recognition and alignment models to locate facial landmarks, expands a face bounding box by 25%, and replaces the corresponding region with uniform noise $\mathbf{U}(-1, 1)$. While this trivially removes biometric identity (Figure 2), it does so by destroying the face at publication time, making it impractical as a defense. Despite its obvious shortcomings, Simple FaceLock still achieves strong scores on several metrics used in prior work (Table 4), exposing the key limitation of current evaluation protocols we talked about.

We additionally report the performance of all evaluated methods under our proposed additional evaluation pipeline in Table 5. Under this framework, FaceLock's behavior can be interpreted more clearly: although it achieves lower defense scores than EditShield in our reproducibility study, it substantially better preserves the

Figure 2: Comparison of defenses and edits on defended images for PhotoGuard, EditShield, FaceLock, and Simple FaceLock.



original image at publication time, making the trade-off between protection and image degradation explicit. However, since PhotoGuard is more effective than FaceLock while also preserving image quality somewhat better, FaceLock does not appear to offer the strongest balance among the evaluated methods. Finally, EoT FaceLock marginally improves image preservation and, together with its higher FR score, establishes itself as a consistent improvement over vanilla FaceLock.

Table 5: Quantitative evaluation on image integrity **prior to editing** (PSNR, SSIM, LPIPS and CLIP-I).

	Visual Similarity			Semantic Similarity
Method	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP-I ↑
PhotoGuard	**38.1396 ± 1.7429**	**0.9508 ± 0.0149**	**0.1341 ± 0.0220**	**0.9808 ± 0.0116**
EditShield	34.4462 ± 1.0001	0.8678 ± 0.0178	0.3106 ± 0.0359	0.9608 ± 0.0148
FaceLock	37.8254 ± 1.6081	0.9447 ± 0.0140	0.1776 ± 0.0270	0.9760 ± 0.0112
Simple FaceLock	18.4474 ± 0.9062	0.8061 ± 0.0182	0.2409 ± 0.0214	0.6914 ± 0.0501
EoT FaceLock	38.0837 ± 1.6880	0.9494 ± 0.0140	0.1611 ± 0.0251	0.9780 ± 0.0110

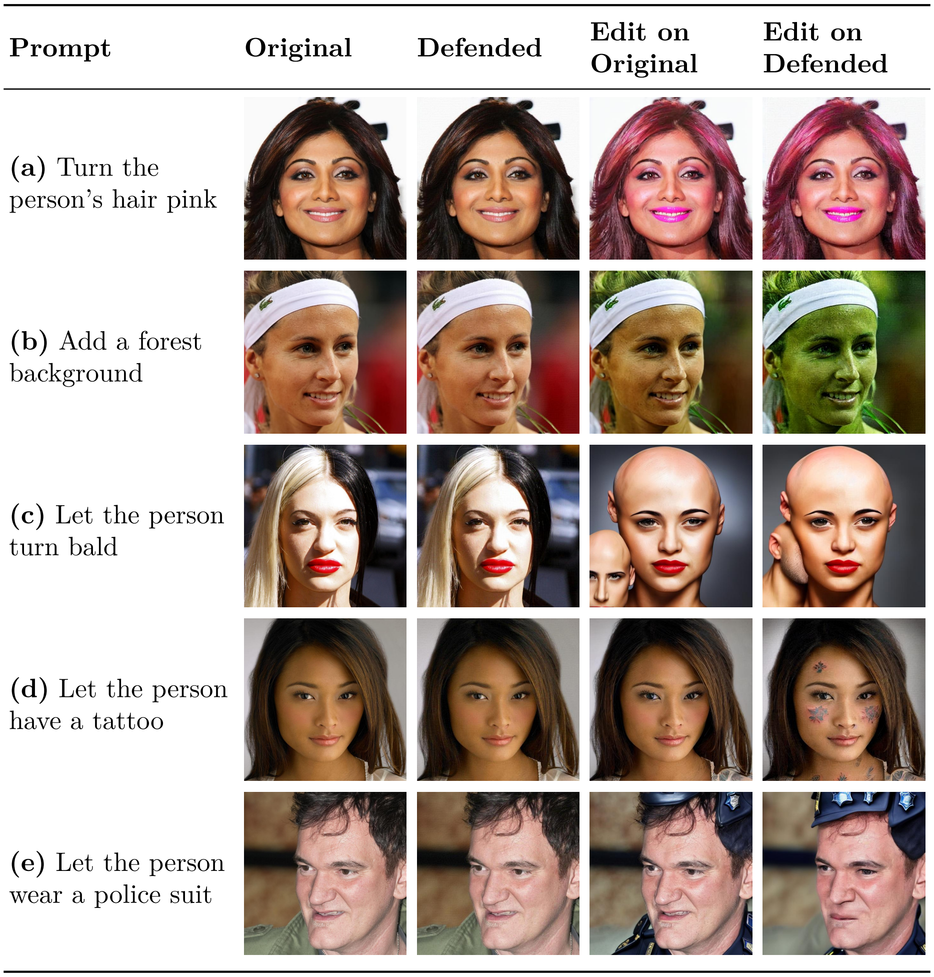
By contrast, Simple FaceLock performs poorly under our additional evaluation, showing that defense effectiveness cannot be assessed in isolation. Our framework captures this failure quantitatively, numerically demonstrating that it is not a viable defense. A similar issue arises for EditShield, which achieves strong defense scores but substantially degrades image quality by adding heavy noise across the entire image.

4.2 Qualitative Results

We complement our quantitative analysis with a qualitative evaluation to better understand FaceLock's behavior. In our reproduction, we observe substantial differences from the original paper's qualitative results, with many failure cases across diverse image–prompt pairs. We analyze these discrepancies systematically by comparing edits on original versus protected images, and contrasting successful and failed cases to contextualize the quantitative gaps.

From an intuitive criterion for successful defense for FaceLock's facial biometric information removal objective, we find that for most prompts across the evaluated dataset, only a very small fraction of images are defended. We present a set of examples in Figure 3, including cases not highlighted in the original paper, to illustrate the dominant behaviors we observed in practice. We also provide a deeper dive into the examples in the appendix (Figures 5, 6, and 7), showcasing much more image-prompt pairs, spanning all prompt categories, and including some of the sporadic success scenarios, to make sure we enable a broad and fair overview of the method's results.

Figure 3: Qualitative comparison of image editing with and without FaceLock defense. Columns show the original image, edited (original) image, defended image, and edited defended image.


Prompt
Original
Defended
Edit on Original
Edit on Defended
(a) Turn the person's hair pink
(b) Add a forest background
(c) Let the person turn bald
(d) Let the person have a tattoo
(e) Let the person wear a police suit

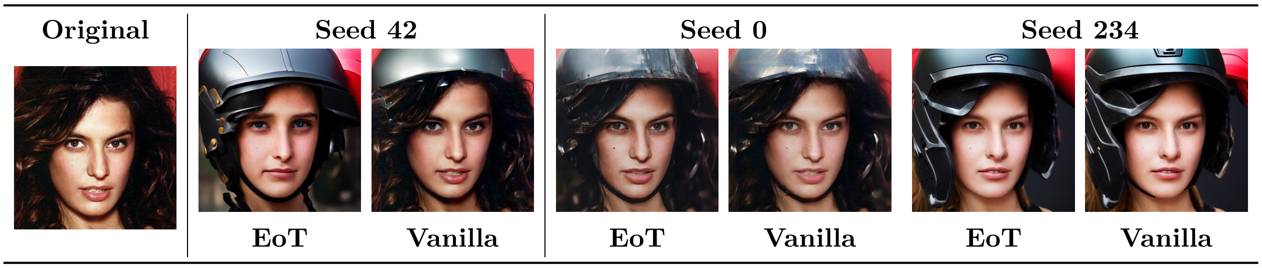
For most images, we observe little difference between edits applied to the original and protected images, and the defense typically fails to introduce changes substantial enough to remove biometric identity. For example, in Figure 3(a), the "pink hair" edit remains effectively unchanged. In other cases, the defense mainly induces a global color shift rather than identity distortion, leaving facial identity intact and resembling edit disruption (as in PhotoGuard) rather than FaceLock's intended mechanism. In the specific instance of Figure 3(b), the defended edit is even more visually altered than the original. Having said that, we've noticed slightly higher success rates for background-alteration prompts, where success often stems from blending the face into the edited background and thereby hiding it.

We observe a recurring link between apparent defense success and failures in the editing process itself. In some cases (Figure 3(c)), the edit on the original image already produces severe artifacts, making the defended output appear successful only because the edit is unstable rather than because FaceLock meaningfully intervenes. We also find cases where the edit on the original image is weak or fails, but the same edit is applied more strongly to the defended image without removing identity (Figure 3(d)), suggesting the defense can amplify edits and produce the opposite of its intended effect.

With all that in mind, we still observe successful cases that align with FaceLock's intended goal. For example, in Figure 3(e), the edit is applied as intended, but the subject's face is distorted to the point that the celebrity doesn't look the same and the image clearly appears AI-generated/edited. We further conducted a reproducibility study (Figure 8) using the exact image–prompt pairs from Figure 6 in Wang et al. (2024), in order to test whether we could replicate their qualitative results. While we were unable to reproduce all of their examples, experiments across multiple random seeds revealed an important trend: even though FaceLock performs poorly overall in our evaluation, it often succeeds for *some* seed. In Figure 8, FaceLock succeeds for at least one seed in 5 out of 9 image–prompt pairs. Although this highlights the potential of the approach, it also demonstrates substantial sensitivity to randomness, which motivated our implementation of EoT FaceLock.

Our experiments show that EoT FaceLock almost always succeeds in the cases where vanilla FaceLock succeeds, while also extending success to a few additional scenarios, although the improvement is incremental rather than transformative. We further evaluated these additional cases across multiple editing seeds, focusing on examples where EoT FaceLock succeeded but vanilla FaceLock failed under the initial seed. As illustrated in Figures 4 and 9, EoT FaceLock also remained more consistent across tested seeds, ratifying the indication of a step towards improved robustness.

Figure 4: Comparison of EoT FaceLock and vanilla FaceLock across different editing seeds. For the initial seed (42), EoT provided protection while vanilla didn't. For seed 0, EoT provided better protection than Vanilla, although not yet ideal. For seed 234, both methods proper (and similar) protection.



4.3 Generalization Experiments

Table 6 reports FaceLock's quantitative results on FaceForensics++ (Rössler et al. (2019)), which are broadly consistent with those obtained on CelebA-HQ (Karras et al. (2017)). While we observe improved prompt fidelity (CLIP-S and LPIPS), this is difficult to interpret given our earlier argument that it is not meaningful for FaceLock's objective. Similarly, the only statistically significant image integrity improvement is in CLIP-I, which we also consider an unreliable indicator. Overall, the results lead to the same conclusion: we cannot support the claim that FaceLock reliably removes biometric identity in edited outputs. On the other hand, FaceForensics++ is a more diverse and challenging dataset, suggesting that FaceLock's (limited) performance generalizes beyond CelebA-HQ, under more realistic conditions and combinations of purification methods in the wild. Figure 10 shows corresponding qualitative examples on FaceForensics++ across prompts from the three categories discussed earlier, and the outcomes closely mirror the ones on CelebA-HQ once again, reinforcing FaceLock's limited reliability, but consistency.

Table 6: Quantitative evaluation on prompt fidelity for FaceLock on the FaceForensics++ dataset. All results are averaged over the 25 prompts for editing.

	Prompt Fidelity				Image Integrity	
Method	CLIP-S ↓	PSNR ↓	SSIM ↓	LPIPS ↑	CLIP-I ↓	FR ↓
FaceLock	0.0720 ± 0.0258	24.0042 ± 2.8763	0.8490 ± 0.0475	0.3050 ± 0.0400	0.8159 ± 0.0522	0.8153 ± 0.1378

5 Conclusion

This work highlights both the relevance of biometric defenses against generative image editing and the conceptual potential of FaceLock (Wang et al. (2024)). The method builds upon general image defenses such as both PhotoGuard and EditShield pipelines (Choi et al. (2024); Chen et al. (2023)) and introduces a shift in perspective by reframing defense as identity distortion in edited outputs rather than edit suppression or complete distortion, what constitutes an interesting and promising direction for future research.

We also acknowledge the Authors' efforts toward reproducibility, including detailed experimental descriptions across the main paper and appendix, as well as the release of an open-source implementation. However, our study reveals substantial room for improvement. In particular, reproducibility is hindered by incomplete specification of dataset splits and random seeds, limited robustness of the codebase to different environments, and reliance on undocumented assumptions regarding image preprocessing. More critically, we identify several inconsistencies between the method as described in the paper and its implementation, as well as a lack of transparency regarding optimization targets, which together pose significant challenges for faithful reproduction.

Moreover, despite following the provided setup as closely as possible, we were unable to reproduce the performance trends reported in the original paper and therefore cannot confirm the claim that FaceLock achieves state-of-the-art performance relative to the proposed baselines. Qualitatively, we observed successful defenses directly attributable to the defense pipeline in only a small fraction of cases, in contrast to the uniformly strong examples presented in Wang et al. (2024). In many instances, apparent successes were instead correlated with failures or instabilities in the editing process itself.

Finally, our work highlights FaceLock's lack of robustness to editing seeds and proposes a promising extension that modestly improves the reliability of biometric distortion without interfering with image quality in publication time. Beyond these findings, our study underscores the need to reassess evaluation practices in this field. Different defense paradigms pursue fundamentally different objectives and should not be evaluated using the same set of metrics. We argue for the importance of contextualizing evaluation metrics according to defense goals and, most importantly, for explicitly measuring the trade-off between protection effectiveness and image distortion at publication time, which we propose in this work with our additional evaluation pipeline.

6 Limitations and Future Work

Although we consider this a comprehensive reproducibility study, examining both quantitative and qualitative aspects of the main claims in the FaceLock paper (Wang et al. (2024)) as well as the surrounding baseline methods, our work still has several limitations and naturally motivates future work. We did not systematically explore hyperparameters for either the defense or editing pipelines, focusing instead on the configurations reported by the authors. Likewise, we did not comprehensively test architectural alternatives included in the codebase, primarily running the setup emphasized as best-performing in the paper: InstructPix2Pix (Brooks et al. (2022)) for editing and the VGG-based backbone (Simonyan & Zisserman (2015)) for LPIPS feature extraction (Zhang et al. (2018)), rather than alternatives such as Stable Diffusion image-to-image (Meng et al. (2022)) or AlexNet-/SqueezeNet-based LPIPS backbones (Krizhevsky et al. (2012); Iandola et al. (2016)). We also did not reproduce all auxiliary experiments from the original paper, such as the full ablation studies.

Despite these limitations, we believe our study remains valuable by reassessing the method in depth, documenting reproducibility issues, and contributing an extension to FaceLock and a proposal for a better evaluation pipeline. For future work, reproducibility efforts could fill the remaining gaps by systematically testing hyperparameter and architectural sensitivity, and completing reproduction of the remaining auxiliary experiments. We also believe that our robustness-oriented extension, EoT FaceLock, could be further investigated and improved through more hyperparameter tuning or by incorporating new strategies to better handle the stochasticity inherent in diffusion-based editing. More broadly, we encourage future work in the field to incorporate our proposed evaluation framework, which explicitly captures the trade-off between biometric protection effectiveness and distortion of the protected image at publication time, enabling more principled comparisons across methods with fundamentally different defense objectives.

References

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples, 2018. URL `https://arxiv.org/abs/1707.07397`.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

Ruoxi Chen, Haibo Jin, Yixin Liu, Jinyin Chen, Haohan Wang, and Lichao Sun. Editshield: Protecting unauthorized image editing by instruction-guided diffusion models. *arXiv preprint arXiv:2311.12066*, 2023.

June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. Diffusionguard: A robust defense against malicious diffusion-based image editing. *arXiv preprint arXiv:2410.05694*, 2024.

Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size, 2016. URL `https://arxiv.org/abs/1602.07360`.

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

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. URL `https://arxiv.org/abs/1812.04948`.

Minchul Kim, Anil K. Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition, 2023. URL `https://arxiv.org/abs/2204.00964`.

Minchul Kim, Yiyang Su, Feng Liu, Anil Jain, and Xiaoming Liu. Keypoint relative position encoding for face recognition, 2024. URL `https://arxiv.org/abs/2403.14852`.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL `https://arxiv.org/abs/1312.6114`.

Pavel Korshunov and Sebastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL `https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022. URL `https://arxiv.org/abs/2108.01073`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL `https://arxiv.org/abs/2103.00020`.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *arXiv preprint arXiv:1901.08971*, 2019.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023.

Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*, 2024.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL `https://arxiv.org/abs/1409.1556`.

Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv preprint arXiv:2001.00179*, 2020.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

Hanhui Wang, Yihua Zhang, Ruizheng Bai, Yue Zhao, Sijia Liu, and Zhengzhong Tu. Edit away and my face will not stay: Personal biometric defense against malicious generative editing. *arXiv preprint arXiv:2411.16832*, 2024.

Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, January 2004. doi: 10.1109/TIP.2003.819861.

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.

Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. Webface260m: A benchmark unveiling the power of million-scale deep face recognition, 2021. URL `https://arxiv.org/abs/2103.04098`.

A Additional Qualitative Results

The figures below present more qualitative results for FaceLock, PhotoGuard and EditShield (Wang et al. (2024); Choi et al. (2024); Chen et al. (2023)) on specific cases of image editing on Celeba-HQ images (Karras et al. (2017)), spanning the three categories reported by the Authors: modification of facial attributes, accessory addition and background modification. The results are separated into the original image, protected images after applying the defense pipelines, and finally the edits on both the original image and the protected images.

A.1 Qualitative Results on Facial Feature Modification

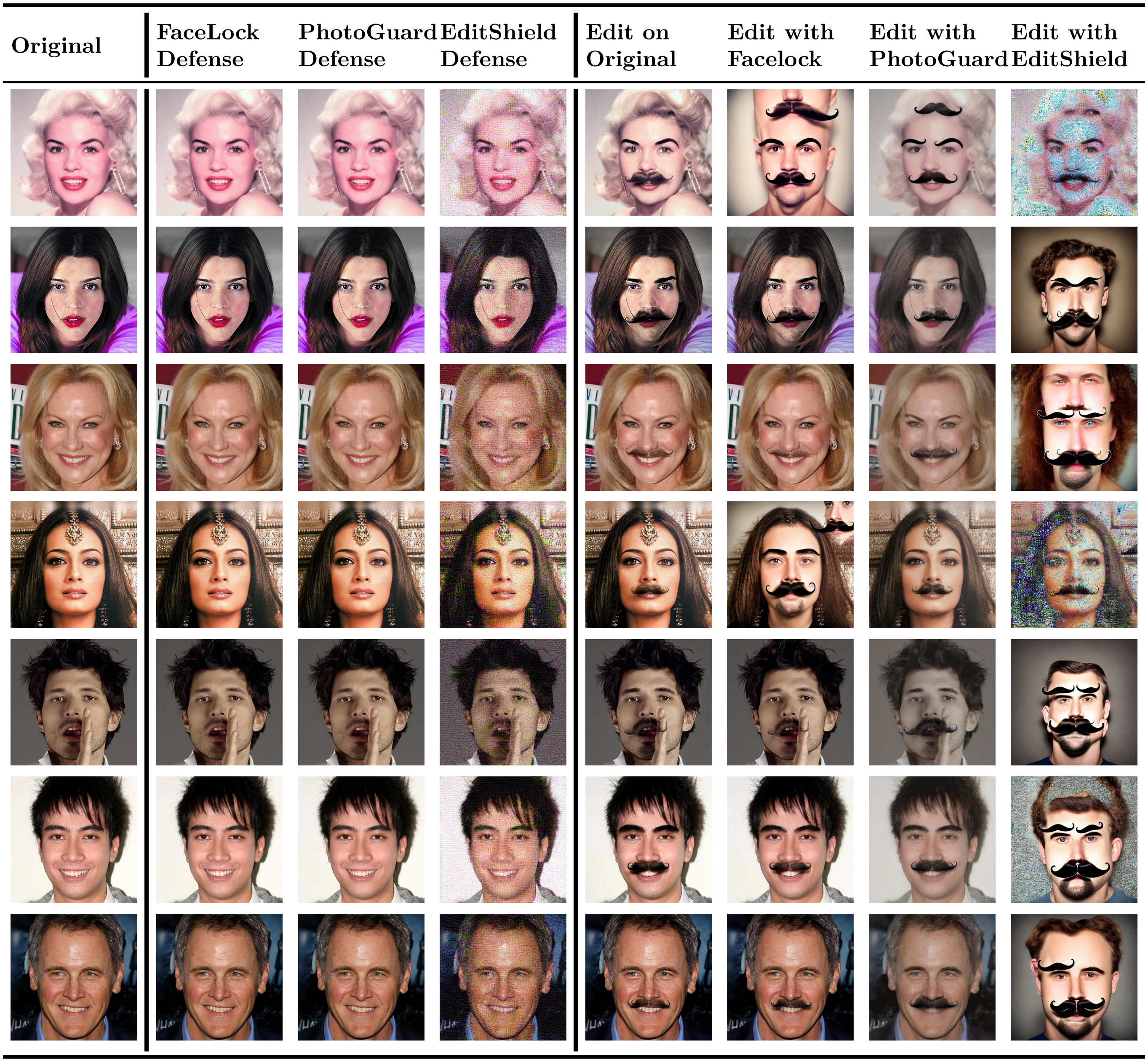


Figure 5: **Prompt: "Let the person grow a mustach"**

A.2 Qualitative Results on Accessory Addition

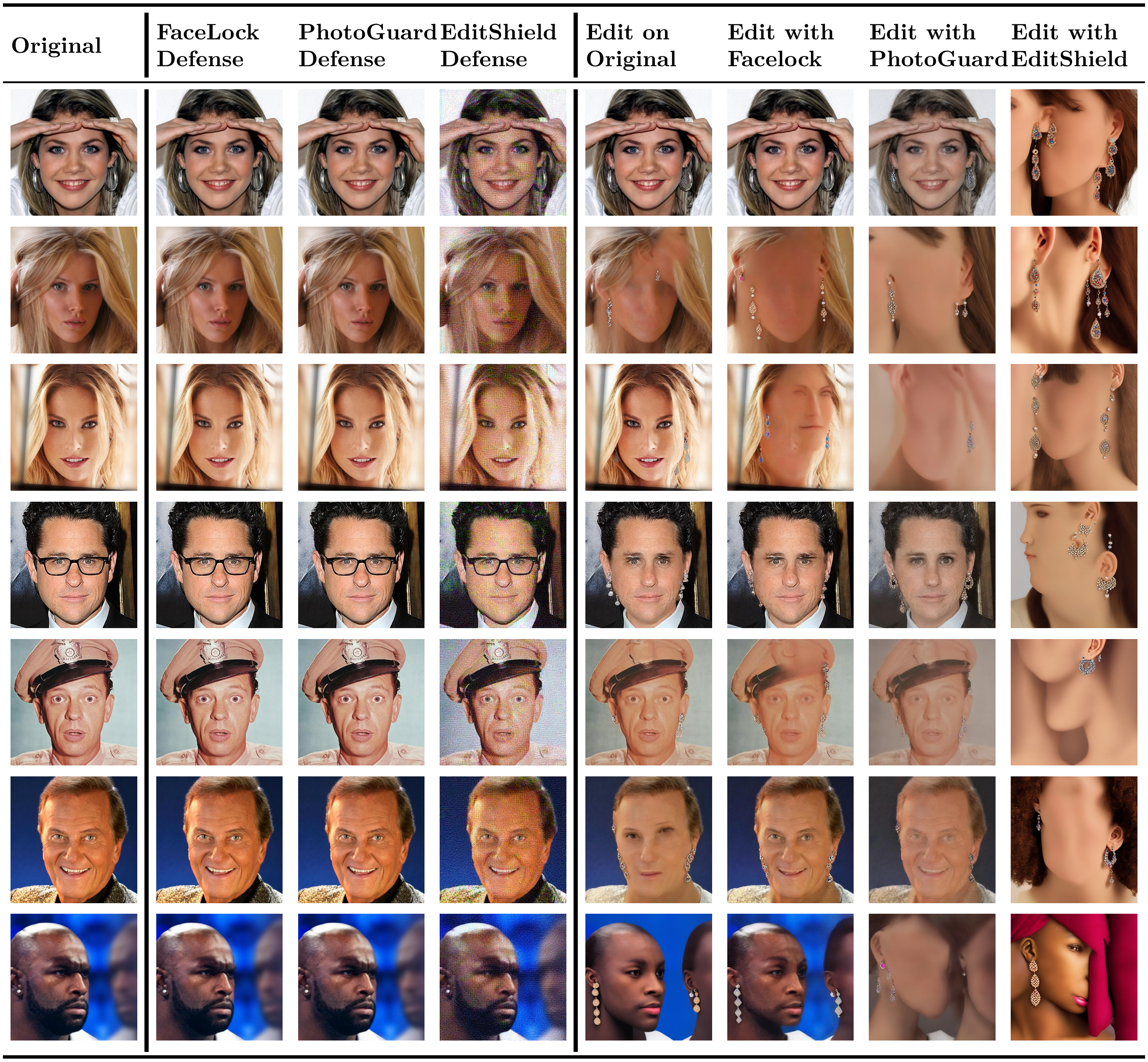


Figure 6: **Prompt: "Let the person wear earrings"**

A.3 Qualitative Results on Background Modification

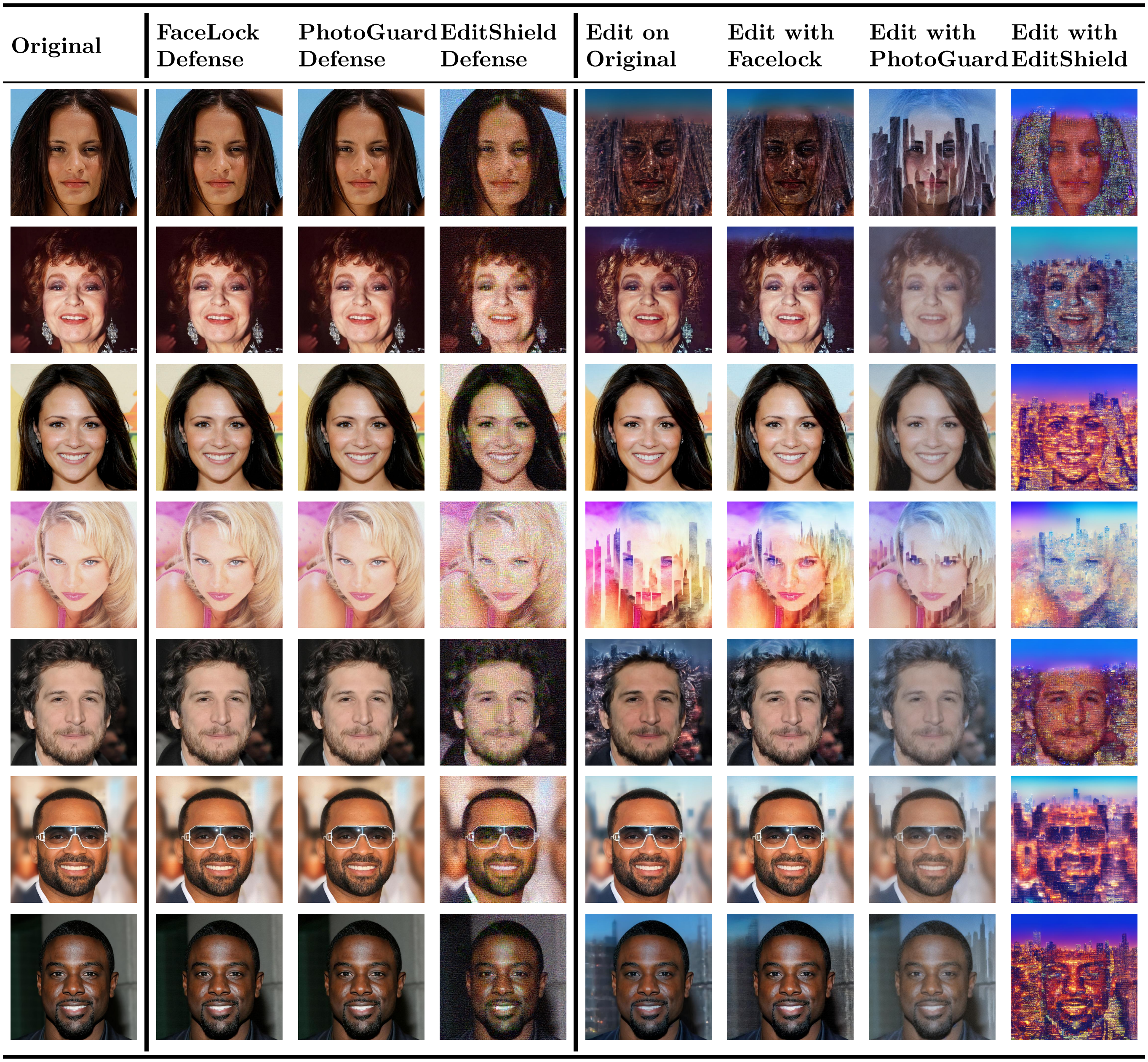


Figure 7: **Prompt: "Add a city skyline background"**

B Reproducibility of Wang et al. (2024) Figure 6

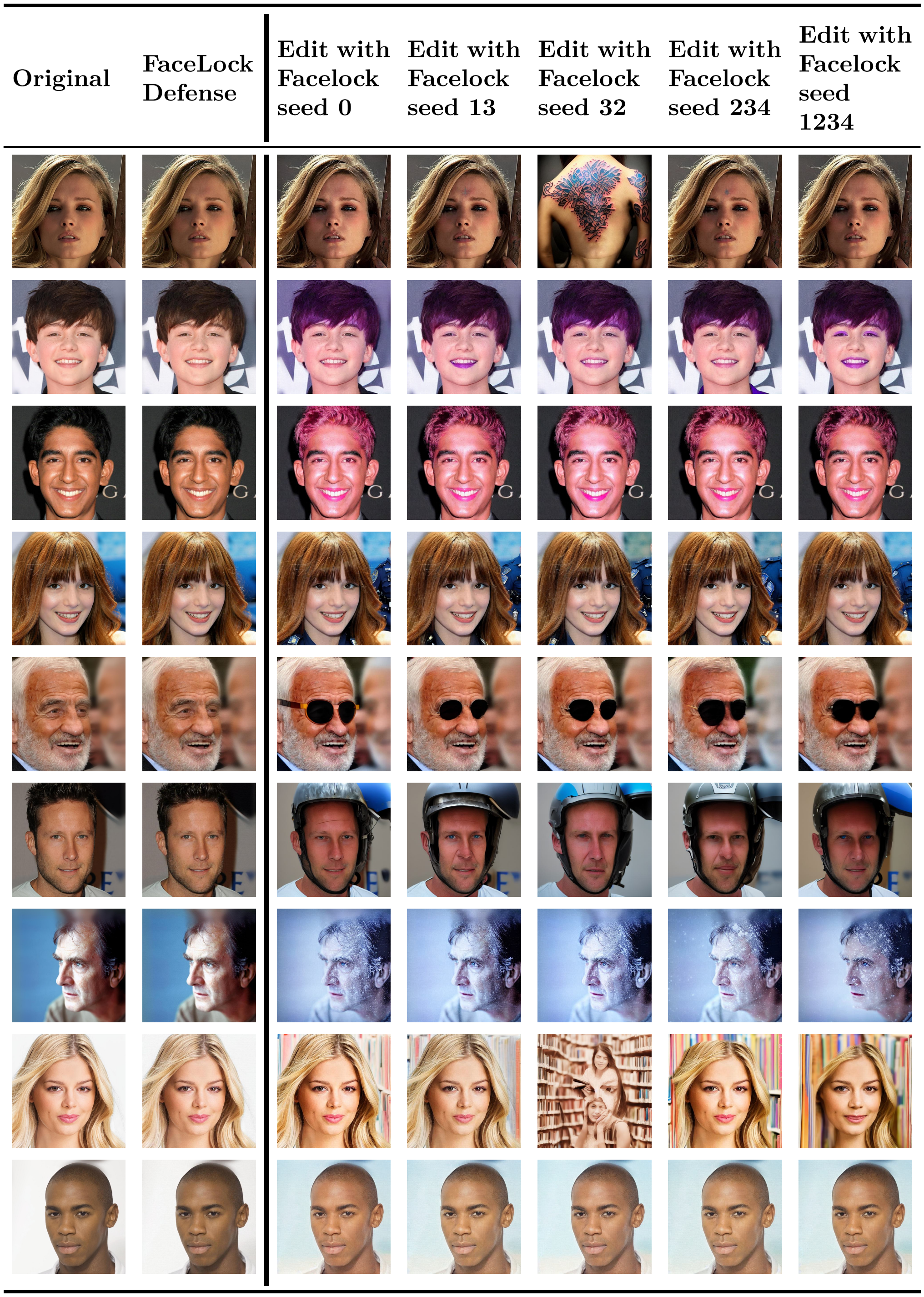

Original
FaceLock Defense
Edit with Facelock seed 0
Edit with Facelock seed 13
Edit with Facelock seed 32
Edit with Facelock seed 234
Edit with Facelock seed 1234

Figure 8: FaceLock on images of Wang et al. (2024) Figure 6 for multiple seeds. From top to bottom, prompts are: 1. "Let the person have a tattoo"; 2. "Let the person wear purple makeup"; 3. "Turn the person's hair pink"; 4. "Let the person wear a police suit"; 5. "Let it be snowy"; 6. "Set the background in a library"; 7. "Change the background to a beach."

C Additional Qualitative Results for EoT FaceLock

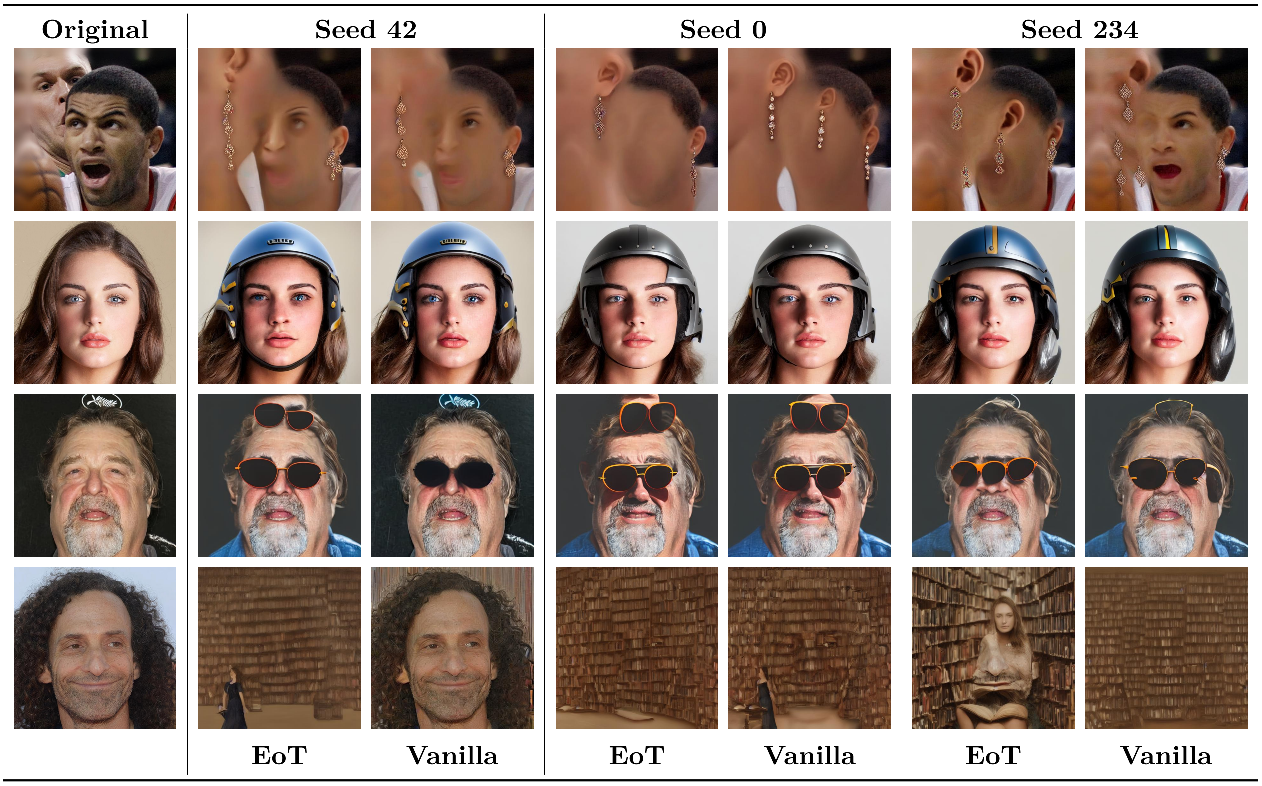


Figure 9: Comparisons of EoT FaceLock and vanilla FaceLock for a few image-prompt pairs for which EoT provided more defense effectiveness, more stability across editing seeds, or both.

D Qualitative Results of Generalization Experiment

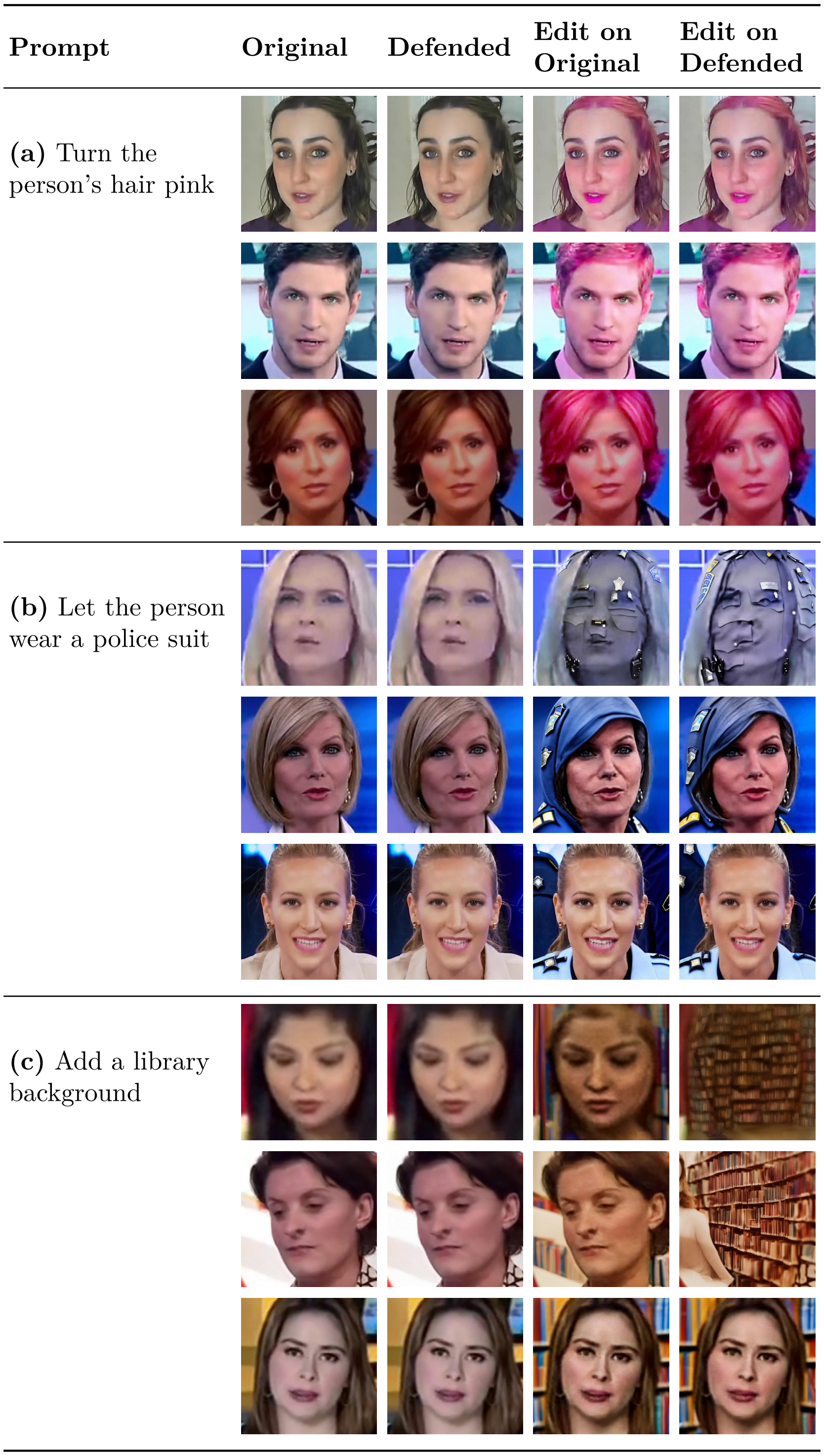

Prompt
Original
Defended
Edit on Original
Edit on Defended
(a) Turn the person's hair pink
(b) Let the person wear a police suit
(c) Add a library background

Figure 10: Qualitative results for FaceLock on FaceForensics++ dataset (Rössler et al. (2019)). Columns show the original image, defended image, edit on the original image, and edit on the defended image. Each prompt is shown for three different source portraits.