Supplementary Material of *Phantom*: A Unified Face-Swap Deepfake Protection Framework with Latent and Spatial Constraints

Jungkon Kim* Cheolseung Jung Jong-Min Choi Juseong Lee
Samsung Electronics, AI Platform Center
{jungkon.kim, cs.jung, jminl.choi, jooseong.lee}@samsung.com

1. Detailed Algorithm of *Phantom*

The full algorithm of ***Phantom*** is outlined in Algorithm 1, summarizing the complete procedure described in the Methodology section in the main paper. As detailed therein, ***Phantom*** is a unified facial privacy protection framework applicable to both dodging and impersonation scenarios. In the dodging setting, perturbations are guided by an adaptively synthesized target (see Fig. 1), whereas in impersonation, a predefined target identity serves as the target. Across both threat scenarios, ***Phantom*** proactively disrupts face recognition (FR) pipelines, offering robust protection against unauthorized FR models and face-swapping deepfakes, while preserving high perceptual quality.

Figure 1. Original input images (top row) and their adaptively synthesized targets (bottom row) used for identity-axis guidance in the dodging setting.

2. Motivation and Region Selection for Masked Adversarial Attack

To identify the optimal region for adversarial perturbation, we analyze the effectiveness of noise placement across facial areas using segmentation maps from MaskGAN [6], which provides 19 fine-grained facial regions. Based on this, we group them into four semantic categories: *all*, *face*, *non-face* (e.g., hair, neck, background), and *face elements* (eyes, nose, mouth, ears, lips; excluding facial skin). Protection effectiveness is measured by Rank-1 paired protection (R1-P), indicating the method's ability to prevent ground-truth identification by face recognition (FR) models. Image quality is evaluated by Structural Similarity Index (SSIM) [10] and Fréchet Inception Distance (FID) [3].

As shown in Fig. 2, our analysis reveals that applying noise to *non-face* regions causes noticeable distortion without yielding meaningful protection. Perturbations restricted to specific *face elements* such as the eyes, nose, and mouth achieve higher protection efficiency per unit area but are limited by their small coverage. In contrast, applying noise to the *face* region, which includes both *face elements* and *skin* while excluding peripheral areas such as hair and background, strikes a more effective balance between visual quality and protection; compared to the *all* region, it improves visual quality by up to 48% while maintaining protection effectiveness in the range of 93.8 to 95.3%. These findings indicate that spatially controlled perturbations confined to semantically relevant regions can enable efficient and robust facial privacy protection.

Motivated by these insights, we propose a masked adversarial attack strategy that confines noise strictly to the *face* region by zeroing gradients outside this area during noise generation. Consequently, our approach maximizes protection against unauthorized FR models and face-swapping deepfakes, minimizing perceptual distortion. We thus define the *face* region as the selective adversarial noise region.

3. Background: Face Editing and Style Transfer Models

To synthesize adversarial target images in our adaptive target strategies, we utilize a generative model originally proposed for face generation, face editing and style transfer.

MaskGAN [6] is a semantic-aware framework that enables localized facial editing by conditioning on both segmentation masks and style references. Given a face image x_u from user u, it extracts a facial segmentation mask $m_u = G_S(x_u)$. For style transfer, a target identity image

*Corresponding author

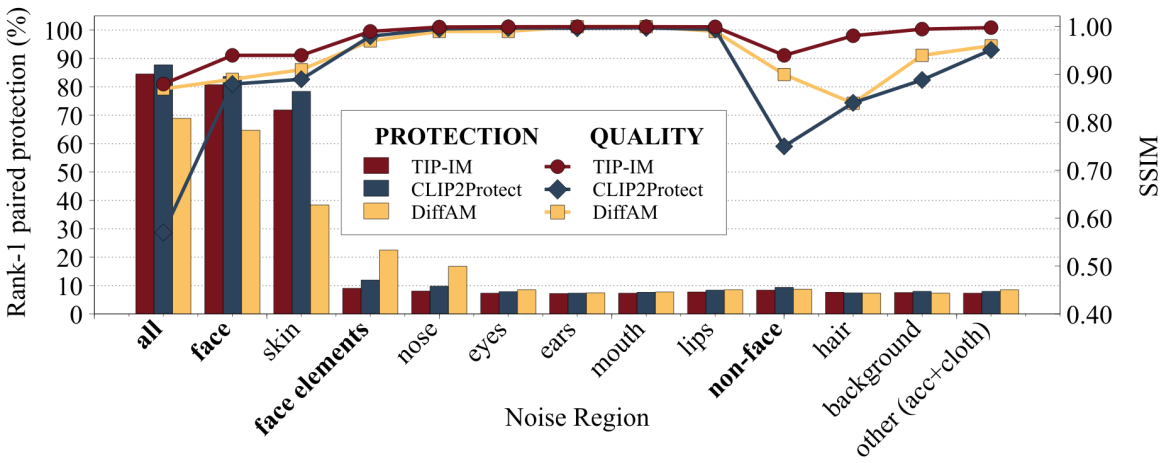

Figure 2. Rank-1 paired protection (R1-P) and SSIM scores for noise regions and individual face segments.

x_w is provided, and the generator synthesizes a new face $x'_w = G_F(m_u, m_w, x_w)$. Alternatively, for face editing, the original mask m_u is manually modified into m'_u, and a new output $x'_u = G_F(m_u, m'_u, x_u)$ is generated, preserving the original appearance while applying localized changes.

Collaborative-Diffusion [5] is a diffusion-based model for multi-modal face editing that separately encodes structure and style. It enables generation from a mask m_u alone, $G_F(m_u)$, or conditioned on a text prompt to control the style, $G_F(m_u, \textit{prompt})$. This allows generation of diverse identities while maintaining the source facial layout.

While both models are non-adversarial in nature, we repurpose them to generate identity-shifted yet attribute-preserving targets for adversarial attack. Specifically, we reapply the source's own style onto its segmentation mask to generate a weakly identity-shifted image $t_u = G_F(G_S(x_u), x_u)$, which serves as the adversarial target in our optimization. This yields perceptually aligned targets with subtle identity shifts, enabling more coherent and effective perturbation guidance (See Fig. 1.).

4. Implementation Details

4.1. Facial Privacy Protection Methods

For all evaluated methods, we configure parameters based on the recommended values in their respective papers and official codes to ensure optimal settings. We adopt the four target identity pairs used in the baselines [4, 7–9], and all reported results are averaged over these targets. The parameter configurations for compared methods are as follows.

TIP-IM [12] is set to maximum perturbation (ϵ) = 16 / 255 and iterations = 100. For dodging, we randomly selected 10 target images from different identities, following the recommendation of the original paper. For impersonation, target images in Fig. 3 was used, consistent with other baseline methods. We use the official implementation of TIP-IM[1] and set γ to $5e{-}4$, which differs in magnitude from the value reported in the paper (Note that the official code and the paper use different notations for γ [4].).

For AMT-GAN [4], we employ the official implementation[2] and training configuration. Specifically, we use λ_{GAN} = 10, λ_{reg} = 10, λ_{adv} = 5, λ_{make} = 2, and λ_{idt} = 5. The model is trained using the Adam optimizer with a learning rate of 0.0002 and decay rates (β_1, β_2) = (0.5, 0.999). We set the batch size to 1 and train for 5 epochs.

For CLIP2Protect [8], we adopt a vision transformer-based CLIP model for adversarial text guidance in both impersonation and dodging scenarios. During latent code initialization, we set the number of iterations to 450 and use $\lambda_2 = 0.5$. To generate protected images, we set the number of iterations to 50 and configure λ_{adv}, λ_{clip}, and λ_{latent} to 1, 0.5, and 0.01 for impersonation, and to 1, 0.5, and 0.0001 for dodging, respectively. For the text guidance, we use three prompts: *matte makeup*", *pink eyeshadows*", and "*tanned makeup with red lipstick*". All evaluation scores are averaged across these three prompts.

For DiffAM [9], we finetune the provided pretrained model for 6 epochs using the Adam optimizer with an initial learning rate of $4\text{e}{-}6$, which is increased linearly by a factor of 1.2 every 50 iterations. The total time step T is set to 60, and the discretization steps for DDIM inversion and sampling, S_{inv} and S_{sam}, are set to 20 and 6, respectively. Additionally, we set the ensemble attack loss weight λ_{adv} to 1.2, and omit the ensemble attack loss $\mathcal{L}_{\text{adv}}$ during the first two iterations, as proposed in their official code[3]. For reference image styles, we use the pretrained styles "XMY-060" and "XYH-045," and report all evaluation results averaged

[1] https://github.com/ShawnXYang/TIP-IM
[2] https://github.com/CGCL-codes/AMT-GAN
[3] https://github.com/HansSunY/DiffAM

Algorithm 1 ***Phantom*** Algorithm

Input:
- x_u: input image of user u
- $M_1, M_2, \ldots, M_r$: surrogate FR models
- G_S: face segmentation model
- G_F: face generation model (for dodging attack only)
- t_u: target image ▷ **given as input for impersonation, synthesized internally for dodging**

Parameters:
- $iter, \epsilon$: max iterations and perturbation bound
- $\lambda_{ori}, \lambda_{tgt}$: weights for identity loss terms
- $\lambda_{id}, \lambda_{qual}$: weights for total loss

Output: x_u^p: protected image

▷ **Step 1: Binary Face Mask Generation**
$B_u \leftarrow$ binary face mask from m_u s.t. $m_u \leftarrow G_S(x_u)$
▷ **Step 2: Target Setting**
if dodging attack **then**
  $t_u \leftarrow G_F(x_u, m_u)$ ▷ **Adaptive Target Synthesis**
else if impersonation attack **then**
  $t_u \leftarrow$ pre-defined target image ▷ **Fixed image**
end if
▷ **Step 3: Masked Adversarial Attack**
Let $n_{adv} \in \mathbb{R}^{C \times H \times W}$ denote the noise map.
Initialize $n_{adv} \leftarrow 0$
while $iter$ times **do**
  ▷ **Identity loss w.r.t. original**
  $\mathcal{L}_{ori} \leftarrow \frac{1}{r} \sum_{i=1}^{r} \cos\left(M_i(x_u + n_{adv}),\ M_i(x_u)\right)$
  ▷ **Identity loss w.r.t. target**
  $\mathcal{L}_{tgt} \leftarrow \frac{1}{r} \sum_{i=1}^{r} \left[1 - \cos\left(M_i(x_u + n_{adv}), M_i(t_u)\right)\right]$
  ▷ **Quality loss**
  $\mathcal{L}_{qual} \leftarrow \text{MSE}(x_u + n_{adv},\ x_u)$
  ▷ **Total loss**
  $\mathcal{L}_{total} \leftarrow \lambda_{id}(\lambda_{ori}\mathcal{L}_{ori} + \lambda_{tgt}\mathcal{L}_{tgt}) + \lambda_{qual}\mathcal{L}_{qual}$
  ▷ **Masked gradient update**
  $g \leftarrow \nabla_{n_{adv}} \mathcal{L}_{total}$
  $n_{adv} \leftarrow n_{adv} - \eta \cdot (B_u \odot g)$
  $n_{adv} \leftarrow \text{clip}(n_{adv}, -\epsilon, \epsilon)$
end while
return $x_u^p \leftarrow x_u + n_{adv}$

over them.

As the official implementation of RMT-GAN [7] is not publicly available[4] , we could not reproduce its full pipeline for extensive evaluation. However, we implemented the proposed adaptive target selection strategy (Algorithm 2) and included it in our ablation studies. Following the original paper, we set the pre-defined ratio (ra) to 20%.

4.2. Face-swapping Deepfakes

Fig. 4 shows randomly selected target images used to measure the Protection Success Rate (PSR) in face-swapping deepfakes. For all PSR scores in the main paper, the reported values represent the average results across three targets per dataset. Additionally, the implementation details of the three deepfake tools used in our evaluation—UniFace, INSwapper, and SimSwap—are as follows.

UniFace [11]. We use the official codes of UniFace[5]. Input images are cropped using a dlib library. For face swapping, the cropped face is resized to (256, 256) before processing and the swapped face is generated at a resolution of (256, 256).

INSwapper [1]. We implement INSwapper in our experiments by official codes of InsightFace[6]. The buffalo_l model pack provided by InsightFace is used for face detection and recognition. During the detection process, images are resized to (320, 320) before detection is performed. For face swapping, the detected face is normalized and cropped to (128, 128) before processing and the swapped face is generated at a resolution of (128, 128).

SimSwap [2]. We use the official codes of SimSwap[7]. Dur-

Figure 3. Target identities. The top row shows images used during training, while the bottom row presents distinct images used for evaluation.

Algorithm 2 Target selection proposed in RMT-GAN [7]

Input: source face image s

Parameters: 68 landmarks of s l_s, landmarks images in target set T, L_2 distance between l_s and l_t $dis_{s,t}$, adaptive threshold τ, pre-defined ratio ra, targets with distances less than τ $T_{<\tau}$

Output: adaptive target t

$l_s \leftarrow lm(s)$
$l_T \leftarrow lm(T)$
for $t \in T$ **do**
  $dist_{s,t} \leftarrow ||l_s, l_t||_2$
end for
$\tau \leftarrow Cal\_thre(ra)$
$T_{<\tau} \leftarrow t \in T$ **where** $dist_{s,t} < \tau$
$random\_select(T_{<\tau})$

[4] https://github.com/tttianyu/RMT-GAN

[5] https://github.com/xc-csc101/UniFace

[6] https://github.com/deepinsight/insightface

[7] https://github.com/neuralchen/SimSwap

ing the detection process, images are resized to (320, 320) before detection is performed. The swapped face is generated at a resolution of (224, 224).

Figure 4. Deepfake target images for measuring the protection success rate (PSR) against three face-swapping deepfakes (UniFace, INSwapper, SimSwap).

5. Transformation Robustness Evaluation

Tbl. 1 reports cosine similarity changes under common post-processing transformations (JPEG compression, Gaussian blur, and resizing, each at five intensity levels). Across all cases, our perturbations remain stable, showing less than 10% degradation in similarity and under 1% for resizing. This confirms the robustness of our method against typical post-processing transformations (See Fig. 5.).

Table 1. Robustness against Post-Processing Transformations.

	Quality factor (%)					
	100	90	70	50	30	**10**
JPEG Compression	0.1260 (baseline)	0.1261 (+0.08%)	0.1264 (+0.32%)	0.1267 (+0.56%)	0.1269 (+0.71%)	0.1332 (+5.71%)
	Scaling ratio					
	1.0	0.9	0.7	0.5	0.3	0.1
Resize	0.1260 (baseline)	0.1260 (+0.00%)	0.1260 (+0.00%)	0.1258 (-0.16%)	0.1257 (-0.24%)	0.1263 (+0.24%)
	Sigma (σ)					
	0.0	1.0	2.0	3.0	4.0	5.0
Gaussian Blur	0.1260 (baseline)	0.1259 (-0.08%)	0.1267 (+0.56%)	0.1289 (+2.30%)	0.1335 (+5.95%)	0.1394 (+9.89%)

Cosine similarity average (cosine similarity variation)

6. Analysis of Controllability Parameters

To investigate the controllability of our approach in adjusting both protection effectiveness and image quality, we conduct experiments by varying ϵ and *iter*. As shown in Fig. 6, protection effectiveness rises sharply with *iter* up to 200, then stabilizes and even slightly decreases. In contrast, image quality steadily declines as *iter* increases further. This suggests that an excessive number of iterations may interfere with achieving optimal adversarial noise generation. Regarding ϵ, higher values enhance protection effectiveness but degrade image quality, as ϵ controls the intensity of noise at specific locations.

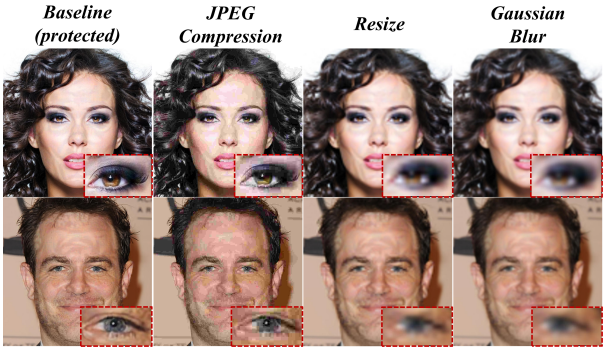

Figure 5. Visual examples of post-processing transformations applied to protected images: baseline, JPEG compression (quality = 10%), resizing (scaling ratio = 0.1), and Gaussian blur (σ = 2.5).

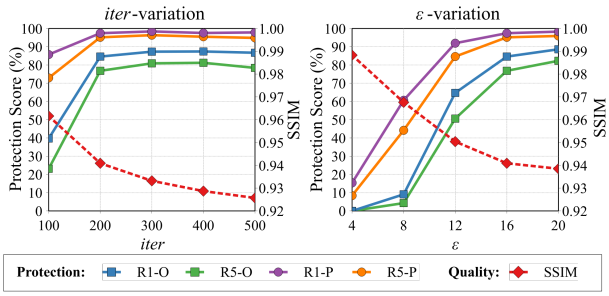

Figure 6. Effect of noise magnitude (ϵ) and iteration ($iter$).

7. Analysis under stressful conditions

Our pipeline, adaptive target synthesis and masked adversarial attack, benefits from accurate face masking, which enables coherent latent and spatial optimization. To assess robustness under imperfect masking conditions, we construct a conservative stress set ($\mathcal{D}_{sts}$) on CelebA-HQ by flagging images with missing face elements (99 images, $\approx$ 10%) and use the remainder as the standard set ($\mathcal{D}_{std}$) for 901 images. For each split, we measure the perceptual shift between original and protected images using FID (lower is better) and assess protection under three face swapping deepfakes using PSR (higher is better).

Table 2. Robustness against Stressful Conditions.

	PSR (↑) (CS (↓))						FID (↓)	
	UniFace [11]		**INSwapper** [1]		**SimSwap** [2]			
	$\mathcal{D}_{sts}$	$\mathcal{D}_{std}$	$\mathcal{D}_{sts}$	$\mathcal{D}_{std}$	$\mathcal{D}_{sts}$	$\mathcal{D}_{std}$	$\mathcal{D}_{sts}$	$\mathcal{D}_{std}$
origin	0.00 (0.78)	0.00 (0.83)	0.01 (0.67)	0.00 (0.74)	0.00 (0.58)	0.00 (0.63)	26.99	19.10
protected	0.74 (0.15)	0.92 (0.05)	0.75 (0.18)	0.84 (0.12)	0.87 (0.10)	0.95 (0.04)		

CS indicates the average cosine similarity on IRSE50 between deepfake and original images.
PSR indicates protection success rate at a 0.241 threshold (FAR@0.01)

As shown in Tab. 2, $\mathcal{D}_{sts}$ exhibits stronger perceptual degradation and weaker protection: FID increases from

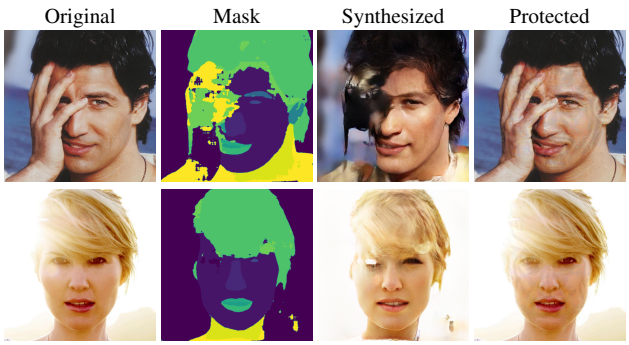

Figure 7. $\mathcal{D}_{sts}$ examples (occlusion and excessive brightness) showing original image, segmentation mask, synthesized target, and protected output. Despite imperfect masking conditions, ***Phantom*** maintains localized and effective perturbations, demonstrating practical robustness.

19.10 on the $\mathcal{D}_{std}$ to 26.99 on the $\mathcal{D}_{sts}$, and swap-level PSR is consistently lower. As shown in Fig. 7, imperfect masks introduce minor spatial misalignment and slightly contaminate geometry cues, which in turn produce mildly off-identity synthesized targets and place some visible noise in localized regions of the protected image.

These effects can steer the optimization slightly away from the identity axis, jointly worsening PSR and FID under stress. Nevertheless, our method still outperforms baselines for two reasons. First, even with imperfect masks, most facial regions remain covered, so the noise budget still concentrates on salient ROIs, limiting the impact of boundary errors. Second, adaptive target synthesis produces targets better aligned in style and geometry with the source than random or fixed targets, which in turn helps preserve identity-axis guidance during optimization. As a result, $\mathcal{D}_{sts}$ PSR remains competitive (0.737 for UniFace, 0.747 for INSwapper, 0.869 for SimSwap), and our method outperforms alternative baselines reported in Tabs. 2 and 3 on the main paper.

In future work, we plan to enhance the pipeline under challenging conditions by deploying a segmentation ensemble to achieve more precise and consistent masking, or by replacing the fixed parser with a lightweight, learnable segmentation module jointly trained with the protection objective in an end-to-end framework. These improvements aim to stabilize masking and consistently achieve higher PSR and lower FID at the same distortion budget.

8. More Visual Results

We provide qualitative visual examples on the CelebA-HQ and LADN datasets, showcasing the results of applying TIP-IM, AMT-GAN, CLIP2Protect, DiffAM, and ***Phantom*** to the dodging scenario of face-swapping deepfakes, as shown in Fig. 8 and Fig. 9.

TIP-IM frequently produces overly visible noise artifacts, while makeup-based methods such as AMT-GAN and CLIP2Protect often fail to align precisely with the intended prompts, resulting in excessive or misplaced makeup. DiffAM achieves perceptually natural modifications but demonstrates limited dodging effectiveness, struggling to sufficiently obfuscate the source identity. In contrast, ***Phantom*** delivers robust protection against face-swapping attacks with minimal visual distortion, highlighting its advantage in balancing perceptual fidelity and adversarial effectiveness.

References

[1] InsightFace. https://github.com/deepinsight/insightface, 2025. (Accessed on 11/11/2025). 3, 4

[2] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 2003–2011, 2020. 3, 4

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017. 1

[4] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15014–15023, 2022. 2

[5] Ziqi Huang, Kelvin CK Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023. 2

[6] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5549–5558, 2020. 1

[7] Xiyao Liu, Junxing Ma, Xinda Wang, Qianyu Lin, Jian Zhang, Gerald Schaefer, Cagatay Turkay, and Hui Fang. Recoverable facial identity protection via adaptive makeup transfer adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 514–522, 2025. 2, 3

[8] Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. 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 (CVPR)*, pages 20595–20605, 2023. 2

[9] Yuhao Sun, Lingyun Yu, Hongtao Xie, Jiaming Li, and Yongdong Zhang. Diffam: Diffusion-based adversarial makeup transfer for facial privacy protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24584–24594, 2024. 2

[10] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1

[11] Chao Xu, Jiangning Zhang, Yue Han, Guanzhong Tian, Xianfang Zeng, Ying Tai, Yabiao Wang, Chengjie Wang, and Yong Liu. Designing one unified framework for high-fidelity face reenactment and swapping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 54–71. Springer, 2022. 3, 4

[12] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3897–3907, 2021. 2

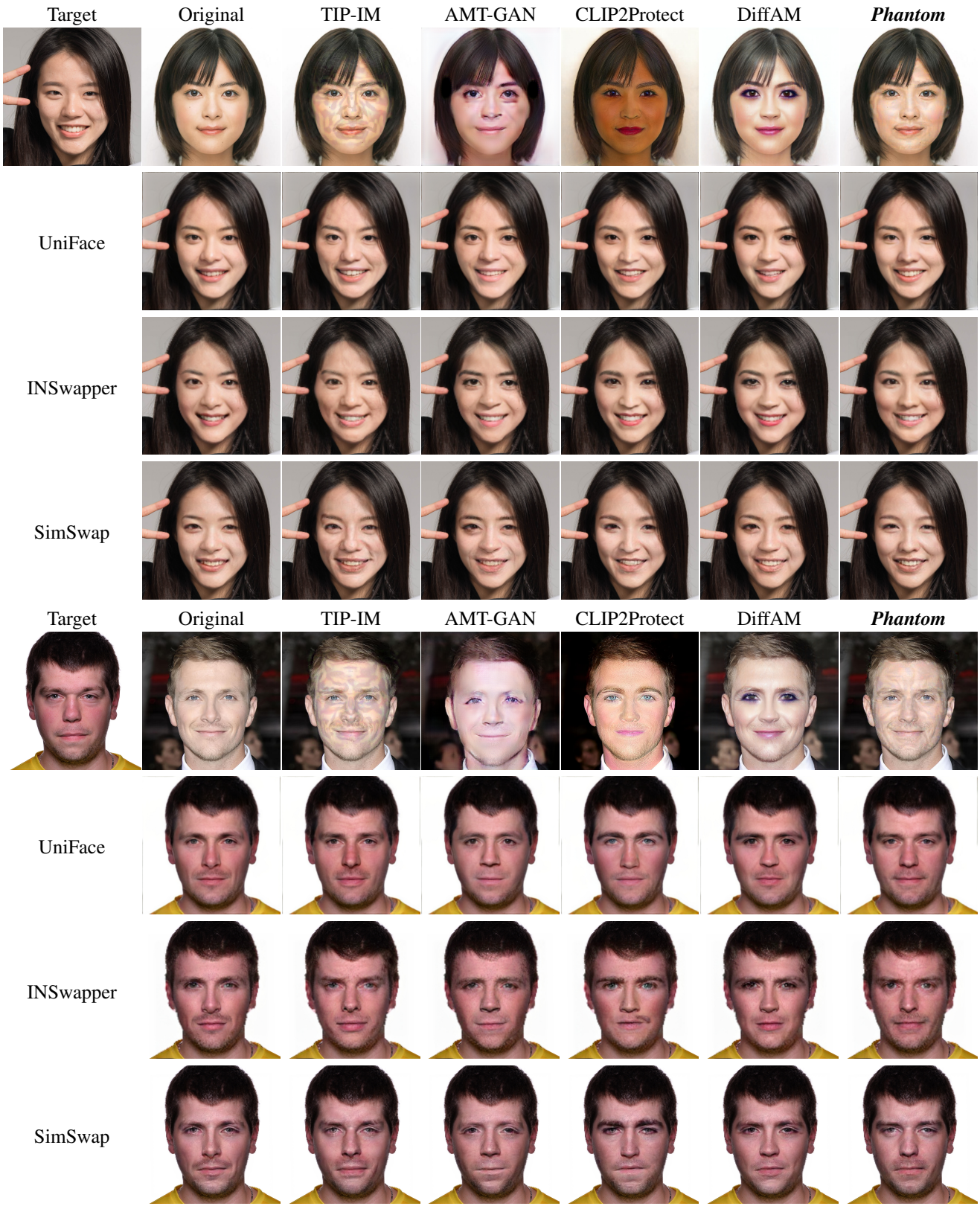

Figure 8. Black-Box Face-Swapping Deepfake Protection Results on the CelebA-HQ Dataset.

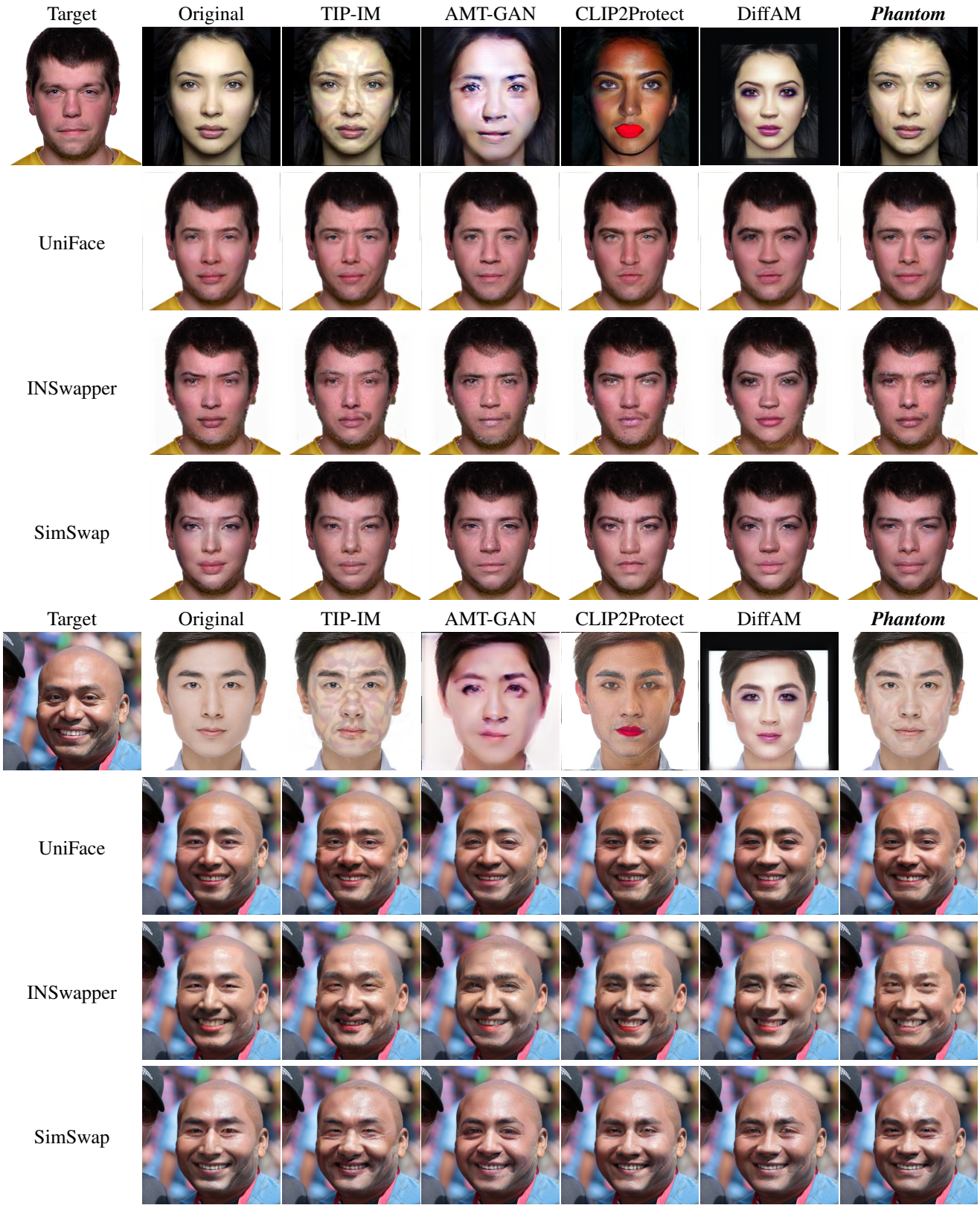

Figure 9. Black-Box Face-Swapping Deepfake Protection Results on the LADN Dataset.