MASQUE: DIFFUSION-BASED LOCALIZED ADVERSARIAL MAKEUP FOR FACIAL PRIVACY

Youngjin Kwon & Xiao Zhang
CISPA Helmholtz Center for Information Security
{youngjin.kwon,xiao.zhang}@cispa.de

ABSTRACT

Facial recognition has been increasingly employed in real-world applications, raising serious privacy concerns over mass surveillance and unauthorized tracking. Existing anti-facial recognition methods perturb face images using generative models to protect privacy but often introduce global artifacts, depend on reference image prompts, or require target identity, compromising both visual quality and anonymity. To address the above limitations, we introduce **MASQUE**, a diffusion-based framework that generates localized adversarial makeup guided by user-defined text prompts. By leveraging precise null-text inversion, targeted cross-attention fusion with masking, and a novel pairwise adversarial guidance mechanism using images of the same individual, **MASQUE** achieves robust dodging performance without the need for an external target identity. Extensive evaluations on open-source FR models and commercial APIs show that **MASQUE** significantly enhances privacy protection over all baselines, achieving average protection success rates of 90% for identification and 87% for verification while preserving high perceptual fidelity.

1 INTRODUCTION

Facial recognition (FR) systems (Parkhi et al., 2015) have been widely adopted across security, biometrics, and commercial applications. However, their unregulated deployment raises serious privacy concerns, as governments and private entities often employ them for mass surveillance and unauthorized tracking. To address these concerns, *anti-facial recognition* (AFR) technologies (Wenger et al., 2023) have emerged to protect users from being identified in online-posted facial images. In particular, adversarial-based methods stand out for their effectiveness in disrupting unauthorized FR models by altering facial images to evade detection. Noise-based approaches (Yang et al., 2021) obscure facial features by crafting norm-bounded global perturbations, while patch-based techniques (Xiao et al., 2021) optimize adversarial patterns in localized image patches. However, these methods usually introduce noticeable visual artifacts, compromising the usability of the produced images.

Built upon state-of-the-art image generative models, makeup-based AFR methods (Yin et al., 2021; Hu et al., 2022; Shamshad et al., 2023; Sun et al., 2024a) strikes a balance by integrating adversarial features into natural and visually appealing facial makeup modifications. Although makeup-based approaches offer promising solutions for safeguarding privacy without compromising aesthetics, they often struggle to preserve the fine-grained details of the original facial images or fail to adhere to the user instructions embedded in different makeup prompts. In addition, these methods require external target identity images to guide the generation process for adversarial makeup transfer, primarily focusing on the impersonation setting. However, when considering the more privacy-relevant dodging scenarios, they often exhibit a significant performance drop in protection success rates, and we argue that the strong reliance on target identity poses additional privacy risks that should be avoided.

Contributions. We propose **MASQUE**, a novel generative method for localized adversarial makeup based on diffusion-based image editing with text guidance, which simultaneously fulfills the following criteria: (i) *Non-Intrusive Identity Protection*: achieve high protection success rates under dodging scenarios while eliminating the need for a specific target identity; (ii) *Localized Perturbation*: preserve fine details of the original facial images while confining adversarial modifications to the intended areas; (iii) *User Control*: strong prompt-following ability and adaptivity to diverse text makeup prompts, offering greater user control and convenience than existing AFR approaches.

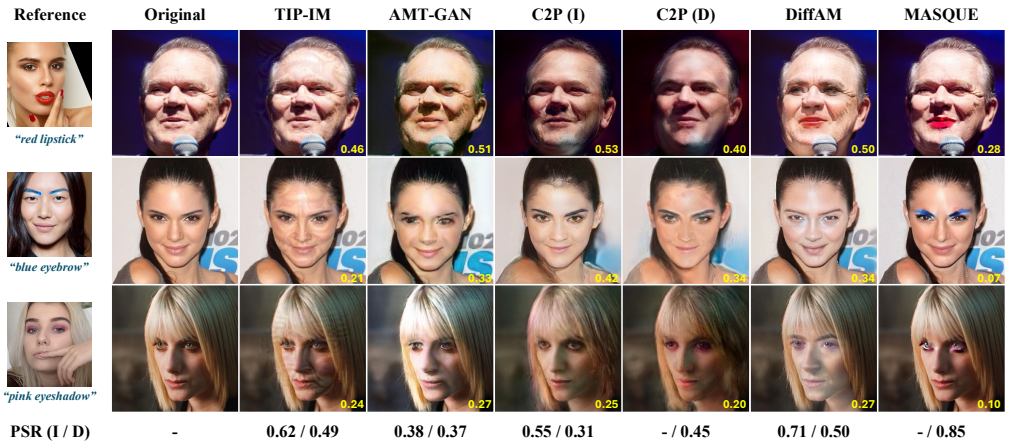

Figure 1: Visualizations of protected images across various AFR methods. The columns (from left to right) stand for the reference image and text makeup prompts, original images, and protected images produced by the respective AFR method. The yellow text denotes the cosine similarity with the corresponding gallery image. Below the images, we demonstrate the average protection success rate (PSR) under impersonation and/or dodging scenarios across four FR models and three makeup styles.

Related Work. We review the prior AFR approaches that are most relevant to ours (see Appendix A for more detailed discussions on related works). Shamshad et al. (2023) proposed CLIP2Protect (C2P), which adopts a vision-language model for text-guided adversarial makeup transfer but struggles with localizing the perturbations, causing global artifacts and loss of fine details such as the background. While C2P supports the functionality of dodging, it still requires a target identity image to guide the generation process to maintain sufficient visual quality. Sun et al. (2024a) proposed DiffAM, which employs diffusion models and two-step makeup transfer, achieving better image quality but relying on reference image prompts, restricting user flexibility compared to text-based methods. Besides, it is designed solely for impersonation, leaving dodging scenarios unaddressed.

2 Limitations of Existing AFR Approaches

In this section, we discuss the limitations of existing AFR approaches, which motivate the design of `MASQUE`. In particular, Figure 1 visualizes and compares the performance of different AFR techniques, including TIP-IM (Yang et al., 2021), AMT-GAN (Hu et al., 2022), C2P (Shamshad et al., 2023), DiffAM (Sun et al., 2024a), and our method (see Appendix B for detailed experimental settings)

Decreased PSR under Dodging Scenario. The goal of AFR is to protect privacy by preventing users from unauthorized facial recognition. Yet most existing works (Yang et al., 2021; Hu et al., 2022; Sun et al., 2024a) only consider impersonation, where FR systems are misled into identifying individuals as specific targets. While Shamshad et al. (2023) provides a dodging variation of their method, it still requires a target identity to ensure the image generation quality. While effective for impersonation, these methods raise ethical concerns, as they allow users to imitate another user's identity without consent, potentially leading to deceptive or harmful misuse. A truly privacy-preserving approach must account for dodging that prevents recognition of the original identity and ensures anonymity without substituting a target identity. In addition, achieving high protection success for impersonation does not necessarily mean effective protection under dodging. As shown in Figure 1, the protection success rate (PSR) achieved by existing AFR methods usually exhibits a noticeable performance drop under dodging scenarios, indicating that focusing solely on impersonation is inadequate.

Limited User Control and Image Quality. Previous AFR methods struggle to accurately apply makeup based on reference text or image images, limiting user control while also struggling to maintain high image quality. As Figure 1 illustrates, AMT-GAN and DiffAM fail to transfer makeup precisely, resulting in noticeable artifacts despite utilizing reference images from a pretrained makeup transfer dataset. C2P, while depending solely on text prompts, lacks fine-grained control over subtle details. Moreover, GAN-based methods tend to modify areas beyond the face, such as altering the background, making them unsuitable for scenarios where users require precise, facial-only editings.

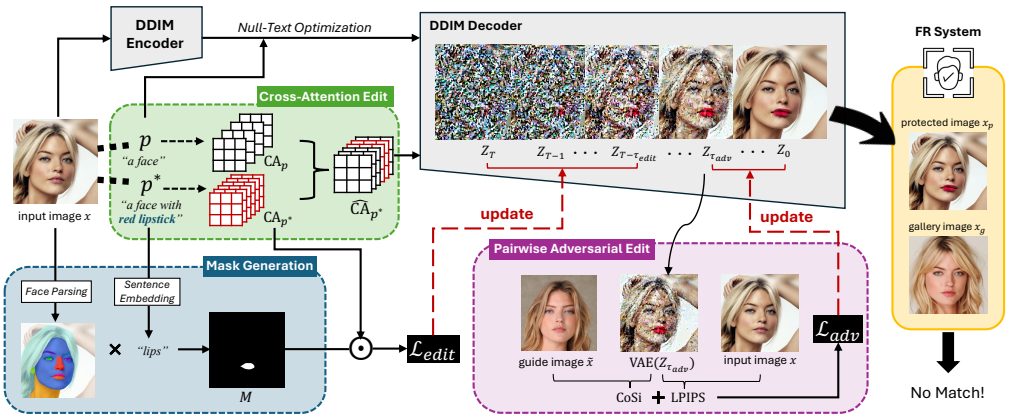

Figure 2: The pipeline of **MASQUE** involves: (1) fusing the editing and reconstruction prompts to produce an updated cross-attention map for diffusion, (2) creating a mask $\mathcal{M}$ to define a target region and optimize an edit loss to maximize makeup-related attention in $\mathcal{M}$, and (3) using pairwise adversarial guidance with same-identity images to enhance identity confusion without external targets.

3 MASQUE: Localized Adversarial Makeup using Diffusion Models

To address the limitations, we introduce **MASQUE**, aiming to disrupt FR models by crafting adversarial makeup guided by a user-defined text prompt p^*. The prompt directs realistic, localized changes, while our novel adversarial guidance ensures these perturbations mislead FR systems without sacrificing visual fidelity. Figure 2 illustrates the overall pipeline of **MASQUE** (see Appendix D for details).

Applying Localized Makeup. Before applying edits, we obtain a faithful latent representation of the original image $\boldsymbol{x}$ using *null-text inversion* (Mokady et al., 2023), which mitigates reconstruction errors common in direct DDIM inversion (Song et al., 2020). By conditioning on an empty prompt, we align the forward and reverse diffusion trajectories, ensuring near-perfect reconstruction of the original image's structure and identity. With an accurate latent representation in hand, we introduce makeup attributes based on the text prompt p^* by manipulating the *cross-attention* (CA) layers of the diffusion model (Rombach et al., 2022). Note that CA layers control how spatial features correspond to semantic tokens. At each diffusion step τ, we extract attention maps A_τ (reconstruction) and A_τ^* (editing) and blend them: preserving CA values from A_τ for shared tokens to maintain structure, while incorporating values from A_τ^* for makeup-specific tokens in p^*:

$$\big(\text{Update}(A_\tau, A_\tau^*)\big)_{i,j} := \begin{cases} (A_\tau)_{i,j}, & \text{if token } j \text{ is in both } p \text{ and } p^*, \\ (A_\tau^*)_{i,j}, & \text{if token } j \text{ is unique to } p^*, \end{cases}$$

where p denotes the original text prompt. This results in $\hat{A}_\tau$, a set of mixed CA maps that preserve the original facial layout while steadily introducing adversarial makeup features (Hertz et al., 2022).

Enhancing Semantic Edits and Locality. To ensure precise localization, we generate a mask $\mathcal{M}$ that defines the region for modification. To achieve this, we embed the prompt p^* using a Sentence Transformer (Reimers & Gurevych, 2019) model and compare it to embeddings of predefined facial regions. The closest match determines the relevant area for the edit. For instance, if the prompt specifies "a face with red lipstick", the model identifies lips as the target and generates a lip-area mask. Once the target region $\mathcal{M}$ is determined, we enhance the influence of makeup-related tokens by maximizing their attention within $\mathcal{M}$. Constraining edits to this region prevents unintended modifications and preserves overall image quality (Mao et al., 2023). Specifically, we optimize:

$$\mathcal{L}_{\text{edit}} := \left(1 - \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \frac{(A_\tau^*)_{i,\text{new}}}{(A_\tau^*)_{i,\text{new}} + \sum_{j \in \text{share}} (A_\tau)_{i,j}}\right)^2,$$

where $(A_\tau^*)_{i,\text{new}}$ denotes the attention weights for the new makeup tokens at spatial index i, and $\sum_{j \in \text{share}} (A_\tau)_{i,j}$ is the sum of attention weights corresponding to tokens shared by both the original and makeup prompts. By prioritizing makeup-specific attention in the masked region $\mathcal{M}$, this loss term ensures the final protected image $\boldsymbol{x}_\text{p}$ has realistically applied localized adversarial makeup.

Table 1: Comparison of PSR and image quality across various AFR methods. PSR is evaluated in a black-box dodging scenario for both identification (first value) and verification (second value) tasks.

Method	Protection Success Rate (PSR) ↑					Image Quality		
	IR152	IRSE50	MobileFace	FaceNet	Face++	LPIPS ↓	PSNR ↑	SSIM ↑
Clean	0.10/0.05	0.13/0.05	0.24/0.10	0.05/0.04	0.01	–	–	–
TIP-IM	0.62/0.44	0.86/0.60	0.74/0.40	0.65/0.51	0.28	0.31	**32.16**	**0.93**
AMT-GAN	0.61/0.40	0.48/0.32	0.57/0.26	0.51/0.52	0.60	0.34	19.51	0.61
C2P (I)	0.31/0.12	0.38/0.11	0.57/0.21	0.22/0.14	0.16	0.46	18.92	0.58
C2P (D)	0.75/0.52	0.76/0.47	0.77/0.42	0.56/0.40	0.74	0.47	17.99	0.56
DiffAM	0.76/0.50	0.73/0.41	0.80/0.45	0.58/0.63	0.64	0.40	18.31	0.77
MASQUE	**0.98 / 0.96**	**0.89 / 0.97**	**0.94 / 0.79**	**0.78 / 0.76**	**0.84**	**0.29**	25.82	0.86

Pariwise Adversarial Guidance. While our makeup edits ensure semantic plausibility, the primary adversarial goal is to disrupt FR models. Previous AFR methods often target another identity, compromising privacy and limiting applicability in dodging scenarios. By contrast, our approach leverages a pair $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ of face images from the same individual, where $\tilde{\boldsymbol{x}}$ serves as the guide image. This strategy highlights a significant issue with naively using the distance from the original image as the adversarial loss. Note that in the standard diffusion process, the objective is to generate images similar to the original. If we merely maximize the distance from the original image as an adversarial loss, it potentially creates conflicting objectives, leading to an unstable performance in both image quality and adversarial effectiveness (see Table 4 in Appendix B for supporting experiments).

Adversarial and Quality Trade-off. We introduce adversarial perturbations during the later stage of the diffusion process, ensuring coarse structure remains intact while subtly altering identity-specific features. To balance adversarial potency with visual fidelity, we incorporate perceptual similarity constraints alongside a *cosine similarity* (CoSi) measure that captures adversarial effectiveness:

$$\mathcal{L}_{\text{adv}} := \lambda_{\text{CoSi}} \cdot \text{CoSi}(\boldsymbol{x}_{\text{p}}, \tilde{\boldsymbol{x}}) + \lambda_{\text{LPIPS}} \cdot \text{LPIPS}(\boldsymbol{x}_{\text{p}}, \boldsymbol{x})$$

where $\boldsymbol{x}_{\text{p}}$ denotes the protected image. CoSi ensures that the adversarial perturbations sufficiently diverge from recognizable identity features, while LPIPS (Zhang et al., 2018) maintain perceptual and structural fidelity, respectively. The parameters $\lambda_{\text{CoSi}}, \lambda_{\text{LPIPS}}$ allow fine-tuning of this trade-off.

4 Experiments

We evaluate `MASQUE` using images from CelebA-HQ and compare it against several AFR baselines on both public and commercial FR models. We report PSR for verification and Rank-1 accuracy for identification. Detailed descriptions of our experimental settings are provided in Appendix B.

Protection Success under Dodging. Table 1 presents the protection success rate (PSR) under dodging scenarios across both face verification and identification settings, evaluated in black-box scenarios using four widely used pretrained FR feature extractors (Deng et al., 2019; Hu et al., 2018; Schroff et al., 2015; Chen et al., 2018). For each target model, the remaining three serve as surrogate models, with results averaged across three makeup styles. Our method significantly outperforms all baselines, achieving an average PSR of 89.67% for identification and 86.92% for verification. In addition, we evaluate our approach against the commercial Face++ API in verification mode, which assigns similarity scores from 0 to 100 with dynamic thresholds. As a proprietary model with unknown training data and parameters, it serves as a realistic testbed for evaluating the effectiveness of AFR methods. Our approach achieves the highest protection success rates, demonstrating effectiveness in both open-source and closed-source settings, reinforcing its real-world applicability.

Advantages in Visual Aspects. Our method achieves superior image quality across multiple evaluation metrics, as summarized in Table 1 and illustrated in Figure 1. While TIP-IM attains the highest PSNR and SSIM due to its small perturbation budget, these pixel-level metrics often fail to reflect perceptual quality. In contrast, our approach prioritizes perceptual consistency, balancing content fidelity and visual realism, as demonstrated by its strong LPIPS performance. We also confirm the advantages of `MASQUE` over existing AFR methods in achieving precise, localized makeup transfer and better alignment in terms of prompt following (see additional experiments in Appendix C).

REFERENCES

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4):1–11, 2023.

Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*, pp. 428–438. Springer, 2018.

Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. *arXiv preprint arXiv:2101.07922*, 2021.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014a.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

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*, pp. 15014–15023, 2022.

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.

Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 645–653, 2018.

Qingjie Liu, Huanyu Zhou, Qizhi Xu, Xiangyu Liu, and Yunhong Wang. Psgan: A generative adversarial network for remote sensing image pan-sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12):10227–10242, 2020.

Qi Mao, Lan Chen, Yuchao Gu, Zhen Fang, and Mike Zheng Shou. Mag-edit: Localized image editing in complex scenarios via mask-based attention-adjusted guidance. *arXiv preprint arXiv:2312.11396*, 2023.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 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.

Thao Nguyen, Anh Tuan Tran, and Minh Hoai. Lipstick ain't enough: beyond color matching for in-the-wild makeup transfer. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 13305–13314, 2021.

Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 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.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

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*, pp. 20595–20605, 2023.

Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*, pp. 1589–1604, 2020.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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*, pp. 24584–24594, 2024a.

Zhaoyang Sun, Shengwu Xiong, Yaxiong Chen, Fei Du, Weihua Chen, Fan Wang, and Yi Rong. Shmt: Self-supervised hierarchical makeup transfer via latent diffusion models. *arXiv preprint arXiv:2412.11058*, 2024b.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1921–1930, 2023.

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.

Emily Wenger, Shawn Shan, Haitao Zheng, and Ben Y Zhao. Sok: Anti-facial recognition technology. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 864–881. IEEE, 2023.

Jianfeng Xiang, Junliang Chen, Wenshuang Liu, Xianxu Hou, and Linlin Shen. Ramgan: region attentive morphing gan for region-level makeup transfer. In *European Conference on Computer Vision*, pp. 719–735. Springer, 2022.

Zihao Xiao, Xianfeng Gao, Chilin Fu, Yinpeng Dong, Wei Gao, Xiaolu Zhang, Jun Zhou, and Jun Zhu. Improving transferability of adversarial patches on face recognition with generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11845–11854, 2021.

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*, pp. 3897–3907, 2021.

Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. *arXiv preprint arXiv:2105.03162*, 2021.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 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.

Yuxuan Zhang, Lifu Wei, Qing Zhang, Yiren Song, Jiaming Liu, Huaxia Li, Xu Tang, Yao Hu, and Haibo Zhao. Stable-makeup: When real-world makeup transfer meets diffusion model. *arXiv preprint arXiv:2403.07764*, 2024.

A Background and Related Work

A.1 Anti-Facial Recognition

The problem task of AFR can be characterized as a two-party security game between an attacker and a defender. The attacker aims to recognize online face images belonging to certain target victims' identities using facial recognition models, whereas the defender adopts AFR techniques to protect the victims' face images from being identified by the attacker. Built on the foundational concept of adversarial examples (Szegedy et al., 2013; Goodfellow et al., 2014b), numerous AFR methods have been proposed, targeting different stages of the facial recognition pipeline (Wenger et al., 2023). In particular, we focus on *adversarial-based AFR methods* that craft naturalistic adversarial perturbations to the victim's facial images to fool black-box FR models:

Definition A.1 (Adversarial-based AFR)**.** Let $\mathrm{FR} : \mathcal{X} \to \mathcal{Y}$ be a facial recognition model, which maps any face image $\boldsymbol{x} \in \mathcal{X}$ to an identity $y \in \mathcal{Y}$. Let $\mathcal{D}$ be a collection of clean face images, each paired with the corresponding ground-truth identity. Then the objective of adversarial-based AFR is to learn a perturbation function $\mathrm{AFR} : \mathcal{X} \to \mathcal{X}$ with respect to the following optimization problem:

$$\max \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x},y^*)\in\mathcal{D}} \mathbb{1}\big\{\mathrm{FR}\big(\mathrm{AFR}(\boldsymbol{x})\big) \neq y^*\big\} \quad \text{s.t.} \ \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x},y^*)\in\mathcal{D}} \Delta\big(\mathrm{AFR}(\boldsymbol{x}); \boldsymbol{x}\big) \leq \gamma, \tag{1}$$

where y^* denotes the ground-truth identity of the input image $\boldsymbol{x}$, $\Delta\big(\mathrm{AFR}(\boldsymbol{x}); \boldsymbol{x}\big)$ captures the visual quality of the perturbed image $\mathrm{AFR}(\boldsymbol{x})$ with reference to the original face image $\boldsymbol{x}$, and $\gamma > 0$ is a threshold parameter indicating how much distortion can be tolerated.

A desirable adversarial-based AFR method is expected to attain a high protection success rate (PSR) on any collection of user face images against unauthorized FR models, specified by the optimization objective in Equation 1. Since the defender does not have the underlying knowledge of the attacker's actual FR model, we consider the black-box FR scenario when evaluating the PSR of an AFR method. From the user's perspective, perturbed images generated by $\mathrm{AFR}(\cdot)$ should look natural or even visually similar to the corresponding original face images, which is characterized by the optimization constraint in Equation 1. Otherwise, if $\mathrm{AFR}(\boldsymbol{x})$ has an unsatisfying image quality, users may not be willing to post the protected face image on their social media, even if a high PSR has been achieved.

Existing AFR Literature. Earlier works on AFR proposed to use obfuscation techniques to obscure the facial identity features or craft ℓ_p-norm bounded perturbations to fool FR models (Yang et al., 2021). While effective, these methods often compromise image quality, limiting their practicality in real-world scenarios. Poisoning-based methods (Cherepanova et al., 2021; Shan et al., 2020) introduced a new approach by injecting subtle adversarial noise into images to degrade the effectiveness of recognition models. These techniques excel at disrupting model training or inference without visibly altering the image, but their reliance on model-specific perturbations limits their generalizability across diverse architectures. Unrestricted adversarial examples, which aim to create natural-looking modifications, marked a significant advancement by integrating adversarial signals into realistic alterations without relying on specific constraints. These strategies leverage the flexibility of generative models to balance visual coherence and adversarial effectiveness.

Adversarial Makeup Transfer. Adversarial makeup transfer (Yin et al., 2021; Hu et al., 2022; Shamshad et al., 2023; Sun et al., 2024a) has emerged as a practical solution to realize the goal of facial privacy protection. These methods leverage the strong generative capability of generative models to embed adversarial perturbations into natural makeup-based facial modifications, deceiving attackers' facial recognition models while largely preserving the aesthetic appeal. For example, Hu et al. (2022) proposed AMT-GAN, which introduces a regularization module and a joint training pipeline for

adversarial makeup transfer within the generative adversarial network (GAN) framework (Goodfellow et al., 2014a). Recent advancements in generative models have further enhanced the performance of adversarial makeup transfer, such as Clip2Protect (Shamshad et al., 2023) and DiffAM (Sun et al., 2024a), which adopt text-guided StyleGAN model and diffusion-based framework respectively to enable seamless adversarial face modifications with much improved visual quality. In this work, we build upon these developments by leveraging diffusion models to generate visually consistent, localized adversarial makeup for privacy protection in dodging scenarios.

A.2 MAKEUP TRANSFER

Recent advancements in makeup transfer have moved from GAN-based methods to more advanced diffusion models. Traditional approaches like BeautyGAN (Li et al., 2018) and PSGAN (Liu et al., 2020) use histogram matching, attention mechanisms, and spatial encodings to transfer makeup while preserving structure. CPM (Nguyen et al., 2021) expands beyond color transfer using UV mapping, while RamGAN (Xiang et al., 2022) ensure component consistency. However, GAN-based models struggle with extreme styles and rely on imprecise pseudo-paired data, limiting fidelity.

To overcome these challenges, diffusion-based methods like SHMT (Sun et al., 2024b) and Stable-Makeup (Zhang et al., 2024) offer improved realism and robustness. SHMT eliminates reliance on pseudo-paired data through a self-supervised "decoupling-and-reconstruction" framework, leveraging a Laplacian pyramid for hierarchical texture transfer. Stable-Makeup employs a diffusion model with a Detail-Preserving encoder and cross-attention layers to ensure precise, structure-preserving makeup application. These approaches set a new standard for high-fidelity, real-world makeup transfer. However, our study takes a different approach—rather than transferring makeup from a reference image, we focus on text-guided makeup generation using diffusion models, enabling intuitive and flexible image editing to create customizable makeup based solely on textual descriptions. This also eliminates the need for model fine-tuning on specific makeup images, allowing us to generate the desired makeup directly during inference, resulting in a more efficient and adaptable approach.

A.3 DIFFUSION-BASED IMAGE EDITING

Stable Diffusion (Rombach et al., 2022) generates high-quality images by denoising a random latent $\boldsymbol{z}_T$ conditioned on a text embedding C. The model is trained to predict the added noise $\boldsymbol{\epsilon}$ via:

$$\min_{\theta} \mathbb{E}_{\boldsymbol{z}_0,\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}),t\sim\mathrm{Unif}(1,T)}\left[\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t,t,C)\|_2^2\right].$$

Localized editing methods extend diffusion models for targeted modifications. Mask-based approaches (Avrahami et al., 2023; Couairon et al., 2022) constrain edits to specific regions using spatial masks, preserving unedited areas but often struggling with structural consistency. Attention-based approaches (Hertz et al., 2022; Tumanyan et al., 2023) guide edits via attention injection, achieving better global structure preservation but suffering from unintended changes (editing leakage). Our proposed method, **MASQUE**, utilizes both mask-based and attention-based guidance without any fine-tuning, combining mask precision with attention flexibility, ensuring semantically consistent and region-specific modifications while addressing the limitations of prior approaches.

B DETAILED EXPERIMENTAL SETTINGS

Dataset. We conduct experiments using 300 face images of 1024×1024 resolution from the CelebA-HQ dataset (Karras et al., 2017) to account for the image uploading situation. We randomly select 100 identities from the dataset, with three images per identity. For each identity, one image serves as the probe image to be protected, another as the reference image for guidance, and the third as the gallery image, which is assumed to be stored in the facial recognition system for comparison.

Configuration. For performance comparison, we evaluate against the following baselines: TIP-IM (Yang et al., 2021), the SOTA method of noised-based AFR, and other recent generative-based AFR methods, including AMT-GAN (Hu et al., 2022), Clip2Protect (Shamshad et al., 2023), and DiffAM (Sun et al., 2024a). For Clip2Protect, we denote it as C2P for simplicity, where both the impersonating version (denoted as C2P (I)) and the dodging version (i.e., C2P (D)) are assessed. We

Table 2: Comparison of perceptual similarity metrics across in-mask and out-mask regions

Metric	Method	In-Mask	Out-Mask	Diff Δ
DISTS (↓)	TIP-IM	2.0125	0.0937	1.9188
	AMT-GAN	7.9930	0.1493	7.8437
	C2P (I)	8.9453	0.1532	8.7921
	C2P (D)	10.0111	0.1688	9.8423
	DiffAM	9.6447	0.1409	9.5038
	MASQUE	12.3058	0.1039	**12.2019**
LPIPS (↓)	TIP-IM	0.3217	0.3046	0.0171
	AMT-GAN	0.3602	0.3159	0.0443
	C2P (I)	0.4002	0.4468	−0.0466
	C2P (D)	0.4355	0.4604	−0.0249
	DiffAM	0.5089	0.4098	0.0991
	MASQUE	0.4343	0.2900	**0.1443**
PieAPP (↓)	TIP-IM	3.6214	0.4623	3.1591
	AMT-GAN	21.3250	0.8136	20.5114
	C2P (I)	17.1863	1.1375	16.0488
	C2P (D)	20.8504	1.3140	19.5364
	DiffAM	17.5233	1.3123	16.2110
	MASQUE	21.4338	0.6445	**20.7893**

compare the performance of `MASQUE` with existing AFR techniques on four public FR models: IR152 (Deng et al., 2019), IRSE50 (Hu et al., 2018), FaceNet (Schroff et al., 2015), and MobileFace (Chen et al., 2018). We also evaluate AFR methods against a commercial FR API: Face++.[1]

Evaluation Metric. In alignment with previous studies, we utilize the *Protection Success Rate (PSR)* as the primary evaluation metric. To compute PSR, we adopt the thresholding strategy for face verification and a closed-set strategy for face identification. For face verification, PSR is defined as:

$$\mathrm{PSR} = \frac{1}{N} \sum_{\boldsymbol{x}} \mathbb{I}\big(\cos(\mathrm{FR}(\boldsymbol{x}_\mathrm{p}), \mathrm{FR}(\boldsymbol{x}_\mathrm{g})) > \tau \big) \times 100\%,$$

where $\mathbb{I}$ is the indicator function, N is the number of face images $\boldsymbol{x}$, FR represents the target face recognition model, $\boldsymbol{x}_\mathrm{p}$ denotes the protected image corresponding to $\boldsymbol{x}$, $\boldsymbol{x}_\mathrm{g}$ stands for the gallery image of the corresponding identity, and τ is the similarity threshold. The value of τ is set at 0.01 False Acceptance Rate (FAR) for each victim model. For face identification, we measure the *Rank-1 Accuracy*, which indicates whether the top-1 candidate list excludes the same identity as the original image $\boldsymbol{x}$, thus signifying a successful dodging attack. We also employ metrics like LPIPS, PSNR, and SSIM (Wang et al., 2004) to assess the quality of the generated AFR-protected images. A lower LPIPS score indicates higher perceptual similarity, while higher PSNR and SSIM values signify better pixel-wise and structural alignment with the original images.

Implementation Detail. `MASQUE` builds on the pre-trained Stable Diffusion v1.4 model, using DDIM denoising over $T = 50$ steps with a fixed guidance scale of 7.5. During the backward diffusion process, CA injection occurs in $[T, T - \tau_\mathrm{attn}]$ ($\tau_\mathrm{attn} = 40$), localize optimization in $[T, T - \tau_\mathrm{edit}]$ ($\tau_\mathrm{edit} = 5$), and adversarial guidance in $[T - \tau_\mathrm{adv}, 0]$ ($\tau_\mathrm{adv} = 45$). We set $\lambda_\mathrm{CoSi} = 0.1$, $\lambda_\mathrm{LPIPS} = 1$, and cap optimization to $\mathrm{max\_adv} = 15$ iterations.

C Additional Experiments

Localized Edit. To evaluate whether edits are confined to the desired region, images are divided into in-mask and out-mask regions using a binary mask, isolating the corresponding areas in both images via element-wise multiplication. The areas are then used to normalize similarity metrics proportionally, ensuring fair comparisons. Specifically, we employ DISTS, LPIPS, and PieAPP as evaluation metrics in this experiment: DISTS measures perceptual dissimilarity based on structure and texture, LPIPS uses deep features, and PieAPP reflects human perceptual preferences.

[1] `https://www.faceplusplus.com/face-comparing/`

Table 3: Evaluation of model prompt adherence using CLIP similarity scores.

Makeup Prompt	AMT-GAN	C2P (I)	C2P (D)	DiffAM	MASQUE
"red lipstick"	0.0115	0.0497	0.0516	0.0424	0.0499
"blue eyebrow"	0.0093	0.0426	0.0437	0.0215	0.0344
"pink eyeshadow"	0.0146	0.0413	0.0390	0.0355	0.0498

Table 4: Comparison of identification confidence, verification similarity, and image quality metrics across different numbers of guide images employed in `MASQUE`.

# Guide Images	Identification Conf. (↑)	Verification Sim. (↓)	LPIPS (↓)	PSNR (↑)
0	0.1143	0.3654	0.3026	24.3865
1	0.1348	0.3488	0.2915	**25.8278**
10	**0.2069**	**0.2658**	**0.2913**	25.8277

Table 2 compares the effectiveness of localized edits, where higher Δ values indicate stronger localization of perturbations. TIP-IM (Yang et al., 2021), as a noise-based method, applies pixel-wise adversarial perturbations uniformly, resulting in the smallest Delta for all metrics due to minimal distinction between in-mask and out-mask regions. In contrast, our method introduces significant perturbations in the in-mask region, leading to poorer metrics there, but achieves the best or second-best results in the out-mask region, indicating minimal disruption to untouched areas.

Prompt Following Ability. To evaluate the model's ability to follow prompts, we conduct experiments to measure whether the specified modifications appeared in the intended regions of the image. Specifically, we utilize the CLIP model to compute the similarity between the image and the textual prompt. The evaluation metric involves comparing the CLIP similarity scores before and after modification. An increase in similarity indicates that the intended change was successfully applied, providing a quantitative measure of the model's effectiveness in generating prompt-aligned alterations. Table 3 demonstrates the advantages of our method compared with existing approaches

Pairwise Adversarial Guidance. `MASQUE` introduces pairwise adversarial guidance as a means to protect identities without introducing external target references, distinguishing it from impersonation attacks that modify images to resemble another individual. By comparing and aligning features between these two images, we identify and manipulate identity-relevant cues directly, injecting subtle adversarial signals that steer the diffusion process away from a recognizable identity manifold. The guide image acts as a proxy for the gallery image, providing guidance on the direction in which the latent representation should be altered. This approach eliminates reliance on external identities, avoiding any ethical concerns surrounding impersonation and ensuring that identity protection does not result in intrusive content. Instead of mimicking another individual, `MASQUE` leverages the guide image to inform the latent adjustments necessary for identity obfuscation.

Leveraging multiple guide images further stabilizes adversarial training, preventing perturbations from being biased toward any single representation and ensuring alignment with the natural identity boundary. To evaluate the effectiveness of our pairwise adversarial guidance, we use 100 identities from CelebA-HQ, varying the number of guide images per identity. The effectiveness is measured by the confidence score, defined as the gap between the similarity score of the original identity's gallery image and that of the misidentified identity. Additionally, we assess similarity with the corresponding gallery image in verification tasks using FaceNet for both tasks.

As reflected in Table 4, increasing the number of guide images $\tilde{\boldsymbol{x}}$ improves adversarial effectiveness, leading to higher confidence in identity obfuscation while reducing similarity with the original identity. Furthermore, images generated with guide images in the pairwise adversarial guidance step exhibit superior quality compared to those without, where adversarial loss is applied only to the original image. This supports our intuition that simply maximizing distance from the original image is ineffective, as it conflicts with the diffusion process's goal of preserving similarity, leading to instability in both image quality and adversarial effectiveness. These results highlight that the pairwise adversarial guidance of `MASQUE` provides a reliable, privacy-centric solution while maintaining high visual quality. In the future, we plan to explore the generalizability of the proposed method across more diverse makeup prompts and other facial recognition benchmarks.

D ALGORITHM PSEUDOCODE

Algorithm 1 Adversarial Makeup Generation with CA Guidance

Input: $\boldsymbol{z}_{\text{edit}}, \boldsymbol{z}_{\text{rec}}$: Edited and original latent, $\boldsymbol{e}_{\text{edit}}, \boldsymbol{e}_{\text{rec}}$: Edited and original text embeddings, $\mathcal{D}$: Stable Diffusion model, T: Number of diffusion steps, $\tau_{\text{attn}}, \tau_{\text{edit}}, \tau_{\text{adv}}$: Step thresholds, $\mathcal{M}$: Binary region mask, $\mathcal{I}_{\text{target}}$: Target token indices, d_k: Key/Query dimensionality, $\lambda_{\text{edit}}, \lambda_{\text{CoSi}}, \lambda_{\text{LPIPS}}$: Loss weights, η: Learning rate, $\sigma(t)$: Noise scale function, $\boldsymbol{x}$: Input image, $\tilde{\boldsymbol{x}}$: Guide image, max_adv: Maximum adversarial iterations

Output: Final protected image $\boldsymbol{x}_{\text{p}}$

for $k \leftarrow T$ **to** 1 **do**
  if $k > T - \tau_{\text{attn}}$ **then**
    `//Perform cross-attention edit`
    $\text{CA}^{\text{refined}} \leftarrow \{\}$
    for $\ell \in$ *CrossAttnLayers* **do**
      $Q_\ell, K_\ell, V_\ell \leftarrow \mathcal{D}.\text{UNet.GetCrossAttentionComponents}(\boldsymbol{z}_{\text{edit}}, \boldsymbol{e}_{\text{edit}}, \ell)$
      $\text{CA}_\ell \leftarrow \text{Softmax}\Big(\frac{Q_\ell K_\ell^\top}{\sqrt{d_k}}\Big)$
      $\text{CA}_\ell^{\text{refined}} \leftarrow \text{CA}_\ell \cdot V_\ell$
      $\text{CA}^{\text{refined}} \leftarrow \text{CA}^{\text{refined}} \cup \{\text{CA}_\ell^{\text{refined}}\}$
  if $k > T - \tau_{\text{edit}}$ **then**
    `//Perform edit loss update`
    $\mathcal{L}_{\text{edit}} \leftarrow 0$
    foreach $\text{CA}_\ell^{\text{refined}} \in \text{CA}^{\text{refined}}$ **do**
      $\text{CA}_{\text{target}} \leftarrow \text{ExtractTargetAttention}\Big(\text{CA}_\ell^{\text{refined}}, \mathcal{I}_{\text{target}}\Big)$
      $\text{CA}_{\text{masked}} \leftarrow \text{CA}_{\text{target}} \odot \mathcal{M}$
      $\mathcal{L}_{\text{edit}} \leftarrow \mathcal{L}_{\text{edit}} + \frac{\left(\sum_{(i,j)\in\mathcal{M}} \text{CA}_{\text{masked}}[i,j]\right)^2}{\sum_{(i,j)\in\mathcal{M}} \mathcal{M}[i,j]}$
    $\mathcal{L}_{\text{edit}} \leftarrow \frac{\mathcal{L}_{\text{edit}}}{|\text{CA}^{\text{refined}}|}$
    $\nabla_{\boldsymbol{z}_{\text{edit}}} \mathcal{L}_{\text{edit}} \leftarrow \text{Backprop}(\mathcal{L}_{\text{edit}})$
    $\boldsymbol{z}_{\text{edit}} \leftarrow \boldsymbol{z}_{\text{edit}} - \eta \cdot \lambda_{\text{edit}} \cdot \nabla_{\boldsymbol{z}_{\text{edit}}} \mathcal{L}_{\text{edit}}$
  `//Apply classifier-free guidance during all diffusion steps`
  $\boldsymbol{\epsilon}_{\text{rec}} \leftarrow \mathcal{D}.\text{UNet}(\boldsymbol{z}_{\text{rec}}, t_k, \boldsymbol{e}_{\text{rec}})$
  $\boldsymbol{\epsilon}_{\text{edit}} \leftarrow \mathcal{D}.\text{UNet}(\boldsymbol{z}_{\text{edit}}, t_k, \boldsymbol{e}_{\text{edit}}, \text{CA}^{\text{refined}})$
  $\boldsymbol{z}_{\text{rec}} \leftarrow \boldsymbol{z}_{\text{rec}} - \sigma(t_k) \cdot \boldsymbol{\epsilon}_{\text{rec}}$
  $\boldsymbol{z}_{\text{edit}} \leftarrow \boldsymbol{z}_{\text{edit}} - \sigma(t_k) \cdot \boldsymbol{\epsilon}_{\text{edit}}$
  $\boldsymbol{z}_{\text{edit}} = \mathcal{M} \cdot \boldsymbol{z}_{\text{edit}} + (1 - \mathcal{M}) \cdot \boldsymbol{z}_{\text{rec}}$
  if $k < T - \tau_{\text{adv}}$ **then**
    `//Perform pairwise adversarial optimization`
    for $i \leftarrow 0$ **to** $max\_adv - 1$ **do**
      $\boldsymbol{x}_{\text{rec}} \leftarrow \mathcal{D}.\text{VAE.Decode}(\boldsymbol{z}_{\text{edit}})$
      $\mathcal{L}_{\text{adv}} \leftarrow \lambda_{\text{CoSi}} \cdot \text{CoSi}\big(\text{FR}(\boldsymbol{x}_{\text{rec}}), \text{FR}(\tilde{\boldsymbol{x}})\big) + \lambda_{\text{LPIPS}} \cdot \text{LPIPS}(\boldsymbol{x}_{\text{rec}}, \boldsymbol{x})$
      $\text{grad} \leftarrow \text{Backprop}(\mathcal{L}_{\text{adv}})$
      $\boldsymbol{z}_{\text{edit}} \leftarrow \boldsymbol{z}_{\text{edit}} - \eta \cdot \text{grad}$
      $\boldsymbol{z}_{\text{edit}} \leftarrow \boldsymbol{z}_{\text{edit}} \odot \mathcal{M} \; + \; \boldsymbol{z}_{\text{rec}} \odot \big(1 - \mathcal{M}\big)$
`//Decode the final latent to produce the protected image`
$\boldsymbol{x}_{\text{p}} \leftarrow \mathcal{D}.\text{VAE.Decode}(\boldsymbol{z}_{\text{edit}})$
return $\boldsymbol{x}_{\text{p}}$