DiffAM: Diffusion-based Adversarial Makeup Transfer for Facial Privacy Protection

Supplementary Material

A. Background: DDPM and DDIM

Denoising Diffusion Probabilistic Model (DDPM)[1, 4] consists of a forward diffusion process and a reverse diffusion process. The forward diffusion process is described as a Markov chain where Gaussian noise is gradually added to the original image x_0 to get the noisy image x_t at every time steps $t \in \{1, ..., T\}$:

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}), \tag{1}$$

where $\beta_t \in (0,1)$ are hyperparameters representing the variance schedule. A good property of this formulation is that we can directly sample x_t given x_0:

$$q(x_t|x_0) = \mathcal{N}(\sqrt{\alpha_t}x_0, (1-\alpha_t)\mathbf{I}), \tag{2}$$

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{3}$$

where $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$.

We can get a new sample from the distribution $q(x_0)$ by following the reverse steps $q(x_{t-1}|x_t)$, starting from $x_T \sim \mathcal{N}(0, \mathbf{I})$. As the posteriors $q(x_{t-1}|x_t)$ is intractable, in the reverse process, a neural network p_θ is trained to approximate it:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t, t), \sigma_t^2 \mathbf{I}), \tag{4}$$

where

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{1-\alpha_t}\epsilon_\theta(x_t, t)\right). \tag{5}$$

A U-net[3] is trained to learn a function $\epsilon_\theta(x_t, t)$ to predict the added noise at time step t by optimizing the objective[1]:

$$\min_\theta \mathcal{L}(\theta) = \mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0,\mathbf{I}), t}||\epsilon - \epsilon_\theta(x_t, t)||_2^2. \tag{6}$$

Then, we can sample the data as follows:

$$x_{t-1} = \mu_\theta(x_t, t) + \sigma_t z, \tag{7}$$

where $z \sim \mathcal{N}(0, \mathbf{I})$.

To accelerate the sampling process, Song *et al.* [5] proposed Denoising Diffusion Implicit Model (DDIM) that has a non-Markovian noising process. The sampling process of DDIM is:

$$x_{t-1} = \sqrt{\alpha_{t-1}}f_\theta(x_t, t) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\epsilon_\theta(x_t, t) + \sigma_t^2 z, \tag{8}$$

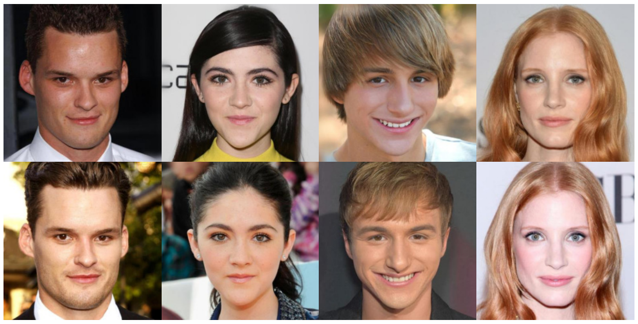


Figure 1. Target identities. The images in the top row are used for training, while the images in the bottom row are used for testing.

where f_θ is the prediction of x_0 at t:

$$f_\theta(x_t, t) = \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}}. \tag{9}$$

By setting $\sigma_t = 0$ in Eq. (8), the sampling process from x_T to x_0 becomes deterministic, which is the principle of DDIM. Also, the operation of DDIM inversion can map x_0 back to x_T by reversing the process, enabling subsequent editing of images. Deterministic DDIM sampling and inversion can be expressed as:

$$\begin{aligned} x_{t-1} &= \sqrt{\alpha_{t-1}}f_\theta(x_t, t) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(x_t, t), \\ x_{t+1} &= \sqrt{\alpha_{t+1}}f_\theta(x_t, t) + \sqrt{1-\alpha_{t+1}}\epsilon_\theta(x_t, t). \end{aligned} \tag{10}$$

B. Target Images

DiffAM aims to generate protected face images which mislead FR models into identifying them as the target identity. So we show the four target identities, provided by [2], used for our experiments in Fig. 1. To simulate real-world attack scenarios, the target images used during training and testing are different.

C. Attack Performance on Tencent API

To fully evaluate attack effects of DiffAM in real-world scenarios, we also shows the quantitative results of attacks on Tencent API[1] here in Fig. 2. We randomly selected 100 images each from CelebA-HQ and LADN datasets to protect and report confidence scores returned from APIs. The

[1] https : / / cloud . tencent . com / product / facerecognition

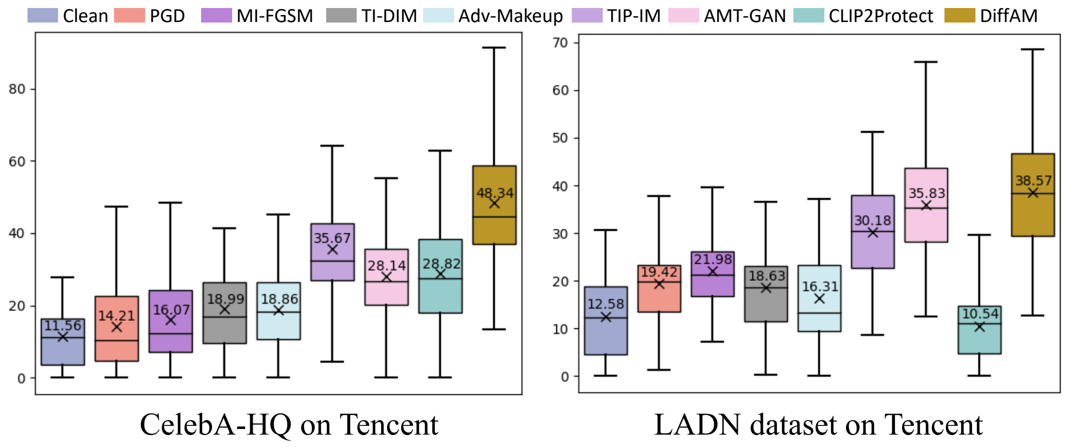


Figure 2. The confidence scores (higher is better) returned from Tencent API.

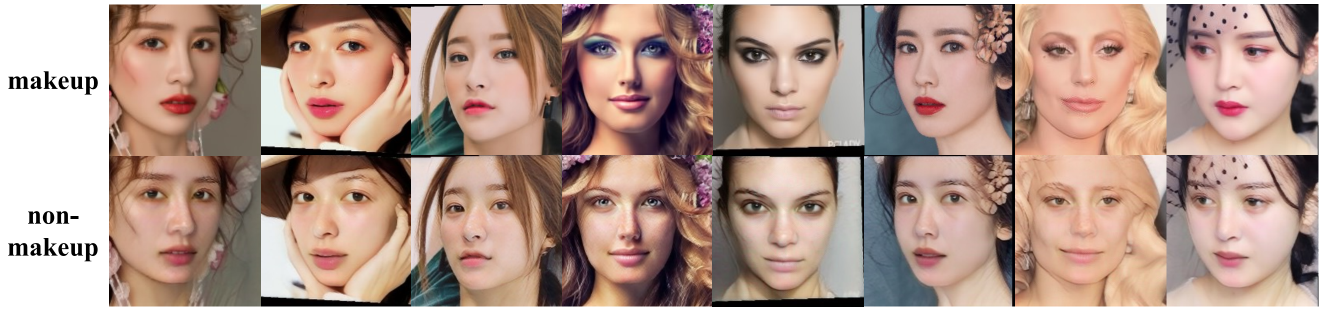


Figure 3. Visualizations of text-guided makeup removal. The top row shows some reference makeup images. The bottom row shows the corresponding non-makeup images generated by text-guided makeup removal module.

confidence scores are between 0 to 100, where the higher score indicates higher similarity between the protected face image and the target image. DiffAM achieves the highest average confidence scores (**48.34** and **38.57**) campared to other methods. This further demonstrates that our precise guidance on adversarial makeup domain and robust adversarial makeup generation ensure high black-box transferability of protected face images generated by DiffAM.

D. More Visual Results

Text-guided Makeup Removal Fig. 3 shows some visual results of text-guided makeup removal. The makeup of reference images, such as lipstick and eyeshadow, are clearly removed, indicating the powerful ability of text guidance in makeup removal. Then the difference in CLIP space between the makeup and non-makeup images can determine accurate makeup direction for subsequent makeup transfer.

Image-guided Makeup Transfer Due to space limitations of the main text, more visual results of image-guided makeup transfer are shown in Fig. 4. DiffAM achieves precise makeup transfer for each given reference image and generates natural-looking protected face images, thanks to our precise control over makeup direction and distance.

References

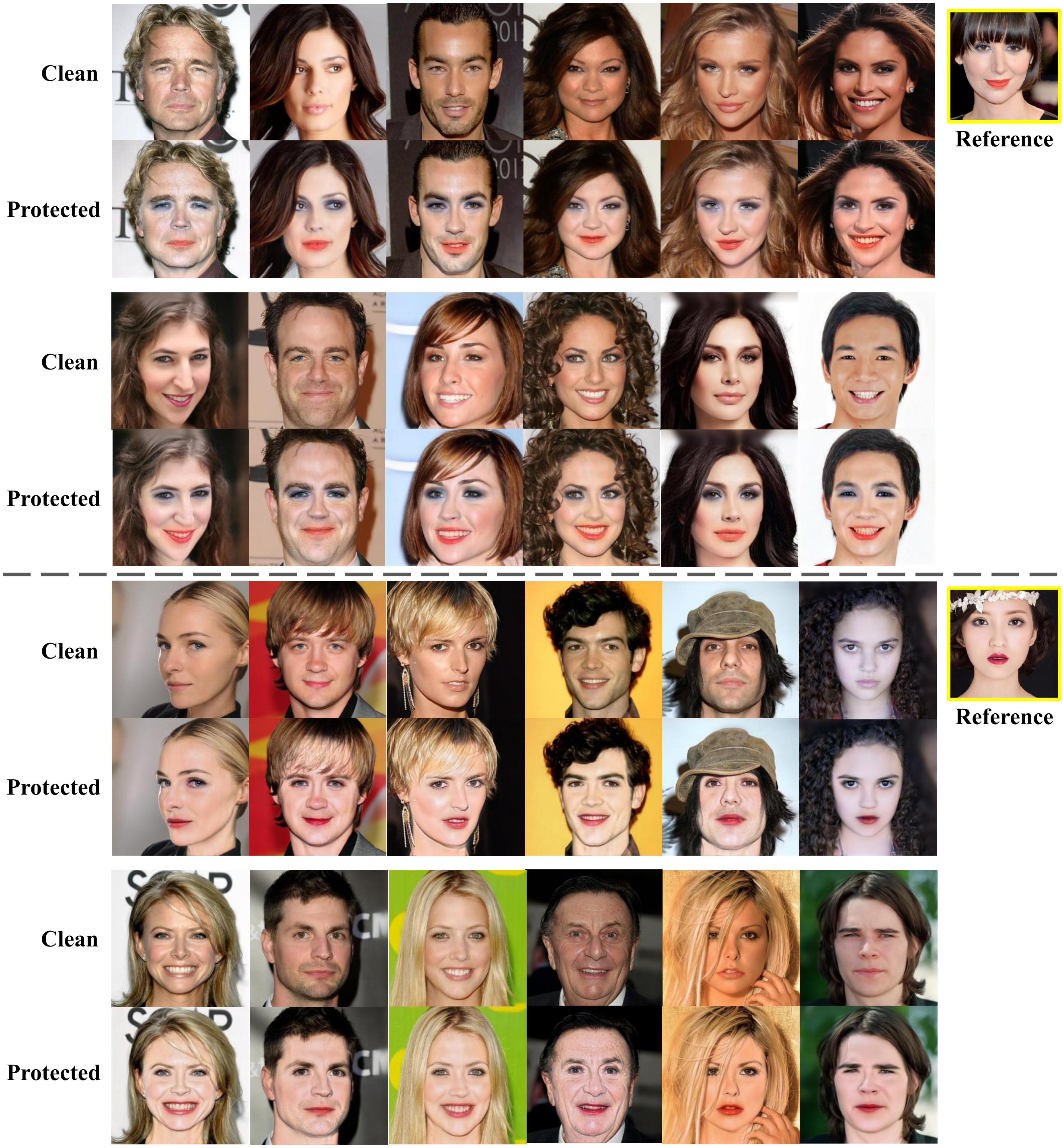
[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

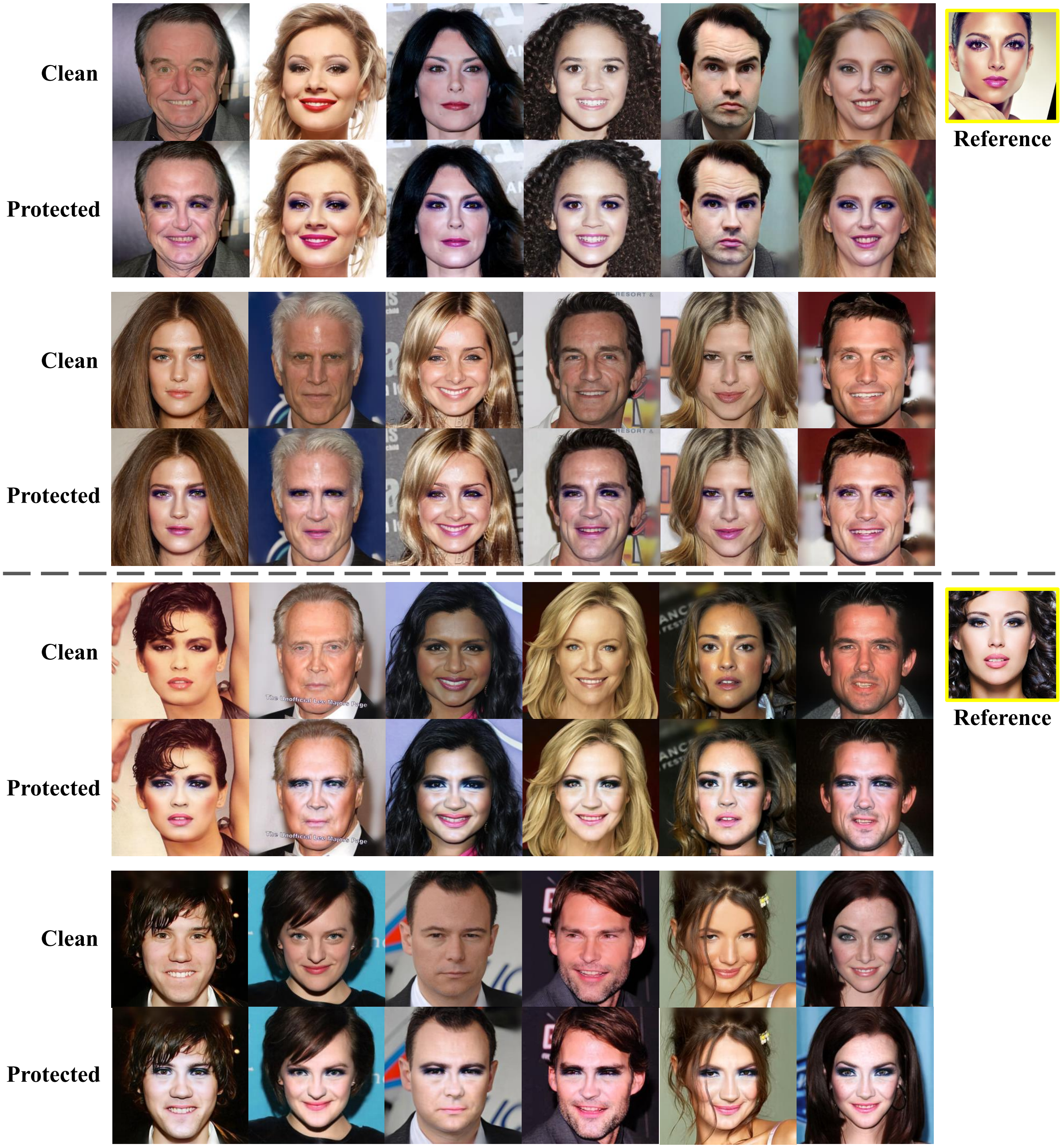
[2] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15014–15023, 2022. 1

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1

[4] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1

[5] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 1


Clean
Protected
Reference
Clean
Protected
Clean
Protected
Reference
Clean
Protected


Clean
Protected
Reference
Clean
Protected
Clean
Protected
Reference
Clean
Protected

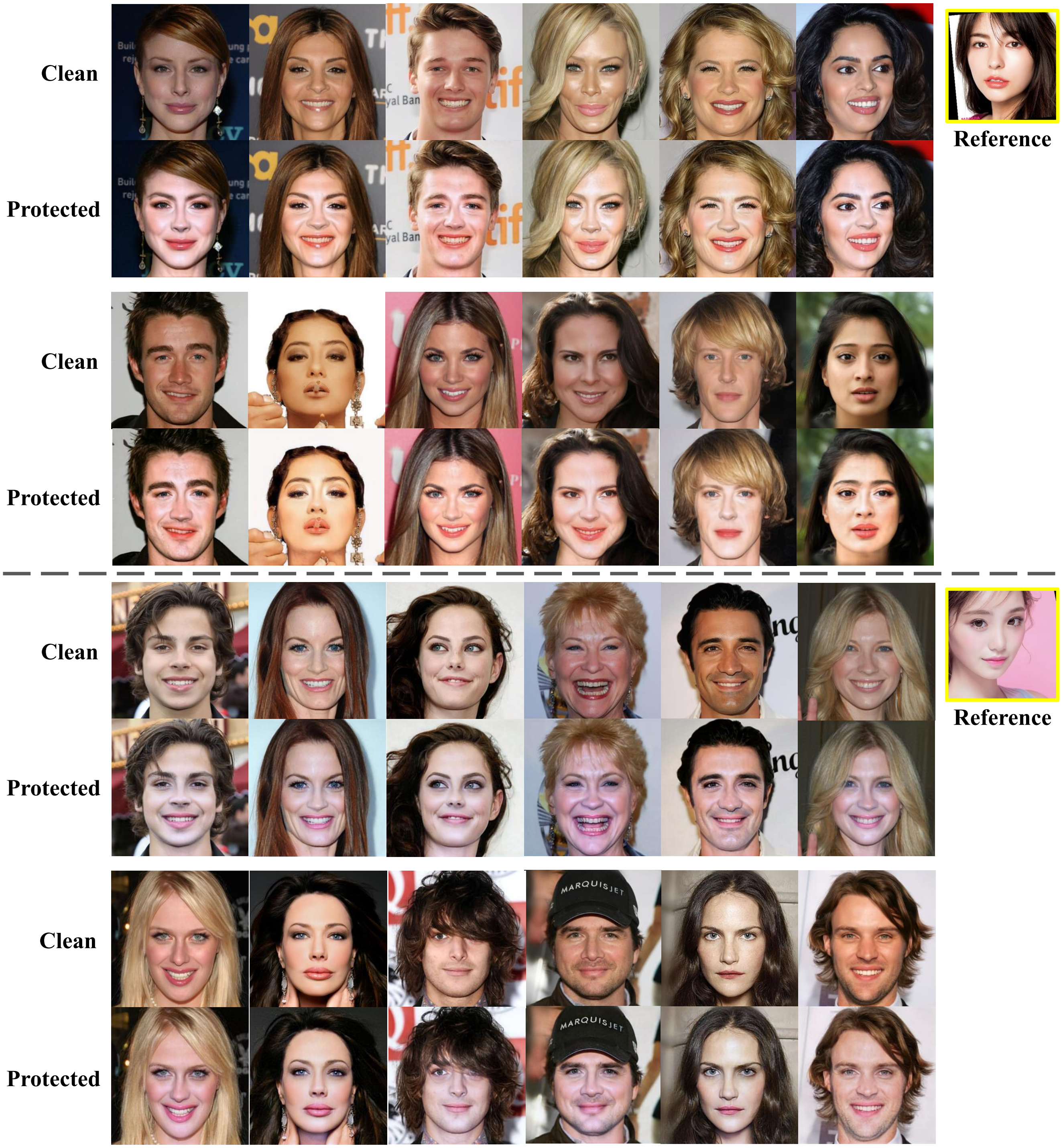


Figure 4. Visualizations of image-guided makeup transfer.