1 SUPPLEMENTARY MATERIAL

1.1 Detailed Network Architecture.

A diagram of our network structure is presented in Fig. 1. The input image can either be a 3-channel synthetic rendering or additional feature maps, which are directly added to the output of the first encoder layer. The unlabeled modules, shown as orange and blue rectangles, represent two convolutional layers with the output channel and resolution labeled accordingly. The encoded multi-scale features are connected to the decoder layers through skip connections. The "UP" and "DOWN" modules correspond to the "To-RGB" and "From-RGB" module of the original StyleGAN, respectively. Our mapping networks consist of four fully connected layers with 64 channels. The style input and noise injection module are the same as those in the original StyleGAN. The output is a 3-channel photo-realistic portrait image with a resolution of 1024×1024. Our discriminator shares the same structure as the original StyleGAN, except for the use of a 6-channel input. Please note that we have substituted the custom operator "fused Leaky ReLU" in StyleGANv2 with the "Leaky ReLU" for the TensorRT accelerating process.

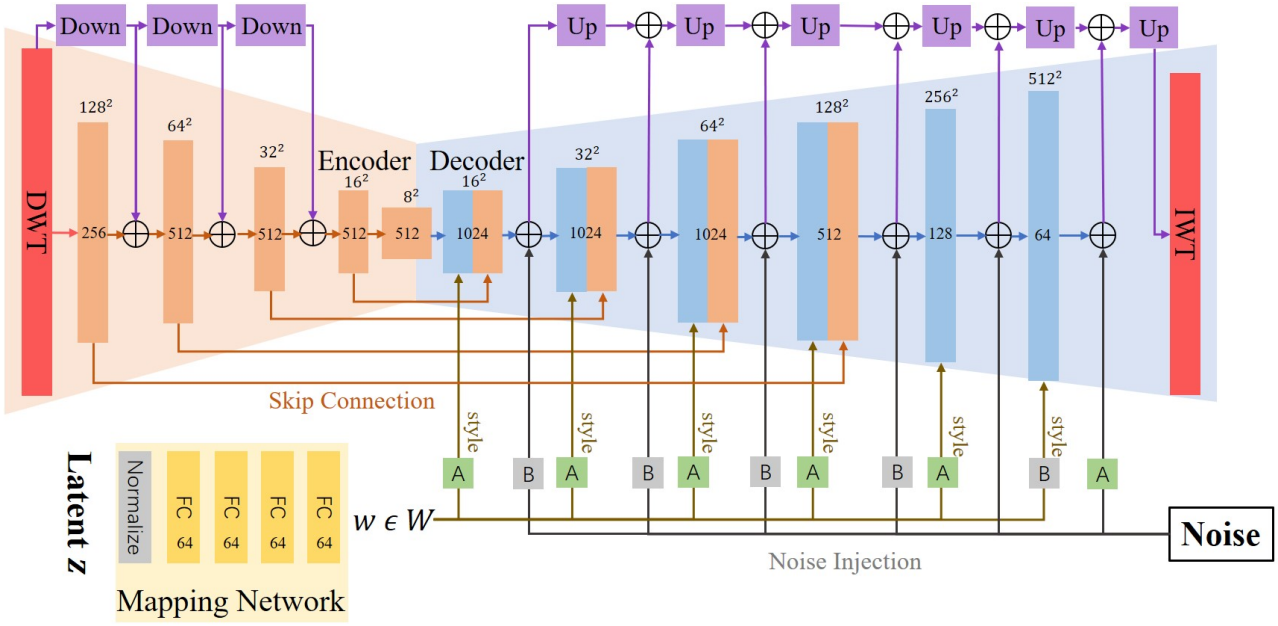

Figure 1: A schematic diagram of our network structure.

1.2 Training Strategy.

During the training of a specific video, we observed that the learning of eye gaze was progressing slowly with the GAN loss, but it could be accelerated by solely training with perceptual loss. Therefore, we first trained our network for 4,000 steps, then removed the GAN loss and continued training for another 3000 steps. Finally, the full loss terms were reintroduced, and thus the learning of eye gaze was accelerated. We compared two pre-training schemes: directly averaging 6 separate models and pre-training on all videos together. Both pre-trained models provided good initial values for subsequent fine-tuning. Although the results may appear similar in the figures (Fig.8), cases without pre-training exhibited flickering.

1.3 Less Training Data.

To investigate the performance of our network with limited training data, we trained our StyleUNet with 50, 300, 1K, and 8K images, which were randomly sampled from the original training set. As depicted in Fig. 2, we observe only a slight decrease in performance when the amount of training data was reduced. This finding supports the claim that our approach can effectively utilize limited training data for neural facial avatar reconstruction, outperforming other methods in this regard.

1.4 Interactive Editting.

We have also included interactive editing results in our video. Since the synthetic renderings are entirely controlled by various parameters such as pose and expression, we can implement interactive editing by directly adjusting these parameters. As shown in our video, we can first use the assigned parameters to generate the corresponding synthetic renderings, and then our method can generate the final photo-realistic portrait images. This implies that our method is versatile and can be applied to a wide range of applications, including audio-driven avatars.

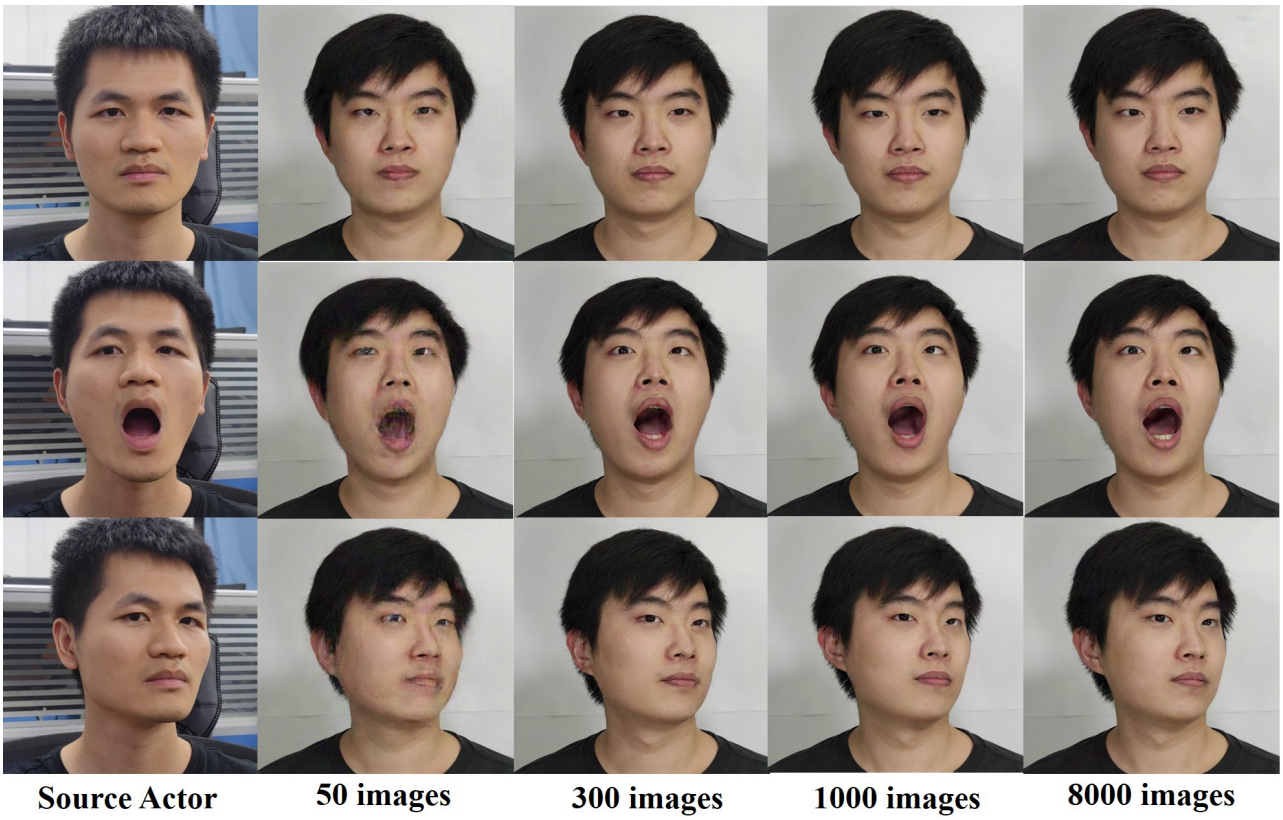

Figure 2: Reenactment results of the networks trained with 50, 300, 1000 and 8000 images.

1.5 Ablation Study for StyleUNet.

To demonstrate the effectiveness of our network and investigate the specific role of each module, we conduct an ablation study using five different network structures: our full StyleUNet, StyleUNet without the mapping network (w/o style), StyleUNet without the noise injection module (w/o noise), StyleUNet without both the mapping network and the noise injection module (w/o style&noise), and the original U-Net. All five networks are trained on the same dataset, consisting of 8,000 training frames and 1,000 testing frames. After training, we evaluate these networks using both self-driven re-animation with the testing set and re-animation driven by another actor. As depicted in Fig. 3, StyleUNet trained without the mapping network (w/o style) shows instability in extreme poses or expressions, such as the closed mouth in the second row, incorrect eyes in the third row, and irrational edge in the last row, with some additional artifacts highlighted by red circles. StyleUNet trained without noise injection (w/o noise) has difficulty in generating fine details, resulting in artifacts in the teeth and eyes.

1.6 Ablation Study for Training Speed.

The convergence diagram in Fig. 4 illustrates that our method achieves faster training speed even without pre-training ("Ours w/o pretrain"). Our network can learn the basic information of the background and other areas faster than DVP and "Single-StyleUNet". The comparison of "Ours" and "Ours w/o pretrain" demonstrates the effectiveness of our pre-training.

1.7 Additional Notes for the Values in Tables.

When calculating the values presented in Tab. 1, Tab. 2, and Tab. 3, we exclude the background and resize all images to a resolution of 512×512. As depicted in Fig. 5, we align all faces and use a smaller box to crop the images when computing the values for Table 1 because the images generated by Next3D have a narrower range. For Tab. 2 and Tab. 3, we show sample images used to compute the values in Fig. 6 and Fig. 7, respectively.

1.8 More facial reenactment Results.

Additional facial reenactment results are showcased in Fig. 8, further illustrating the high quality of our generated images.

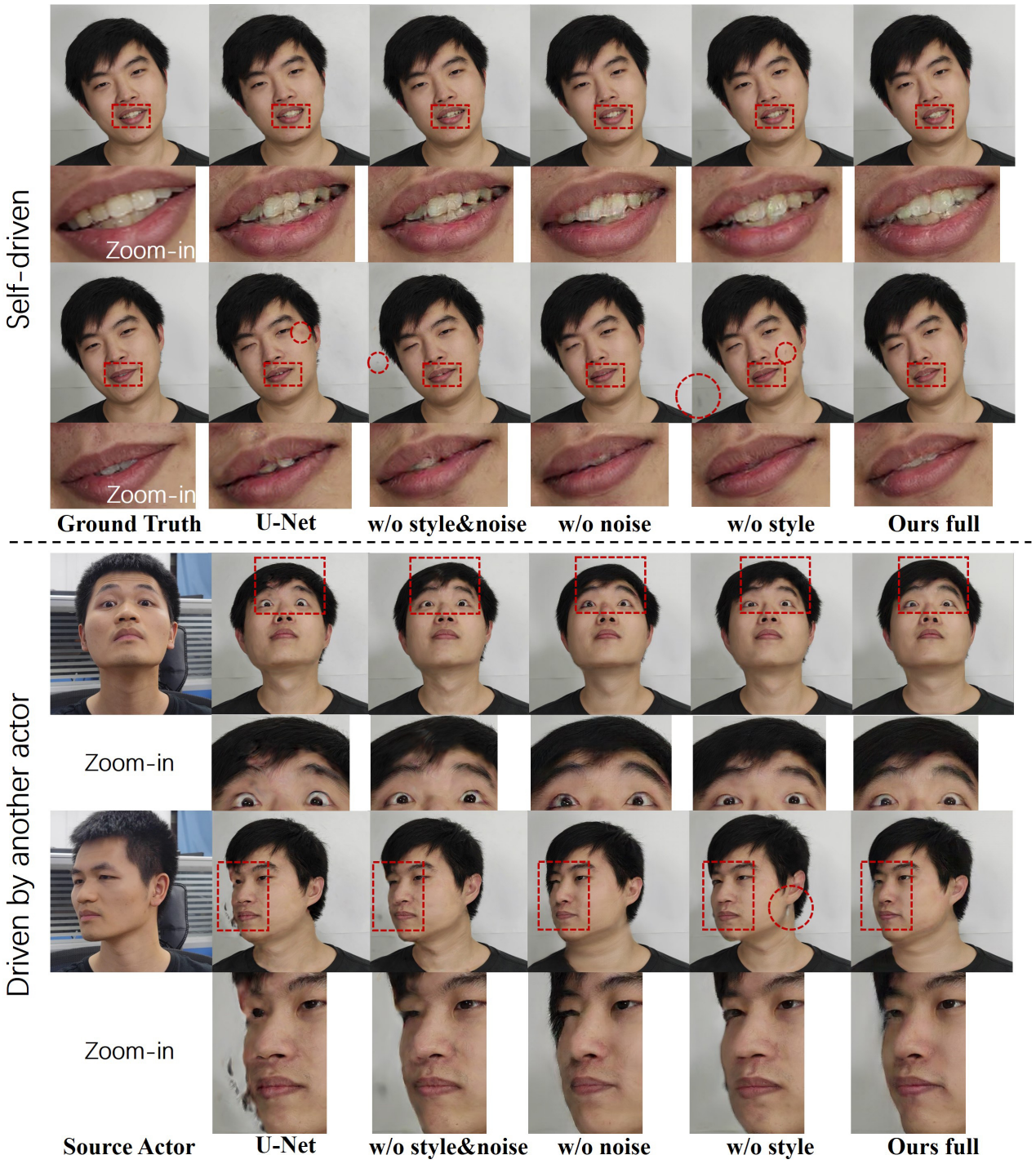

Figure 3: Ablation Study for StyleUNet.

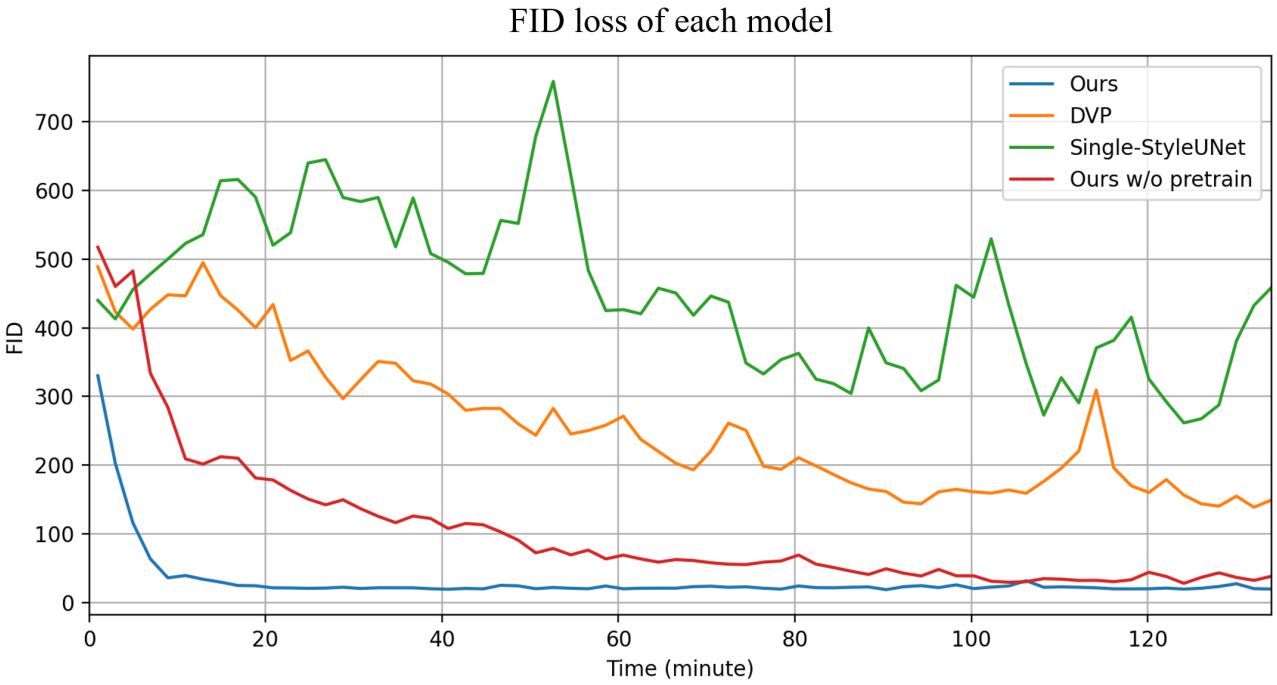

Figure 4: The FID curves of different network structures during the training stage.

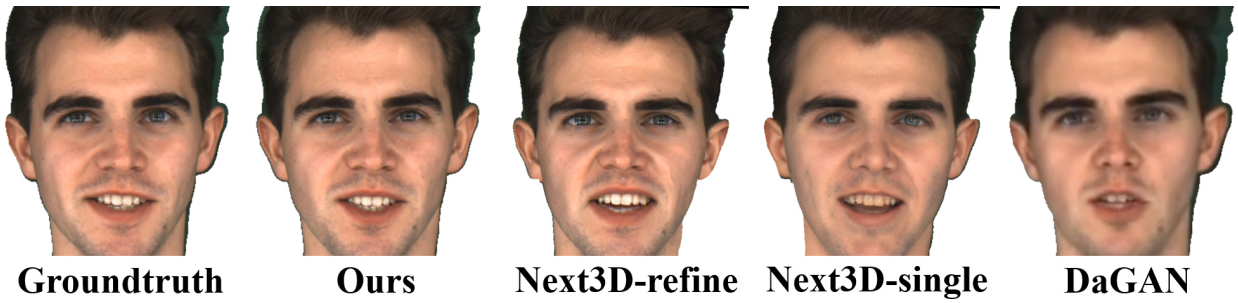

Figure 5: Sampled images for calculating the values in Tab. 1

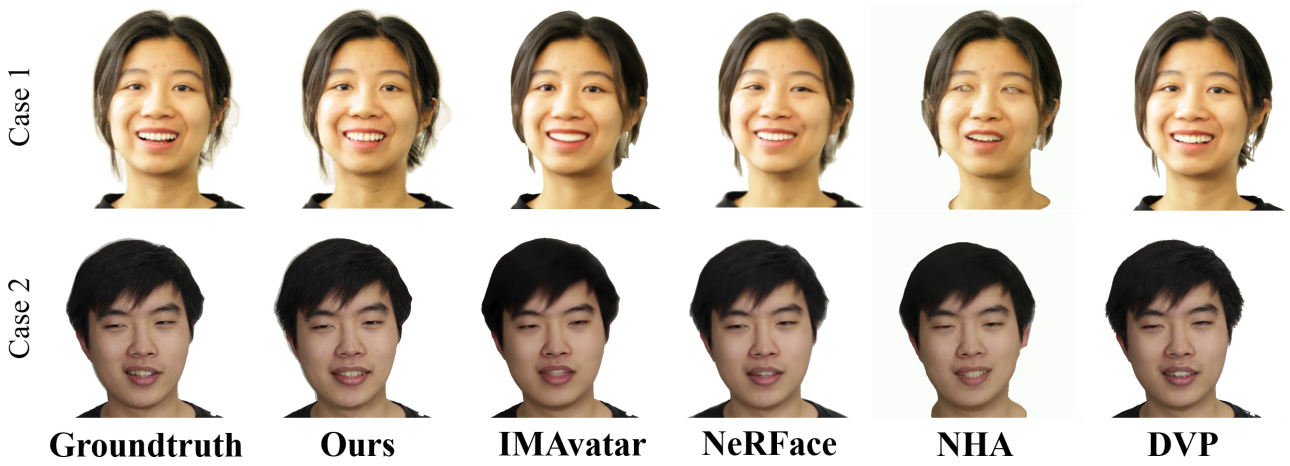

Figure 6: Sampled images for calculating the values in Tab. 2

Figure 7: Sampled images for calculating the values in Tab. 3

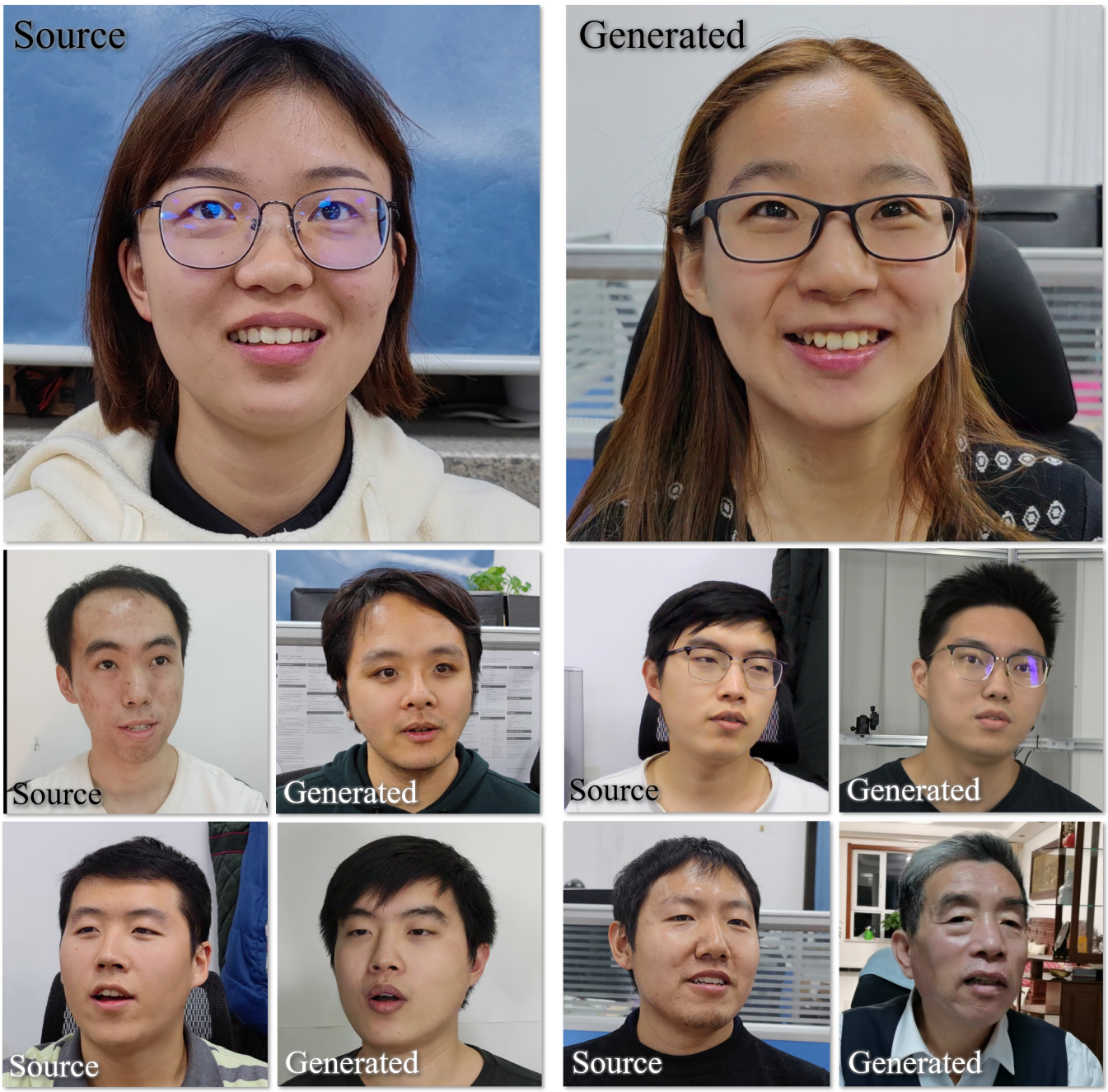

Figure 8: More facial reenactment results.