Supplementary Materials for Controllable 3D Face Generation with Conditional Style Code Diffusion

Qualitative Results

Conditional style code diffusion. Figure 3 and Figure 4 show our qualitative results of text-to-3D face generation. Figure 5 shows our qualitative results of expression & text-to-3D face generation. All the results demonstrate our model can produce photorealistic faces that align with the given conditions.

3D GAN Inversion. Figure 6 and Figure 7 show qualitative comparison with the previous method. The results show that our method mitigates 3D inconsistency while ensuring visual quality.

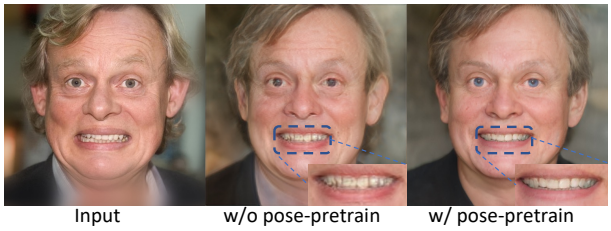

Figure 1: Ablation studies of pose-pretrain.

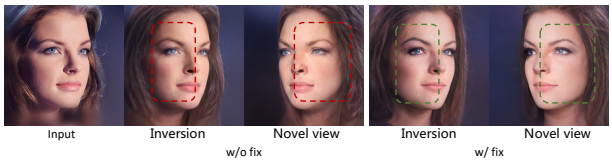

Figure 2: Ablation studies of pose-pretrain.

Method	MSE ↓	LPIPS ↓	ID ↓
w/o pose-pretrain	0.0266	0.1754	0.1524
w/ pose-pretrain w/ fix	0.0404	0.2128	0.2156
w/ pose-pretrain w/o fix	0.0289	0.1790	0.1632

Table 1: Ablation study of our proposed methods.

Layer	FID ↓	CLIP score ↑
4	56.89	26.44
5	**55.00**	26.53
6	56.77	26.62
7	56.80	**26.64**

(a) Numbers of Decoding block.

Ratio(%)	FID ↓	CLIP score ↑
25	**56.00**	**26.63**
50	56.89	26.44
75	63.08	25.77

(b) Mask ratio.

Table 2: Ablation studies of style code diffusion

More ablation studies

3D GAN inversion. As shown in Table 1, we provide ablation studies in Pose-guided inversion pretraining. 'w/ pose-pretrain' means using synthetic data to learn a mapping that projects style codes of different views onto a single code. Although w/o pose-pretrain achieves the best results, the visualization in Figure 1 shows the model does not model the pose-condition, leading to image unalignment, *e.g.*, teeth unalignment. 'w/ fix' represents freezing the PoM module. While the results of 'w/o fix' are better than 'w/ fix', 'w/ fix' suffers from 3D inconsistency as shown in Figure 2. Considering facial symmetric, 'w/o fix' fails to model this property thereby overfitting to the input view, leading to undesirable results under the mirrored camera view, *e.g.*, eye, mouth, eyebrow.

Conditional 3D face generation. We first examine the effect of the mask ratio in Table 2b. We observe that when applying 25%, the model achieves the best performance. Then we investigate the impact of different layers in the style code denoiser as shown in Table 2a. The experiment revealed that employing 5 layers achieves the best FID metric while using 7 layers results in the highest CLIP score.

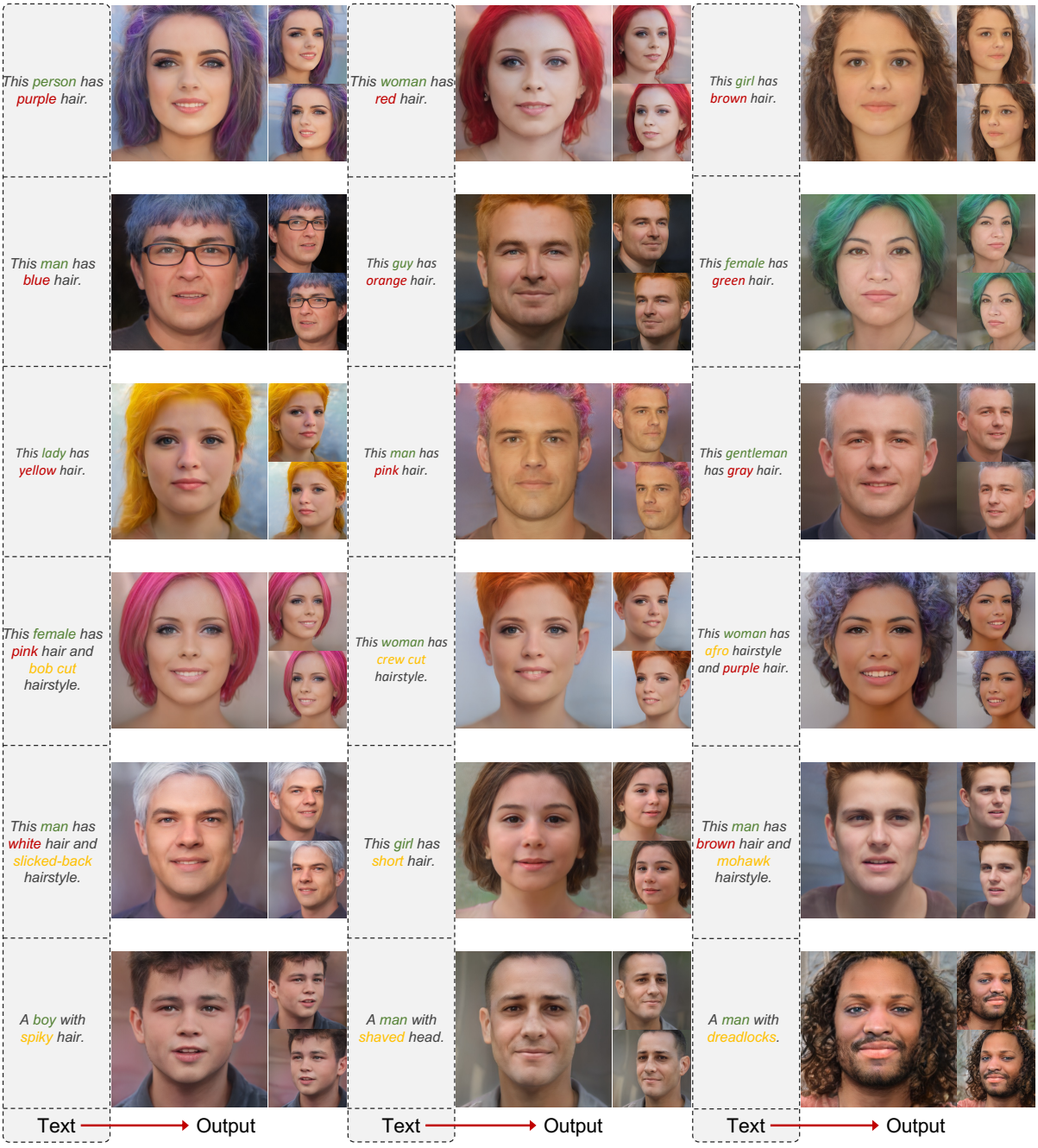

Figure 3: Qualitative results of hair attributes.

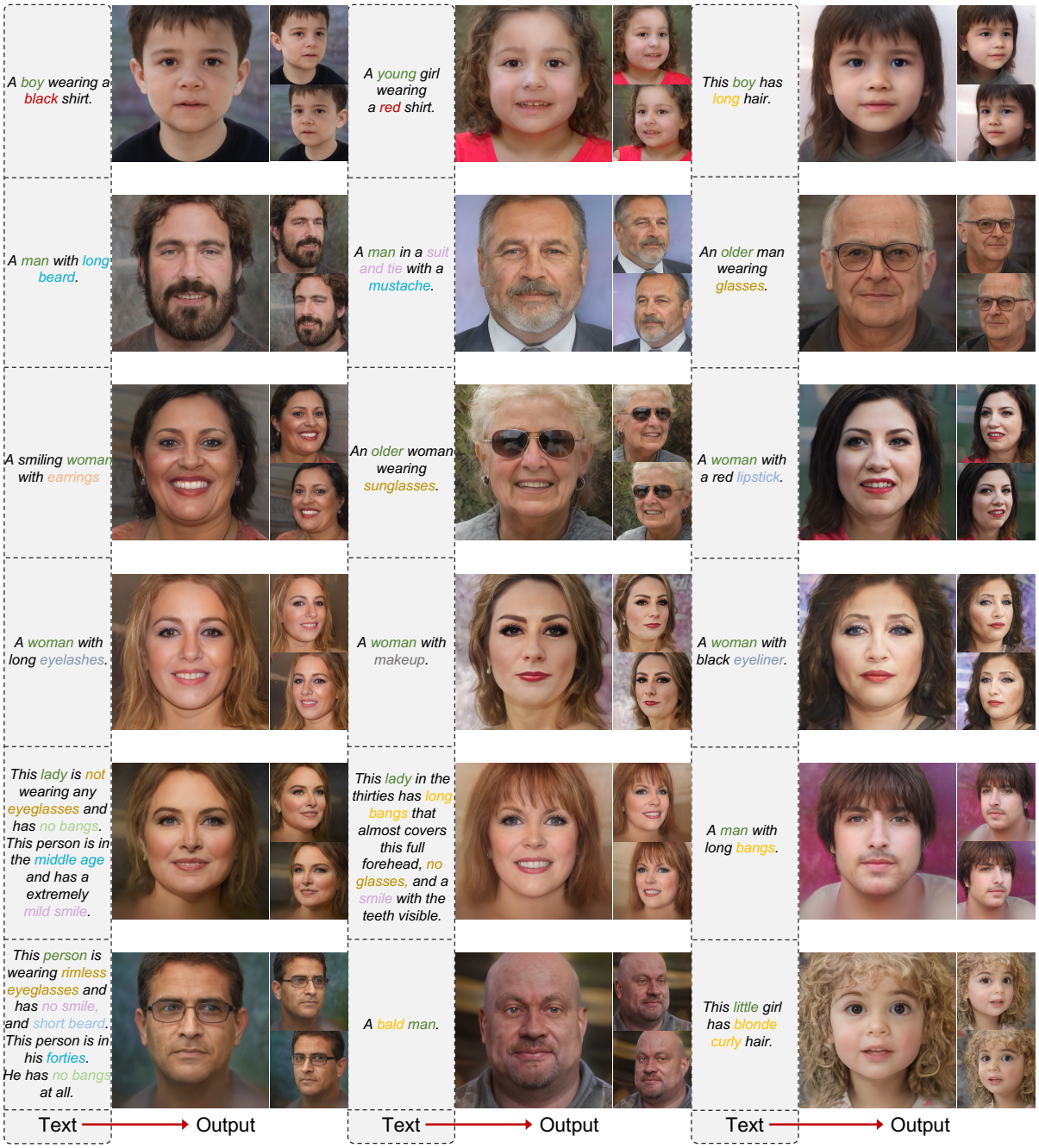

Figure 4: Qualitative results of text-to-3D face generation.

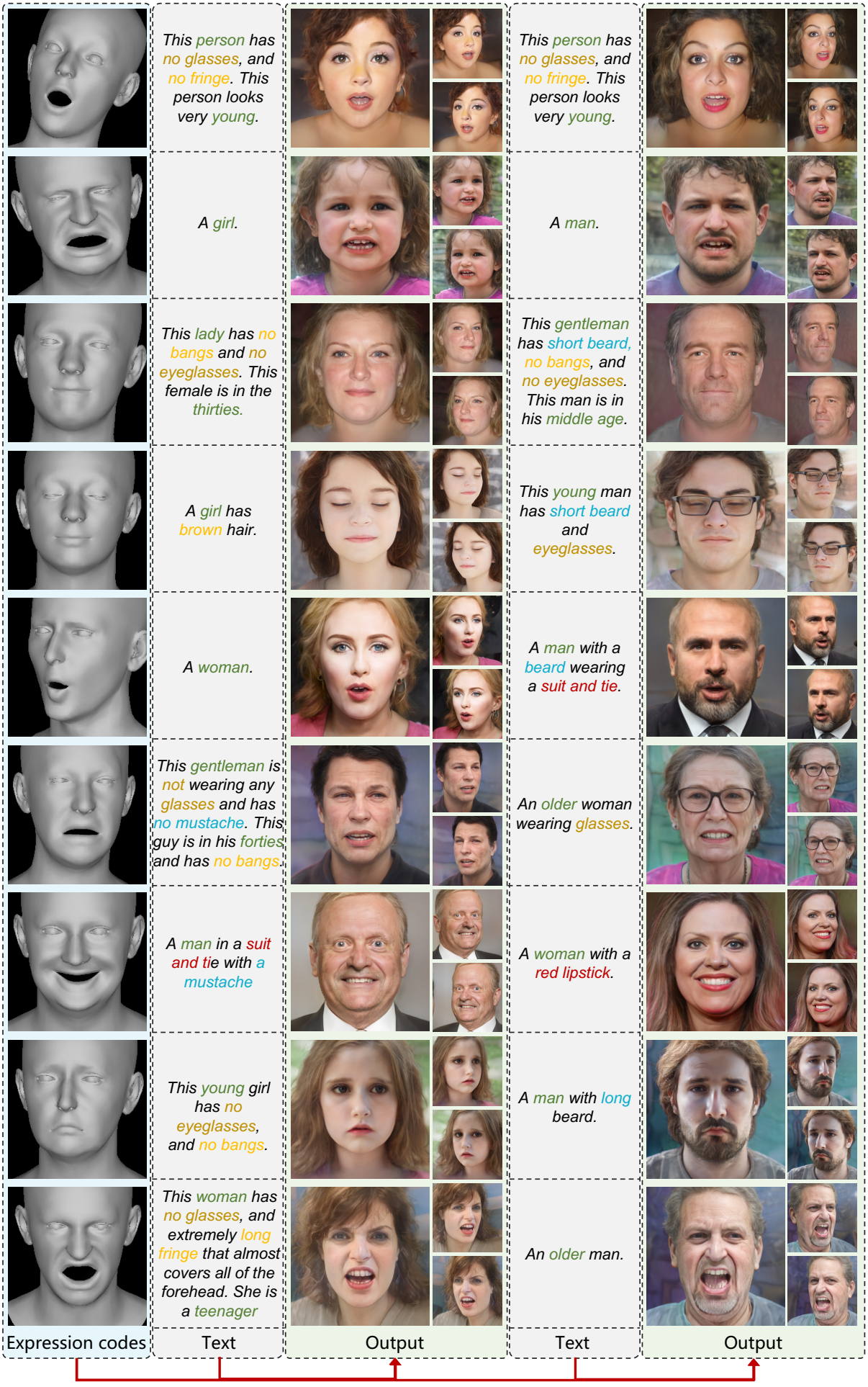

Figure 5: Qualitative results of text & expression-to-3D face generation.

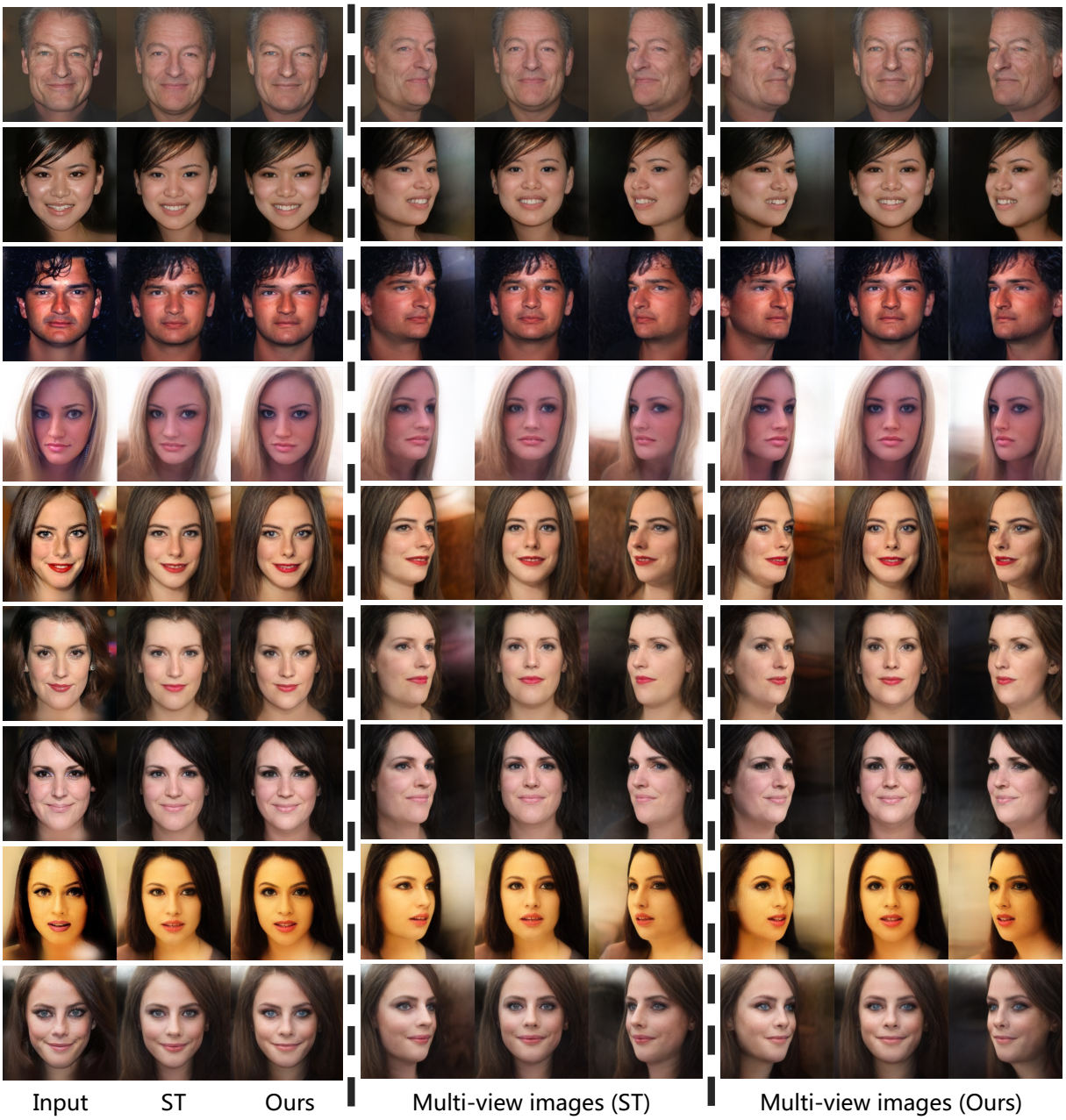

Figure 6: Qualitative comparison of 3D GAN inversion with the previous method.

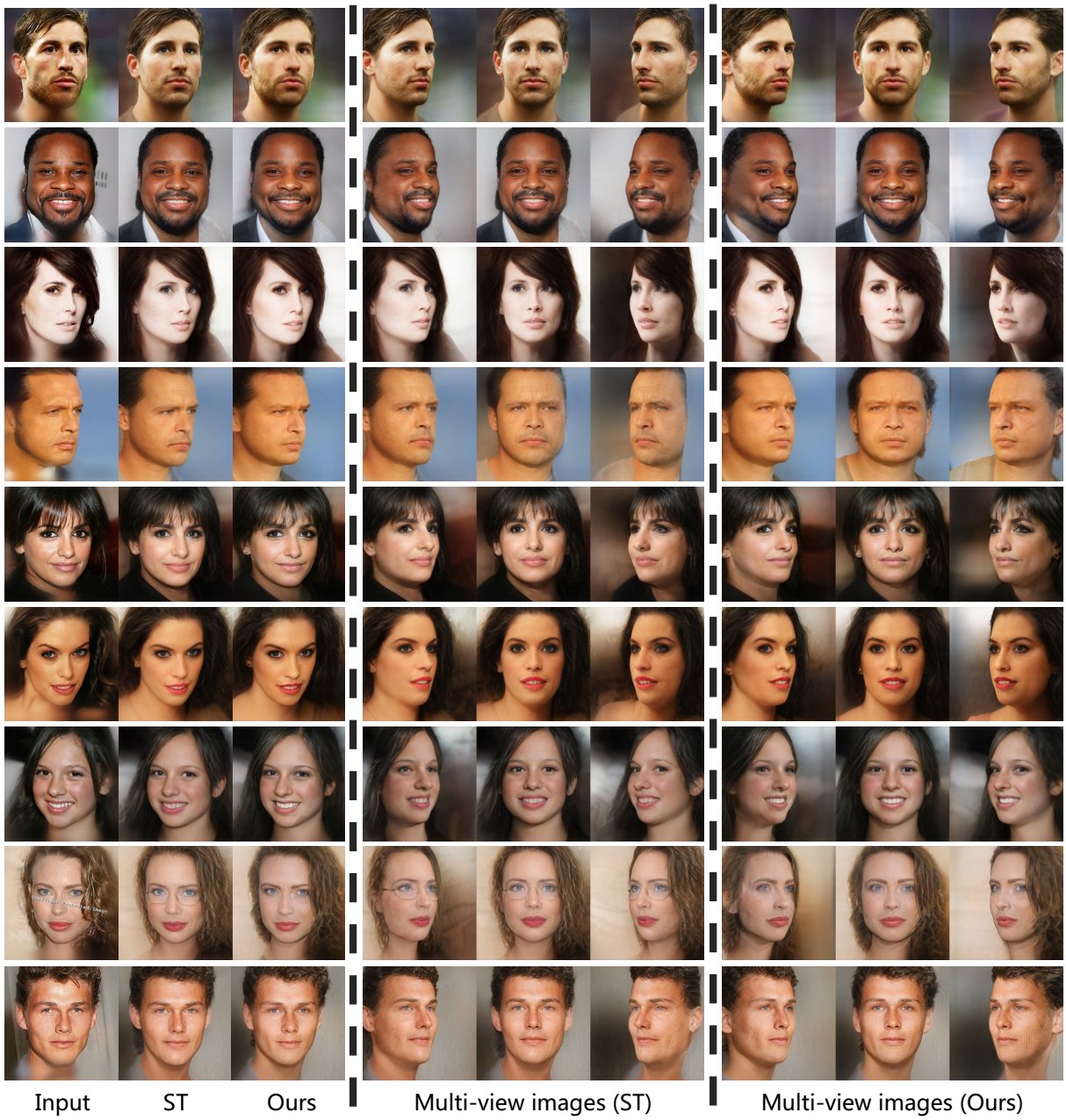

Figure 7: Qualitative comparison of 3D GAN inversion with the previous method.