Eurographics Symposium on Rendering 2023
T. Ritschel and A. Weidlich
(Guest Editors)

Volume 42 (2023), Number 4

PVP Supplementary Materials

1. Implementation Details

We discuss the implementation details of our algorithm in this section.

$\mathcal{F}_{\text{rot}}$ is a 2-layer MLP with leaky ReLU as the nonlinear activation layer in the middle and Tanh as the output layer. There are 128 channels for the hidden layer. Instead of outputting the blending weights directly, the network predicts the residual weights from the centroid, which has equal weights from all pivots, to ensure the initialization is around the centroid (since network output is small initially). Then it linearly combines the blending weights into a latent code in the $\mathcal{W}+$ space. Additionally, to prevent the MLP from generating values lower than negative beta, we shift the output values by beta and apply a SoftPlus function. This would effectively bound the values.

$\mathcal{F}_{\text{expr}}$ is composed of the same 2-layer MLPs, except that there are 8 of them to control the first 8 layers of the latent code in the $\mathcal{W}+$ space. The predicted values are added to the latent code output from $\mathcal{F}_{\text{rot}}$.

Another implementation detail is that we perform a normalization stage for reenactment. The idea is that the distributions of expressions are different for each video. Since we only train on a single input video, it is possible to overfit to the given sequence. As a result, we normalize the driving features with standard deviation and mean of the expressions from the source and driving video before input it to the $\mathcal{F}_{\text{expr}}$ for reenactment.

For the clustering stage discussed in Sec. 4.1, we use K=200 in our experiments. It is important when the expression distribution in the data is uneven (e.g. same expressions for a long time) as the clustering can select more representative frames.

In terms of rendering resolution, we downsample our results from 1024×1024 to 512×512 for fair comparison with other baseline methods. However, our method can run in 1024×1024 resolution with similar efficiency (removing the downsampling stage), given input videos with 1024×1024 resolution.

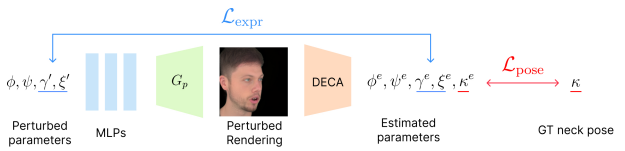

Figure 1: **Illustration of our supervision.**

2. Dataset Composition

We show the range of head poses in Table 1. Dataset names starting with "id" denote the subjects from the NeRFBlendShape dataset, whereas names starting with "person" denote subjects from the NHA dataset.

It is difficult to determine the number of expressions as they are often transitioning from one to another (e.g. from smiling to grinning). However, in the NHA and NeRFBlendShape dataset, there are different expressions like smiling, winking and puffing.

Table 1: **Composition of our dataset.** We show the range of head poses (in degrees) of each video sequence. Our evaluation dataset contains a wide range of head poses.

Dataset	Min Yaw	Max Yaw	Min Pitch	Max Pitch
id1	-45.36	32.29	-23.25	17.06
id2	-57.74	35.85	-39.35	15.70
id3	-13.43	14.38	-20.60	12.44
id4	-29.09	12.39	-14.98	14.95
id5	-17.22	12.27	-5.42	11.41
id6	-9.62	2.31	-9.67	-1.38
id7	-39.96	36.23	-22.38	13.24
id8	-4.98	10.09	-20.16	-4.51
person0000	-75.56	75.92	-35.00	10.49
person0004	-70.62	72.65	-37.2	17.85

3. Perturbed Renderings for $\mathcal{L}_{\text{expr}}$

We show the renderings from the perturbed parameters γ'_i in Fig. 2. The expression matching loss $\mathcal{L}_{\text{expr}}$ helps the network learn unseen expressions through perturbing the parameters and matching the predictions from DECA.

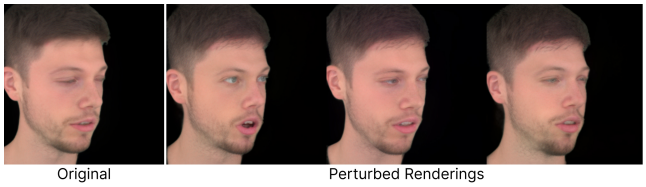

Figure 2: **Renderings from the perturbed parameters γ'_i.**

4. Extra Visual Results

We include a video demonstrating the visual quality of our method. Moreover, we include extra visual results in this section. This is an expanded version of Figure 5 in the main paper, where we showed only one pose/expression for each subject. In this supplement, we show multiple frames for each subject, and also include a couple of additional subjects. Please refer to **Figs.** 3 to 12.

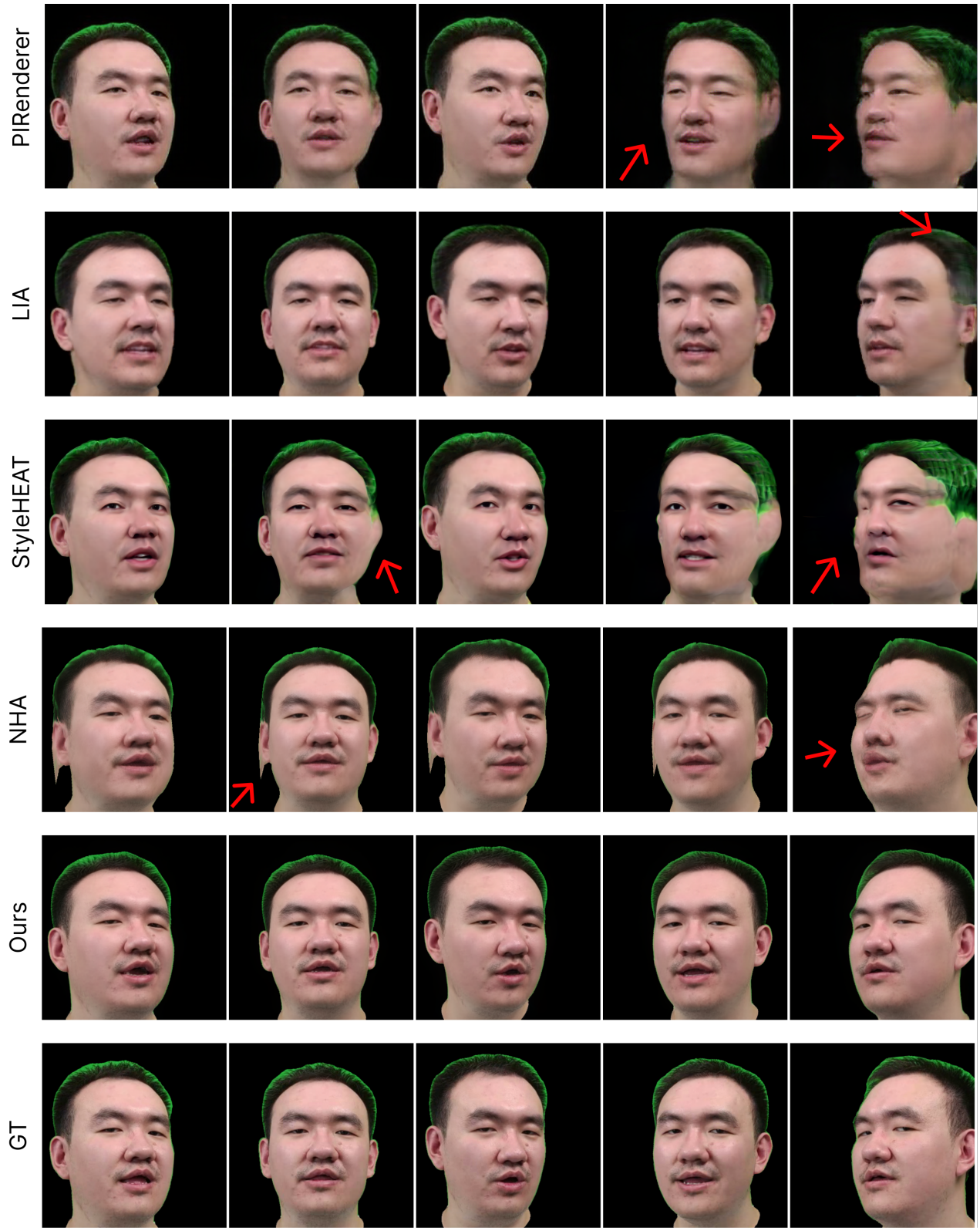

Figure 3: **Qualitative comparison of our method.**

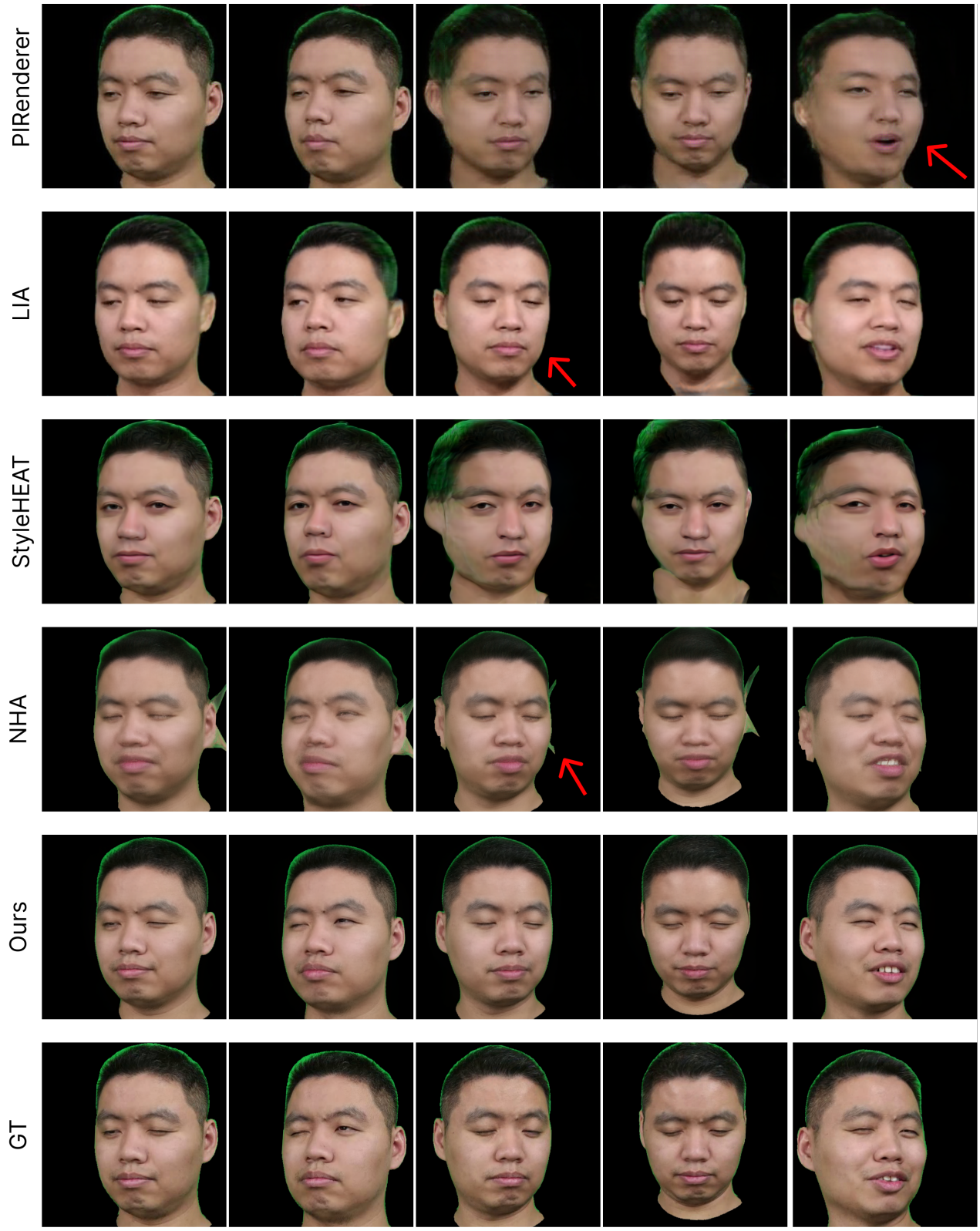

Figure 4: **Qualitative comparison of our method.**

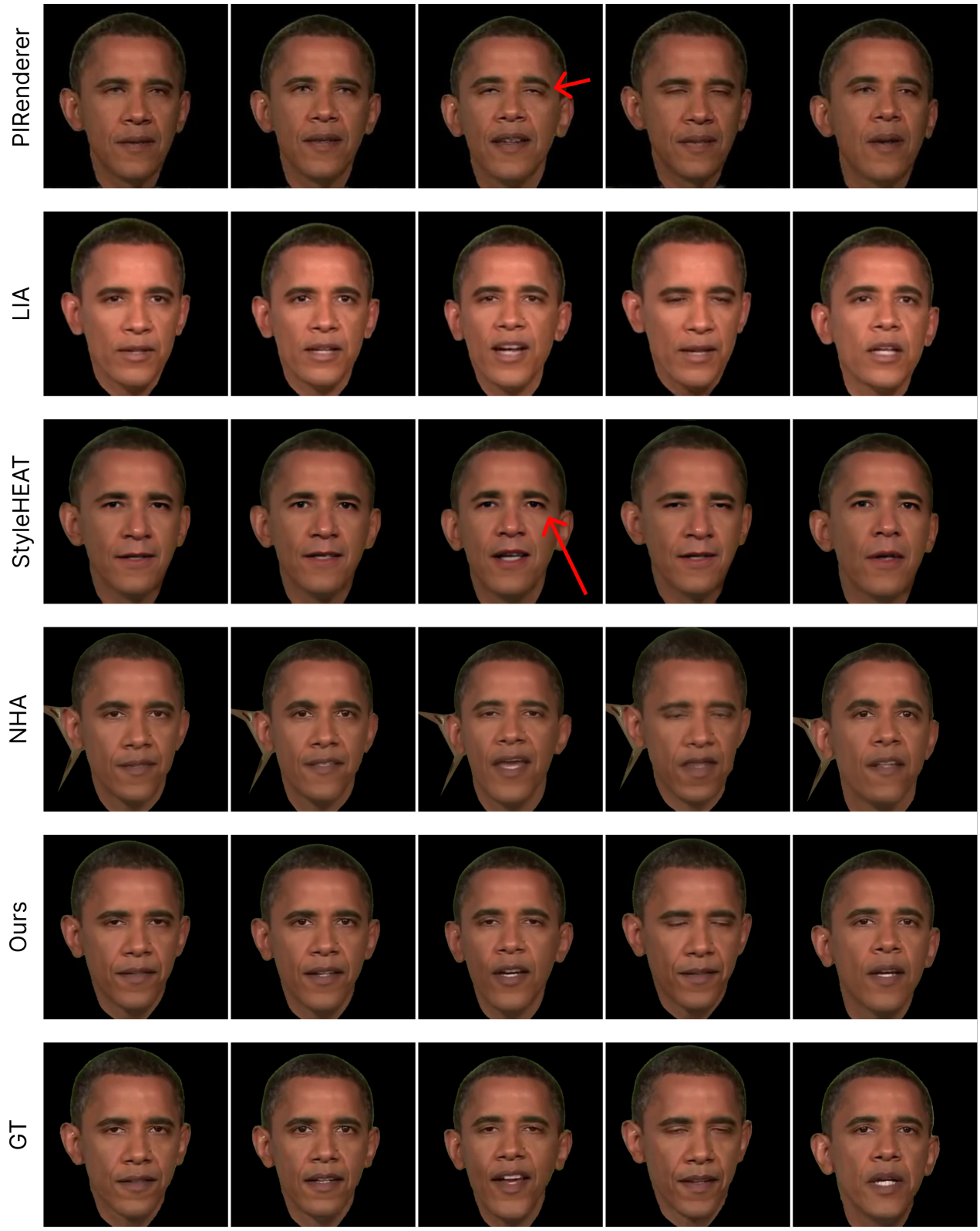

Figure 5: **Qualitative comparison of our method.**

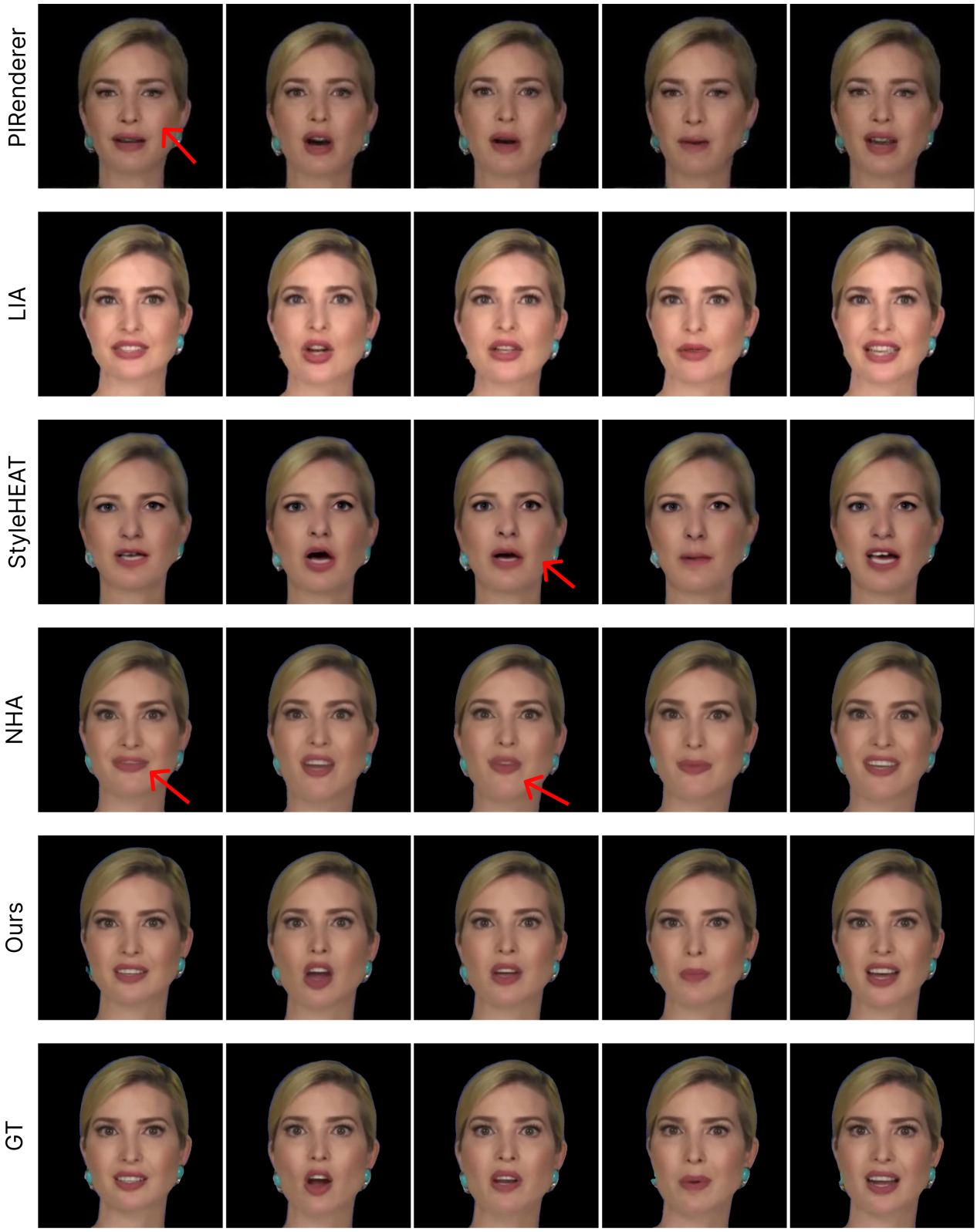

Figure 6: **Qualitative comparison of our method.**

Figure 7: **Qualitative comparison of our method.**

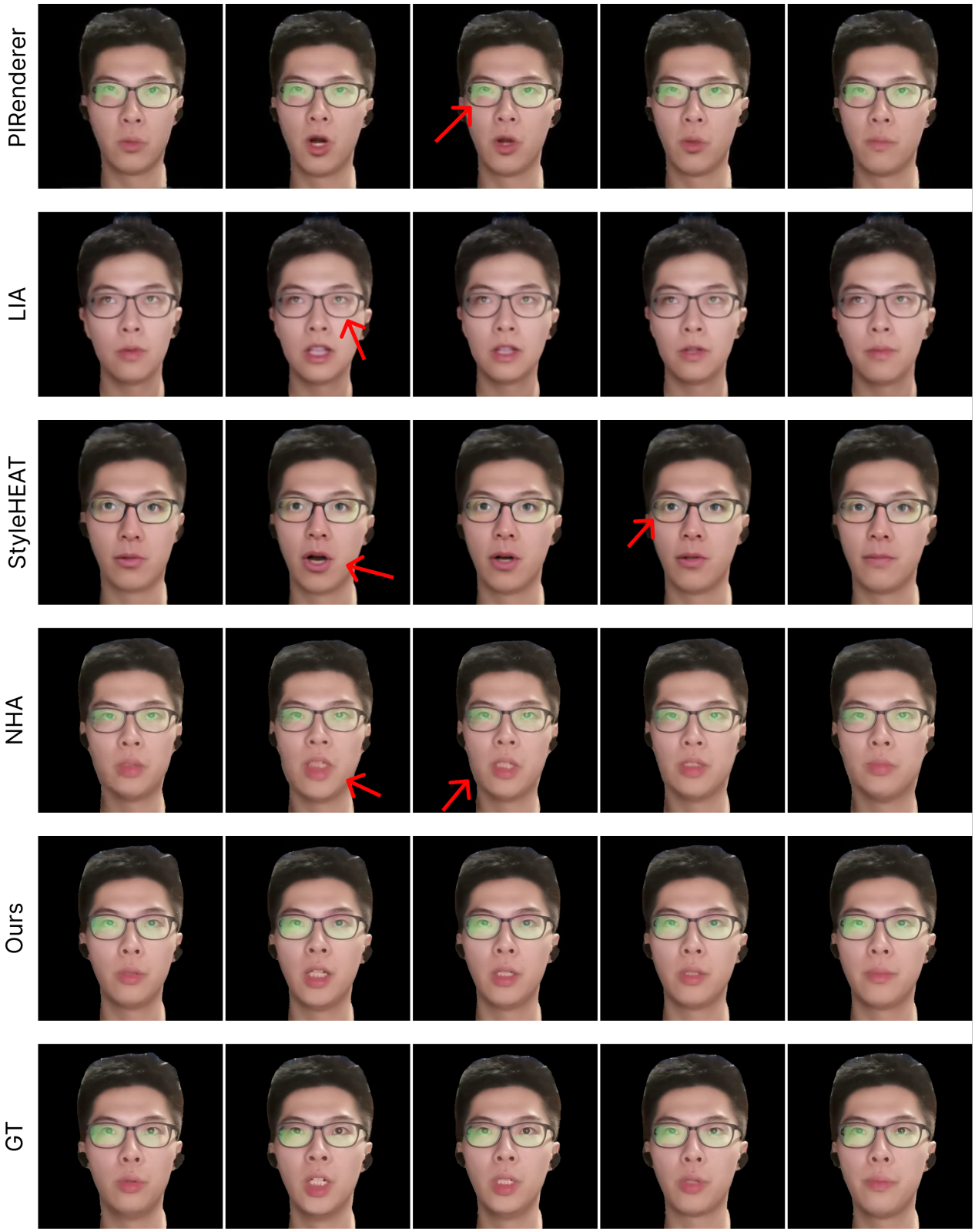

Figure 8: **Qualitative comparison of our method.**

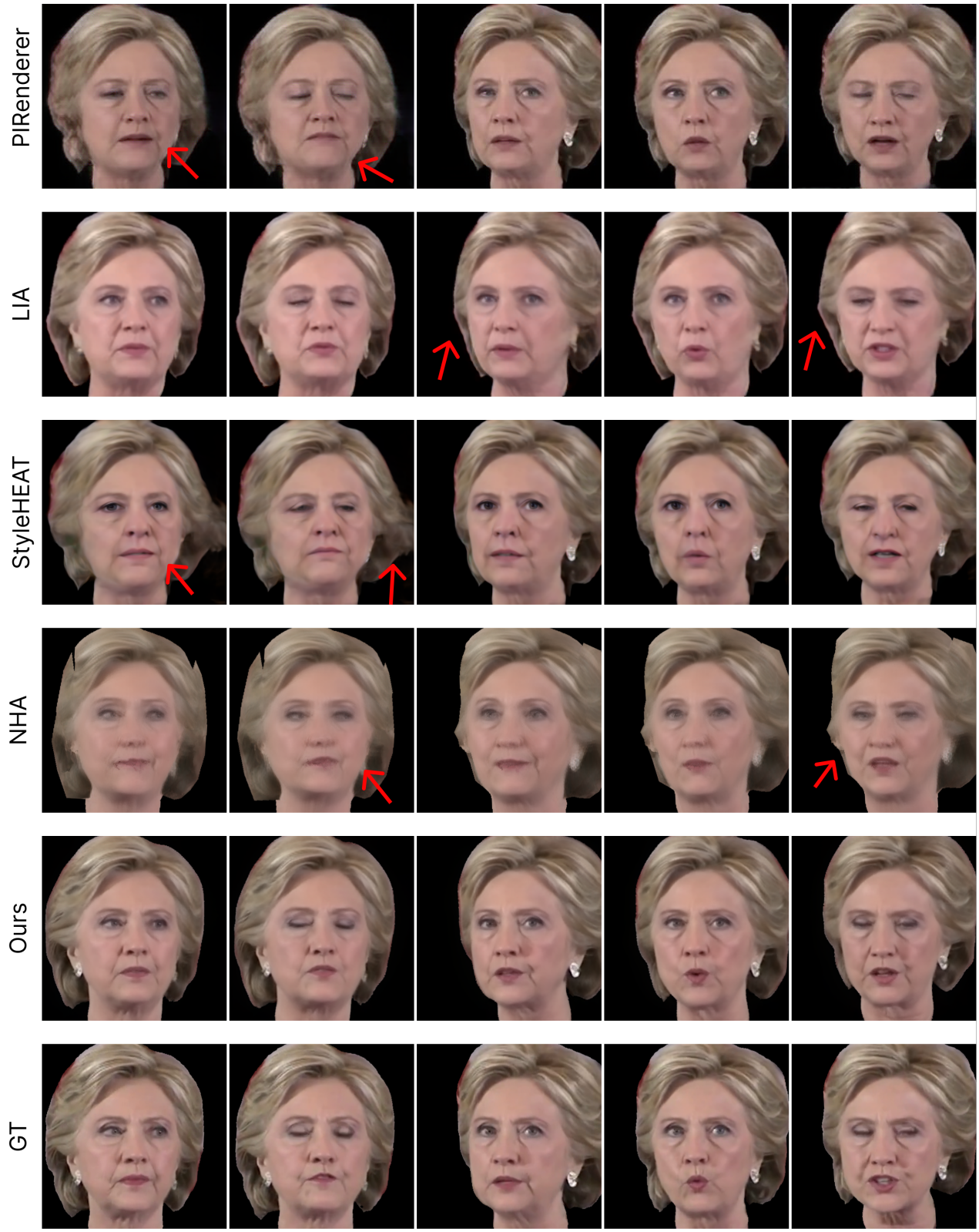

Figure 9: **Qualitative comparison of our method.**

Figure 10: **Qualitative comparison of our method.**

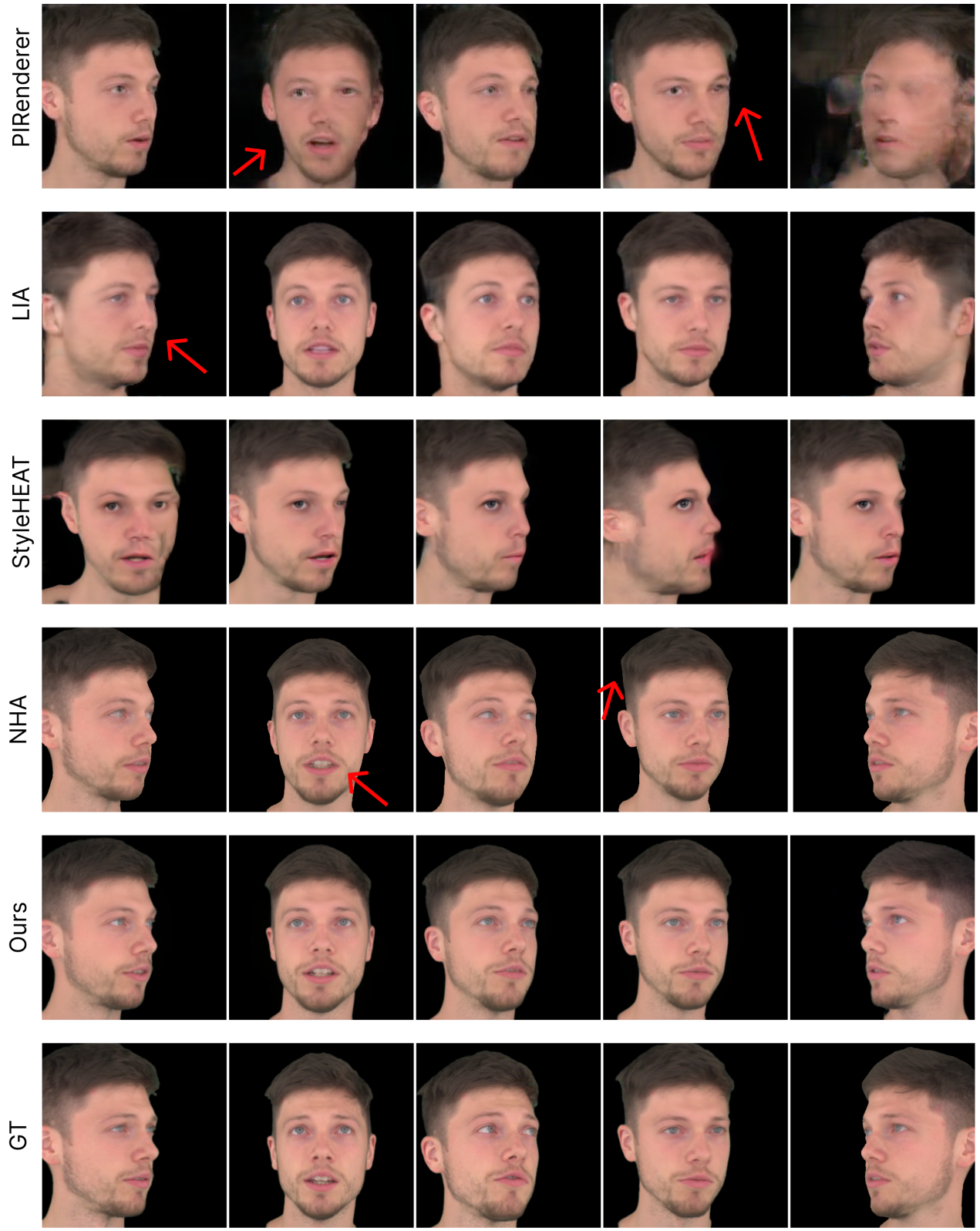

Figure 11: **Qualitative comparison of our method.**

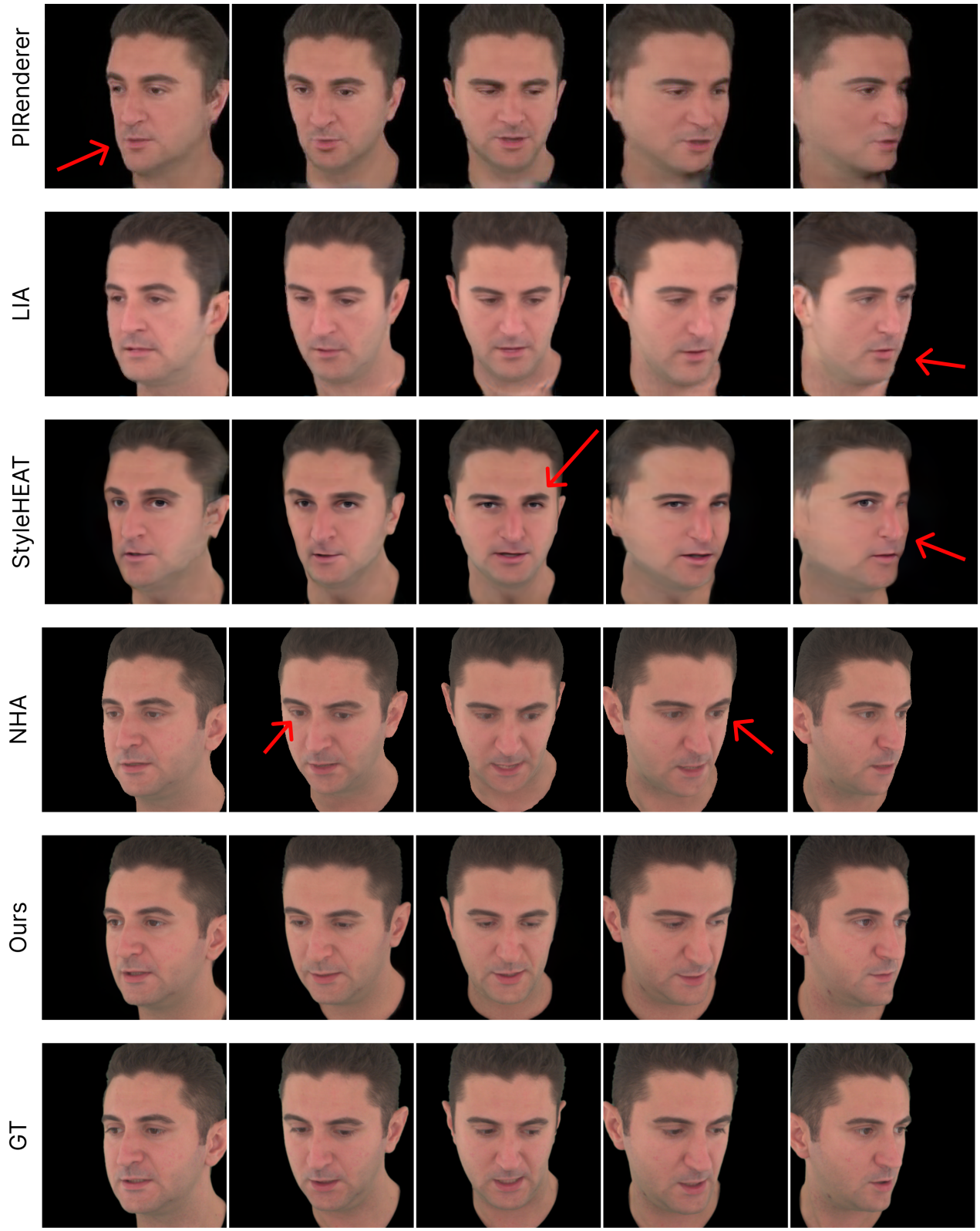

Figure 12: **Qualitative comparison of our method.**