FaceCLIPNeRF: Text-driven 3D Face Manipulation using Deformable Neural Radiance Fields

Supplementary Material

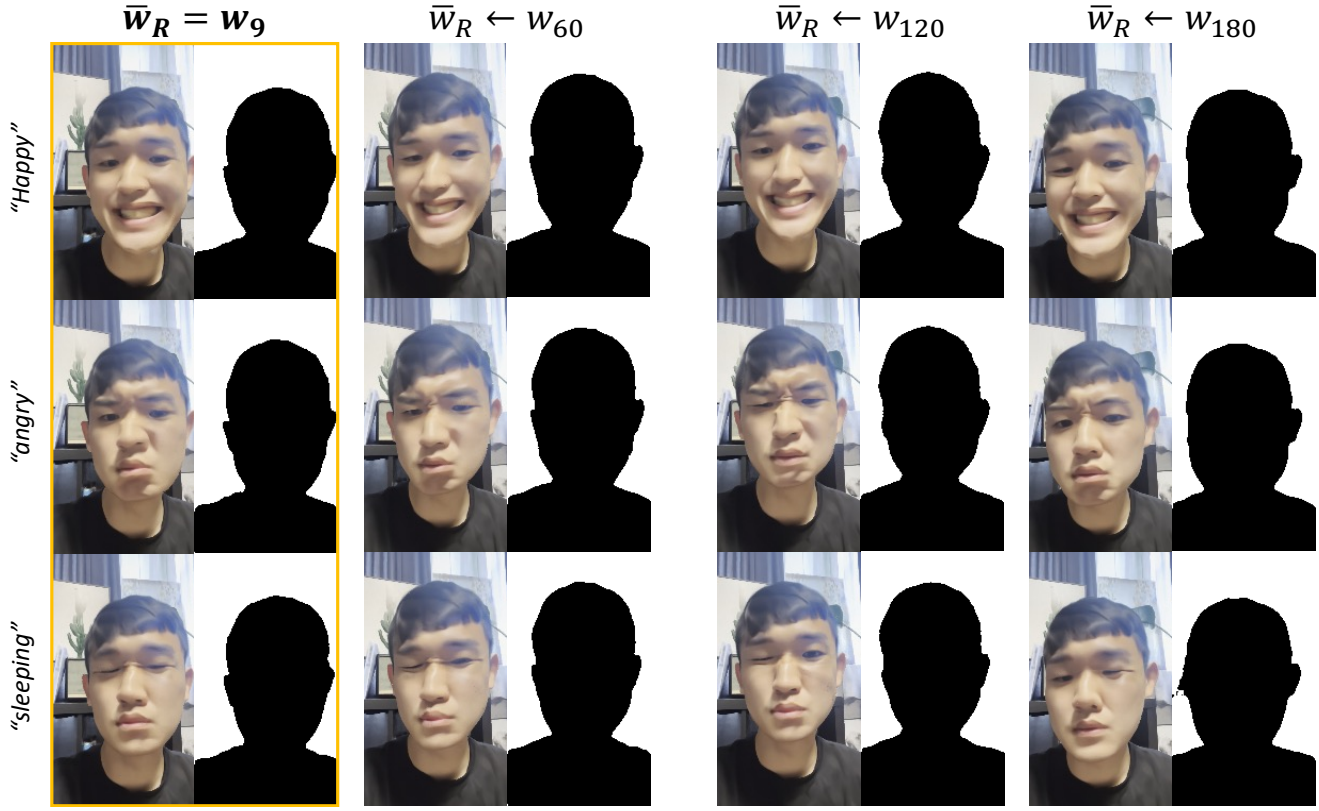

Figure 1: Manipulations rendered with different $\bar{w}_R$ for deformation field T. Our manipulation pipelines for all texts were trained with learned latent code of 9-th training frame, $\bar{w}_9$. Then, we replaced $\bar{w}_9$ with $\bar{w}_{60}$, $\bar{w}_{120}$ and $\bar{w}_{180}$ to observe any difference in renderings. Human segmentation masks are also provided for clearer visualization of head poses.

1. Constant Latent Code for Deformation Field

We learned from experiments that fixing $\bar{w}_R$, a latent code to deformation field T, creates strong controllability of head pose. For example, we conducted text-driven manipulation tasks by allocating fixed $\bar{w}_9$, a trained latent code of 9-th training frame, to $\bar{w}_R$. After training PAC for manipulations, we replaced $\bar{w}_R$ with $\bar{w}_{60}$, $\bar{w}_{120}$, and $\bar{w}_{180}$ during inference over the manipulated scenes to observe any difference. Following results are shown in Fig. 1, where we provide human segmentation masks for clearer visualizations of head poses. One could observe that head poses are relatively constant given a constant $\bar{w}_R$ regardless of the text used for our manipulation method, whereas head poses do change when using different latent code for $\bar{w}_R$. We made a qualitative measure to verify such observations by measuring two metrics: Intra-mIoU and Inter-mIoU. Intra-mIoU is measured by calculating mIoU between all possible pairs of human segmentation masks extracted from the renderings of the same latent code for $\bar{w}_R$. We calculate the average on manipulation networks trained with all 7 proposed text prompts, and the images used to extract human segmentation mask are all rendered in 200 novel views. Meanwhile, Inter-mIoU is calculated using pairs of segmentation masks

Intra-mIoU	Inter-mIoU
0.986	0.816

Table 1: Intra-mIoU and Inter-mIoU calculated with human segmentation masks extracted from images rendered with 4 different latent codes for $\bar{w}_R$ fed to deformation network T. Higher Intra-mIoU means that more constant head poses are rendered over the same $\bar{w}_R$, whereas lower Inter-mIoU means that head poses changes more for different $\bar{w}_R$.

extracted from different latent codes for $\bar{w}_R$. Following results are reported in Table. 1. Intra-mIoU is close to 1, meaning that the head poses are almost constant for constant $\bar{w}_R$, while Inter-mIoU is smaller than Intra-mIoU by large amount (0.170), which means that head poses change more drastically when using different codes for $\bar{w}_R$.

Another observation noticed in Fig. 1 is that facial expressions slightly change when replacing different latent code for $\bar{w}_R$, even when there is no change in PAC network and anchor codes. We may conclude from the observation that T intervenes over both rigid deformations such as head pose and detailed deformations such as facial expressions, whereas H only controls detailed facial expressions. We may leave the problem of selecting optimal latent code for $\bar{w}_R$ or explicit head control fully disentangled from facial expression for future research topics.

2. Observed facial deformations in our Dataset

In this section, we report the types of facial deformations observed during the scene manipulator training in Fig. 2. We show that only 4 types of facial deformations are set to be available for each scene. As one of the key motivations of our work is to make the manipulation user-friendly, requiring the subjects to make only a few types of deformation and to collect approximately 300 frames (10 seconds video for 30fps) greatly reduces the amount of labor for non-expert users during data collection phase.

However, a recommended requirement of the observed facial deformations of a scene is that each face part is provided with at least two oppositely extreme deformations, so that locally composited latent codes can express wider range of interpolated, thus unseen, types of local deformations. For example, a data is required to include at least closed eyes and widely opened eyes to express as many types of eye deformations in between. Same applies to other face parts such as mouth, eyebrows and skin between the eyebrows. We learned that the 4 types of the selected facial deformations are enough to cover the wide range of deformations for each local face part, yet further researches can be done to find optimal set of facial deformation types.

Figure 2: Observed facial deformations in our dataset used for the experiments. All volunteers were asked to make the following facial expressions during the data collection phases.

3. Quantitative Results by Texts

Here, we report qualitative metrics on each text prompt selected for text-driven manipulation tasks. For R-Precision [1], our approach outperformed baseline methods except for the text *"fearful"*. However, considering that NeRF + *FT* fine-tunes the whole network to approach its renderings to target text in CLIP embedding, it has more capacity to over-fit to text attribute reflection at the cost of visual quality and face identity preservation. In addition, our approach out-performed all baselines in LPIPS[2], which implies the contribution of Lipschitz regularization and alpha total variation loss for high visual quality of rendered images. Ours outperformed in Cosine Face Similarity in most text prompts. However, ours showed a relatively low performance in *"crying"*. Such result can be attributed to the ability of our method to render unseen facial deformations by compositing local face deformations observed in different instances. Since the word *"crying"* requires complicated compositions of local deformations during manipulation, the face identity may have varied slightly. However, not only the difference is minute, but ours outperformed in most text prompts in all metrics. As so, we may conclude

	"happy"	*"surprised"*	*"fearful"*	*"angry"*	*"crying"*
NeRF + *FT*	*0.908*	*0.809*	**0.503**	*0.768*	*0.828*
Nerfies + *I*	0.235	0.166	0.205	0.193	0.264
HyperNeRF + *I*	0.771	0.263	0.154	0.233	0.288
Ours	**1.000**	**0.843**	*0.313*	**0.849**	**0.898**

Table 2: R-Precision[1] by different text prompts for manipulation

	"happy"	*"disappointed"*	*"surprised"*	*"scared"*	*"angry"*	*"crying"*	*"sleeping"*
NeRF + *FT*	0.337	0.349	0.342	0.356	0.351	0.342	0.376
Nerfies + *I*	*0.192*	0.218	*0.188*	0.246	0.243	0.232	0.234
HyperNeRF + *I*	0.222	*0.171*	0.192	*0.206*	*0.196*	*0.208*	*0.191*
Ours	**0.090**	**0.052**	**0.107**	**0.089**	**0.070**	**0.097**	**0.070**

Table 3: LPIPS[2] by different text prompts for manipulation

	"happy"	*"disappointed"*	*"surprised"*	*"scared"*	*"angry"*	*"crying"*	*"sleeping"*
NeRF + *FT*	0.466	0.306	0.361	0.296	0.246	0.364	0.410
Nerfies + *I*	*0.696*	0.722	0.655	0.678	*0.668*	**0.690**	0.676
HyperNeRF + *I*	0.672	**0.859**	*0.690*	*0.695*	0.650	0.685	*0.793*
Ours	**0.785**	*0.837*	**0.712**	**0.704**	**0.787**	*0.613*	**0.806**

Table 4: Cosine Face Similarity by different text prompts for manipulation

that our approach is the best option for a text-driven manipulation task of a face in NeRF.

4. More Qualitative Results and Comparisons to Baselines

We report all text-driven manipulation results over all scenes captured from six volunteers in Fig 3 - Fig 8. We may also conclude from the extensive qualitative comparisons that our manipulation approach reflects target attributes most faithfully while preserving the visual quality and face identity on most results.

However, there are a few failure cases to reflect on. The face manipulated with *"angry"* in Fig. 7 shows a face identity that is slightly different from the original, whereas the face manipulated with *"crying"* in Fig. 7 exhibits a noticeable degradation in visual quality. We may conclude from the observation that such a few failure cases are drawbacks of (i) Anchor Composition Network (ACN) that regularizes the natural renderings of face parts by assimilating adjacent latent codes, and of (ii) Lipschitz regularization to increase visual quality of interpolated latent codes. When ACN smoothens the adjacet latent codes, the network is not provided with prior information on facial parts, meaning that even the smoothing can be applied to unwanted locations such as boundaries of two different facial parts. Also, Lipshitz regularization is an implicit regularization method, meaning that the network is not provided with ground truth renderings of interpolated codes to be supervised. As so, further research may focus on compositing deformation given information over face priors on in order to prevent such failure cases.

References

[1] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2, 3

[2] 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*, pages 586–595, 2018. 2, 3

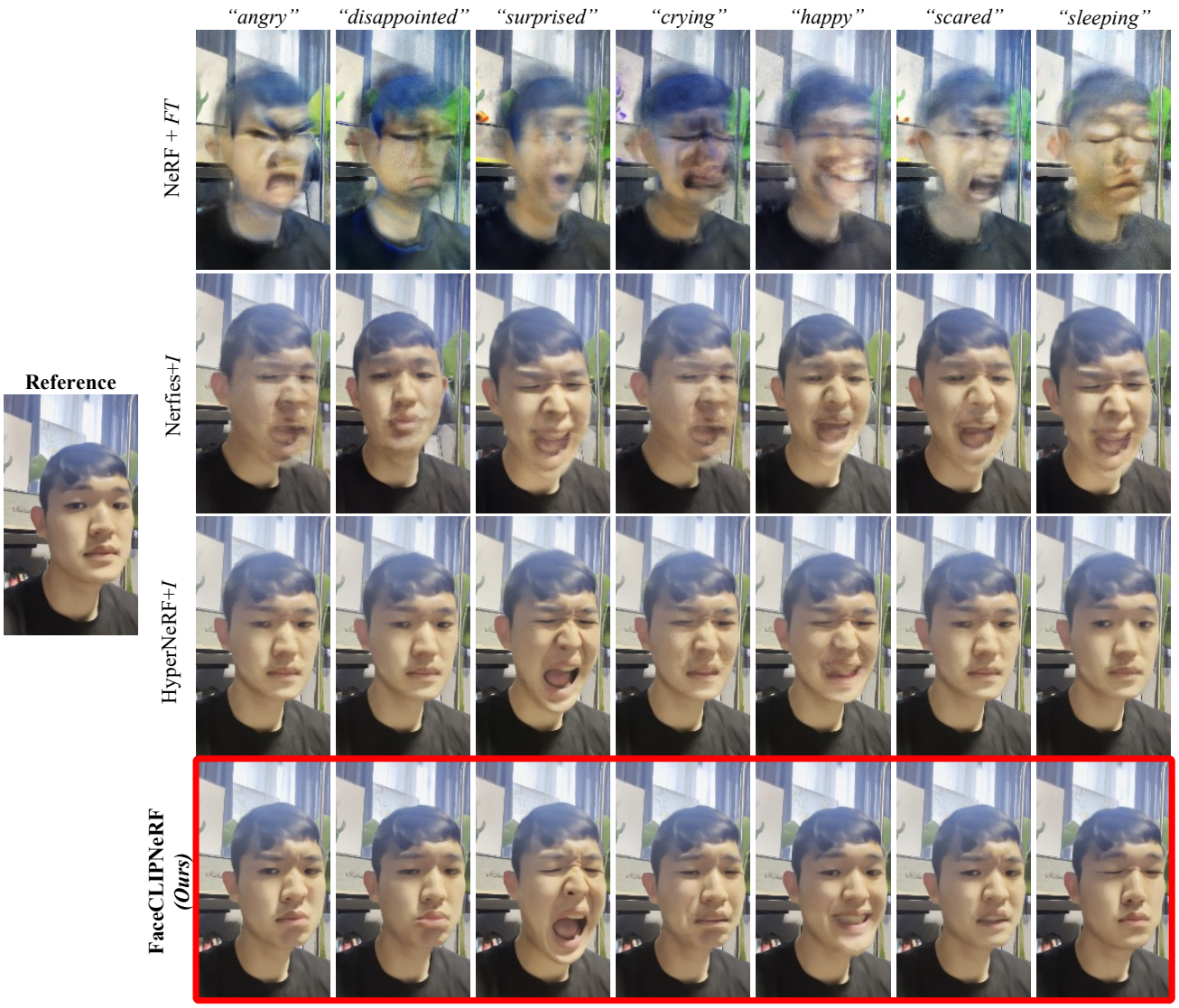

Figure 3: Text-driven manipulation results on volunteer #1 using baseline methods and our approach.

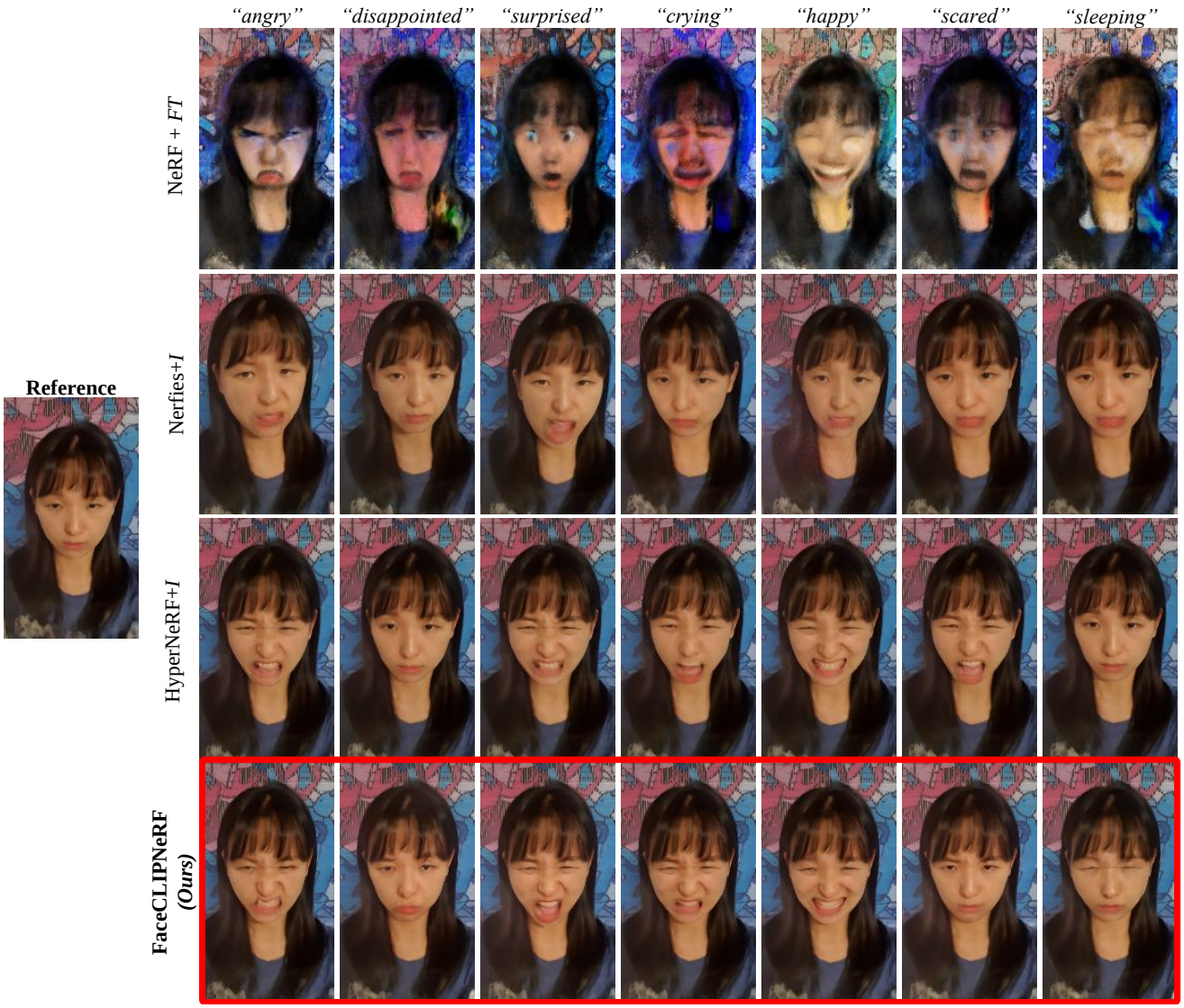

Figure 4: Text-driven manipulation results on volunteer #2 using baseline methods and our approach.

Figure 5: Text-driven manipulation results on volunteer #3 using baseline methods and our approach.

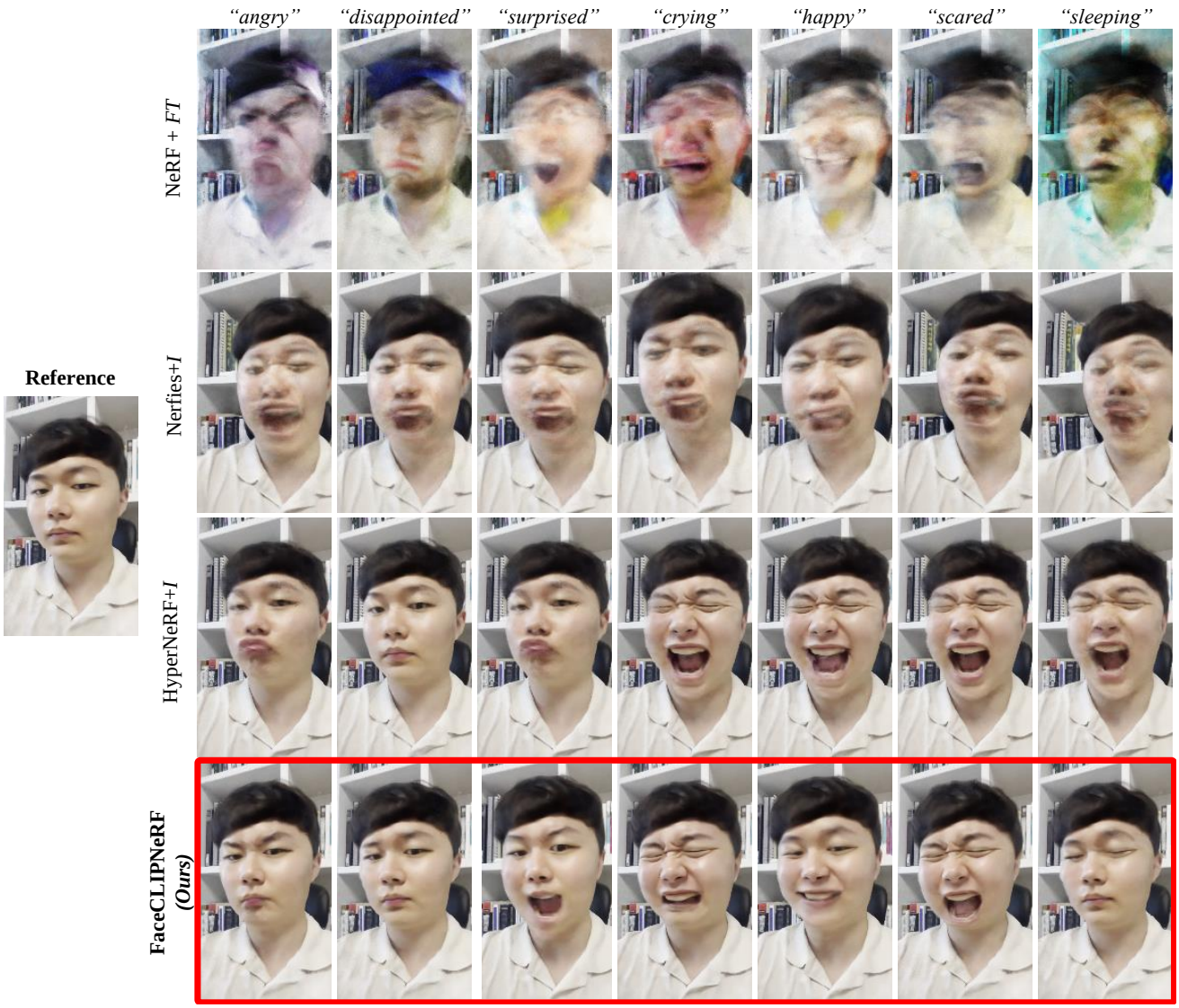

Figure 6: Text-driven manipulation results on volunteer #4 using baseline methods and our approach.

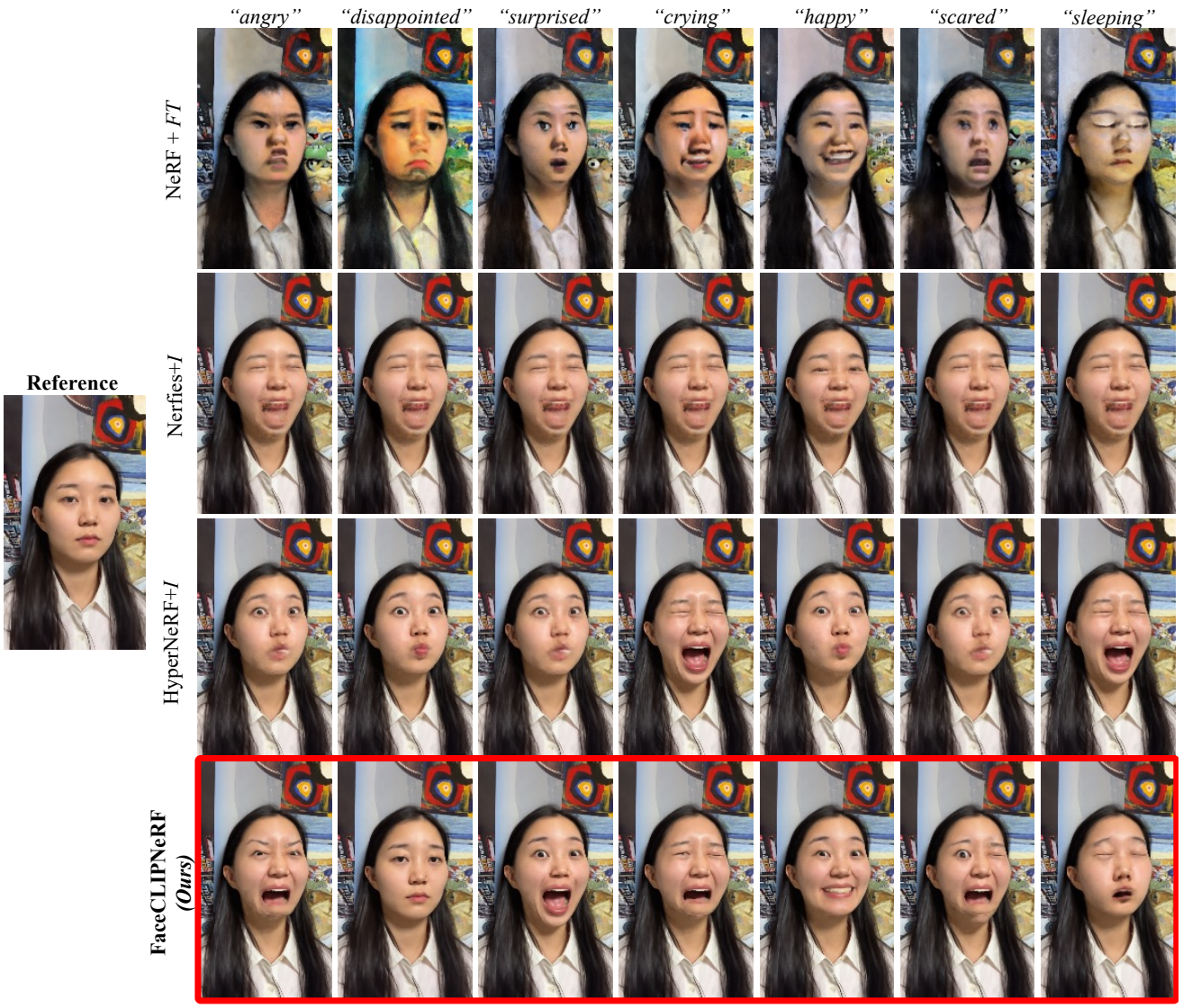

Figure 7: Text-driven manipulation results on volunteer #5 using baseline methods and our approach.

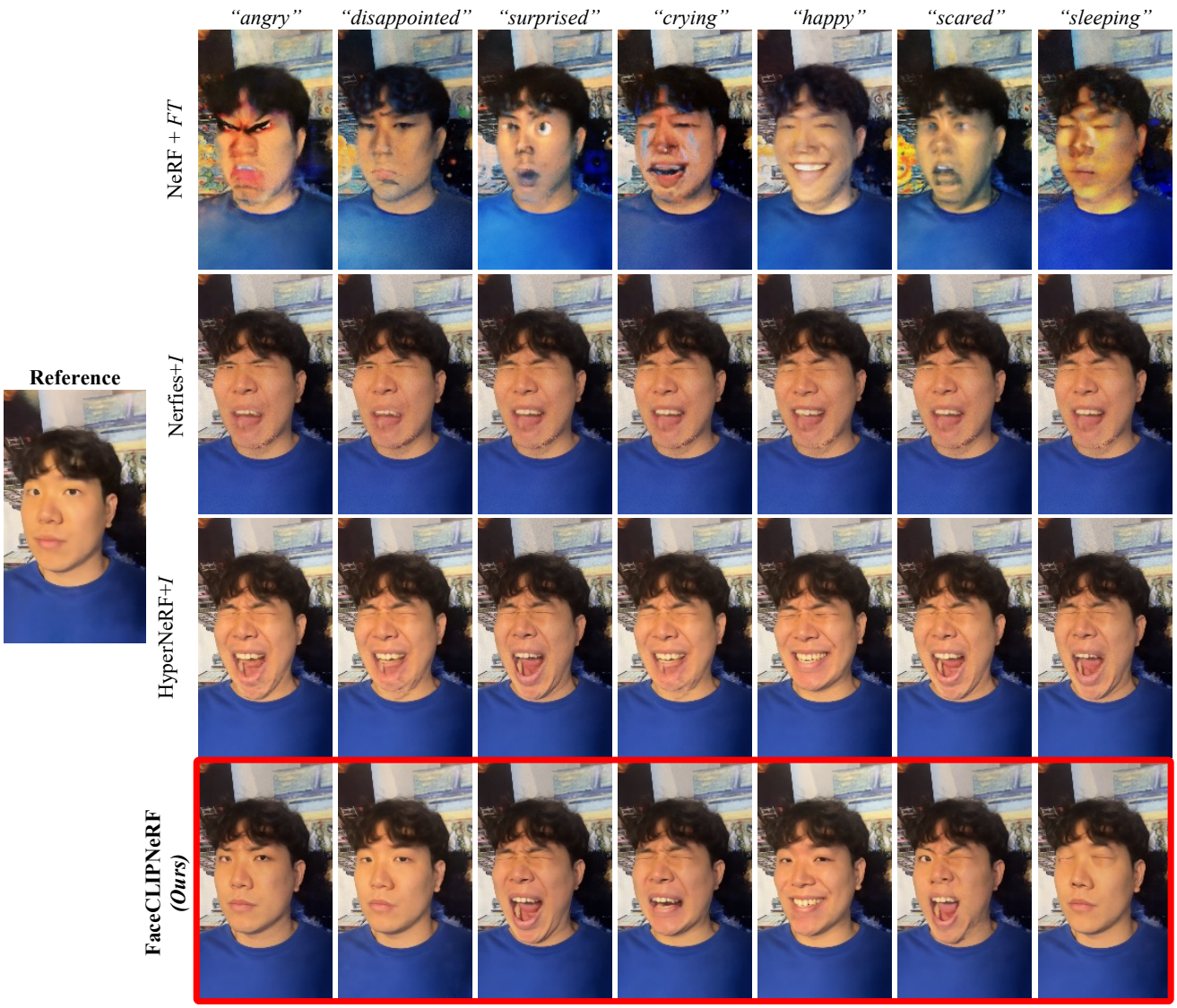

Figure 8: Text-driven manipulation results on volunteer #6 using baseline methods and our approach.