Supplementary Materials of Controllable and Expressive One-Shot Video Head Swapping

Chaonan Ji Jinwei Qi Peng Zhang Bang Zhang Liefeng Bo
Alibaba Group

1. Training Details

The training stage consists of two stages. We first train our model without MotionModule and train it for 10w steps with a batch size of 16 using 8 A100 GPUs. Then we freeze the pretrained model weights and train MotionModule for 2w steps with video length of 24 and a batch size of 8. The MotionModule are initialized with AnimateDiff [3] similar to AnimeAnyone [6]. Referring to the Hallo2 [1] long-duration generation, we improved the temporal consistency of generated video clips by integrating 3 previously generated frames as temporal reference frames. During inference, we utilize the DDIM sampler [8] and set the classifier-free guidance scale [5] to 3.5.

2. One-Step Reconstruction

Given the latent embedding z_0, the noised latent embedding z_t at timestep t, the learned conditional score ϵ_θ, we can recover latent embedding $\hat{z_0}$:

$$\hat{z_0} = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta}{\sqrt{\bar{\alpha}_t}} \tag{1}$$

Then it is decoded into an denoised video sequence $\hat{I}_d$ using VAE decoder. Benefiting from pixel-level supervision, the generated images exhibit improved identity consistency.

3. Application

Owing to the framework design that decouples ID, expression, and background, our approach facilitates not only head swapping but also extends to various video editing capabilities, such as video background replacement, expression editing, and hair generation control.

Video background replacement. By replacing the background of the driving video during the inference stage, our method can generate seamless driving videos against the target background, achieving cohesive foreground-background integration. The results is shown in Fig.1.

Expression editing. Benefiting from the design of the expression-aware retarget module, we are able to modulate the intensity of expressions by adjusting the proportions of face features, as shown in Fig.3. Moreover, this design facilitates the decoupling of identity, expression, and pose, enabling the transfer of expressions from a third person to produce the final output. The results are shown in Fig.2.

Hair generation control. By adjusting the size of the shoulder mask, we can control the hair generation scope for long-haired portraits. The results are shown in Fig.4.

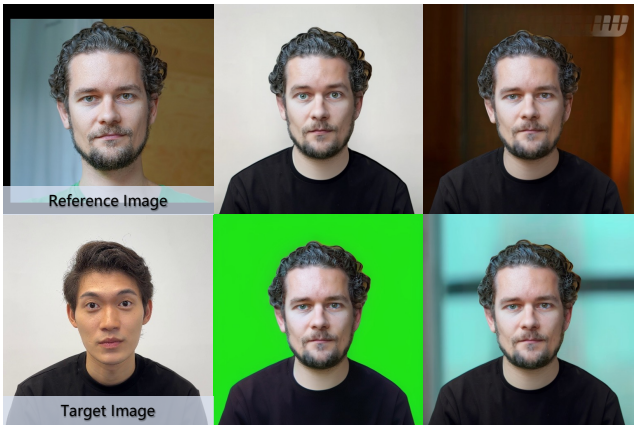

Figure 1. The results of video background replacement.

Figure 2. The results of id, pose and expression composition.

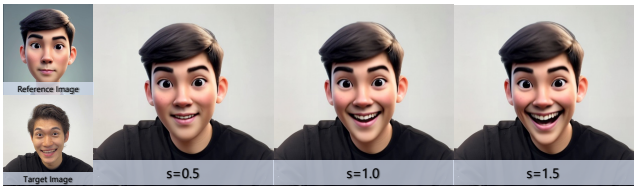

Figure 3. The results of expression editing.

Figure 4. The results of hair generation control.

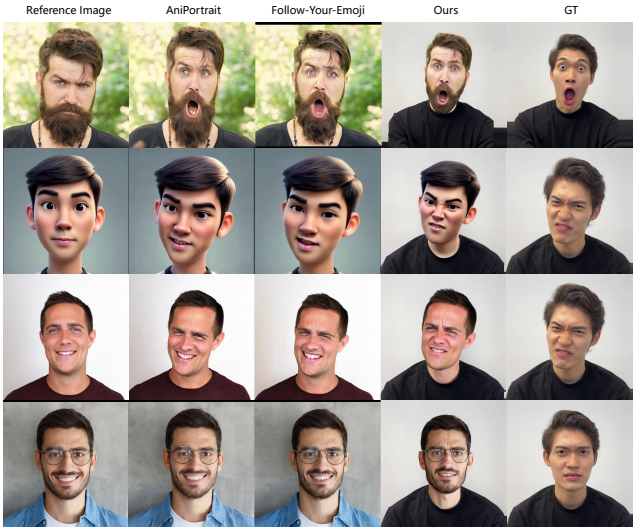

Figure 5. The comparison results with portrait animation methods

4. More Results

4.1. Portrait Animation

Fig.5 shows the portrait animation results comparing our method to others. Our method outperforms Follow-Your-Emoji [7] and AniPortrait [9] in terms of expression transfer. The scale-aware retargeting strategy allows better preservation of identity under extreme expressions. Furthermore, the neutralization of the input image's expression enhances the accuracy of expression transfer.

4.2. Consistent Lighting

With the aid of data augmentation strategies, our method demonstrates a certain level of capability in harmonizing lighting, leading to visually plausible and natural-looking synthesis results, as shown in the supplementary material.

Figure 6. The reference image is relit using IC-Light to match the lighting conditions of the target image.

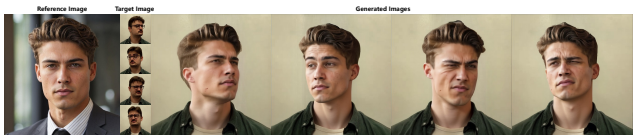

Figure 7. The results of large head movement.

Figure 8. The results of dynamic background.

Method	ID Similarity ↑	Pose ↓	Expression ↓
LivePortrait	0.890	20.1	0.029
X-Portrait	0.842	9.92	0.017
FaceAdapter	0.604	3.51	0.011
Ours	0.895	9.83	0.014

Table 1. The quantitative results.

While not perfect, the outputs are generally acceptable under most conditions. Additionally, state-of-the-art relighting techniques, such as IC-Light [11], can be used to modify the lighting conditions of reference images to ensure consistency with driving video, as illustrated in Fig.6.

4.3. Dynamic Background

Fig.7 and Fig.8 present the generated results under large poses and dynamic backgrounds, indicating that our method can reasonably transfer facial expressions under large poses and produce plausible head-swapping results in dynamic background settings.

4.4. Comparison with SOTA methods

We have included comparative experimental results with LivePortrait [2], X-Portrait [10] and FaceAdapter [4], as illustrated in Tab.1. Our method shows promising results in both facial expression transfer accuracy and the naturalness of head-swapping results.

References

[1] Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *CoRR*, abs/2410.07718, 2024. 1

[2] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *CoRR*, abs/2407.03168, 2024. 2

[3] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 1

[4] Yue Han, Junwei Zhu, Keke He, Xu Chen, Yanhao Ge, Wei Li, Xiangtai Li, Jiangning Zhang, Chengjie Wang, and Yong Liu. Face-adapter for pre-trained diffusion models with fine-grained ID and attribute control. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part L*, pages 20–36. Springer, 2024. 2

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. 1

[6] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 8153–8163. IEEE, 2024. 1

[7] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, and Qifeng Chen. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers, SA 2024, Tokyo, Japan, December 3-6, 2024*, pages 110:1–110:12. ACM, 2024. 2

[8] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1

[9] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *CoRR*, abs/2403.17694, 2024. 2

[10] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers, SIGGRAPH 2024, Denver, CO, USA, 27 July 2024- 1 August 2024*, page 115. ACM, 2024. 2

[11] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 2