Supplementary Material for "PerformRecast: Expression and Head Pose Disentanglement for Portrait Video Editing"

Jiadong Liang* Bojun Xiong* Jie Tian Hua Li Xiao Long Yong Zheng Huan Fu†

HUJING Digital Media & Entertainment Group

*Equal contribution †Corresponding author

Table of Contents

We first provide a brief overview of our supplementary material. This supplementary material consists of the following sections and contents:

- Sec. 1: LLM usage claim.
- Sec. 2: The detailed process of our used 3DMM-based face tracking method Pixel3DMM [10].
- Sec. 3: Definitions of all training loss terms used in our model.
- Sec. 4: Describes the facial and non-facial masks calculation process of each frame.
- Sec. 5: Provides the inference process of portrait animation task.
- Sec. 6: Shows our keypoints selection strategy.
- Sec. 7: Contains the specific training settings to train our model.
- Sec. 8: The detailed definition of each evaluation metric.
- Sec. 9: Construction pipeline of our test benchmark.
- Sec. 10: How we modify four diffusion-based methods to support the portrait video editing task.
- Sec. 11: Provides more generated results of our PerformRecast, including both portrait video expression editing and portrait animation.
- Sec. 12: Discusses the limitations of our method and future plans.
- Sec. 13: Ethics statement of our method to avoid malicious use.

1. LLM Use Claim

We employ a large language model (LLM) to assist with the language polishing and revision of certain sections of our paper, including the supplementary material. The LLM is used solely to enhance grammar, clarity, and overall readability by rephrasing sentences, correcting linguistic errors, and ensuring stylistic consistency. All authors have carefully reviewed and approved the final manuscript and take full responsibility for its content.

2. 3DMM-based Face Tracking

To obtain temporally continuous FLAME [18] model reconstruction results from input portrait videos, We adopt a recently-proposed 3D face tracking method, Pixel3DMM [10] to predict FLAME parameters of each frame from input portrait videos. Pixel3DMM firstly trains two expert networks: $\mathcal{N}$ and $\mathcal{U}$, which are built on the top of the pretrained DINOv2 [21] backbone to predict surface normal $\mathcal{N}(I)$ and UV-space coordinate $\mathcal{U}(I)$ given a portrait image I.

Then, it optimizes for FLAME parameters [18], including face identity $\beta \in \mathbb{R}^{300}$, expression $\psi \in \mathbb{R}^{100}$, head pose $\theta \in \mathbb{R}^{3*4+3=15}$ and other camera parameters. The head pose θ contains four 3D rotation vectors for four joints: $\theta_{\text{neck}}, \theta_{\text{jaw}}, \theta_{\text{left-eyeball}}, \theta_{\text{right-eyeball}}$ and one global rotation θ_{head} in axis-angle. Specifically, Pixel3DMM directly uses MICA's [38] identity prediction as β. The remaining parameters are optimized via minimizing a 2D vertex loss and a normal rendering loss between the projection of current estimated FLAME model and predicted UV-space coordinate $\mathcal{U}(I)$ as well as surface normal $\mathcal{N}(I)$.

For monocular video tracking, Pixel3DMM freezes $\mathbf{z}_{\text{id}}$ using the average result of MICA's [38] identity predictions across all frames. Then, it sequentially optimize for the remaining parameters for each frame. Finally, it adds a smoothness term to ensure smoothness across all frames.

As a result, Pixel3DMM is capable of reconstructing temporally continuous FLAME parameters and fixed face identity of each input portrait video.

3. Definitions of Training Loss Terms

We utilize $\mathcal{L}_{\text{animate}}$ which is described in the main manuscript to train our teacher and student models. $\mathcal{L}_{\text{animate}}$ is formulated as:

$$\mathcal{L}_{\text{animate}} = \mathcal{L}_{\text{FLAME}} + \mathcal{L}_{P,\text{cascade}} + \mathcal{L}_{1,\text{cascade}} + \mathcal{L}_{G,\text{cascade}} + \mathcal{L}_{\text{faceid}}, \tag{1}$$

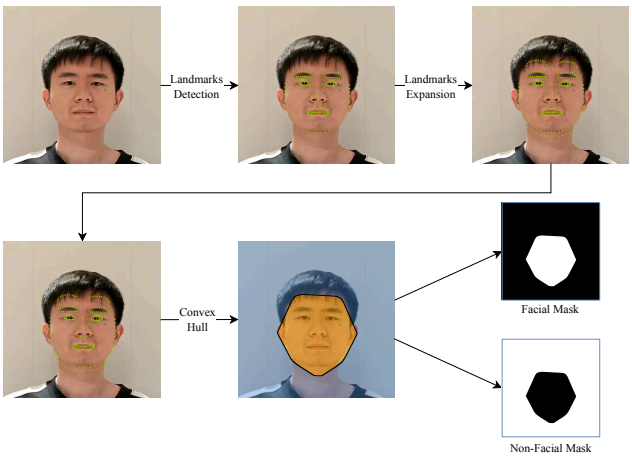

Figure 1. The facial mask calculation process of each frame in training dataset.

To calculate the difference between the reconstructed frame $\hat{I}_d$ and driving frame I_d, we utilize three commonly-used loss term: the perceptual loss, L_1 loss and GAN-loss. To further improve the texture quality, the perceptual loss, L_1 loss and GAN loss are applied on both global region and local regions of face and lip, which are denoted as a cascaded perceptual loss $\mathcal{L}_{P,\text{cascade}}$, a cascaded L_1 loss $\mathcal{L}_{1,\text{cascade}}$ and a cascaded GAN loss $\mathcal{L}_{G,\text{cascade}}$. $\mathcal{L}_{G,\text{cascade}}$ consists of $\mathcal{L}_{\text{GAN,global}}$, $\mathcal{L}_{\text{GAN,face}}$ and $\mathcal{L}_{\text{GAN,lip}}$, which depend on the corresponding discriminators $\mathcal{D}_{\text{global}}$, $\mathcal{D}_{\text{face}}$ and $\mathcal{D}_{\text{lip}}$ training from scratch. The face and lip regions are defined using the 2D semantic facial landmarks which are extracted by a pre-trained landmark detector in LivePortrait [11]. And the face-id [5] loss is used to preserve the identity of source image I_s.

4. Facial Mask Calculation

As shown in Fig. 1, to obtain masks of facial and non-facial regions, we also utilize the pre-trained 2D facial landmark detector in LivePortrait [11] to extract 203 landmarks of each frame from our dataset. Then, we expand the detected 2D facial landmarks of source frame I_s outward and compute their convex hull as the facial region, while the remaining area in I_s is regarded as the non-facial region.

5. Inference Process of Portrait Animation

In the inference phase of portrait animation task, we first extract the appearance feature volume $f_s = \mathcal{F}(I_s)$ from the source image I_s. Given a driving video sequence $\{I_{d,i}|i = 0, 1, ..., N-1\}$, the source and driving explicit keypoints are transformed as follows:

$$\begin{cases} x_s = s_s \cdot \big((x_{c,s} + \delta_s)R_s\big) + t_s, \\ x_{d,i} = s_{d,i} \cdot \big((x_{c,s} + \delta_{d,i})R_{d,i}\big) + t_{d,i}, \end{cases} \tag{2}$$

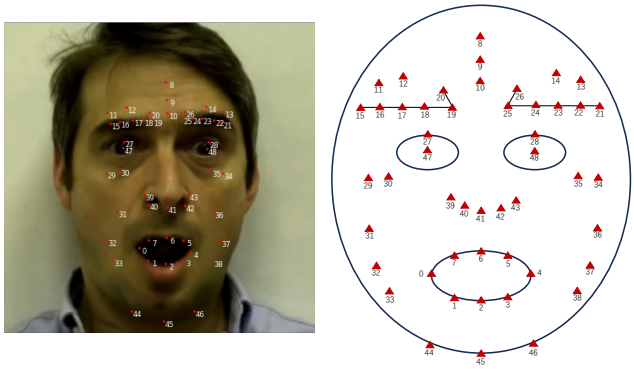

Figure 2. The specific location of each keypoint used in our method. Please zoom in for better inspection.

which utilizes the same formula as training stage.

6. Keypoints Selection

We select $K = 49$ keypoints from the reconstructed FLAME face mesh in total to supervise our motion extractor. Fig. 2 shows the specific location of each keypoint. We select as few keypoints as possible, covering important facial regions such as the forehead, eyebrows, eye sockets, eyeballs, nose, lip, and jaw.

7. Training Settings

We train our model from scratch using 128 NVIDIA H20 GPUs for approximately one week with a batch size of 8 per GPU. We adopt the Adam [16] optimizer with different learning rates for different modules. Specifically, the appearance feature extractor is trained with a learning rate of 5×10^{-5}, while the motion extractor, warping module, and decoder are assigned a higher learning rate of 1.2×10^{-4}. To further stabilize adversarial training, we set the learning rates of the image, face, and lip discriminators to 1×10^{-4}, 2.5×10^{-5}, and 1.5×10^{-5}, respectively. To improve the robustness of training process, we further add random gaussian noise with small variance on extracted keypoints x_s and x_d, but not during inference stage.

8. Evaluation Metrics Details

LPIPS. For potrait video expression editing and self-reenactment, we calculate the perceptual similarity metric LPIPS [36] based on AlexNet [17] between the animated images and ground truth images.

Fréchet Inception Distance. For portrait video expression editing and self-reenactment, FID compares the distribution of generated images with the distribution of ground truth images. The formula for FID is defined as:

$$\text{FID} = ||\mu_g - \mu_r||^2 + \text{Tr}(\Sigma_g + \Sigma_r - 2(\Sigma_r\Sigma_g)^{1/2}), \tag{3}$$

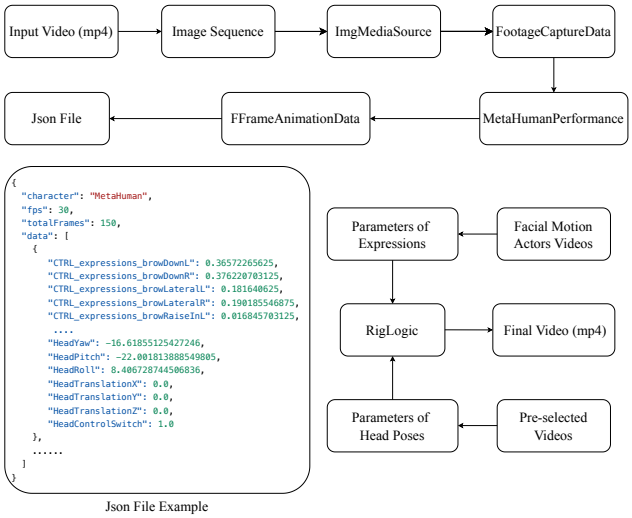

Figure 3. Construction pipeline of our proposed test benchmark.

where g and r denote the features of the generated image and ground truth images, which is extracted by Inception-v3 model [25]. μ and Σ denote the mean and covariance matrices of each image set. A lower FID indicates better generation quality.

Fréchet Video Distance. For portrait video expression editing and self-reenactment, FVD compares the distribution of generated videos with the distribution of ground truth videos. The formula for FVD is similar to FID, which is defined as:

$$\text{FVD} = ||\mu_g - \mu_r||^2 + \text{Tr}(\Sigma_g + \Sigma_r - 2(\Sigma_r\Sigma_g)^{1/2}), \quad (4)$$

where g and r denote the features of the generated videos and ground truth videos, which is extract by the pre-trained Inflated 3D ConvNet [2]. μ and Σ denote the mean and covariance matrices of each video set. A lower FVD indicates better generation quality.

Cosine SIMilarity of identity features. We utilize CSIM to measure the identity preservation between two images, through the cosine similarity of two embeddings from a recently proposed pretrained face recognition network AdaFace [15]. For portrait video expression editing and self-reenactment, the CSIM is calculated between the animated image and ground truth image. For cross-reenactment, the CSIM is calculated between the animated and the source images.

Average Expression Distance. AED is the mean L_1 distance of the expression parameters between the edited and driving images in expression editing task as well as the animated and driving images in portrait animation task. These parameters, which include expression coefficient, eyelid and jaw pose parameters, are extracted by the state-of-the-art 3D face reconstruction method SMIRK [23].

Average Pose Distance. APD is the mean L_1 distance of the pose parameters between the edited and source images in expression editing task as well as the animated and driving images in portrait animation task. The pose parameters are also extract by SMIRK [23].

Mean Angular Error. The mean angular error is used to measure the eyeball direction error between the edited and driving images in expression editing task as well as the animated and driving images in portrait animation task. It is adopted as: $\text{MAE}(I_g, I_d) = \arccos(\frac{\mathbf{b}_g \cdot \mathbf{b}_d}{\|\mathbf{b}_g\| \cdot \|\mathbf{b}_d\|})$, where $\mathbf{b}_g$ and $\mathbf{b}_d$ are the eyeball direction vectors of the generated image I_g (including the edited image and animated image) and the driving image I_d respectively. Both of them are predicted by a pre-trained eyeball direction prediction network [1].

9. Construction of Our Test Benchmark

Fig. 3 visualizes the construction pipeline of our proposed test benchmark. Given a input video, our pipeline first utilize MetaHuman [9] to extract the expression and head pose parameters of each frame. The detailed information of this process are shown at the top of Fig. 3. The extracted parameters of each frame are saved in a json file. Among them, the keys of parameters related to facial expressions start with "CTRL_expressions". The three keys "HeadYaw", "HeadPitch" and "HeadRoll" describe the head pose rotation. Then, we combine the parameters with keys starting with "CTRL_expressions" in the json files extracted from facial motion actors and "HeadYaw", "HeadPitch", "HeadRoll" in the json files extracted from our pre-selected videos containing large head pose rotation to RigLogic system to drive the ditigal human in MetaHuman and create final videos. For enhancement mode, all parameters with keys starting with "CTRL_expressions" are set to zero.

The resolution of original videos synthesized from MetaHuman are set to 2560×1440, which is the default setting. Each video contains 150 frames and is recorded with 30 frames per second (FPS). We crop all the videos into squares to maintain the face at the center and resize them to the resolution of 512×512 for further training.

10. Modification of Diffusion-based Methods

We modify several diffusion-based portrait animation methods to make them support the task of editing the facial expression of source video according to the driving video. All these four methods leverage large-scale pre-trained video diffusion models to animate the input static portrait image

from the driving video. However, our portrait video expression editing task needs to utilize the i-th frame $I_{d,i}$ in driving video to edit the expression of i-th frame $I_{s,i}$ in source video. Therefore, we expand each frame of the driving video into a short static video clip, which is then used to animate the i-th frame of the source video, thus conforming to the video input formula required by video diffusion models. For the source and driving video of N frames, we repeat this animation process for N times, and concatenate N animated images to form the edited video.

To realize expression editing instead of portrait animation, the key idea is to combine the facial expression of driving frame with the head pose of source frame, and use this combined signal to animate the source frame. We then describe the detailed modification of each method.

SkyReels-A1. SkyReels-A1 [22] utilizes SMIRK [23] to extract FLAME [18] parameters of each frame in driving video. In our task, we replace the head pose parameters in FLAME model of driving frame with that of source frame to animate the source frame.

Hunyuan-Portrait. Hunyuan-Portrait [35] utilizes pre-trained motion encoder MegaPortraits [7] to extract facial motion representations of driving video. Specifically, these representations consist of the explicit head rotations R, translations t, and the latent expression descriptors z. Therefore, we replace the head rotations R and translations t of driving frame with those of source frame to animate the source frame.

FantasyPortrait. FantasyPortrait employs a pre-trained implicit expression motion extractor PD-FGC [27] to encode the driving frame into latent features. These latent features include lip motion e_{lip}, eye gaze and blink e_{eye}, head pose e_{head} and emotional expression e_{emo}. And we replace the head pose parameters e_{head} of driving frame with that of source frame to animate the source frame.

Wan-Animate. Wan-Animate uses VitPose [34] to extract the facial skeleton for the character in portrait video as head pose representations. Then, it adopts an encoder structure identical to that of LIA [30] to extract expression features from driving frame. Therefore, we combine the facial skeleton of source frame with expression features from driving frame to animate the source frame.

11. More Results

We provide more generated results of our PerformRecast in this section.

Figure 4. Qualitative comparison of portrait video expression editing on enhancement mode. The top of the figure shows editing results of different methods. The bottom presents our ablation studies and analysis. The bottom-right insets exhibit driving frames. The light green circles highlight the misalignment between the facial and non-facial regions. Please zoom in for better inspection.

11.1. Portrait Video Expression Editing

We first compensate for the missing visual comparisons on the enhancement mode in Fig. 4 as mentioned in the main manuscript. LivePortrait [11] generates inaccurate lip movements on the enhancement mode. Act-Two [24] tends to synthesize exaggerated mouth movements, leading to less realistic facial animations. On the contrary, our method succeeds in enhancing the facial expressions via adding the expressions of driving video on the top of that of source video.

We then show qualitative results of all the compared methods on our proposed test benchmark in Fig. 5. The four modified diffusion-based methods perform extremely poorly on portrait video expression editing task. They all produce incorrect facial expressions with severe artifacts and distortions. As a result, our carefully designed PerformRecast achieves the best performance on both replacement and enhancement modes compared to all previous approaches. We also provide all the original videos in Fig. 5 under `Comparison` folder in our supplementary material.

Furthermore, we also uncuratedly select 10 well-known movie clips from different countries, and use the driving videos in our test benchmark captured from facial motion actors to edit the expressions of them. Among them, five are edited with replacement mode and the remaining five are edited with enhancement mode. These results are included in `Movies` folder in our supplementary material. The source video, the edited results of LivePortrait and

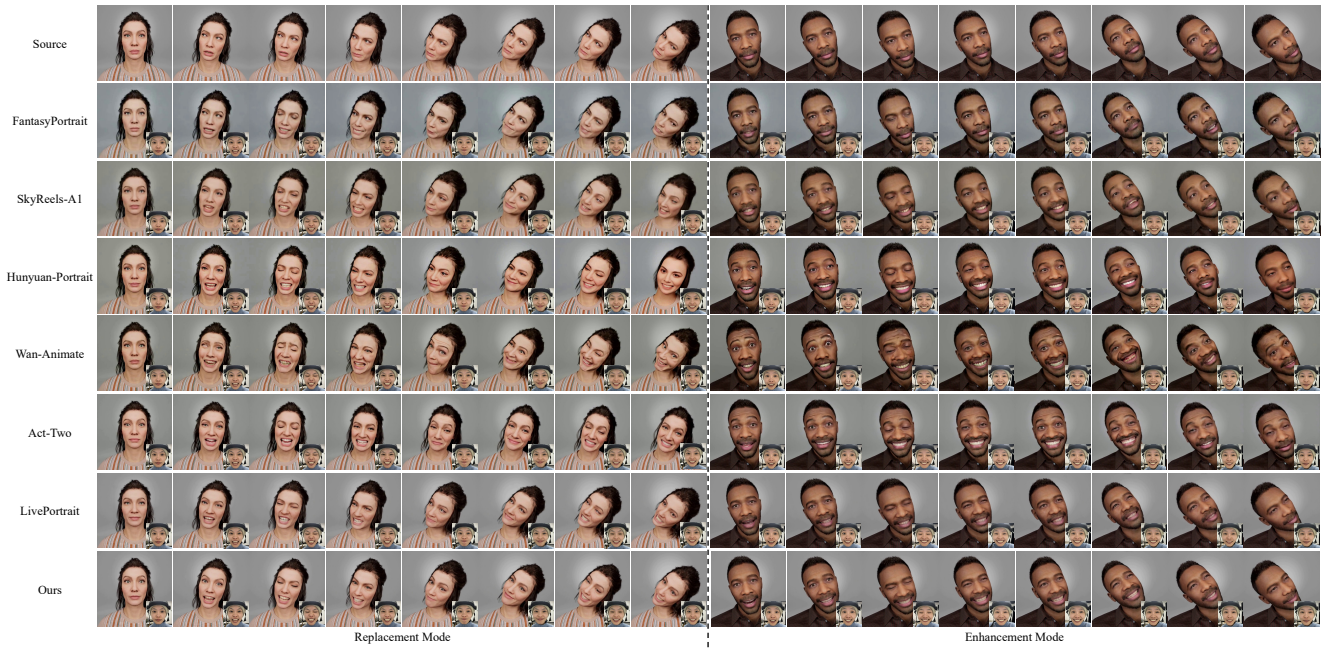

Figure 5. Full qualitative comparison with all the methods mentioned in the main manuscript on our proposed test benchmark. The bottom-right insets exhibit driving frames. Please zoom in for better inspection.

Table 1. Quantitative comparisons of self-reenactment portrait animation on MEAD [28] dataset. The top of the table shows the results of non-diffusion-based methods, while the bottom presents diffusion-based methods.

Method	PSNR↑	SSIM↑	LPIPS↓	$\mathcal{L}_1$↓	CSIM↑	MAE(°)↓	AED↓	APD↓	FID↓	FVD↓
GAGAvatar [4]	-	-	-	-	0.8946	5.1074	0.3781	0.0101	-	-
Portrait4D-v2 [6]	20.0907	0.7746	0.3358	0.0617	0.8793	5.4329	0.4353	0.0149	85.3589	460.7595
PD-FGC [27]	20.8419	0.7824	0.341	0.0573	0.3256	8.3772	0.7156	0.0202	92.8886	1276.3855
EMOPortrait [8]	26.1748	0.8729	0.1544	0.0287	0.5959	6.6994	0.4992	0.0128	37.5216	443.7428
EDTalk [26]	26.9246	0.8964	0.1443	0.0319	0.8592	6.218	0.4333	0.0077	43.4199	343.9973
LIA-X [31]	22.6439	0.8232	0.1816	0.0386	0.8957	5.3919	0.4633	0.0636	32.4902	323.2831
LivePortrait [11]	<u>32.9063</u>	<u>0.9464</u>	<u>0.0527</u>	<u>0.0148</u>	<u>0.9379</u>	<u>3.5497</u>	<u>0.2471</u>	<u>0.0041</u>	<u>10.4759</u>	<u>84.3131</u>
FYE [20]	27.1819	0.8963	0.1039	0.0243	0.8767	5.6658	0.527	0.0109	30.5002	350.4705
AniPortrait [32]	29.0281	0.9125	0.081	0.0198	0.8904	4.8224	0.3989	0.0077	19.9857	191.8255
X-NeMo [37]	22.4136	0.7313	0.1916	0.0551	0.8594	10.4812	0.4168	0.0097	50.6639	409.2123
ReliPA [12]	24.0052	0.8525	0.1601	0.0409	0.8212	6.3512	0.5174	0.0117	35.1439	455.0235
SkyReels-A1 [22]	25.9931	0.8825	0.1182	0.032	0.8668	5.7577	0.594	0.0105	22.6852	278.6554
Hunyuan-Portrait [35]	26.4138	0.8779	0.0941	0.0309	0.922	4.7961	0.3348	0.0101	18.896	139.2772
FantasyPortrait [29]	22.6155	0.7789	0.1498	0.0634	0.8622	6.606	0.5147	0.0116	35.7586	258.8542
Wan-Animate [3]	21.9017	0.8105	0.2159	0.054	0.827	5.7592	0.5307	0.0144	24.9683	465.8136
VACE [13]	15.0009	0.5046	0.4083	0.1587	0.5472	10.0836	0.745	0.021	118.5134	870.7917
AvatarArtist [19]	18.7405	0.7173	0.3891	0.0774	0.7402	6.5339	0.5571	0.0194	83.5354	815.3565
Ours	**33.7235**	**0.9501**	**0.0491**	**0.0125**	**0.9521**	**3.1576**	**0.1971**	**0.0038**	**10.132**	**71.58**

the edited results of our PerformRecast are concatenated together. It can be clearly observed that our method can effectively edit facial expressions of characters with various styles, generating temporally continuous and smooth videos, and markedly outperforms existing approaches such as LivePortrait [11]

In addition, to verify the ability of our method to be applied to humanoid characters, we randomly choose a 3D animation and use our model to edit the facial expressions of characters in it. We select 20 video clips in total. The source video clips and the edited results are concatenated together and placed in `3D_Animation.mp4`. After edit-

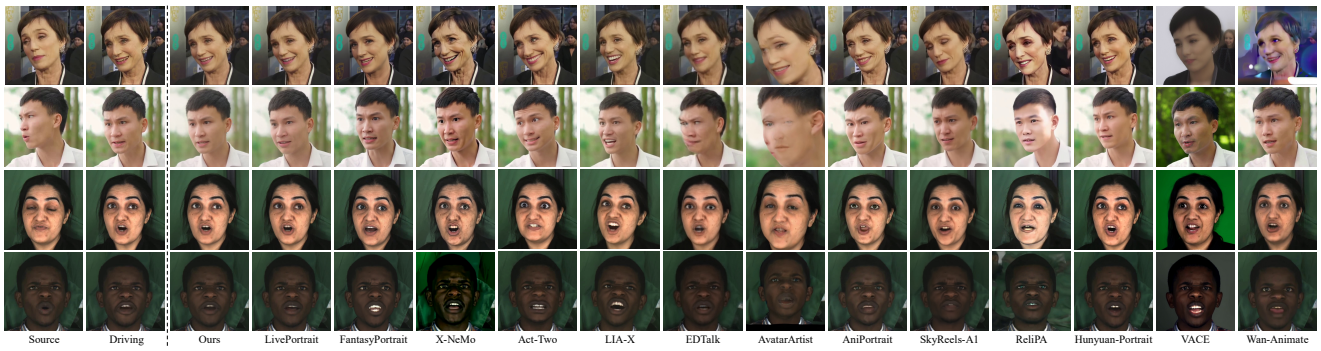

Figure 6. More generated results on self-reenactment task of different methods. The first two source-driving paired images are from VFHQ dataset [33] and the last two source-driving paired images are from MEAD dataset [28].

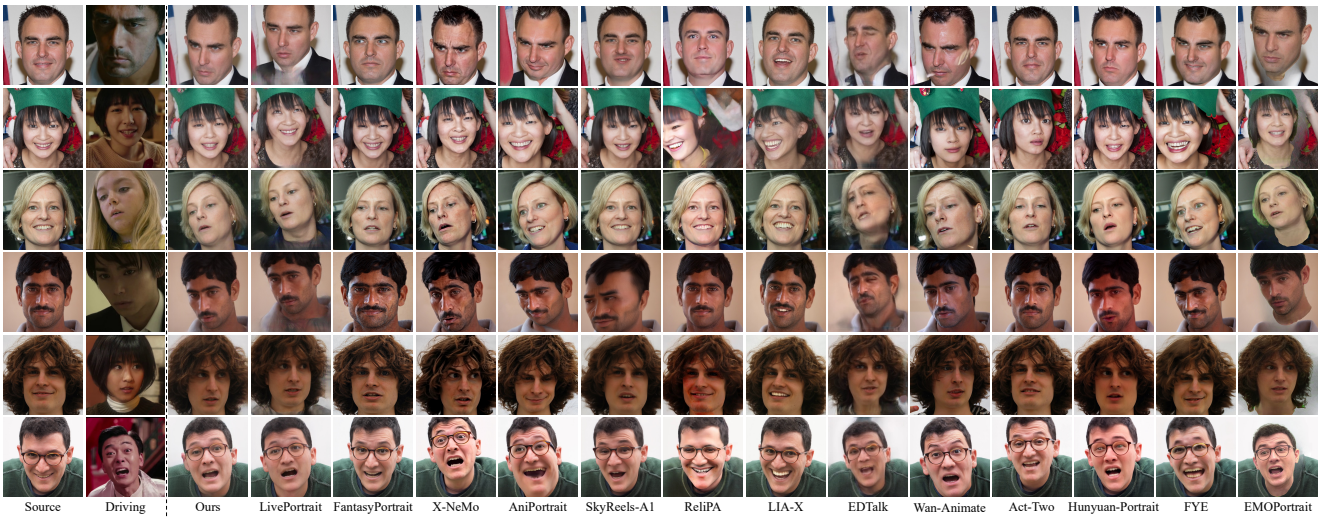

Figure 7. More generated results on cross-reenactment task of different methods. The source images are from FFHQ dataset and the driving frames are from famous films and television clips.

ing with our model, the expressions of characters in the 3D animation become more vivid, demonstrating the applicability of our model in practical scenarios.

11.2. Self-reenactment

We also report the quantitative results of self-reenactment portrait animation on MEAD dataset [28] in Tab. 1. All the methods are evaluated on a random split of MEAD dataset, which consists of 70 videos. As shown in Tab. 1, our method achieves the best performance across all metrics on MEAD dataset [28], highlighting its superiority over other existing approaches.

What's more, we also present more qualitative results of different compared methods in Fig. 6. LivePortrait [11] tends to generate blurred results around the eyes in the first and third cases. It also struggles to preserve the subtle expressions in the second and fourth cases. Other diffusion-based methods are prone to generating unstable results and exaggerated facial expressions. On the contrary, our PerformRecast faithfully recovers the driving frames with fine-grained details.

11.3. Cross-reenactment

We provide more cross-reenactment portrait animation results generated by our PerformRecast and some other methods in Fig. 7. The source images are from FFHQ dataset [14] and we use some famous films and television clips as driving frames. From which we can conclude that our method is capable of preserving the head pose, facial expressions and eyeball directions in the driving frames with high fidelity, while generating clear and high-quality images. Although our method does not achieve best performance on all evaluation metrics as reported in the main manuscript, it markedly outperforms all other methods in visual effects. This is most likely because our used quantitative evaluation metrics mainly rely on some pre-trained

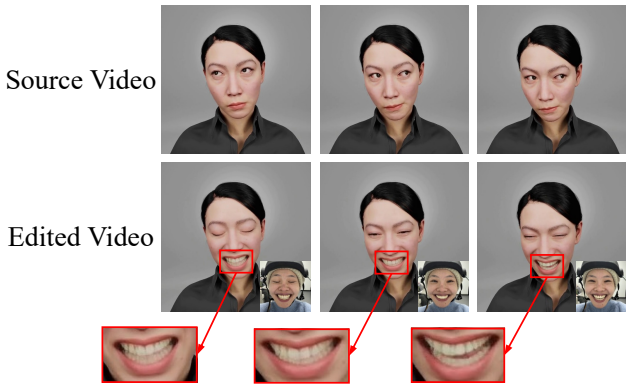

Figure 8. A typical failure case when our method generating teeth while the mouth is closed in source video.

networks, whose inherent priors may limit their ability to faithfully reflect the actual performance of each method in some scenarios. Developing more evaluation metrics which are capable of accurately assessing the accuracy of head pose, facial expressions and gaze direction is an interesting research direction in the future.

12. Limitations and Discussions

In portrait video expression editing task, our method tends to produce blurry results in regions that are not visible in the source video, especially when generating teeth while the mouth is closed in source video. Fig. 8 presents a typical failure case of this scenario. This is mainly because our model is GAN-based, and unlike diffusion-based models, it has limited ability to imagine and synthesize unseen objects. In the future, we are planning to combine the disentangling capability of 3D Morphable Face Model with the generative power of large-scale pre-trained image diffusion models or video diffusion models, aiming to further improve the fidelity and clarity of synthesized videos.

13. Ethics Statement

This work advances portrait animation and portrait video facial expression editing for virtual avatars. Our methods are not intended for malicious use, and all synthesized content should clearly indicate its artificial nature. We acknowledge potential misuse, such as deepfakes, and are developing tools to help detect synthetic videos. At the same time, our technology can support education, communication assistance, and therapeutic applications, reflecting our commitment to responsible and ethical AI development.

References

[1] Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, Ayoub Al-Hamadi, and Laslo Dinges. L2cs-net: Fine-grained gaze estimation in unconstrained environments. In *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, pages 98–102. IEEE, 2023. 3

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3

[3] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025. 5

[4] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 5

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2

[6] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. In *European Conference on Computer Vision*, pages 316–333. Springer, 2024. 5

[7] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. 2022. 4

[8] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars, 2024. 5

[9] Epic Games. Metahuman creator. https://www.unrealengine.com/en-US/digital-humans, 2021. 3

[10] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction, 2025. 1

[11] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 2, 4, 5, 6

[12] Mingtao Guo, Guanyu Xing, and Yanli Liu. High-fidelity relightable monocular portrait animation with lighting-controllable video diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 228–238, 2025. 5

[13] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17191–17202, 2025. 5

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 6

[15] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[16] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[18] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 1, 4

[19] Hongyu Liu, Xuan Wang, Ziyu Wan, Yue Ma, Jingye Chen, Yanbo Fan, Yujun Shen, Yibing Song, and Qifeng Chen. Avatarartist: Open-domain 4d avatarization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10758–10769, 2025. 5

[20] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. *arXiv preprint arXiv:2406.01900*, 2024. 5

[21] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[22] Di Qiu, Zhengcong Fei, Rui Wang, Jialin Bai, Changqian Yu, Mingyuan Fan, Guibin Chen, and Xiang Wen. Skyreels-a1: Expressive portrait animation in video diffusion transformers. *arXiv preprint arXiv:2502.10841*, 2025. 4, 5

[23] George Retsinas, Panagiotis P Filntisis, Radek Danecek, Victoria F Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2490–2501, 2024. 3, 4

[24] Runway. Creating with act-two. https://help.runwayml.com/hc/en-us/articles/42311337895827-Creating-with-Act-Two, 2025. 4

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

[26] Shuai Tan, Bin Ji, Mengxiao Bi, and Ye Pan. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *European Conference on Computer Vision*, pages 398–416. Springer, 2024. 5

[27] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4, 5

[28] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. 5, 6

[29] Qiang Wang, Mengchao Wang, Fan Jiang, Yaqi Fan, Yonggang Qi, and Mu Xu. Fantasyportrait: Enhancing multi-character portrait animation with expression-augmented diffusion transformers. *arXiv preprint arXiv:2507.12956*, 2025. 5

[30] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations*, 2022. 4

[31] Yaohui Wang, Di Yang, Xinyuan Chen, Francois Bremond, Yu Qiao, and Antitza Dantcheva. Lia-x: Interpretable latent portrait animator. *arXiv preprint arXiv:2508.09959*, 2025. 5

[32] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 5

[33] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 6

[34] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 4

[35] Zunnan Xu, Zhentao Yu, Zixiang Zhou, Jun Zhou, Xiaoyu Jin, Fa-Ting Hong, Xiaozhong Ji, Junwei Zhu, Chengfei Cai, Shiyu Tang, et al. Hunyuanportrait: Implicit condition control for enhanced portrait animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15909–15919, 2025. 4, 5

[36] 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

[37] Xiaochen Zhao, Hongyi Xu, Guoxian Song, You Xie, Chenxu Zhang, Xiu Li, Linjie Luo, Jinli Suo, and Yebin Liu. X-nemo: Expressive neural motion reenactment via disentangled latent attention. *arXiv preprint arXiv:2507.23143*, 2025. 5

[38] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022. 1