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

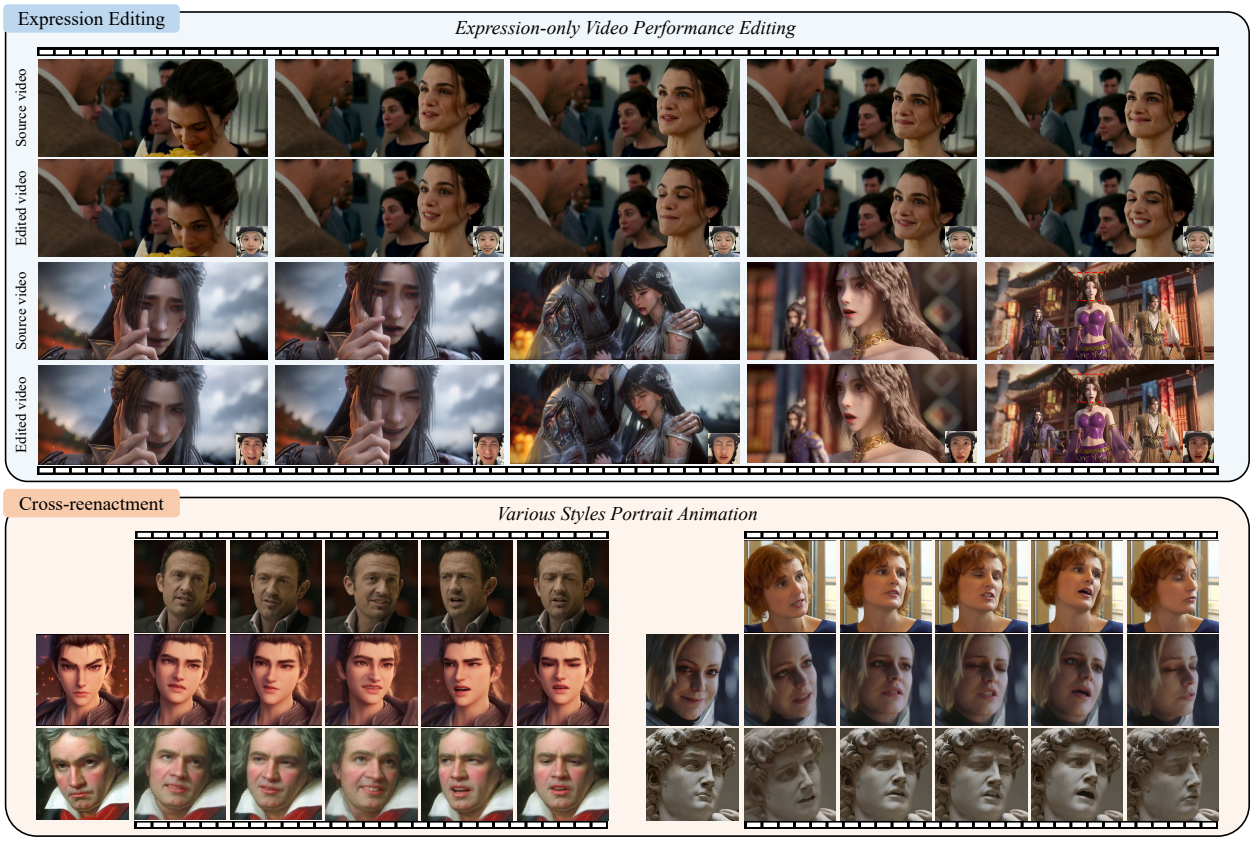

Figure 1. Our proposed PerformRecast is capable of editing the facial expression of a source portrait video as well as animating a static portrait image according to a driving video. The top part of this figure shows the expression editing results of a movie clip and a 3D animation. The generated results exhibit high fidelity to the driving video, facilitating the production processes in film and animation industries. For the shot with multiple characters, we can select the specific person whose facial expression we want to edit, which is indicated by a red dashed box. The bottom-right insets on the top part show the driving frames. Please zoom in for better inspection.

Abstract

This paper primarily investigates the task of expression-only portrait video performance editing based on a driving video, which plays a crucial role in animation and film industries. Most existing research mainly focuses on portrait animation, which aims to animate a static portrait image according to the facial motion from the driving video. As a consequence, it remains challenging for them to disentangle the facial expression from head pose rotation and thus lack the ability to edit facial expression independently. In this paper, we propose PerformRecast, a versatile expression-only video editing method which is dedicated to recast the performance in existing film and animation. The key insight of our method comes from the characteristics of 3D Morphable Face Model (3DMM), which models the face identity, facial expression and head pose of 3D face mesh with separate parameters. Therefore, we improve the key-

points transformation formula in previous methods to make it more consistent with 3DMM model, which achieves a better disentanglement and provides users with much more fine-grained control. Furthermore, to avoid the misalignment around the boundary of face in generated results, we decouple the facial and non-facial regions of input portrait images and pre-train a teacher model to provide separate supervision for them. Extensive experiments show that our method produces high-quality results which are more faithful to the driving video, outperforming existing methods in both controllability and efficiency. Our code, data and trained models are available at https://youku-aigc.github.io/PerformRecast.

1. Introduction

Character animation with vivid facial expression is of vital importance in the film and computer animation industries [39]. However, it is quite challenging for artists to create lifelike expression on 3D face models, and for film actors to consistently perform satisfying facial expressions, whether in a single take or with minimal cost. Therefore, developing an automatic algorithm on expression-only portrait video performance editing would be highly meaningful and important in the field of computer animation.

With the rapid development of deep generative models, such as GANs [18, 28] and Diffusion Models [23, 45, 50], most existing studies mainly focus on animating a static portrait image following the facial motion of a driving video, also termed portrait animation, which is a little different from our target. Given a video (e.g., a movie clip), our goal is to edit or enhance only the facial expression based on a real actor's time-varying facial expression, while strictly preserving all other factors, including the original face ID, head pose, camera motion and background. Any other change outside facial expression is considered as a failure. Diffusion-based portrait animation methods are typically built upon pre-trained image diffusion models [2, 45] or video diffusion models [4, 25, 57, 72] via attaching additional modules. But they struggle to fully disentangle the facial expression from head pose rotation, making it difficult to accurately perform our expression editing task.

On the other hand, recent GAN-based portrait animation methods typically construct a 3D feature volume given source and driving portrait images [13, 14, 19, 61], which is further sent into the generator to produce the final image. These methods employ motion encoder to extract corresponding features [13, 14] or implicit keypoints [19, 61] to guide the computation of feature volume. However, since the extracted features are obtained directly from raw images, they still have difficulties in disentangling face identity, facial expression and head pose. Moreover, the implicit keypoints lack explicit physical meaning and direct supervision, thereby compromising the precision of facial motion control and leading to suboptimal results.

In this paper, we propose PerformRecast, a well-designed and effective GAN-based method tailored to our expression-only video performance editing task. Traditional 3DMM [3] represents face identity, facial expression, and head pose with separate parameters, naturally disentangling these factors. Therefore, we improve the keypoints transformation used in previous method to make it more consistent with the forward process of FLAME [34], a representative 3D Morphable Face Model. Specifically, we employ a 3D face tracking method [17] to extract temporally continuous FLAME [34] parameters from input portrait videos, and select explicit 3D keypoints on face mesh vertices to supervise the motion extractor in our model. To further avoid the misalignment around the boundary of the face in the generative results, we introduce a boundary alignment module which segments each frame into facial and non-facial regions. An additional teacher model is pre-trained to provide the training loss for the facial region, enabling separate supervision for both regions. As a result, our method can not only edit and enhance facial expressions in existing videos due to its disentanglement from head pose, but also outperform many previous methods on the traditional portrait animation task. In addition, to better evaluate expression editing performance, we construct a benchmark using digital humans from MetaHuman [16]. For each digital human, we render multiple portrait videos with the same head pose rotation but different facial expressions. In summary, the contributions of our paper are fourfold:

- We modify the keypoints transformation formula and utilize explicit 3D keypoints on FLAME face mesh to directly supervise the motion extractor.
- We adopt a boundary alignment module to alleviate misalignment between the facial and non-facial regions.
- We present a benchmark, tailored for the assessment of portrait video expression editing, which will be publicly available to advance the evaluation of future research.
- We propose PerformRecast, a versatile and effective method which is expert at both expression-only video performance editing and portrait animation tasks. Qualitative and quantitative experiments have been conducted to verify the superiority of our method over other existing approaches in controllability and quality.

2. Related Work

In this section, we mainly summarize different types of portrait animation methods, for they are quite similar to our expression editing task from a methodological perspective.

2.1. Non-Diffusion-based Portrait Animation

Early 3D face model-based methods [36, 54] reconstruct high-quality geometry and appearance for rendering. With the development of 3D Morphable Face Models [3, 34]

and neural rendering [29, 38, 55], methods such as Portrait4D [10, 11] and GAGAvatar [8] adopt triplane or 3D Gaussian representations for animatable head synthesis. However, pure 3D-based representations often struggle to capture fine details, leading to blurry results.

A great deal of other methods are built upon the Generative Adversarial Networks [18, 27, 28], which provide stronger image synthesis capabilities. Early approaches directly decode latent appearance and motion features [5, 73], while later works focus on disentangling identity and motion via specialized designs [52, 53, 58, 75]. Nevertheless, these methods rely heavily on complex loss functions and still face challenges in achieving complete disentanglement.

More recent portrait animation models are predominantly warping-based [47–49] methods, which estimate motion fields using learned landmarks/keypoints and warp source features [12, 24, 43, 66, 71, 74]. Representative works such as LIA-X [62, 63], EMOPortrait [13, 14], Face Vid2vid [61] and LivePortrait [19] improve motion modeling and controllability. However, their implicit motion representations lack explicit physical meaning and supervision, which restricts the control flexibility and accuracy.

2.2. Diffusion-based Portrait Animation

With the rapid development of diffusion models [23, 50], recent works leverage large-scale pre-trained image and video diffusion models [2, 4, 25, 45, 57, 72] for portrait animation. These methods typically incorporate several plug-and-play modules to capture identity, background content as well as facial motion and maintain the cross-frame coherence.

Some representative approaches include Follow-Your-Emoji [37], X-NeMo [78], Wan-Animate [7], VACE [26], Hunyuan-Portrait [70], AniPortrait [65], SkyReels-A1 [42], and AvatarArtist [35]. They adopt diverse motion representations such as landmarks, latent motion codes, 3D priors, or implicit features. However, despite the various motion representations they used, they can hardly disentangle the face identity, facial expression and head pose. What's more, they struggle to guarantee temporal consistency and require much more inference time.

3. Method

In this section, we provide a detailed explanation of our method, PerformRecast. Our method is built upon LivePortrait [19], a typical warping-based portrait animation model. The overall training pipeline of our method is shown in Fig. 2. We change the original keypoints transformation formula used in LivePortrait to make it more consistent with 3D face parametric model which is naturally capable of disentangling the face identity, facial expression and head pose. Then, we present our boundary alignment module which alleviates the misalignment between the facial and non-facial regions in generated images. Finally, we elaborate on the inference process of our method.

3.1. Preliminary

We begin with a brief review of LivePortrait [19] and 3DMM-based face tracking [17], upon which our method builds. LivePortrait [19] utilizes the self-reenactment training pipeline, which takes source and driving frames within the same subject and video. We denote the source and driving frames as $I_s \in \mathbb{R}^{3\times H\times W}$ and $I_d \in \mathbb{R}^{3\times H\times W}$. The model is learned to reconstruct the driving frame I_d, and the synthesized frame is denoted as $\hat{I}_d$. The original framework consists of an appearance feature extractor $\mathcal{F}$, a motion extractor $\mathcal{M}$, a warping field estimator $\mathcal{W}$ and a SPADE decoder [41] based generator $\mathcal{G}$. $\mathcal{F}$ maps the source portrait image I_s to a 3D appearance feature volume f_s. The motion extractor $\mathcal{M}$, with ConvNext-V2-Tiny [67] backbone, directly predicts the canonical keypoints $x_c \in \mathbb{R}^{K\times 3}$, head pose $R \in \mathbb{R}^{3\times 3}$, expression deformation $\delta \in \mathbb{R}^{K\times 3}$, scale factor $s \in \mathbb{R}^3$ and translation $t \in \mathbb{R}^3$ from both source and driving frames. K represents the number of implicit keypoints. Then, the source 3D keypoints x_s and the driving 3D keypoints x_d are transformed as follows:

$$\begin{cases} x_s = s_s \cdot (x_{c,s} R_s + \delta_s) + t_s, \\ x_d = s_d \cdot (x_{c,s} R_d + \delta_d) + t_d, \end{cases} \tag{1}$$

where the subscripts s and d denote the source and driving, respectively. Next, $\mathcal{W}$ generates the warping field using the implicit keypoints x_s and x_d, and employs this flow field to warp the source feature volume f_s. Finally, the warped features pass through the generator $\mathcal{G}$, which translates them into image space and generates a target image.

What's more, We adopt a recently-proposed 3D face tracking method, Pixel3DMM [17] to predict FLAME parameters of each frame from input portrait videos. The FLAME parameters include 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. We describe the details of Pixel3DMM [17] in the supplementary material.

3.2. FLAME-based Keypoints Transformation

Similar to LivePortrait [19], our method also utilize the self-reenactment training pipeline, which reenacts both the facial expression and head pose to reconstruct the driving frame I_d from source frame I_s. We argue that the implicit keypoints transformation used in LivePortrait can not fully disentangle the face identity, facial expression and head pose. The canonical keypoints x_c are first multiplied by the head pose R and then added with the expression deformation δ. As a consequence, the learned expression deformation would contain some residual head pose information. On the contrary, the traditional 3D Morphable Face Model (3DMM) [3], such as FLAME [34], uses separate parameters to represent identity, expression, and head pose, achieving a natural disentanglement. The transformation process

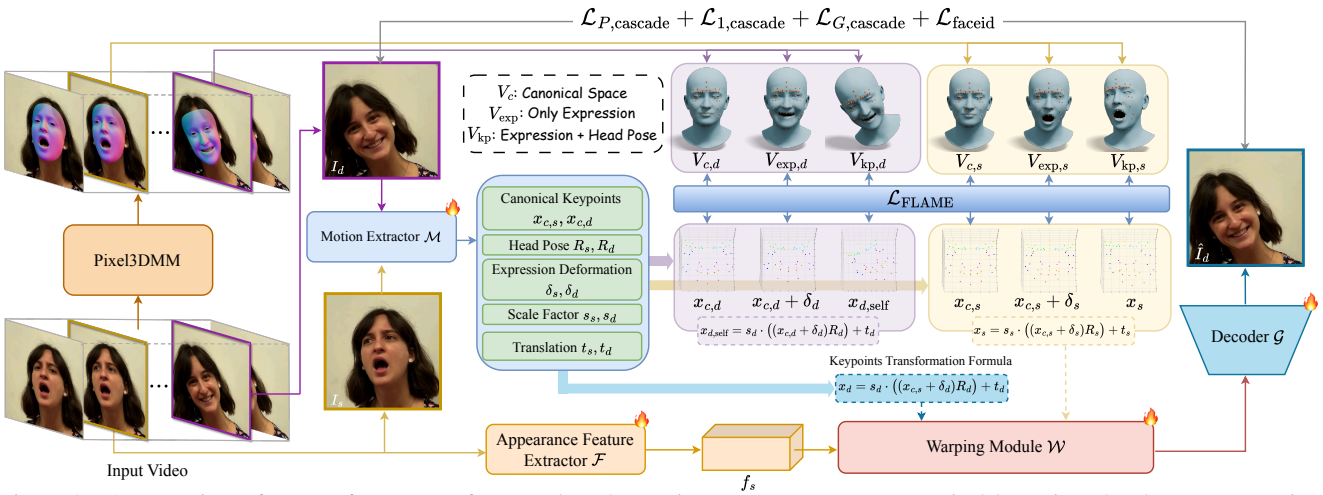

Figure 2. An overview of our PerformRecast framework. The motion extractor extracts canonical keypoints, head pose, expression deformation, scale factor and translation from source and driving frames. The facial keypoints are obtained via our improved keypoints transformation formula and compared with the tracking results from Pixel3DMM [17] to calculate the FLAME loss. Finally, the appearance feature volume, source and driving keypoints are sent to the warping module, followed by the decoder to reconstruct the driving frame.

of FLAME [34] is defined as:

$$\begin{aligned} M(\beta,\theta,\psi) &= W(T_P(\beta,\theta,\psi),\mathbf{J}(\beta),\theta,\mathcal{W}),\\ T_P(\beta,\theta,\psi) &= \mathbf{T} + B_S(\beta;\mathcal{S}) + B_P(\theta;\mathcal{P}) + B_E(\psi;\mathcal{E}), \end{aligned} \tag{2}$$

This function takes different coefficients to describe shape β, head pose θ, expression ψ and returns N vertices. Eq. (2) illustrates that the template mesh $\mathbf{T}$, in the zero pose and expression, is first added with identity related shape variation $B_S(\beta;\mathcal{S})$ and expression blendshapes $B_E(\psi;\mathcal{E})$, then multiplied by head pose rotation θ (the pose blendshapes term $B_P(\theta;\mathcal{P})$ is not necessary in our task).

Therefore, we modify the keypoints transformation formula used in LivePortrait to the scale orthographic projection [19], which is formulated as:

$$\begin{cases} x_s = s_s \cdot \big((x_{c,s} + \delta_s)R_s\big) + t_s,\\ x_d = s_d \cdot \big((x_{c,s} + \delta_d)R_d\big) + t_d, \end{cases} \tag{3}$$

The scale orthographic projection is more similar with FLAME model, whose canonical keypoints are first added with the expression deformation δ and then multiplied by the head pose R, effectively avoiding information leakage between head pose and facial expression.

What's more, instead of treating x_s and x_d as implicit keypoints, we select several explicit keypoints from vertices of FLAME face mesh tracked by Pixel3DMM [17] to directly supervise the keypoints transformation process in Eq. (3). Specifically, we derive three sets of explicit keypoints from the reconstructed FLAME model of source and driving frame. The first set of explicit keypoints $V_{c,i}$ is obtained from the canonical FLAME face mesh $T_{c,i}$, which encodes only shape blendshapes by setting both head pose θ expression ψ to zero. The second set $V_{\text{exp},i}$, which is used to supervise the expression deformation, is extracted from the FLAME face mesh $T_{\text{exp},i}$. $T_{\text{exp},i}$ adds expression blendshapes and three joint rotations $\theta_{\text{jaw}}, \theta_{\text{left-eyeball}}, \theta_{\text{right-eyeball}}$ to T_c and keeps neck pose θ_{neck} as well as global head pose θ_{head} to zero, for facial expression often contains the orientation of eyeballs and the movement of jaw, while excluding the rotation of neck in most scenarios. The third set, $V_{\text{kp},i}$ is derived from $T_{\text{kp},i}$, where all FLAME parameters are enabled. The subscripts $i \in \{s,d\}$ denotes the source and driving frames, respectively. We then introduce the FLAME loss which utilizes these three sets of explicit keypoints to directly supervise the motion extractor. The FLAME loss is formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{FLAME}} = &\ \text{Wing}(x_{c,s}, V_{c,s}) + \text{Wing}(x_{c,d}, V_{c,d})\\ &+ \text{Wing}(x_{c,s} + \delta, V_{\text{exp},s}) + \text{Wing}(x_{c,d} + \delta, V_{\text{exp},d})\\ &+ \text{Wing}(x_s, V_{\text{kp},s}) + \text{Wing}(x_{d,\text{self}}, V_{\text{kp},d}), \end{aligned} \tag{4}$$

where $x_{d,\text{self}} = s_d \cdot \big((x_{c,d} + \delta_d)R_d\big) + t_d$, which is additionally computed to accelerate training and Wing loss is adopted following [15]. The FLAME loss provides a much stronger supervision to motion extractor, enabling more accurate learning of keypoint transformations. It also prevents the model from learning overly flexible expressions δ as mentioned in [19]. Therefore, we discard the implicit keypoints equivariance loss, keypoint prior loss, deformation prior loss and head pose loss used in previous methods [19, 61] for head pose R can be effectively learned via self supervised learning. The overall training loss of our model is similar to that of previous portrait animation methods, which is formulated as:

$$\mathcal{L}_{\text{animate}} = \mathcal{L}_{\text{FLAME}} + \mathcal{L}_{P,\text{cascade}} + \mathcal{L}_{1,\text{cascade}} + \mathcal{L}_{G,\text{cascade}} + \mathcal{L}_{\text{faceid}}, \tag{5}$$

where the cascaded perceptual loss $\mathcal{L}_{P,\text{cascade}}$, the cascaded L_1 loss $\mathcal{L}_{1,\text{cascade}}$, the cascaded GAN loss $\mathcal{L}_{G,\text{cascade}}$ and face-id [9] loss $\mathcal{L}_{\text{faceid}}$ are calculated between the generated frame $\hat{I}_d$ and target frame I_d. The detailed definitions of these loss terms are available in the supplementary mate-

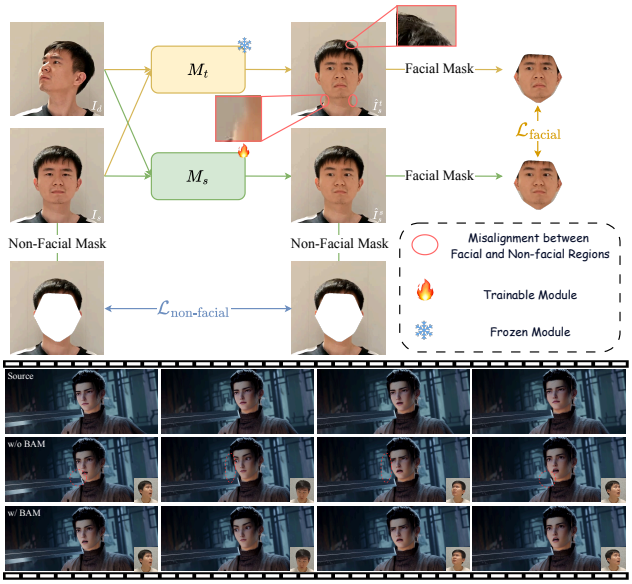

Figure 3. Top: we adopt the boundary alignment module to alleviate the misalignment between the facial and non-facial regions. Bottom: we provide some visual comparisons to demonstrate the necessity of using boundary alignment module (denoted as "BAM"). The red dashed circles highlight the misalignment between the facial and non-facial regions. The bottom-right insets show the driving frames. Please zoom in for better inspection.

rial. Moreover, due to the better disentanglement of facial expression and head pose, we do not need to train the second stage in LivePortrait [19], including stitching and retargeting modules.

3.3. Boundary Alignment Module

Another important objective of portrait video expression editing is to maintain the smooth alignment around the boundary of face. However, to minimize the animation loss in Eq. (5) on global region, the learned warping field of explicit 3D keypoints will also influence the non-facial region, resulting in misalignment on the boundary when performing facial expression editing task. To solve this problem, we design a boundary alignment module to mitigate the warping impact of explicit 3D keypoints on non-facial region, which is shown in the top of Fig. 3. Specifically, we firstly train a teacher model M_t using the animation loss in Eq. (5) between the generated image $\hat{I}_d^t$ and driving image I_d. Despite the misalignment between the facial and non-facial regions in generative results of M_t, it is capable of synthesizing precise and clear facial expression on the facial region. Therefore, given the source frame I_s and target frame I_d, we first employ the trained teacher model M_t to edit only the facial expression of I_s, obtaining an intermediate result $\hat{I}_s^t$. The keypoints of this process is calculated as:

$$\begin{cases} x_s = s_s \cdot \big((x_{c,s} + \delta_s)R_s\big) + t_s, \\ x_d = s_s \cdot \big((x_{c,s} + \delta_d)R_s\big) + t_s, \end{cases} \tag{6}$$

which only replaces the expression deformation δ with that of I_d while keeping all other parameters unchanged.

The student model M_s is then trained to synthesize both the facial expression replacement results $\hat{I}_s^s$ and driving frame reconstruction result $\hat{I}_d^s$ from I_s. Additional loss terms are applied separately on each region of $\hat{I}_s^s$: for facial region, we compute a perceptual loss $\mathcal{L}_{P,\text{facial}}$ and an L_1 loss $\mathcal{L}_{1,\text{facial}}$ between $\hat{I}_s^s$ and $\hat{I}_s^t$; for non-facial region, we also calculate a perceptual loss $\mathcal{L}_{P,\text{non-facial}}$ and an L_1 loss $\mathcal{L}_{1,\text{non-facial}}$ between $\hat{I}_s^s$ and the ground truth I_s. These two losses can be formulated as follows:

$$\mathcal{L}_{\text{facial}} = \mathcal{L}_{P,\text{facial}} + \mathcal{L}_{1,\text{facial}}, \tag{7}$$

$$\mathcal{L}_{\text{non-facial}} = \mathcal{L}_{P,\text{non-facial}} + \mathcal{L}_{1,\text{non-facial}}, \tag{8}$$

Meanwhile, the animation loss in Eq. (5) is also applied between the generated frame $\hat{I}_d^s$ and driving frame I_d. We provide the detailed calculation process for the mask of facial and non-facial regions in the supplementary material. The bottom part of Fig. 3 shows some visual comparisons of our model trained with or without boundary alignment module, highlighting its importance in training pipeline.

3.4. Inference

Given a source video sequence $\{I_{s,i}|i = 0, 1, ..., N-1\}$ and a driving video sequence $\{I_{d,i}|i = 0, 1, ..., N-1\}$, we propose two modes: replacement mode and enhancement mode for portrait video expression editing in the inference stage. Replacement mode directly replace the facial expression of the i-th frame $I_{s,i}$ in source video with that of the i-th frame $I_{d,i}$ in driving video. The keypoints transformation of this mode is the same as Eq. (6).

Enhancement mode aims to enhance the facial expression in source video, whose keypoint transformation process of the i-th frame is modified to:

$$\begin{cases} x_{s,i} = s_{s,i} \cdot \big((x_{c,s} + \delta_{s,i})R_{s,i}\big) + t_{s,i}, \\ x_{d,i} = s_{s,i} \cdot \big((x_{c,s} + \delta_{s,i} + \delta_{d,i} - \delta_{d,0})R_{s,i}\big) + t_{s,i}. \end{cases} \tag{9}$$

which adds the facial expression of driving video on the top of that of source video. The appearance feature volume $f_{s,i} = \mathcal{F}(I_{s,i})$ is extracted from the i-th frame of source video in both modes. For portrait animation task, the keypoints transformation in inference stage is the same as Eq. (3) in training stage. We provide more details in the supplementary material.

4. Experiments

Implementation Details. We adopt a similar architecture to LivePortrait [19], except that our appearance feature extractor $\mathcal{F}$ is built upon the pretrained DINOv2 [40] backbone. To calculate the FLAME loss $\mathcal{L}_{\text{FLAME}}$, the number of explicit 3D keypoints K is set to 49, which cover most of the key regions of the face. The resolution of input frames and output image of our model is set to 512×512. The

Figure 4. The typical usage scenarios for our proposed replacement mode and enhancement mode. The bottom-right insets exhibit driving frames. Please zoom in for better inspection.

details of keypoints selection and training settings are presented in the supplementary material.

Dataset. We utilized a combination of several publicly-available datasets, including VFHQ[68], MEAD[59], Nersemble[32], FEED[14], and ETH-XGaze[77]. We also incorporate a large amount of portrait videos from the Internet, including high-definition anime and film, to further enhance the diversity and quality of our in-house dataset. Finally, this results in a total of 597,331 video clips, which include a variety of expressions and emotional intensities.

Evaluation Metrics. To measure the generation quality and controllability of our PerformRecast on both video expression editing and portrait animation task, we adopt PSNR, SSIM [64], LPIPS [76], $\mathcal{L}_1$ distance, Fréchet Inception Distance (FID) [22], Fréchet Video Distance (FVD) [56], Cosine SIMilarity of identity features (CSIM) [30], Average Expression Distance (AED) [48], Average Pose Distance (APD) [48] and Mean Angular Error (MAE) of eyeball direction [21]. Details of these metrics are provided in the supplementary material. All these metrics are calculated under the resolution of 512×512 for each compared method.

4.1. Portrait Video Expression Editing

We first would like to clarify the two inference modes and discuss their respective usage scenarios. When the facial expression in a shot does not fully satisfy the director's creative expectations, two inference modes arise. Replacement mode is appropriate when the actor's overall performance is compelling and remains consistent with the narrative context. Enhancement mode is more suitable when only localized improvements are needed without compromising the performance plausibility. Fig. 4 shows the typical usage scenarios for our proposed two inference modes.

We then conduct experiments to compare the ability of our proposed PerformRecast with other methods on portrait video expression editing task. Besides LivePortrait [19], we modify several diffusion-based portrait animation methods, including SkyReels-A1 [42], Hunyuan-Portrait [70], FantasyPortrait [60] and Wan-Animate [7] to make them capable of editing only the expressions of source video according to

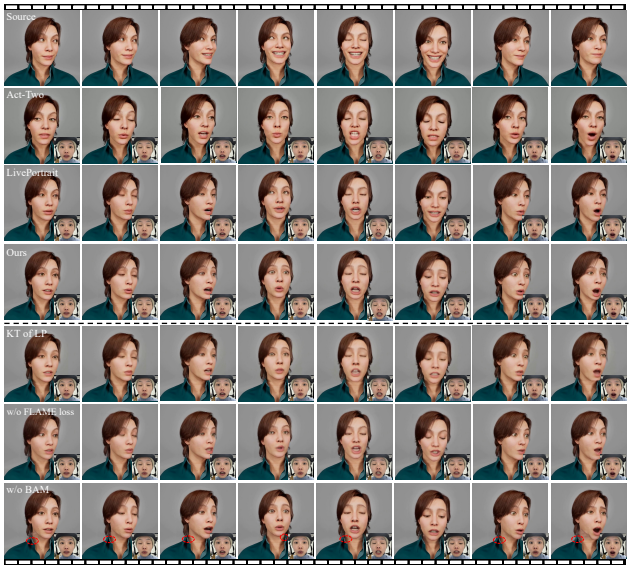

Figure 5. Qualitative comparison of portrait video expression editing on replacement mode. The top of the figure shows editing results of different methods. The bottom presents our ablation studies and analysis. The bottom-right insets exhibit driving frames. The red dashed circles highlight the misalignment between the facial and non-facial regions. Please zoom in for better inspection.

driving video. Specifically, we combine the expression information of driving frame with the head pose information of source frame to generate the editing results. The detailed modification of each method is described in the supplementary material. We also compare with the closed-source commercial product Runway Act-Two [46].

To more accurately evaluate the performance of different methods, we construct a test benchmark utilizing MetaHuman [16]. Our test benchmark contains 18 portrait videos with diverse expression and without head pose rotation recorded from professional facial motion actors. Then, we select 20 digital humans from MetaHuman, and for each digital human, we render 19 videos, of which 18 are using the expressions of facial motion actors, and the last one is without expression. In addition, all the videos are added with our pre-defined head pose rotation. Further details of our test benchmark can be found in the supplementary ma-

Table 1. Quantitative results of facial expression editing on our test benchmark on both replacement and enhancement modes. The top of the table shows the comparison with other methods while the bottom presents our ablation studies and analysis.

Method	Replacement Mode										Enhancement Mode									
	PSNR↑	SSIM↑	LPIPS↓	$\mathcal{L}_1$↓	CSIM↑	MAE(°)↓	AED↓	APD↓	FID↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	$\mathcal{L}_1$↓	CSIM↑	MAE(°)↓	AED↓	APD↓	FID↓	FVD ↓
SkyReels-A1 [42]	24.9051	0.8586	0.1622	0.0418	0.7163	13.0791	0.7157	0.0155	50.2007	1249.7021	24.9206	0.8584	0.1612	0.0417	0.7176	11.3797	0.676	0.016	51.47	1196.1767
Hunyuan-Portrait [70]	22.4287	0.7924	0.1691	0.0423	0.7364	10.4001	0.6608	0.0348	38.5997	1925.0654	22.7968	0.7984	0.1649	0.0406	0.736	9.6267	0.6073	0.0332	39.3112	1861.9558
FantasyPortrait [60]	23.9456	0.8204	0.1883	0.0349	0.7338	16.7765	0.7953	0.0168	60.4381	606.6303	24.2032	0.8239	0.1848	0.0339	0.7387	12.2154	0.6614	0.0155	62.0027	551.448
Wan-Animate [7]	22.8196	0.8021	0.1319	0.0467	0.614	11.9722	0.7002	0.0238	28.1003	849.2188	22.6417	0.8006	0.1402	0.0491	0.6062	11.2752	0.6847	0.0225	31.2186	885.4313
Act-Two [46]	20.8344	0.7913	0.1634	0.0459	0.6819	15.9618	0.7901	0.0723	36.1933	322.0726	20.8066	0.7916	0.165	0.0458	0.685	15.4109	0.8019	0.0727	38.0579	330.5131
LivePortrait [19]	27.7296	0.8989	0.0591	0.0183	0.7494	10.4746	0.6098	0.0162	14.3562	165.1011	28.0103	0.9024	0.0479	0.0172	0.796	7.6325	0.4915	0.0112	12.2974	114.2545
Ours	**29.2724**	**0.9141**	**0.0474**	**0.014**	**0.7613**	9.1217	**0.4986**	**0.0122**	**12.0138**	**102.9898**	**30.2665**	**0.9216**	**0.0394**	**0.0128**	0.8191	**6.8211**	0.4472	**0.0102**	10.7694	**90.2483**
Ours (KT of LP)	27.0623	0.8908	0.0968	0.018	0.7512	**8.7438**	0.5733	0.0144	27.6785	288.8396	28.3379	0.903	0.0548	0.0151	0.785	7.2571	0.473	0.013	13.8178	139.0848
Ours (w/o FLAME loss)	24.9869	0.8726	0.0805	0.0224	0.7268	9.9748	0.6625	0.0324	18.3957	188.1082	28.4225	0.8977	0.0838	0.0159	0.8063	9.728	0.5384	0.0132	19.5149	132.4458
Ours (w/o T-S)	27.7346	0.8976	0.0583	0.0166	0.7395	8.847	0.5749	0.0147	14.4061	136.4097	29.6156	0.9148	0.0421	0.0135	**0.821**	7.025	**0.4453**	0.0106	**10.7291**	100.7387

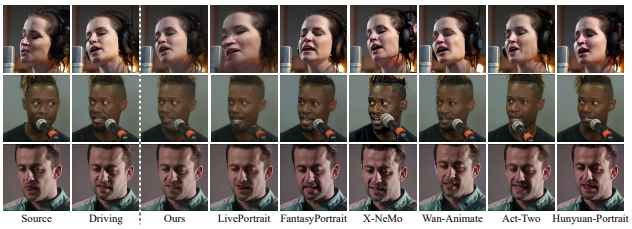

Figure 6. Qualitative comparison of self-reenactment portrait animation. We show the source frame, driving frame and generative results. These source-driving paired images are from official test split of VFHQ dataset [68]. Please zoom in for better inspection.

terial. For replacement mode, we randomly select one of the 18 expressions as source video, and select another different expression from facial motion actors as driving video for each digital human. For enhancement mode, we select the video without expression as source video. As a result, for each mode, we obtain 20 source-driving-ground truth video triplets, one for each digital human. We will release our constructed test benchmark to facilitate the future research.

Qualitative results. The top of Fig. 5 presents some qualitative results generated by different methods on replacement modes. LivePortrait [19] performs poorly and tends to generate inaccurate eyeballs direction. Act-Two [46] fails to accurately preserve the head pose of source video and struggles to synthesize the fine-grained expression details from the driving video. On the contrary, our method is capable of faithfully preserving both the head pose of source video and facial expressions in driving video. More visual comparisons on the enhancement mode are provided in the supplementary material.

Quantitative results. We provide the quantitative comparison with different methods at the top part of Tab. 1. Due to the delicate design of expression and head pose disentanglement, our method outperforms all other methods on all metrics. Diffusion-based methods inherently lack the capability of expression-only video editing, and their performances are still of low quality even after our modification. Therefore, we do not conduct qualitative comparison with them in the main manuscript.

4.2. Ablation Studies and Analysis

Keypoints Transformation Formula. We first analyze the importance of using our improved keypoints transformation instead of that of LivePortrait. We train an additional model using the keypoints transformation of LivePortrait on expression editing task. The third-to-last row of Fig. 5 and Tab. 1 (denoted as "KT of LP") shows the generated results. This variant tends to synthesize relatively blurry videos, and exhibits suboptimal quantitative metrics.

Essentials of FLAME Loss. We then verify the effectiveness of FLAME loss $\mathcal{L}_{\text{FLAME}}$ in our training pipeline by training a PerformRecast variant without FLAME loss, which is reported at the second-to-last row of Fig. 5 and Tab. 1. Without the supervision from FLAME loss, the performance of our model drops markedly, due to the inaccuracy of the motion extractor.

Essentials of Boundary Alignment Module. Finally, we provide the performance of our model trained without boundary alignment module at the last row of Tab. 1 and Fig. 5 (denoted as "w/o BAM"). From which we can conclude that boundary alignment module effectively alleviates the misalignment between the facial and non-facial regions in generated images and leads to the improvement on pixel-level evaluation metrics (such as PSNR, FID, etc.).

4.3. Portrait Animation

We further investigate our model's ability on portrait animation task, which aims to animate a static portrait using both the facial expression and head pose from a driving video. We compare our model with several non-diffusion-based methods, including GAGAvatar [8], EDTalk [52], LivePortrait [19], together with other approaches mentioned in Sec. 2.1. We also compare against recent diffusion-based models, such as AniPortrait [65], X-NeMo [78], Wan-Animate [7], as well as some other methods in Sec. 2.2. Finally, we compare our model with Act-Two [46].

4.3.1. Self-reenactment

Self-reenactment portrait animation task uses the first frame as the source image and animate it using the whole frames in the same video. All the methods are evaluated on official test split of VFHQ dataset [68], which consists of 50 videos.

Qualitative results. Fig. 6 provides some qualitative results on the same source-driving paired frames by different methods. LivePortrait [19] is likely to move other object in the source frame, such as the microphone in the second case. Other diffusion-based methods tend to produce unstable results and struggle to capture subtle facial expression,

Table 2. Quantitative comparisons of different methods on portrait animation task. The top of the table shows the results of non-diffusion-based methods while the bottom presents diffusion-based methods.

Method	Self-reenactment										Cross-reenactment			
	PSNR↑	SSIM↑	LPIPS↓	$\mathcal{L}_1$↓	CSIM↑	MAE(°)↓	AED↓	APD↓	FID↓	FVD↓	CSIM↑	MAE(°)↓	AED↓	APD↓
GAGAvatar [8]	-	-	-	-	0.7878	8.1317	0.4057	0.014	-	-	0.6871	12.2879	0.8144	0.0256
Portrait4D-v2 [11]	14.3781	0.5601	0.4877	0.1268	0.7525	8.4331	0.4434	0.0262	55.0744	506.1298	0.6702	13.4384	0.8385	0.0321
PD-FGC [58]	16.4477	0.62	0.4252	0.0936	0.3282	11.9368	0.6345	0.0322	76.2318	628.5279	0.2181	15.0314	0.8988	0.0417
EMOPortrait [14]	18.7677	0.6979	0.2711	0.0648	0.6297	7.7967	0.4676	0.0179	43.2991	444.5448	0.3483	**10.3905**	0.8012	**0.0245**
EDTalk [52]	20.0771	0.7272	0.3054	0.0648	0.674	16.3315	0.572	0.0252	60.5285	369.9713	0.5435	22.5909	1.018	0.0531
LIA-X [63]	18.249	0.6711	0.2763	0.0731	0.7416	10.6918	0.5797	0.0768	35.1192	317.5413	**0.8270**	27.6757	1.2581	0.2083
LivePortrait [19]	**22.8809**	0.7891	0.165	0.0433	0.8008	6.595	0.3419	0.0095	21.3192	192.0196	0.6595	12.6259	0.8264	0.0295
FYE [37]	20.1905	0.7168	0.2118	0.0564	0.7618	11.8316	0.5676	0.0273	30.0686	343.32	0.7187	15.4403	1.1178	0.0482
AniPortrait [65]	21.0342	0.7334	0.1809	0.0499	0.7654	10.0084	0.4143	0.015	26.6765	210.9606	0.6894	18.6912	1.117	0.0444
X-NeMo [78]	16.5234	0.5666	0.3404	0.096	0.7472	10.3959	0.4102	0.0152	34.3514	373.0014	0.6555	12.8173	0.7997	0.028
ReliPA [20]	16.1173	0.6264	0.3702	0.103	0.5317	12.5351	0.5572	0.021	41.1397	470.7546	0.553	28.4037	1.2146	0.199
SkyReels-A1 [42]	17.4286	0.6644	0.3332	0.087	0.7103	13.1925	0.5291	0.0306	34.9209	363.1019	0.5856	21.6549	1.0562	0.1058
Hunyuan-Portrait [70]	16.97	0.6366	0.3291	0.0873	0.7741	9.9504	0.4218	0.0355	28.0526	266.6914	0.5939	17.7903	0.9142	0.0899
FantasyPortrait [60]	16.8794	0.6321	0.3394	0.0963	0.7368	9.9752	0.4743	0.0393	42.5995	446.281	0.7694	18.581	0.9745	0.1475
Wan-Animate [7]	18.3191	0.6505	0.2825	0.0768	0.7327	9.5602	0.4797	0.0186	26.8861	302.6864	0.5812	14.7004	0.9231	0.0405
VACE [26]	11.1789	0.5398	0.5259	0.2117	0.4311	13.0378	0.6314	0.0248	99.9335	918.8369	0.4001	19.5828	0.9737	0.0396
AvatarArtist [35]	13.3046	0.5414	0.5791	0.1524	0.4025	16.3325	0.6983	0.0443	91.3534	1043.3879	0.5073	14.8549	0.9374	0.0338
Ours	22.7117	**0.7895**	**0.1593**	**0.0409**	**0.8434**	**4.9976**	**0.2606**	**0.009**	**20.1612**	**164.1895**	0.6966	10.9564	**0.7025**	0.0303

Figure 7. Qualitative comparison of cross-reenactment portrait animation. We show the source image, driving image and the generative results. The source images are from FFHQ dataset and driving frames are from VFHQ dataset.

such as eye gazes and lip movements. On the contrary, our model faithfully reconstructs the fine-grained expressions and head pose of the driving frame.

Quantitative results. The left part of Tab. 2 reports the comparison of each metric between PerformRecast and other methods on self-reenactment. Since GAGAvatar [8] can only produce results with black background, we do not calculate pixel-level metrics for it. From Tab. 2 we can observe that our method obtains the best performance on all metrics except PSNR, outperforming all other methods.

4.3.2. Cross-reenactment

For cross-reenactment portrait animation, we select 50 images from FFHQ dataset [27] as source portraits. The official test split of VFHQ dataset [68] are used as driving videos. We only adopt CSIM, MAE, AED and APD as evaluation metrics due to the lack of ground truth target images.

Qualitative results. Fig. 7 visualizes the generated results of different compared methods on the same source-driving paired images. LivePortrait [19] exhibits severe artifacts on the third case. Other diffusion-based methods generally produce overly exaggerated facial expression, such as wrinkles on the forehead and the wide smiles around the lips. In contrast, our method is capable of faithfully transferring the subtle facial expression, direction of eyeballs and head pose rotations from the driving image to source image.

Quantitative results. The right part of Tab. 2 presents the quantitative results of the cross-reenactment comparisons. Our model outperforms all previous methods in terms of facial expression accuracy. While EMOPortrait [14] reports lower MAE and APD than ours, its generated results suffer from severe background blur and distortion as shown in Fig. 7. What's more, although FantasyPortrait [60] and FYE [37] attain higher facial identity similarity, they lag significantly behind on all other metrics. Therefore, we can conclude that our PerformRecast achieves state-of-the-art overall performance on the cross-reenactment task. In addition, our method is able to generate six images per second on the consumer-grade GPU device, which significantly reduces the complexity of deployment in practical scenarios.

5. Conclusion

In this paper, we propose PerformRecast, a versatile and effective method tailored for expression-only video performance editing and portrait animation tasks. We improve the original keypoints transformation formula in LivePortrait [19] to make it more consistent with 3DMM. As a result, our PerformRecast is much better at disentangling the face identity, facial expression and head pose rotation. Experimental results demonstrate that our method is capable of editing the expression of a portrait video as well as animating a static portrait image, offering great convenience in the film and computer animation industries.

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. 14

[2] Stability AI. Stable diffusion v1.5 model card. https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5, 2022. 2, 3

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 2, 3

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 3

[5] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13786–13795, 2020. 3

[6] 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. 14

[7] 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. 3, 6, 7, 8, 16

[8] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3, 7, 8, 16

[9] 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. 4, 13

[10] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7119–7130, 2024. 3

[11] 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. 3, 8, 16

[12] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14398–14407, 2021. 3

[13] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. 2022. 2, 3, 15

[14] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars, 2024. 2, 3, 6, 8, 16

[15] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2235–2245, 2018. 4

[16] Epic Games. Metahuman creator. https://www.unrealengine.com/en-US/digital-humans, 2021. 2, 6, 14

[17] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction, 2025. 2, 3, 4, 13

[18] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2, 3

[19] 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, 3, 4, 5, 6, 7, 8, 13, 16, 17

[20] 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. 8, 16

[21] 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 *European Conference on Computer Vision*, pages 20–36. Springer, 2024. 6

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[24] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3397–3406, 2022. 3

[25] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2, 3

[26] 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. 3, 8, 16

[27] 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. 3, 8, 17

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2, 3

[29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 3

[30] 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. 6, 14

[31] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 14

[32] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. 6

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 14

[34] 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. 2, 3, 4, 13, 15

[35] 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. 3, 8, 16

[36] Luming Ma and Zhigang Deng. Real-time hierarchical facial performance capture. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pages 1–10, 2019. 2

[37] 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. 3, 8, 16

[38] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

[39] Jacek Naruniec, Leonhard Helminger, Christopher Schroers, and Romann M Weber. High-resolution neural face swapping for visual effects. In *Computer Graphics Forum*, pages 173–184. Wiley Online Library, 2020. 2

[40] 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. 5, 13

[41] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 3

[42] 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. 3, 6, 7, 8, 15, 16

[43] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13759–13768, 2021. 3

[44] 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. 14, 15

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[46] Runway. Creating with act-two. https://help.runwayml.com/hc/en-us/articles/42311337895827-Creating-with-Act-Two, 2025. 6, 7, 16

[47] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2377–2386, 2019. 3

[48] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 6

[49] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13653–13662, 2021. 3

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 2, 3

[51] 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. 14

[52] 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. 3, 7, 8, 16

[53] Shuai Tan, Bill Gong, Bin Ji, and Ye Pan. Fixtalk: Taming identity leakage for high-quality talking head generation in extreme cases. *arXiv preprint arXiv:2507.01390*, 2025. 3

[54] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2

[55] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3

[56] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 6

[57] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3

[58] 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. 3, 8, 15, 16

[59] 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. 6, 16, 17

[60] 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. 6, 7, 8, 16

[61] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3, 4

[62] 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. 3, 15

[63] 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. 3, 8, 16

[64] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[65] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 3, 7, 8, 16

[66] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018. 3

[67] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142, 2023. 3

[68] 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, 7, 8, 17

[69] 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. 15

[70] 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. 3, 6, 7, 8, 15, 16

[71] Kewei Yang, Kang Chen, Daoliang Guo, Song-Hai Zhang, Yuan-Chen Guo, and Weidong Zhang. Face2face ρ: Real-time high-resolution one-shot face reenactment. In *European conference on computer vision*, pages 55–71. Springer, 2022. 3

[72] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 3

[73] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019. 3

[74] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020. 3

[75] Xianfang Zeng, Yusu Pan, Mengmeng Wang, Jiangning Zhang, and Yong Liu. Realistic face reenactment via self-supervised disentangling of identity and pose. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12757–12764, 2020. 3

[76] 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. 6, 14

[77] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose

and gaze variation. In *European Conference on Computer Vision (ECCV)*, 2020. 6

[78] 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. 3, 7, 8, 16

[79] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022. 13

Table of Contents

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

A. 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.

B. 3DMM-based Face Tracking

To obtain temporally continuous FLAME [34] model reconstruction results from input portrait videos, We adopt a recently-proposed 3D face tracking method, Pixel3DMM [17] 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 [40] 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 [34], 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 [79] 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 [79] 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.

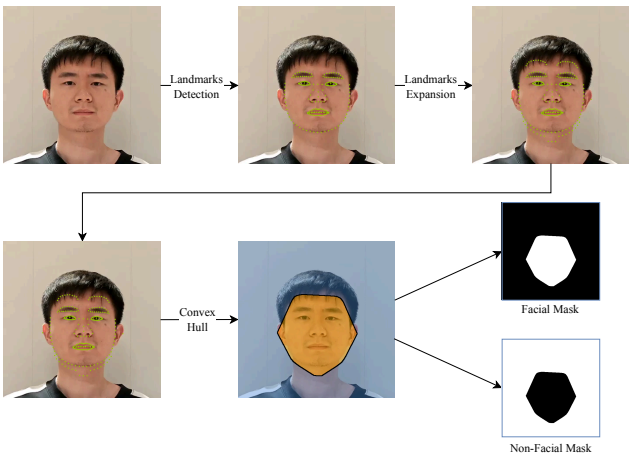

Figure 8. The facial mask calculation process of each frame in training dataset.

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

C. 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:

$$\mathcal{L}_{\text{animate}} = \mathcal{L}_{\text{FLAME}} + \mathcal{L}_{P,\text{cascade}} + \mathcal{L}_{1,\text{cascade}} + \mathcal{L}_{G,\text{cascade}} + \mathcal{L}_{\text{faceid}}, \tag{10}$$

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 [19]. And the face-id [9] loss is used to preserve the identity of source image I_s.

D. Facial Mask Calculation

As shown in Fig. 8, to obtain masks of facial and non-facial regions, we also utilize the pre-trained 2D facial landmark detector in LivePortrait [19] 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.

E. 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

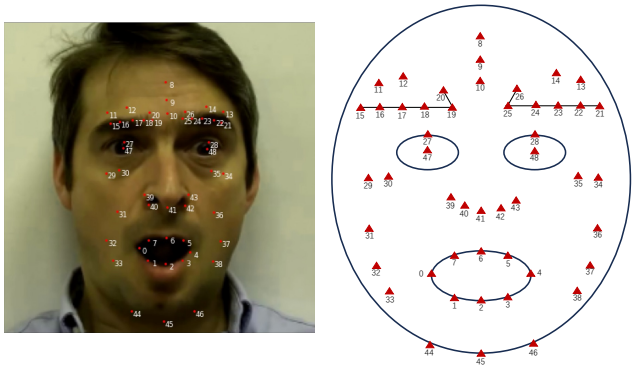

Figure 9. The specific location of each keypoint used in our method. Please zoom in for better inspection.

source and driving explicit keypoints are transformed as follows:

$$\begin{cases} x_s = s_s \cdot \big((x_{c,s} + \delta_s)R_s\big) + t_s, \\ x_{d,i} = s_{d,i} \cdot \big((x_{c,s} + \delta_{d,i})R_{d,i}\big) + t_{d,i}, \end{cases} \tag{11}$$

which utilizes the same formula as training stage.

F. Keypoints Selection

We select $K = 49$ keypoints from the reconstructed FLAME face mesh in total to supervise our motion extractor. Fig. 9 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.

G. 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 [31] 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.

H. Evaluation Metrics Details

LPIPS. For potrait video expression editing and self-reenactment, we calculate the perceptual similarity metric LPIPS [76] based on AlexNet [33] 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:

$$\text{FID} = ||\mu_g - \mu_r||^2 + \text{Tr}(\Sigma_g + \Sigma_r - 2(\Sigma_r\Sigma_g)^{1/2}), \tag{12}$$

where g and r denote the features of the generated image and ground truth images, which is extracted by Inception-v3 model [51]. 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:

$$\text{FVD} = ||\mu_g - \mu_r||^2 + \text{Tr}(\Sigma_g + \Sigma_r - 2(\Sigma_r\Sigma_g)^{1/2}), \tag{13}$$

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 [6]. 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 [30]. 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 [44].

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 [44].

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].

I. Construction of Our Test Benchmark

Fig. 10 visualizes the construction pipeline of our proposed test benchmark. Given a input video, our pipeline first utilize MetaHuman [16] to extract the expression and head pose parameters of each frame. The detailed information of this process are shown at the top of Fig. 10. 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

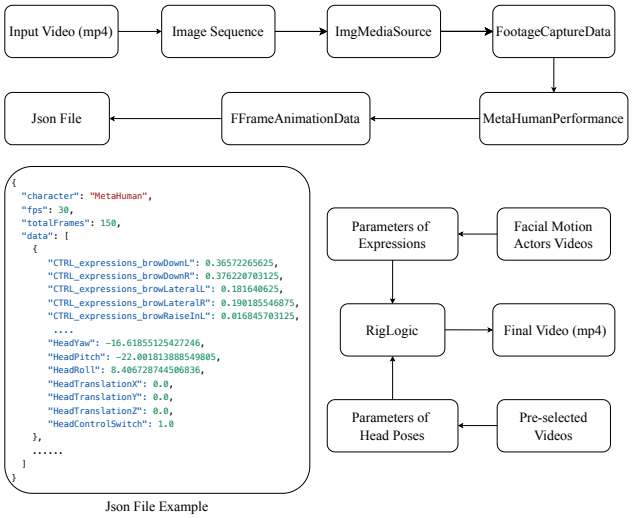

Figure 10. Construction pipeline of our proposed test benchmark.

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.

J. 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 [42] utilizes SMIRK [44] to extract FLAME [34] 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.

Figure 11. 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.

Hunyuan-Portrait. Hunyuan-Portrait [70] utilizes pre-trained motion encoder MegaPortraits [13] 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 [58] 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 [69] 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 [62] 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.

K. More Results

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

K.1. Portrait Video Expression Editing

We first compensate for the missing visual comparisons on the enhancement mode in Fig. 11 as mentioned in the main manuscript.

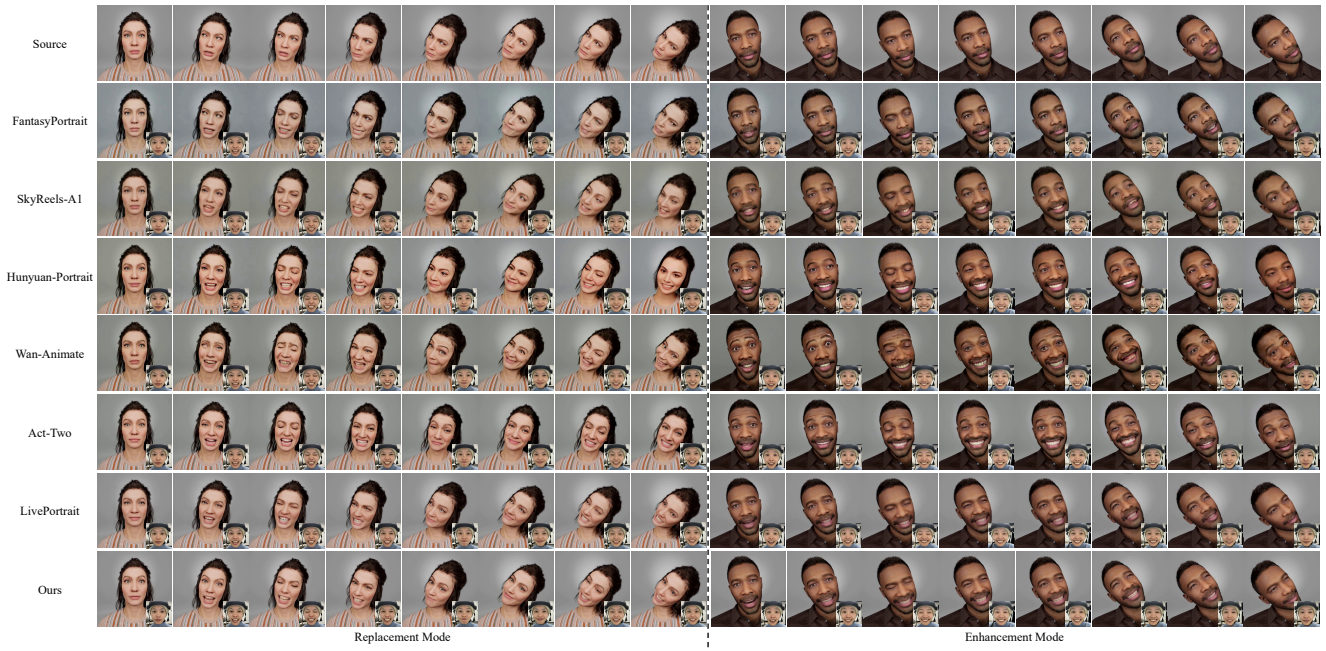

Figure 12. Full qualitative comparison with all the methods mentioned in the main manuscript on our proposed test benchmark. The bottom-right insets exhibit driving frames. Please zoom in for better inspection.

Table 3. Quantitative comparisons of self-reenactment portrait animation on MEAD [59] 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 [8]	-	-	-	-	0.8946	5.1074	0.3781	0.0101	-	-
Portrait4D-v2 [11]	20.0907	0.7746	0.3358	0.0617	0.8793	5.4329	0.4353	0.0149	85.3589	460.7595
PD-FGC [58]	20.8419	0.7824	0.341	0.0573	0.3256	8.3772	0.7156	0.0202	92.8886	1276.3855
EMOPortrait [14]	26.1748	0.8729	0.1544	0.0287	0.5959	6.6994	0.4992	0.0128	37.5216	443.7428
EDTalk [52]	26.9246	0.8964	0.1443	0.0319	0.8592	6.218	0.4333	0.0077	43.4199	343.9973
LIA-X [63]	22.6439	0.8232	0.1816	0.0386	0.8957	5.3919	0.4633	0.0636	32.4902	323.2831
LivePortrait [19]	<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 [37]	27.1819	0.8963	0.1039	0.0243	0.8767	5.6658	0.527	0.0109	30.5002	350.4705
AniPortrait [65]	29.0281	0.9125	0.081	0.0198	0.8904	4.8224	0.3989	0.0077	19.9857	191.8255
X-NeMo [78]	22.4136	0.7313	0.1916	0.0551	0.8594	10.4812	0.4168	0.0097	50.6639	409.2123
ReliPA [20]	24.0052	0.8525	0.1601	0.0409	0.8212	6.3512	0.5174	0.0117	35.1439	455.0235
SkyReels-A1 [42]	25.9931	0.8825	0.1182	0.032	0.8668	5.7577	0.594	0.0105	22.6852	278.6554
Hunyuan-Portrait [70]	26.4138	0.8779	0.0941	0.0309	0.922	4.7961	0.3348	0.0101	18.896	139.2772
FantasyPortrait [60]	22.6155	0.7789	0.1498	0.0634	0.8622	6.606	0.5147	0.0116	35.7586	258.8542
Wan-Animate [7]	21.9017	0.8105	0.2159	0.054	0.827	5.7592	0.5307	0.0144	24.9683	465.8136
VACE [26]	15.0009	0.5046	0.4083	0.1587	0.5472	10.0836	0.745	0.021	118.5134	870.7917
AvatarArtist [35]	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**

LivePortrait [19] generates inaccurate lip movements on the enhancement mode. Act-Two [46] 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. 12. 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

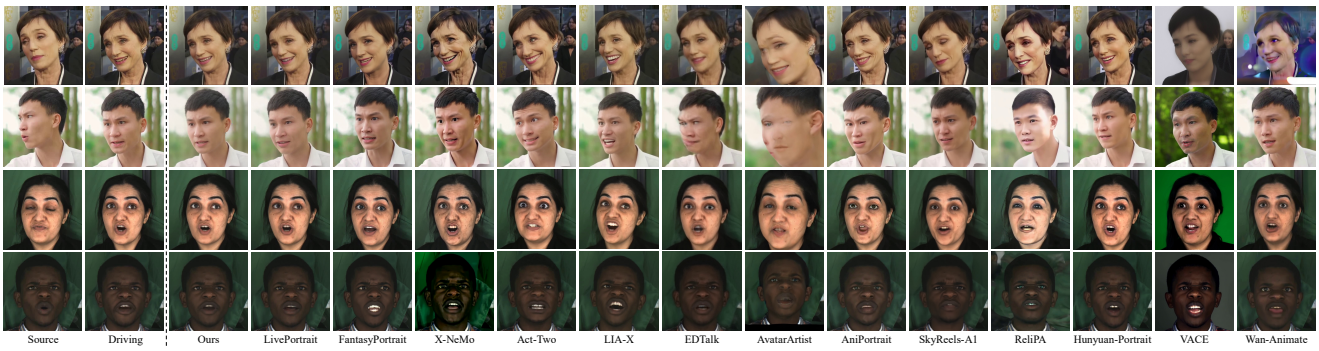

Figure 13. More generated results on self-reenactment task of different methods. The first two source-driving paired images are from VFHQ dataset [68] and the last two source-driving paired images are from MEAD dataset [59].

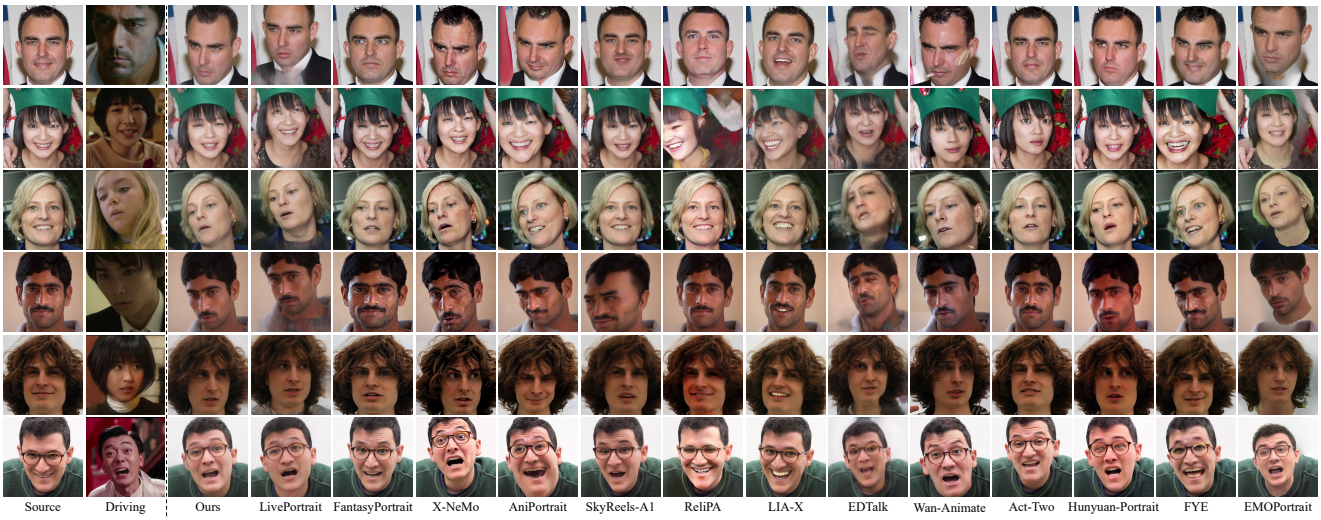

Figure 14. More generated results on cross-reenactment task of different methods. The source images are from FFHQ dataset and the driving frames are from famous films and television clips.

on both replacement and enhancement modes compared to all previous approaches.

K.2. Self-reenactment

We also report the quantitative results of self-reenactment portrait animation on MEAD dataset [59] in Tab. 3. All the methods are evaluated on a random split of MEAD dataset, which consists of 70 videos. As shown in Tab. 3, our method achieves the best performance across all metrics on MEAD dataset [59], highlighting its superiority over other existing approaches.

What's more, we also present more qualitative results of different compared methods in Fig. 13. LivePortrait [19] 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.

K.3. Cross-reenactment

We provide more cross-reenactment portrait animation results generated by our PerformRecast and some other methods in Fig. 14. The source images are from FFHQ dataset [27] 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 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.

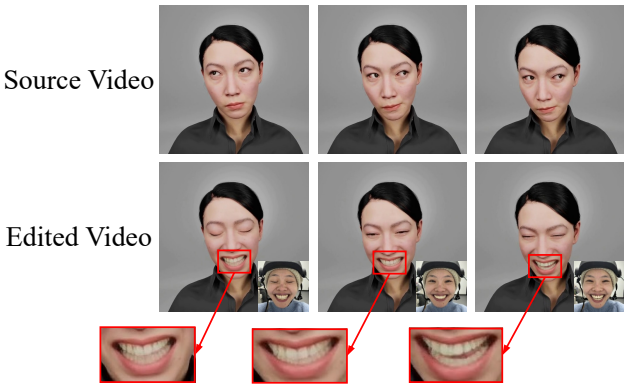

Figure 15. A typical failure case when our method generating teeth while the mouth is closed in source video.

L. 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. 15 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.

M. 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.