Computers & Graphics (2025)

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Navigating Large-Pose Challenge for High-Fidelity Face Reenactment with Video Diffusion Model

Mingtao Guo[a], Guanyu Xing[b], Yanci Zhang[c], Yanli Liu[a,c,*]

[a]National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, 610065, China
[b]School of Cyber Science and Engineering, Sichuan University, Chengdu, 610065, China
[c]College of Computer Science, Sichuan University, Chengdu, 610065, China

ARTICLE INFO

Article history:

Face Reenactment, Motion Extractor, Warping Feature Mapper, Motion-aware Latent Space, Video Diffusion Model.

ABSTRACT

Face reenactment aims to generate realistic talking head videos by transferring motion from a driving video to a static source image while preserving the source identity. Although existing methods based on either implicit or explicit keypoints have shown promise, they struggle with large pose variations due to warping artifacts or the limitations of coarse facial landmarks. In this paper, we present the Face Reenactment Video Diffusion model (FRVD), a novel framework for high-fidelity face reenactment under large pose changes. Our method first employs a motion extractor to extract implicit facial keypoints from the source and driving images to represent fine-grained motion and to perform motion alignment through a warping module. To address the degradation introduced by warping, we introduce a Warping Feature Mapper (WFM) that maps the warped source image into the motion-aware latent space of a pretrained image-to-video (I2V) model. This latent space encodes rich priors of facial dynamics learned from large-scale video data, enabling effective warping correction and enhancing temporal coherence. Extensive experiments show that FRVD achieves superior performance over existing methods in terms of pose accuracy, identity preservation, and visual quality, especially in challenging scenarios with extreme pose variations.

1. Introduction

Face reenactment is the process of synthesizing a lifelike talking head video using a provided source image as a reference, guided by a driving video. In this synthesis, the resulting face maintains the identity attributes of the source image while adopting the pose and expressions from the driving video. Face reenactment has many valuable applications, including character role-playing, digital avatars, online education, video conferencing, etc.

Existing face reenactment methods have demonstrated remarkable capabilities in generating talking faces. Keypoints are typically employed by these methods to represent facial motion. According to whether the keypoints are learned automatically (implicit) or predefined (explicit), these methods can be categorized into two types: implicit keypoint-based methods and explicit keypoint-based methods. Among them, implicit keypoint-based methods [22, 12, 37, 26, 8] learn keypoints from the source and driving images to estimate dense motion fields (e.g., optical flow). These fields are then used to warp the source image toward the pose and expression of the driving image. Finally, a generator [7, 15, 14] inpaints occluded

*Corresponding author: Email: yanliliu@scu.edu.cn
e-mail: mingtaoguo@stu.scu.edu.cn (Mingtao Guo), xingguanyu@scu.edu.cn (Guanyu Xing), yczhang@scu.edu.cn (Yanci Zhang), yanliliu@scu.edu.cn (Yanli Liu)

or degraded regions to produce the final frame. By providing dense and flexible motion guidance, these methods effectively capture fine-grained facial deformations. However, when there is a large pose discrepancy between the source and driving images, the limited identity and appearance information in a single source image often fails to support effective inpainting in severely warped regions, leading to degraded synthesis quality. In contrast, explicit keypoint-based methods [20, 31, 4] rely on facial landmarks extracted from the driving video to generate pose maps as spatial conditions for the pre-trained Stable Diffusion model [21]. To maintain identity and appearance consistency with the source image, they further incorporate fine-grained texture and appearance features from the source using ReferenceNet [13]. Although these approaches can produce high-quality facial textures, they are fundamentally limited by the coarse nature of explicit facial landmarks [19, 3], which primarily capture rigid facial contours. When handling large head poses (e.g., profile views), these rigid contours often result in overlapping or collapsed keypoints, causing the generated pose maps to lose meaningful facial structure and identity cues. Therefore, handling large pose variations remains a significant challenge for existing face reenactment methods.

To achieve high-fidelity face reenactment under large pose variations, we leverage implicit facial keypoints to represent facial motion and use them to warp the source image toward the target pose and expression. To address warping-induced degradation, we observe that image-to-video (I2V) models trained on large-scale video datasets are highly effective at synthesizing realistic and temporally coherent facial dynamics—such as head movements, speech, and blinking—while reliably preserving identity and appearance consistency, even under extreme pose variations. Therefore, our key insight is to exploit the I2V model's motion-aware latent feature space to reconstruct regions degraded by warping, enabling temporally coherent video generation that faithfully preserves source identity while recovering fine-grained details lost during the warping process.

In this paper, we propose a **F**ace **R**eenactment **V**ideo **D**iffusion model (**FRVD**) for high-fidelity face reenactment under large pose variations. Our model first employs a Motion Extractor to extract implicit keypoints from both the source and driving images, which serve as fine-grained representations of facial motion. These keypoints are then used in a warping module to align the motion of the source image with that of the driving image. To recover regions degraded during the warping process, we introduce a Warping Feature Mapper (WFM) that maps features from the warped source image into the motion-aware latent space of a pretrained I2V model. This latent space, learned from large-scale video data, encodes rich spatiotemporal priors of facial dynamics, enabling the model to perform effective warping correction. By leveraging these priors, the WFM facilitates high-quality reconstruction of facial details while preserving both identity and temporal coherence.

Our main contributions are summarized as follows:

- We propose a **F**ace **R**eenactment **V**ideo **D**iffusion model (**FRVD**) to address the challenge of face reenactment under large pose variations, overcoming the limitation of existing methods, which typically produce satisfactory results only when the pose of the source image closely matches that of the driving image.
- We introduce the Warping Feature Mapper (WFM), which maps the warped source image into the motion-aware latent space of a pretrained image-to-video (I2V) model. This allows the model to leverage its prior knowledge to reconstruct degraded regions, thereby enabling high-fidelity face reenactment under large pose variations.
- Extensive experiments demonstrate that FRVD outperforms state-of-the-art methods, achieving significant improvements in pose accuracy, identity preservation, and overall video quality.

2. Related Work

Implicit-Keypoints-Based Face Reenactment Methods. Contemporary face reenactment approaches leveraging implicit keypoint representation [22, 12, 37, 26, 8] eliminate the need for prior knowledge of driving subjects during model training. Notably, the FOMM [22] establishes a theoretical framework through first-order Taylor expansions around latent keypoints, implementing local affine transformations to approximate complex facial motions. This foundational work demonstrates significant performance improvements in motion transfer fidelity. Building upon this foundation, DaGAN [12] introduces a self-supervised paradigm for fine-grained pixel-level depth estimation, enabling enhanced 3D facial structure perception and high-frequency spatial detail preservation. In contrast, LivePortrait [8] proposes a novel motion cue disentanglement mechanism that implicitly captures and transfers holistic facial dynamics—including head pose and expression variations—from driving videos while maintaining content consistency. Despite these advancements, fundamental limitations remain: existing methods still suffer from performance degradation under extreme pose variations, often resulting in geometric distortions and unrealistic texture artifacts caused by the warping module.

Explicit-Keypoints-Based Face Reenactment Methods. In contrast to implicit-keypoints-based methods, explicit keypoint-driven approaches [20, 31, 4, 9] typically rely on existing facial landmark detection models [19, 3] to extract keypoints from each frame of the driving video. These keypoints are then used to construct facial contour maps, which serve as pose guidance for the generation model. The pose guider directs the synthesis of facial images with corresponding head poses. Meanwhile, identity and appearance information from the source image is preserved by leveraging spatial attention mechanisms between features extracted by the ReferenceNet [13] and those in the UNet of the Stable Diffusion model [21]. Additionally, diffusion-based face reenactment methods [17, 2] also leverage facial contours derived from explicit keypoints to guide the diffusion model in generating faces with the desired poses. However, under extreme poses, explicit-keypoints-based approach often fails to preserve facial structure in the contour map, resulting in ineffective guidance and noticeable distortions in the reenacted face.

Image-to-Video Diffusion Models. Image-to-video diffusion models [1, 32, 25] aim to generate a video from a single reference image, where the first frame is identical to the input image, and subsequent frames maintain consistent foreground and background while the motion is guided by user-provided textual descriptions. To ensure temporal consistency across frames, the Stable Video Diffusion (SVD) [1] model extends Stable Diffusion [21] by introducing a temporal attention module. To address the limitation of separate spatial and temporal attention—particularly the failure to track fast-moving objects—CogVideoX [32] integrates spatial attention, cross-attention, and temporal attention [10] into a unified self-attention mechanism, enabling stronger semantic understanding and temporal coherence. However, this unified attention design significantly increases the number of tokens, leading to high computational costs and low inference efficiency. Considering these trade-offs, we adopt SVD as the backbone for our image-to-video generation model.

3. Methodology

3.1. Overview

Our method takes a source image and a driving video as input, and reenacts the source face to match the pose in each frame of the driving video. As illustrated in Fig. 1, our face reenactment framework consists of two stages. During the training stage, we first employ a Motion Extractor to extract pose and expression coefficients from both the source image and the driving video. The source image is then warped to match the pose and expression of the driving video, ensuring spatial alignment with each frame (Sec. 3.2). Next, the warped source image is encoded by a Warping Feature Mapper (WFM). At each layer of the WFM, the extracted features are modulated by an Internal Feature Modulator (IFM). These modulated features are then fed into the Stable Video Diffusion (SVD) model [1]. By leveraging the identity and appearance priors of the source image embedded in SVD, we perform warping correction within its motion-aware latent space to reconstruct the missing regions of the warped source image (Sec. 3.3). During the inference stage, we introduce an additional Motion Alignment Module into the Motion Extractor to support cross-identity face reenactment (Sec. 3.4).

3.2. Motion Extraction and Warping

The core objective of face reenactment is to transfer the pose and expression from a driving video to a source image, thereby transforming a static source image into a dynamic face that maintains the identity of the source image while aligning with the pose and expression of the driving video. To achieve this, the primary task is to disentangle the pose, expression, identity, and appearance from the source image.

As illustrated in Fig. 1, the training stage of the motion extractor, we first employ the appearance and identity feature extractor $\mathcal{F}$ to extract features from the source image, denoted as $\mathbf{F}_s$. Subsequently, the canonical keypoint estimator $\mathcal{L}$ is used to extract facial keypoints in the canonical space from the source image. Next, the head pose estimator $\mathcal{H}$ estimates the pose of the source image's face, represented by a rotation matrix $\mathbf{R}_s$ and a translation matrix $\mathbf{t}_s$. For each frame of the driving video, the facial pose is also estimated and denoted as $\left\{\mathbf{R}_d^i\right\}_{i=1}^m$ and $\left\{\mathbf{t}_d^i\right\}_{i=1}^m$, where m represents the total number of frames in the driving video. Finally, the expression estimator $\mathcal{E}$ is utilized to estimate the facial expression coefficients for both the driving video and the source image, denoted as $\left\{\boldsymbol{\delta}_d^i\right\}_{i=1}^m$ and $\boldsymbol{\delta}_s$, respectively.

Based on these estimated parameters, the facial keypoints for the source image and each frame of the driving video are computed using Eq. (1),

$$\begin{cases}\mathbf{x}_s = \mathbf{R}_s\mathbf{x}_c + \mathbf{t}_s + \boldsymbol{\delta}_s \\ \mathbf{x}_d^i = \mathbf{R}_d^i\mathbf{x}_c + \mathbf{t}_d^i + \boldsymbol{\delta}_d^i\end{cases} \tag{1}$$

finally, following OSFV [26], we utilize the source image keypoints $\mathbf{x}_s$ and the driving video frame keypoints $\mathbf{x}_d^i$ in the warping module to warp the source image features $\mathbf{F}_s$. This process is formulated as $f_w : (\mathbf{x}_s, \mathbf{x}_d^i, \mathbf{F}_s) \rightarrow \mathbf{F}_s^w$, where $\mathbf{F}_s^w$ represents the warped version of $\mathbf{F}_s$, ensuring that its pose is aligned with the driving frame.

3.3. Warping Correction in Motion-aware Latent Space

Using the Motion Extractor, we warp the source image in the feature space to align its pose and expression with those of the driving image. However, the warping process inevitably introduces information loss in the affected regions of the feature map. This problem becomes particularly pronounced when there is a significant pose discrepancy between the source and driving images, leading to large-scale facial distortions and severe identity degradation.

To address the distortion and identity loss introduced by warping, our key idea is to extract identity and appearance features from the source image and utilize them to restore the regions degraded during the warping process. We observe that image-to-video (I2V) models are inherently capable of synthesizing temporally continuous frames from a single input image, enabling realistic image-to-video generation. Given a facial image, I2V models can predict natural motions such as head turns, speech, and blinking, while maintaining consistent identity and appearance across frames. This makes them particularly well-suited for handling faces under varying poses. We exploit this property by leveraging the perceptual capabilities of a pre-trained I2V model to restore warped regions at the feature level, ensuring that the reconstructed face remains consistent with the source identity while generating temporally coherent driving videos.

Specifically, we propose a Warping Feature Mapper (WFM), as illustrated in Fig. 1. We first feed the warped feature $\mathbf{F}_s^w \in \mathbb{R}^{C\times H\times W}$ into WFM, where it is encoded by the WFM Encoder (WFMEnc), denoted as $[\mathbf{F}_s^{(1)}, \ldots, \mathbf{F}_s^{(i)}, \ldots] = \text{WFMEnc}(\mathbf{F}_s^w)$, where each $\mathbf{F}_s^{(i)} \in \mathbb{R}^{C_s^{(i)}\times H_s^{(i)}\times W_s^{(i)}}$ denotes the feature representation at the i-th scale. The encoded features $\mathbf{F}_s^{(i)}$ from each encoder layer are then passed to the corresponding Internal Feature Modulator (IFM) for feature modulation: $\mathbf{F}_{s\_m}^{(i)} = \text{IFM}_i(\mathbf{F}_s^{(i)})$, where $\mathbf{F}_{s\_m}^{(i)} \in \mathbb{R}^{C^{(i)}\times H^{(i)}\times W^{(i)}}$ represents the modulated features at the i-th scale. The modulated features $\mathbf{F}_{s\_m}^{(i)}$ are subsequently fed into a pre-trained I2V model. Leveraging the I2V

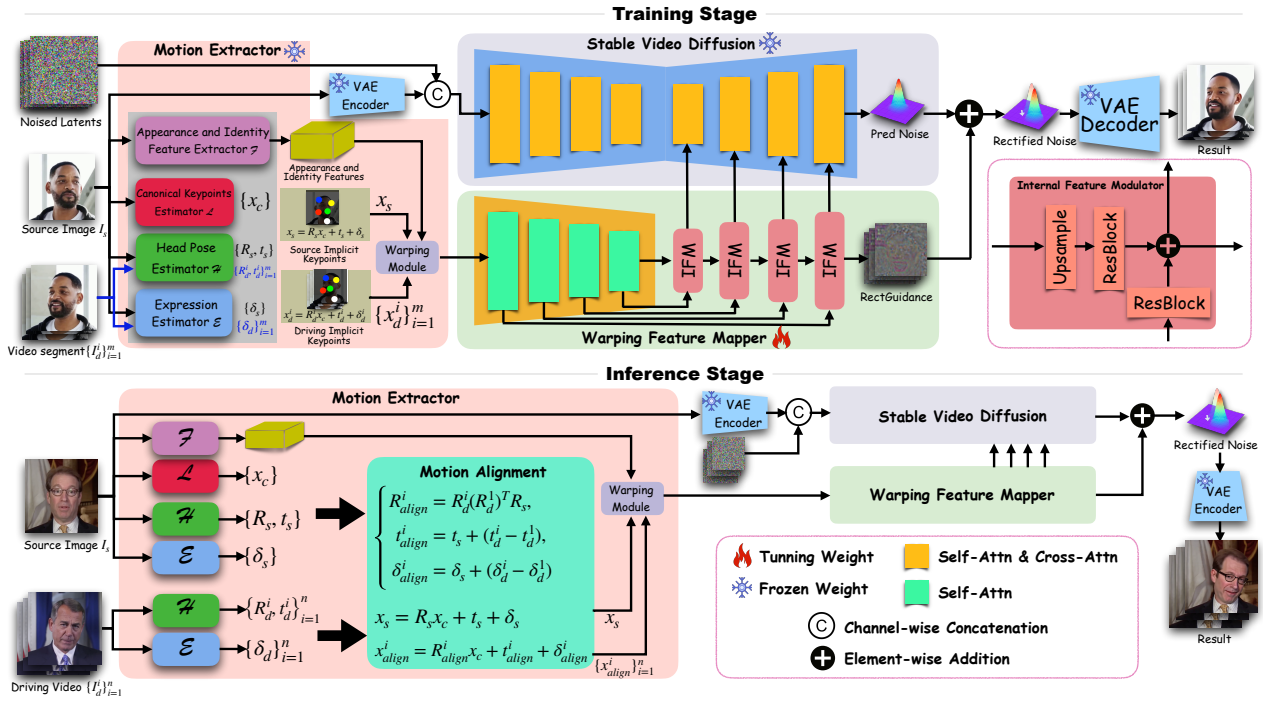

Fig. 1. Our face reenactment framework comprises two stages: (1) Training stage: We begin by employing the Motion Extractor to extract pose and expression coefficients from both the source image and the driving video, while simultaneously encoding appearance and identity features from the source image. These motion coefficients are then used to warp the source image features via the Warping Module, aligning them with the motion of the driving video. The warped features are further encoded by the Warping Feature Mapper and modulated by the Internal Feature Modulator before being passed to the Stable Video Diffusion (SVD) model. Leveraging the identity and appearance priors of the source image inherent in SVD, the model performs warping correction in the motion-aware latent space to reconstruct regions distorted or missing due to warping. (2) Inference stage: To support cross-identity face reenactment, a Motion Alignment Module is introduced into the pipeline, enabling the model to generalize to unseen identities.

model's perceptual ability to encode identity and appearance from the source image, we use it to restore the features degraded by warping. Let $\mathbf{F}^{(j)}$ denote the feature map from the j-th layer of the I2V model, which contains rich identity and appearance cues from the source image. The tensor $\mathbf{F}^{(i)}_{s\_m}$ shares the same shape as $\mathbf{F}^{(j)}$. We fuse $\mathbf{F}^{(j)}$ with $\mathbf{F}^{(i)}_{s\_m}$ via element-wise addition to inject identity-aware information into the warped features: $\mathbf{F}^{(j)}_{fuse} = \mathbf{F}^{(j)} + \mathbf{F}^{(i)}_{s\_m}$. The fused feature $\mathbf{F}^{(j)}_{fuse}$ is then fed into the next layer of the I2V model, and this process continues recursively. In our framework, the I2V model is implemented using the SVD model [1].

Additionally, to enhance SVD's ability to capture the global appearance characteristics of the source image, we introduce a rectified guidance signal predicted by the final IFM layer. This signal shifts the mean of the Gaussian distribution output by SVD, enabling more accurate modeling of the source image's appearance. The training objective is defined in Eq. (2).

$$\mathbf{Loss} = \mathbb{E}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\sqrt{\bar{\alpha}_t}\mathbf{z}_0 - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \mathbf{F}_c, t\right) - \mathbf{r}\|\right] \quad (2)$$

Here, $\boldsymbol{\epsilon}$ represents noise sampled from a standard normal distribution, while $\mathbf{z}_0$ denotes the latent representation of the driving image (i.e., the target image) obtained from the VAE [6]. $\boldsymbol{\epsilon}_\theta$ refers to the backbone of the SVD model. The operation $[\mathbf{F}_c, \mathbf{r}] = WFM(\mathbf{F}^w_s)$ denotes the output of the WFM, where $\mathbf{F}_c = [\mathbf{F}^{(1)}_{s\_m}, ..., \mathbf{F}^{(i)}_{s\_m}, ...]$ is the modulated feature output from WFM, and $\mathbf{r}$ is the rectified guidance used to shift the mean of the Gaussian distribution predicted by SVD. The term $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$, where β_i represents the noise strength coefficient.

3.4. Cross-identity Face Reenactment

Face reenactment aims to drive an image with a video of another identity. During training, a frame is randomly selected as the source image, and a video segment is cropped as the driving video. The model processes the source image to generate a video matching the driving video, enabling self-supervised learning.

However, the goal of face reenactment is to drive a face image of one identity using a face video of another identity. To prevent identity leakage from the driving video, we employ the Motion Alignment Module to align the pose and expression of the driving frames with those of the source image. Our approach computes the facial motion for each frame of the driving video, using the first frame as a reference. We then calculate the motion differences between each frame and the reference, obtaining a relative motion sequence. This sequence is applied to the source image, producing a face reenactment video where the initial motion matches the source image, and subsequent pose and expression changes follow the relative sequence smoothly and consistently. Specifically, we first use the Motion Extractor to compute the rotation matrices $\{\mathbf{R}^i_d\}^m_{i=i}$, translation vectors $\{\mathbf{t}^i_d\}^m_{i=i}$, and expression coefficients $\{\boldsymbol{\delta}^i_d\}^m_{i=i}$ for each frame of the driving video. Similarly, we extract the source image's rotation matrix $\mathbf{R}_s$, translation vector $\mathbf{t}_s$, and expression coefficients $\boldsymbol{\delta}_s$.

Using the first frame's rotation matrix as a reference, we compute the aligned rotation matrix, translation vector and expression coefficients as Eq. (3).

$$\begin{cases} \mathbf{R}^i_{align} = \mathbf{R}^i_d \left(\mathbf{R}^1_d\right)^T \mathbf{R}_s \\ \mathbf{t}^i_{align} = \mathbf{t}_s + \left(\mathbf{t}^i_d - \mathbf{t}^1_d\right) \\ \boldsymbol{\delta}^i_{align} = \boldsymbol{\delta}^i_s + \left(\boldsymbol{\delta}^i_d - \boldsymbol{\delta}^1_d\right) \end{cases} \tag{3}$$

Using these aligned pose and expression parameters, we compute the facial keypoints for both the source image and the driving video, as described in Eq (4).

$$\begin{cases} \mathbf{x}_s = \mathbf{R}_s \mathbf{x}_c + \mathbf{t}_s + \boldsymbol{\delta}_s \\ \mathbf{x}^i_{align} = \mathbf{R}^i_{align} \mathbf{x}_c + \mathbf{t}^i_{align} + \boldsymbol{\delta}^i_{align} \end{cases} \tag{4}$$

We then employ these facial keypoints to warp the source image features, denoted as $\mathbf{F}^w_s = f_w(\mathbf{x}_s, \mathbf{x}_{align}, \mathbf{F}_s)$. The warped features are subsequently encoded by the WFM, $[\mathbf{F}_c, \mathbf{r}] = WFM(\mathbf{F}^w_s)$, and further processed by the SVD model. SVD leverages the DDIM reverse diffusion process [23] to perform warping correction, effectively restoring regions lost during the warping operation. Additionally, we adopt a classifier-free guidance mechanism to control the strength of the correction, as formulated in Eq. (5), where adjusting the value of w modulates the degree of restoration in the warped regions.

$$\boldsymbol{\epsilon}_\theta(\mathbf{z}_t) = w \cdot (\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{F}_c) - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \phi)) + \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{F}_c) + \mathbf{r} \tag{5}$$

4. Experiments

4.1. Implementation Details

Datasets. We train our model using the VFHQ [30] datasets. For fair evaluation, we conduct self-reenactment and cross-identity reenactment experiments on the HDTF [36] and CelebV-HQ [38] dataset, analyzing results both quantitatively and qualitatively.

Training Details. During training, we sample a 14-frame video sequence to ensure temporal consistency within SVD's temporal attention layer, with each frame at a resolution of 512×512. The weights of the Motion Extractor and SVD are kept fixed, while only the Warping Feature Mapper and Internal Feature Modulator are updated. The model is trained for 30,000 iterations with a batch size of 8, using gradient accumulation and gradient checkpointing to manage memory consumption. Optimization is performed using 8-bit Adam [16] with a learning rate of 1×10^{-5} on a single A6000 GPU.

Inference Details. During inference, we input 14-frame sequences into the model with a 6-frame overlap, following [35] to ensure temporal consistency. We use DDIM sampling with 30 steps and a guidance scale of 2.5. On an RTX 4090, generating a 100-frame video takes about 4 minutes.

4.2. Metrics and Comparisons

Evaluation Metrics. To evaluate our method, we conduct self-reenactment and cross-identity reenactment experiments on the HDTF [36] dataset. For self-reenactment, we assess the similarity between the reenacted results and the driving video using Mean Absolute Error (L1), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity (SSIM) [27]. Perceptual error is measured with LPIPS [34], which uses a pre-trained AlexNet [18] model. Additionally, identity preservation (ID) is assessed using an ArcFace-based [5] face recognition model. For cross-identity reenactment, we use ID to compare the reenacted results with the source image. Average Pose Distance (POSE) is computed by detecting facial keypoints in both the reenacted results and the driving video and calculating the keypoint error. To evaluate expression score (EXP), we follow [12] to measure expression similarity between the reenacted results and the driving video. Additionally, we utilize a no-reference video quality assessment model [29] to evaluate the video quality (VQ) of the reenacted results, and adopt Fréchet Inception Distance (FID) [11] and Fréchet Video Distance (FVD) [24] to measure visual fidelity.Beyond these objective metrics, we also conduct a user study to further evaluate the quality of the face reenactment results. The study assesses four key aspects: pose accuracy (POSE-User), expression realism (EXP-User), identity preservation (ID-User), and video quality (VQ-User). Each dimension is rated on a five-point Likert scale: 1 (Poor), 2 (Fair), 3 (Average), 4 (Good), and 5 (Excellent). A total of 11 participants took part in the user study.

Comparative Methods. We conduct a comparative analysis between our method and seven state-of-the-art face reenactment methods: OSFV [26], TPSMM [37], LivePortrait [8], FADM [33], AniPortrait [28], Echomimic [4], and FollowYourEmoji [20]. All methods are evaluated on the HDTF and CelebV-HQ datasets.

4.3. Quantitative Evaluation

We compare our method with seven state-of-the-art face reenactment approaches: OSFV [26], TPSMM [37], LivePortrait [8], FADM [33], AniPortrait [28], Echomimic [4], and FollowYourEmoji [20]. To provide a comprehensive evaluation, we conduct experiments under both self-reenactment and cross-identity reenactment settings.

For self-reenactment, we use the first frame of each video as the source image and the subsequent frames as driving frames. The objective is to generate a reenacted frame that aligns with the corresponding driving frame, allowing us to use the driving frame as ground truth for evaluation. Additionally, to assess the robustness of our method, we conduct cross-identity reenactment, where a face video of one identity is used to drive a face image of another identity.

As shown in Table 1, in the self-reenactment, our method outperforms all other approaches across multiple metrics, including L1, PSNR, SSIM, and LPIPS. Notably, our method achieves an 11.7% lower pixel-wise error (L1) and a 25.5% lower perceptual error (LPIPS) compared to the second-best method. However, in terms of identity preservation (ID), our method is slightly inferior to LivePortrait. For the cross-identity reenactment, our method achieves the best performance in POSE, EXP, FID, and FVD metrics. Specifically, it reduces POSE error by 4.0% and FID by 6.5% compared to the second-best method. Additionally, LivePortrait outperforms our model in terms of the ID metric, primarily due to its use of a large-scale private facial video dataset exceeding 16 million frames and the explicit

Table 1. Quantitative comparison of self-reenactment and cross-identity reenactment on the HDTF video dataset. We evaluate seven state-of-the-art face reenactment methods: OSFV, TPSMM, LivePortrait, FADM, AniPortrait, Echomimic, and FollowYourEmoji. The best scores are highlighted in bold, and the second-best are underlined.

	Self-Reenactment					Cross-identity Reenactment					
Methods	L1↓	PSNR↑	SSIM↑	LPIPS↓	ID↑	POSE↓	EXP↑	ID↑	FID↓	FVD↓	VQ↑
OSFV [26]	0.0303	26.221	0.8476	0.1531	0.8584	3.239	0.6456	<u>0.9127</u>	20.85	169.4	0.6318
TPSMM [37]	<u>0.0256</u>	<u>27.582</u>	<u>0.8692</u>	0.1550	0.8662	<u>2.806</u>	0.6495	0.8962	22.39	175.2	0.6173
LivePortrait [8]	0.0418	22.788	0.7746	0.1222	**0.8967**	6.159	0.6384	**0.9294**	25.07	<u>142.8</u>	0.7739
FADM [33]	0.0359	25.330	0.8348	0.1765	0.8463	18.25	0.6256	0.8994	22.29	164.6	0.6275
AniPortrait [28]	0.0412	21.518	0.7702	0.1680	0.8506	4.375	0.6558	0.8854	18.35	277.5	<u>0.7954</u>
Echomimic [4]	0.0353	24.166	0.8096	<u>0.1020</u>	<u>0.8674</u>	19.29	<u>0.6685</u>	0.7635	19.88	399.4	0.7192
FollowYourEm. [20]	0.0358	23.753	0.7951	0.1077	0.8535	4.341	0.6625	0.8976	<u>15.75</u>	164.1	0.7577
Ours	**0.0226**	**27.708**	**0.8702**	**0.0760**	0.8570	**2.695**	**0.6710**	0.8975	**14.73**	**140.8**	**0.8061**

Table 2. User study comparison of cross-identity reenactment on the HDTF video dataset. We conduct a user study to evaluate the perceptual quality of seven state-of-the-art face reenactment methods. Participants were asked to score the results of each method based on pose accuracy, expression realism, identity preservation, and overall video quality. The highest-rated scores are highlighted in bold, and the second-highest are underlined.

Methods	POSE-User↑	EXP-User↑	ID-User↑	VQ-User↑
OSFV [26]	3.246	<u>3.560</u>	3.812	3.650
TPSMM [37]	2.833	2.626	3.020	3.293
LivePortrait [8]	<u>4.211</u>	3.478	<u>4.118</u>	<u>3.800</u>
FADM [33]	2.833	2.502	2.823	2.414
AniPortrait [28]	3.522	2.375	3.011	2.325
Echomimic [4]	3.190	2.612	2.820	2.375
FollowYourEm. [20]	3.012	2.713	2.491	2.582
Ours	**4.372**	**4.023**	**4.375**	**3.835**

identity supervision provided by the ArcFace [5] face recognition model, which significantly enhances identity preservation. However, except for the ID metric, our method achieves the best performance across all other metrics, demonstrating overall superiority over other state-of-the-art methods.

Since cross-identity reenactment lacks ground-truth data for direct evaluation, relying solely on no-reference image quality metrics may lead to an incomplete assessment. While objective metrics such as PSNR, FID, and LPIPS are commonly reported, the quality of facial generation often depends heavily on human perception—particularly with respect to expression realism, identity consistency, and overall naturalness. A user study is therefore essential, especially in cases where metric differences are small but perceptual differences are significant.

To this end, we conducted a comprehensive user study to evaluate the effectiveness of our method. As shown in Table 2, our approach achieves the highest ratings across all four perceptual dimensions: pose (POSE-User), expression (EXP-User), identity preservation (ID-User), and video quality (VQ-User). Notably, our method outperforms the second-best method by 13% in expression realism and by 6.2% in identity preservation. These results indicate that, although our method may not lead in all objective identity metrics, it consistently produces superior perceptual quality, demonstrating its effectiveness in generating visually compelling reenactment results.

4.4. Qualitative Evaluation

We present a qualitative comparison between our method and recent state-of-the-art face reenactment approaches on the CelebV-HQ [38] dataset, covering both self-reenactment and cross-identity reenactment settings. The comparison includes OSFV [26], TPSMM [37], LivePortrait [8], FADM [33], AniPortrait [28], Echomimic [4], and FollowYourEmoji [20].

4.4.1. Qualitative Evaluation of Self-Reenactment

The qualitative results of the self-reenactment experiment are presented in Fig. 2. As shown, methods such as OSFV and TPSMM produce lower image quality, particularly when there is a large pose discrepancy between the driving and source images (e.g., the third and fourth columns of the first, fourth, and fifth rows). These methods suffer from significant background detail loss, facial distortions, and inconsistencies in identity and appearance. Moreover, they exhibit poor perception of depth cues when the subject interacts with objects—for instance, the cup filled with yellow liquid in the second row and the microphone in the third row are both occluded by the face. In comparison, LivePortrait benefits from a high-quality, large-scale dataset of talking heads, enabling improved image generation. It better preserves background details and facial identity, producing high-resolution results. Nevertheless, LivePortrait still faces challenges with occlusions involving objects such as microphones or cups, indicating limitations in handling complex spatial interactions.

Additionally, FADM yields relatively low-quality reenactment results and exhibits limited awareness of occlusions, as shown in the second and third rows of the sixth column. AniPortrait generates visually appealing outputs but struggles to preserve facial identity, particularly in the first and fourth rows of the seventh column. Echomimic shows weaknesses in object perception, leading to artifacts such as missing hair in the first-row image of the eighth column, the disappearance of a cup in the second row, and the occlusion of a microphone by the face in the third row. FollowYourEmoji, meanwhile, reveals mismatches between the reenacted facial poses and those in the driving video, as seen in the first and second rows of the ninth column. Overall, in the qualitative comparison for self-reenactment, our method demonstrates superior performance across several critical dimensions, including image quality, pose accuracy, expression consistency, and robust handling of occlusions involving surrounding objects. These results highlight our method's consistent advantage over existing state-of-the-art approaches.

Fig. 2. Self-reenactment qualitative comparison with state-of-the-art methods including OSFV [26], TPSMM [37], LivePortrait [8], FADM [33], AniPortrait [28], Echomimic [4], and FollowYourEmoji [20]. The first column shows the source image, the second column presents the driving image (ground truth), and the remaining columns display the reenacted results. Our method delivers more realistic outcomes, especially under challenging conditions such as extreme facial poses.

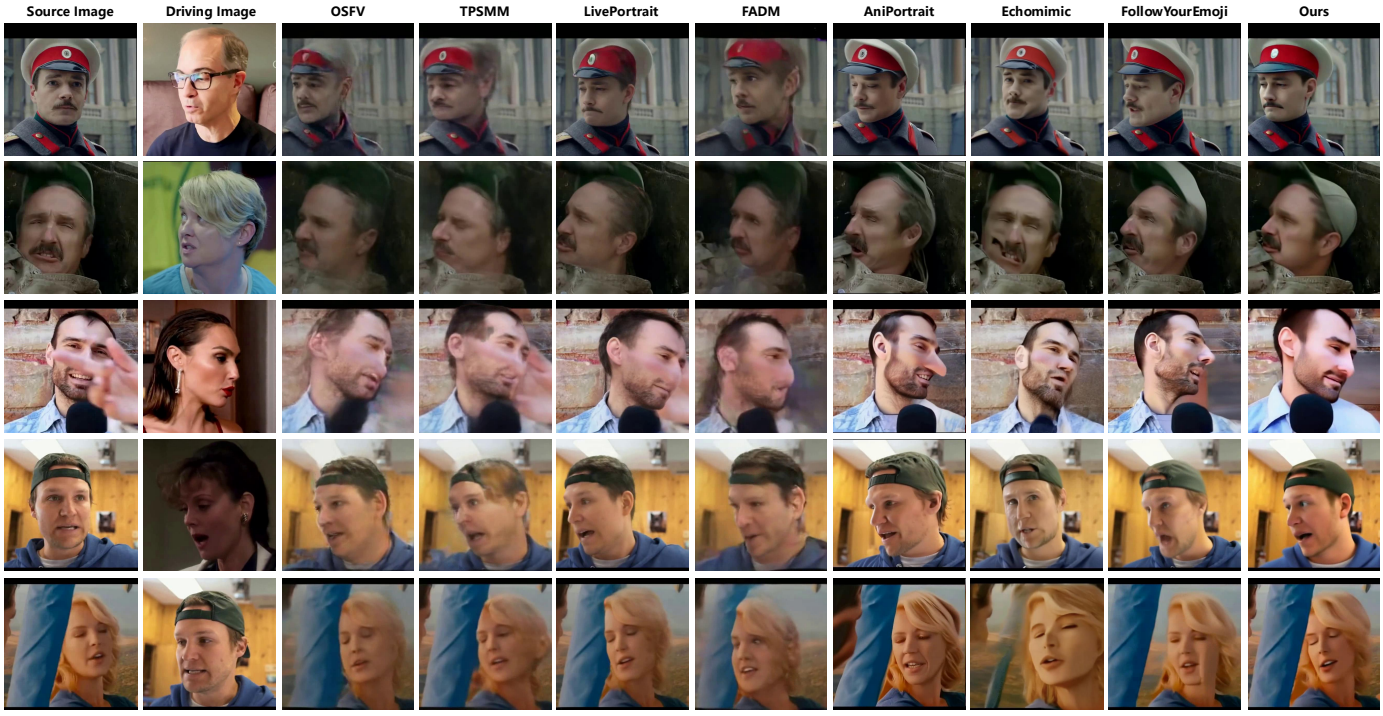

Fig. 3. Cross-identity reenactment qualitative comparison with state-of-the-art methods, including OSFV [26], TPSMM [37], LivePortrait [8], FADM [33], AniPortrait [28], Echomimic [4], and FollowYourEmoji [20]. The first column shows the source image, the second presents the driving image, and the remaining columns display reenacted results. Our method produces more realistic outputs

4.4.2. Qualitative Evaluation of Cross-identity Reenactment

Fig. 3 presents the qualitative results for the cross-identity reenactment experiment, where the source and driving images belong to different individuals. The objective is to match the facial pose in the driving image while maintaining the identity and appearance of the source. We compare our method with recent state-of-the-art face reenactment approaches, demonstrating its effectiveness in preserving identity and achieving accurate pose transfer.

As observed in the figure, OSFV and TPSMM generally produce lower-quality reenacted images and struggle with generating non-facial elements accurately. For instance, in the reenactment results of the first, second, and fourth rows, these methods introduce severe distortions to headwear. Additionally, in the third row, where the source image contains an occlusion (a hand covering part of the face), they fail to reconstruct the occluded regions properly, leading to severe facial distortions, such as a missing realistic nose and mouth. Similarly, in the fifth row, where the source face is partially obscured by a blue pillar, the reenacted face incorrectly appears in front of the pillar, indicating that these methods struggle to handle occlusions and depth relationships between objects and faces effectively.

In contrast, LivePortrait, FADM, AniPortrait, Echomimic, and FollowYourEmoji generally achieve higher overall image quality than OSFV and TPSMM. However, these methods still face challenges in handling headwear distortions. As observed in the first, second, and fourth rows, although they can synthesize hats, the resulting structures often exhibit visible deformation. When dealing with occlusions, these methods show enhanced reconstruction capabilities for missing facial regions—for instance, in the third row, the occluded nose and mouth are plausibly recovered. Nevertheless, this often comes at the cost of identity inconsistency, where the reenacted face diverges noticeably from the source identity. For the blue pillar occlusion in the fifth row, these methods successfully render the face behind the pillar, preserving correct spatial relationships. However, background artifacts remain an issue—for example, in the seventh column of the fifth row, a black object appears unnaturally where the face should be partially visible, reflecting limitations in background-foreground reasoning.

Overall, our method surpasses state-of-the-art approaches in both image quality and identity preservation, particularly in handling challenges such as headwear deformation, facial occlusions (e.g., hands), and maintaining correct depth ordering in occluded scenes.

5. Ablation Study

5.1. Effectiveness of Warping Feature Mapper

We validate the effectiveness of the proposed Warping Feature Mapper (WFM) through an ablation study, as shown in Table 3. "w/" denotes the use of WFM, while "w/o" represents the baseline without it. Our method consistently outperforms the baseline across all five evaluation metrics—L1, PSNR, SSIM, LPIPS, and ID. Specifically, L1 and LPIPS are reduced by 81.9% and 82.2%, respectively, while PSNR, SSIM, and ID are improved by 88.6%, 57.9%, and 60.0%, demonstrating the significant contribution of WFM to the overall performance.

Table 3. Quantitative results of the ablation study on the Warping Feature Mapper (WFM). "w/" indicates the inclusion of WFM, while "w/o" denotes its removal. The best performance is highlighted in bold, and the second-best is underlined for clarity.

Methods	L1↓	PSNR↑	SSIM↑	LPIPS↓	ID↑
w/o WFM	<u>0.1247</u>	<u>14.694</u>	<u>0.5512</u>	<u>0.4270</u>	<u>0.5358</u>
w/ WFM	**0.0226**	**27.708**	**0.8702**	**0.0760**	**0.8570**

Fig. 4. Qualitative results of the ablation study with and without the Warping Feature Mapper (WFM). The first column presents the source image, while the second column shows the ground-truth driving image. The third column illustrates the reenactment results without WFM, and the fourth column shows the results with WFM. Incorporating WFM leads to more realistic and visually faithful reenactments, demonstrating its effectiveness in improving generation quality.

Without WFM, the SVD model loses the motion constraints provided by WFM, effectively degenerating into a random image-to-video generation model. As a result, the generated outputs deviate significantly from the ground truth, which is clearly reflected in the performance drop shown in Table 3.

We further visualize the reenactment results with and without WFM in Fig. 4. Without WFM, the reenacted faces exhibit misalignment in motion compared to the ground truth and suffer from severe background degradation. For example, in the third column of the first row, the mouth remains closed and the head scale is incorrect; in the second row, the head pose is noticeably wrong; and in the third row, both the facial pose and scale are inaccurate, accompanied by visible background changes. In contrast, with WFM, the reenactment results show more accurate facial pose and scale, as well as better background preservation. This improvement is mainly attributed to WFM's ability not only to provide motion information to the I2V model but also to spatially constrain the generation process, ensuring that the synthesized faces match the motion patterns of the driving video.

5.2. Effectiveness of Rectified Guidance

We validate the effectiveness of the proposed rectified guidance, as shown in Table 4. During training, we incorporate rec-

Table 4. Quantitative comparison of the ablation study with rectified guidance used during training. "w/" denotes the use of rectified guidance, while "w/o" indicates its absence. The best scores are highlighted in bold, and the second-best are underlined.

Methods	L1↓	PSNR↑	SSIM↑	LPIPS↓
w/o rectified guidance	<u>0.0343</u>	<u>25.527</u>	<u>0.8295</u>	<u>0.0967</u>
w/ rectified guidance	**0.0226**	**27.708**	**0.8702**	**0.0760**

Fig. 5. Qualitative comparison of the ablation study with and without rectified guidance during training. The first column shows the source image, and the second column presents the ground truth (driving image). The third column displays the reenactment results without rectified guidance, while the fourth column shows the results with rectified guidance. Our method produces more realistic and visually faithful reenactments when rectified guidance is used.

tified guidance into the loss function, denoted as "w/" in the table, while "w/o" indicates training without rectified guidance. Across all four metrics—L1, PSNR, SSIM, and LPIPS—our method achieves consistent improvements. Notably, L1 and LPIPS are reduced by 34.1% and 21.4%, respectively, compared to the model trained without rectified guidance.

We further visualize the reenactment results with and without rectified guidance in Fig. 5. When rectified guidance is not used, the reenacted faces exhibit noticeable color deviations. For example, in the third column of the first row, the lips appear overly dark; in the second row, the skin tone deviates significantly; and in the third row, the woman's hair appears darker than in the source image. In contrast, with rectified guidance, the reenacted results exhibit more natural and faithful appearance, including improved consistency in skin tone, hair color, and lip color, making the overall result more realistic.

5.3. Effectiveness of Classifier-free Guidance Strength

Our method leverages the prior knowledge of a pre-trained SVD model to recover the warped regions of the source image, thereby achieving face reenactment. From the perspective of conditional image generation, the warped image can also be viewed as a condition guiding the SVD model. Therefore, the classifier-free guidance strength (denoted as w) directly influences the final reenactment performance.

To validate this, we conduct a comparative study across different w values. As shown in Table 5, when the guidance strength exceeds 1, the performance metrics—L1, PSNR, SSIM, and LPIPS—consistently degrade as w increases. The best results are achieved at $w = 2$. When $w = 1$, the metrics are slightly worse than those at $w = 2$, but still relatively close. We also present qualitative comparisons of facial reenactment under different w values in Fig. 6. When $w = 1$, significant detail loss is observed in the reenacted faces, leading to a hazy or blurred appearance. In contrast, when w exceeds 2, facial details become overly pronounced. For example, in the first and second rows, the age spots on the faces become increasingly blocky and prominent, deviating noticeably from the ground truth, which results in a decline in performance metrics. At $w = 2$, the reenacted faces are visually closest to the ground truth. Overall, classifier-free guidance strength significantly affects the detail level of reenacted faces: higher values enhance detail but may introduce artifacts and distortions if the strength surpasses a certain threshold.

Table 5. Quantitative comparison of the ablation study on classifier-free guidance strength. Different values of w represent varying guidance strengths, with larger w indicating stronger classifier-free guidance. The best scores are highlighted in bold, and the second-best scores are underlined.

Metrics	$w = 1$	$w = 2$	$w = 3$	$w = 4$	$w = 5$	$w = 6$
L1↓	<u>0.0235</u>	**0.0226**	0.0289	0.0306	0.0354	0.0471
PSNR↑	<u>27.447</u>	**27.708**	26.167	25.854	24.883	22.786
SSIM↑	<u>0.8702</u>	**0.8730**	0.8325	0.8222	0.7853	0.7084
LPIPS↓	<u>0.1089</u>	**0.0760**	0.1697	0.2271	0.3395	0.4977

Fig. 6. Qualitative comparison of the ablation study on classifier-free guidance strength. The first column shows the source image, and the second column presents the ground truth (driving image). Columns three to six display the reenacted face results under different guidance strengths. As the guidance strength increases (i.e., larger w), facial textures become more pronounced. However, overly strong guidance leads to excessively enhanced high-frequency details, resulting in blocky artifacts on the face.

6. Conclusion

In this work, we presented FRVD, a novel framework for high-fidelity face reenactment under large pose variations. By leveraging implicit facial keypoints to model fine-grained motion and employing a warping module for motion alignment, our method effectively transfers facial dynamics from the driving video to the source image. To address the degradation introduced by warping, we proposed a Warping Feature Mapper (WFM) that maps the warped source image into the motion-aware latent space of a pretrained image-to-video model. This

design enables perceptually accurate reconstruction of facial details and ensures temporal consistency across frames. Extensive experiments demonstrate that FRVD outperforms state-of-the-art methods in terms of pose accuracy, identity preservation, and visual quality, especially in scenarios with extreme head movements. Our approach highlights the potential of combining implicit motion representations with pretrained video priors for robust and expressive face reenactment.

While FRVD demonstrates strong performance in high-fidelity face reenactment under large pose variations, we acknowledge certain limitations. Specifically, the current inference speed (approximately 4 minutes per 100 frames) may hinder real-time applications. In addition, although FRVD effectively handles significant pose changes, further evaluation on extreme non-frontal views (e.g., back-facing poses) is warranted. In future work, we plan to address these issues by exploring model distillation techniques to improve inference efficiency and by extending the evaluation to a broader spectrum of head orientations.

References

[1] 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] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. Diffusionact: Controllable diffusion autoencoder for one-shot face reenactment. *arXiv preprint arXiv:2403.17217*, 2024.

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[4] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024.

[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.

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[7] Ian 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.

[8] 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.

[9] 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.

[10] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*.

[11] 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.

[12] 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.

[13] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024.

[14] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

[15] 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.

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

[17] Idan Kligvasser, Regev Cohen, George Leifman, Ehud Rivlin, and Michael Elad. Anchored diffusion for video face reenactment. *arXiv preprint arXiv:2407.15153*, 2024.

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

[19] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[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. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024.

[21] 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.

[22] 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.

[23] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020.

[24] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[25] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[26] 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.

[27] 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.

[28] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animations, 2024.

[29] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment, 2022.

[30] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022.

[31] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[32] 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.

[33] Bohan Zeng, Xuhui Liu, Sicheng Gao, Boyu Liu, Hong Li, Jianzhuang Liu, and Baochang Zhang. Face animation with an attribute-guided diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 628–637, 2023.

[34] 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.

[35] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024.

[36] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3661–3670, 2021.

[37] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022.

[38] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022.