RealPortrait: Realistic Portrait Animation with Diffusion Transformers

Zejun Yang*, **Huawei Wei***, **Zhisheng Wang**

Tencent
{zejunyang, huaweiwei, plorywang}@tencent.com

Abstract

We introduce RealPortrait, a framework based on Diffusion Transformers (DiT), designed to generate highly expressive and visually appealing portrait animations. Given a static portrait image, our method can transfer complex facial expressions and head pose movements extracted from a driving video onto the portrait, transforming it into a lifelike video. Specifically, we exploit the robust spatial-temporal modeling capabilities of DiT, enabling the generation of portrait videos that maintain high-fidelity visual details and ensure temporal coherence. In contrast to conventional image-to-video generation frameworks that necessitate a separate reference network, we incorporate an efficient reference attention within the DiT backbone, thereby obviating the computational overhead and achieving superior reference appearance preservation. Concurrently, we integrate a parallel ControlNet to precisely regulate intricate facial expressions and head poses. Diverging from prior methods that utilize explicit sparse motion representations, such as facial landmarks or 3DMM coefficients, we adopt a dense implicit motion representation as the control guidance. This implicit motion representation excels in capturing nuanced emotional facial expressions and subtle non-rigid dynamics of the lips. To further enhance the generalization capability of the model, we augment the training dataset by incorporating a substantial volume of facial image data through random crop augmentation. This strategy ensures the model's robustness across a wide variety of facial appearances and expressions. Empirical evaluations demonstrate that RealPortrait excels in generating portrait animations with highly-realistic quality and exceptional temporal coherence in appearance retention.

Introduction

Portrait animation involves the task of transferring motion and facial expressions from a driving video to a target portrait. This technology has a wide range of potential applications, such as creating digital avatars for video conferencing, visual effects, and interactive digital agents. However, existing methods often struggle to produce realistic and natural faces, frequently resulting in artifacts like blurring and flickering. This paper aims to generate highly realistic and vivid portrait animations that are indistinguishable from real ones, thereby enhancing the quality and applicability of portrait animation.

Current methods can be broadly categorized into two types based on their motion representation techniques. The first type employs explicit and sparse motion representations, such as facial landmarks (Chang et al. 2023a; Wei, Yang, and Wang 2024) or 3D Morphable Model (3DMM) coefficients (Ren et al. 2021), to drive portraits. These representations, however, exhibit notable limitations in capturing the intricate and non-rigid motions of the human face, often resulting in animations that appear stiff and unnatural. The second type of methods utilizes implicit warping fields (Wang, Mallya, and Liu 2021; Siarohin et al. 2019; Zhao and Zhang 2022; Hong et al. 2022; Guo et al. 2024) to represent facial motions. This data-driven representation offers a dense and detailed depiction of motion, enabling the generation of more expressive and nuanced portrait animations. However, these methods typically rely on simple GAN-based decoders for video rendering, which leads to animations with suboptimal resolution and perceptual quality.

Recently, diffusion models have demonstrated unprecedented diversity and stability in image generation (Rombach et al. 2022; Zeng et al. 2023; Xie et al. 2024; Chang et al. 2023a; Hu 2024; Tian et al. 2024), attributed to increased model capacity and large-scale pretraining datasets. The advent of these models has addressed the limitations of GAN-based renderers in producing high-quality animations.

In this paper, we propose RealPortrait, an innovative portrait animation generation framework that combines the advantages of dense motion representations and diffusion models. Specifically, we leverage Diffusion Transformers (DiT) (Peebles and Xie 2023; Zheng et al. 2024; Chen et al. 2023) for their superior spatial-temporal modeling capabilities and enhanced scalability. Compared to Unet-based diffusion models such as Stable Diffusion (SD), DiT excels at generating highly realistic and temporally coherent portrait animations. In our framework, DiT serves as the backbone responsible for video rendering. Additionally, we integrate a parallel ControlNet (Zhang, Rao, and Agrawala 2023; Chen et al. 2024), which is responsible for receiving the driving motions to guide the generation of portrait animations. We employ a dense implicit motion representation, akin to warping fields, as the control guidance. However, instead of using traditional warping fields, we convert this relative motion

*These authors contributed equally.

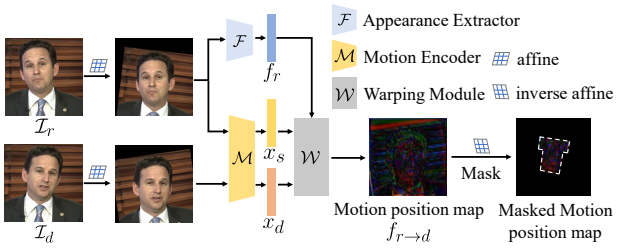

Figure 1: Process to obtain the motion position maps.

field into an absolute motion position map. This modified motion representation proves to be highly suitable for the ControlNet framework. By providing direct pixel-level spatial position guidance, motion position maps enable us to effectively capture subtle emotional facial expressions and the nuanced non-rigid dynamics of the lips, ensuring the naturalness and expressiveness of facial movements.

Furthermore, we optimize the framework by eliminating the overhead of the separate reference network , which is commonly used in SD-based image-to-video generation frameworks (Xie et al. 2024; Chang et al. 2023a; Hu 2024; Tian et al. 2024; Wei, Yang, and Wang 2024). Instead, we employ an efficient reference attention mechanism to effectively capture and preserve reference appearance features. The core of this mechanism lies in allowing the reference image to share the DiT backbone, followed by the use of a modified spatial attention to seamlessly integrate the appearance information into the video generation process. This approach not only reduces the network burden but also enhances visual consistency and fidelity in the generated videos.

Given the limited availability of high-quality portrait video data, the generalization capability of the model may be constrained. To address this issue, we expand the training dataset by incorporating a large volume of facial image data and applying random crop augmentation. This random augmentation simulates static single images as motion-varied data, significantly enhancing the model's ability to generalize to unseen appearances.

Empirical evaluations demonstrate that RealPortrait excels in generating portrait animations with highly-realistic quality and exceptional temporal coherence.

Related Work

Motion Representations for Portrait Animation

Motion representation is a critical aspect of portrait animation, as it directly influences the realism and expressiveness of the generated animations. Various approaches have been proposed in the literature to capture and represent motion for this purpose. A common approach is to represent facial movements using 2D facial landmarks (Chang et al. 2023a; Wei, Yang, and Wang 2024). This type of representation is easy to obtain and highly interpretable. However, its sparsity limits its precision. A direct extension is to represent motion using dense 3D facial landmarks or another parameterized method, such as 3D Morphable Model (3DMM) parameters (Chu et al. 2024; Ren et al. 2021). While these methods overcome the limitation of sparsity, the current techniques for extracting 3D facial landmarks or 3DMM parameters still exhibit significant errors, making them insufficient for capturing fine facial movements. Another category of methods leverages the idea of disentangling appearance and motion, learning implicit motion representations from large datasets (Deng, Wang, and Wang 2024; Drobyshev et al. 2024). These methods have achieved notable success in the domain of talking heads. Additionally, there is a class of methods that use implicit keypoints (Siarohin et al. 2019; Zhao and Zhang 2022; Hong et al. 2022; Guo et al. 2024; Hong and Xu 2023; Wang, Mallya, and Liu 2021) to represent motion, which are then converted into deformation fields to warp the face. Due to their robustness, these methods are increasingly becoming mainstream.

Non-diffusion-based Portrait Animation

Non-diffusion methods typically follow a framework that involves two main steps: first, obtaining a reenacted facial representation, and second, rendering it into a facial image. The initial step generally employs various representations, such as deformed latents (Siarohin et al. 2019; Zhao and Zhang 2022; Hong et al. 2022; Guo et al. 2024), triplane features (Chu et al. 2024), NeRF representations (Ye et al. 2023), or volumes (Deng, Wang, and Wang 2024; Drobyshev et al. 2024), to store the reenacted face. The subsequent step usually utilizes a GAN-based decoder or a differentiable rendering network to convert the facial representation into an RGB image. However, due to the relatively weak performance of these decoders or renderers, the resulting facial images often exhibit poor quality and are generally confined to the head region, with subpar rendering of shoulders and hair.

Diffusion-based Portrait Animation

Diffusion models (Song, Meng, and Ermon 2020; Song et al. 2020) have recently shown remarkable capabilities in image editing and generation, with latent diffusion models (Rombach et al. 2022) pushing these boundaries even further. These advancements have also been applied to portrait animation and human body animation generation (Xie et al. 2024; Hu 2024; Chang et al. 2023a; Tian et al. 2024; Wei, Yang, and Wang 2024; Chang et al. 2023b; Xu et al. 2024). Typically, these methods utilize a Unet-based Stable Diffusion (SD) network as the backbone, a plug-and-play reference network to incorporate appearance information, and a ControlNet (Zhang, Rao, and Agrawala 2023) to provide motion guidance. However, these approaches encounter two significant challenges. Firstly, they often depend on explicit landmarks for motion guidance, which can be inaccurate and insufficient for producing vivid and realistic videos. Secondly, the computational demands of the network, coupled with limited training resources, make it difficult to scale up the framework, resulting in videos that lack coherence and quality. To address these challenges, this paper introduces the use of implicit dense motion representations (Wang, Mallya, and Liu 2021; Guo et al. 2024) to guide animation generation, enabling fine-grained facial motion synthesis. Additionally, we employ DiT (Zheng et al. 2024; Peebles and Xie 2023) as the network backbone, which excels

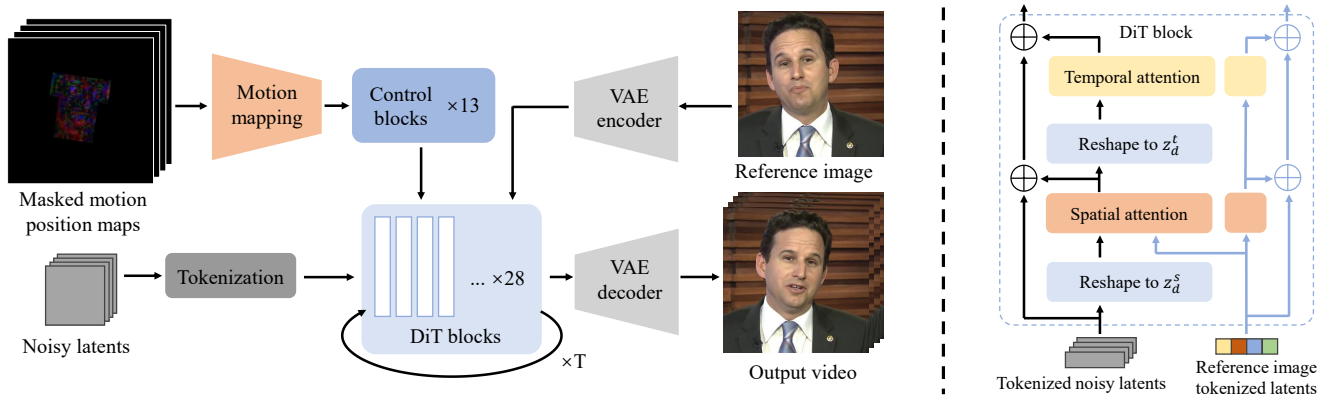

Figure 2: **Left**: An overview of our framework. The DiT backbone is responsible for video rendering, while the ControNet handles the motion position maps. **Right**: Details of our DiT block, illustrating how reference attention is executed.

at modeling spatial-temporal relationships and is more scalable. This approach facilitates the creation of more realistic and lifelike portrait animations.

Method

In this section, we begin with a brief introduction to Face Vid2vid (Wang, Mallya, and Liu 2021), which serves as a prerequisite for extracting the motion representation. Next, we propose the use of a motion position map as a control signal to guide the animation generation. Following this, we provide a detailed explanation of the DiT (Peebles and Xie 2023) and ControlNet (Zhang, Rao, and Agrawala 2023) frameworks for generating portrait videos. Finally, we discuss how leveraging a large dataset of images enhances the model's generalization capabilities.

Preliminary

Face Vid2vid is a facial reenactment method. Given a reference face image $\mathcal{I}_r$ and a driving image $\mathcal{I}_d$, it first employs an appearance extractor $\mathcal{F}$ to extract appearance features f_r from the reference image. Subsequently, a motion encoder $\mathcal{M}$ is used to extract implicit keypoints from both the reference and driving image, denoted as x_r and x_d, which are then utilized to generate a deformation field. Following this, a warping module $\mathcal{W}$ use the deformation field to warp the appearance features. Finally, a generation module decodes the deformed feature $f_{r\to d}$ into a face image that retains the appearance of the reference image while adopting the motion of the driving image.

Motion Position Map

A straightforward idea is to use implicit keypoints x_d to guide the generation of portrait animations. However, we find that due to the sparsity of implicit keypoints and their lack of fine-grained spatial correspondence with the image, they can only provide relatively coarse semantic guidance. This level of guidance is insufficient for tasks such as video generation, which demand a high degree of spatial motion coherence. We will discuss this further in the experimental section. To address this issue, we employ the deformed feature $f_{s\to d}$ as the motion representation, which maintain a pixel-level spatial correspondence with the image. In this paper, we refer to this representation as the *motion position map*.

It is worthy to note that the input portrait image for Face Vid2vid undergo cropping and alignment operations. These steps ensure that the motion information extracted by the network excludes any background information unrelated to the face. However, these operations result in generated portraits that only contain a compact facial region, making it impossible to capture details such as hair and shoulder movement. This limitation reduces the realism of the portrait videos and restricts applications that require a more comprehensive view. To overcome this limitation, we transform the motion position map back to the original image's size and position. For regions outside the cropped area, we pad with 0. By using this full-image motion representation, we can generate half-body portrait animations, significantly enhancing the expressiveness and vividness of the animations. To enable the model to focus on learning the motion of the facial region, we mask out all parts except for the facial region. Our processing steps are detailed in Figure 1.

Conditional Spatial-Temporal Modeling

Our framework is divided into three components. First, we employ DiT as the backbone, leveraging its robust spatial-temporal modeling capability for video rendering. Second, the reference attention embedded within the backbone is used to incorporate appearance information into the temporal latents, facilitating the generation of background and portrait appearance in the generated video. Additionally, a parallel ControlNet is utilized to handle motion guidance, specifically responsible for generating head movements and facial expressions. Details of our framework are shown in Figure 2.

DiT backbone We follow the network architecture of Open-Sora (Zheng et al. 2024), which consists of 28 transformer blocks (Vaswani 2017), each containing a spatial attention layer and a temporal attention layer. The former is dedicated to extracting spatial information among tokens that share the same temporal index, while the latter gathers temporal information across different time steps.

Given a driving video clip $F_d = \{f_d^1, f_d^2, \ldots, f_d^L\}$, we encode it into the latent space $V_d = \{v_d^1, v_d^2, \ldots, v_d^L\}$ using a VAE (Rombach et al. 2022), where each $v_d^i \in \mathbb{R}^{H \times W \times C}$. We then translate V_d into a sequence of tokens, denoted as $\mathbf{z}_d \in \mathbb{R}^{(L \times n_h \times n_w) \times c}$. Here L, H, W, and C represent the number of video frames, the height, width, and channel of video frames in the latent space, respectively. The total number of tokens within a video clip in the latent space is $L \times n_h \times n_w$, and c represents the dimension of each token. We reshape $\mathbf{z}_d$ into $\mathbf{z}_d^s \in \mathbb{R}^{L \times S \times c}$ as the input of the spatial attention layer. Here, $S = n_h \times n_w$ denotes the token count of each temporal index. Subsequently, $\mathbf{z}_d^s$ is reshaped into $\mathbf{z}_d^t \in \mathbb{R}^{S \times L \times c}$ to serve as the input for the temporal attention layer. The attention operations are then applied as follows:

$$\text{SpaAttn}(z_d^s) = \text{Attention}(Q_d^s, K_d^s, V_d^s); \quad (1)$$

$$\text{TempAttn}(z_d^t) = \text{Attention}(Q_d^t, K_d^t, V_d^t). \quad (2)$$

Here, $\text{SpaAttn}(z_d^s)$ and $\text{TempAttn}(z_d^t)$ compute the attention scores using:

$$\begin{aligned} Q_d^{\{s,t\}} &= W_Q^{\{s,t\}} \cdot z_d^{\{s,t\}}, & W_Q^{\{s,t\}} &\in \mathbb{R}^{c \times c} \\ K_d^{\{s,t\}} &= W_K^{\{s,t\}} \cdot z_d^{\{s,t\}}, & W_K^{\{s,t\}} &\in \mathbb{R}^{c \times c} \\ V_d^{\{s,t\}} &= W_V^{\{s,t\}} \cdot z_d^{\{s,t\}}, & W_V^{\{s,t\}} &\in \mathbb{R}^{c \times c} \end{aligned} \quad (3)$$

Reference attention To integrate appearance information from the reference image, recent works typically employ a separate reference network (Hu 2024), which shares the same architecture as the backbone, to extract identity attributes and background context from the reference image and incorporate them into the backbone using an attention layer. We have verified that using a separate reference network is redundant, as directly sharing weights with the backbone can equally capture the appearance information.

Specifically, we obtain the latent encoding v_r of the reference image f_r through the VAE, and tokenize it into z_r. It is important to note that we do not add noise to z_r in the diffusion training process, thereby fully preserving its details. z_r and z_d are processed through the shared backbone. Within the network, the spatial attention and temporal attention are modified to treat z_r and z_d differently. We adjust the spatial attention such that for each frame z_{d_i}, its key and value are concatenated with the key and value of z_r. This allows the appearance information to be incorporated. To ensure that the reference features remain in the same feature space as z_d for following operations, z_r also undergoes a separate spatial attention process. The modification is depicted as follows:

$$\text{SpaAttn}(z_{d_i}^s) = \text{Attention}(Q_{d_i}^s, K_{d_i}^s \odot K_r^s, V_{d_i}^s \odot V_r^s) \quad (4)$$

$$\text{SpaAttn}(z_r^s) = \text{Attention}(Q_r^s, K_r^s, V_r^s) \quad (5)$$

Here, $\odot$ represents concat operation, $\text{SpaAttn}(z_r^s)$ compute the attention scores using:

$$Q_r^s = W_Q^s \cdot z_r^s, \quad K_r^s = W_K^s \cdot z_r^s, \quad V_r^s = W_V^s \cdot z_r^s,$$

Regarding the temporal attention, since there is no temporal relationship between the reference image and the driving video clip, we still use Equation 2 to fusion temporal information among z_d. Similarly, to ensure that z_r remains in the same feature space, we also perform a separate temporal attention on it.

$$\text{TempAttn}(z_r^t) = \text{Attention}(Q_r^t, K_r^t, V_r^t). \quad (6)$$

The essence of the reference attention mechanism lies in sharing a common backbone between the reference image and the driving video, ensuring their latent representations remain within the same semantic space. By employing a modified spatial attention layer, appearance information from the reference image is seamlessly integrated into the latents of the driving video. This method not only alleviates the burden of model size but also retains detailed appearance information effectively.

Motion ControlNet Given a clip of motion position maps $F_{r \to d} = \{f_{r \to d}^1, f_{r \to d}^2, \ldots, f_{r \to d}^L\}$, we first transform it using a motion mapping module composed of a few simple convolutional layers, resulting in a latent tensor with the same dimensions as V_d. We denote this tensor as $V_{r \to d} = \{v_{r \to d}^1, v_{r \to d}^2, \ldots, v_{r \to d}^L\}$, where each $v_{r \to d}^i \in \mathbb{R}^{H \times W \times C}$. Then, $V_{r \to d}$ is fed into the ControlNet for subsequent processing. We adopt the ControlNet structure from PixArt-δ (Chen et al. 2024), which consists of 13 DiT blocks. The output of the i-th ControlNet block is added to the input of the $(i+1)$-th block of the backbone. This combined input is then processed by the backbone. Unlike PixArt-δ, where the backbone is frozen during the training of the ControlNet, we train both the ControlNet and the backbone simultaneously.

Due to the density of motion position maps and their precise pixel-level spatial correspondence with the driving video, these maps have a strong guiding capability. As a result, the generated video closely matches the driving video in terms of facial expressions and head movements. Furthermore, our motion position maps mask the area outside the facial region, offering two key benefits: first, it allows the network to concentrate more on generating facial movements; second, it enables the network to 'imagine' the motion of areas outside the face based on the appearance provided by the reference image and the physical laws learned from large datasets. This simple masking mechanism activates the "associative" ability of the diffusion model, enabling the generation of videos that exhibit realistic hair movement and shoulder dynamics.

Image-Video Mixed Training

Our training process is divided into two stages. In the first stage, we adopt a single-frame training mode, utilizing both image and video data. For the video data, we randomly sample two frames: one as the reference image and the other as the driving image. For the image data, directly using the same image as both the reference and driving images would allow the network to find the shortcut, meaning it could easily learn to replicate the reference image as the output. To address this issue, we obtain the driving image by applying random cropping to the original image. Although random cropping does not alter the facial expression or pose, the resulting changes in scale and translation are sufficient to prevent the network from taking the shortcut. We have observed

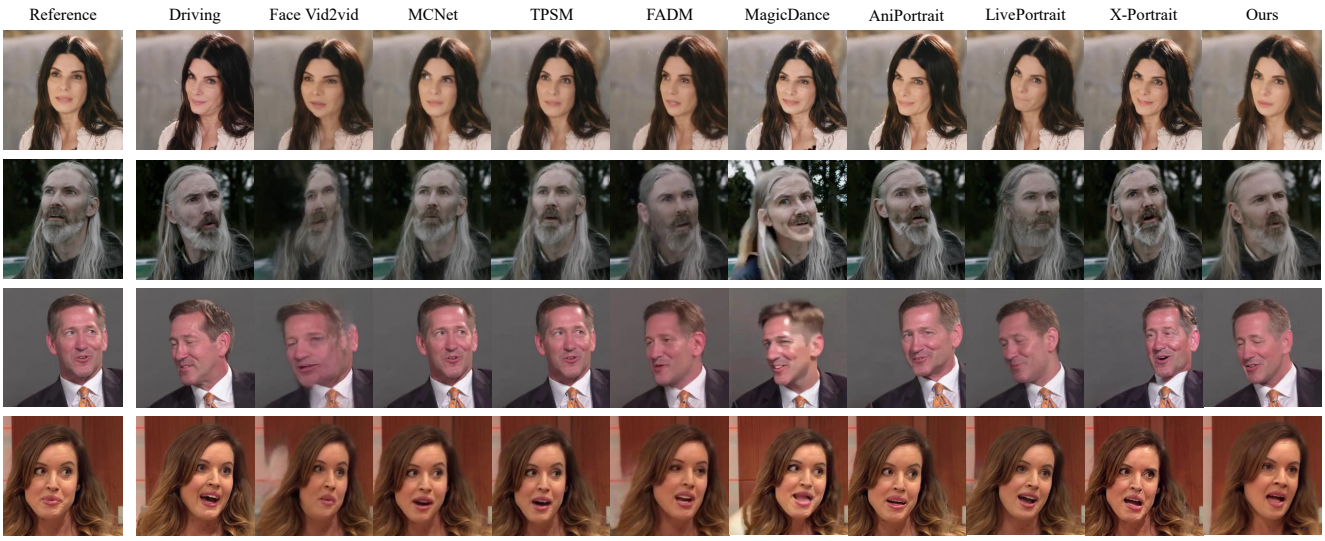

Figure 3: **Qualitative comparisons of self-reenactment.** Our method exhibits stronger appearance preservation, even under significant movements. It achieves more accurate consistency in facial expressions and head poses compared to other methods.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Face Vid2vid	30.653	0.773	0.093	83.507
MCNet	31.737	0.806	0.099	87.304
TPSM	31.198	0.786	0.095	82.353
LivePortrait	32.154	0.821	0.071	57.896
FADM	31.586	0.771	0.103	96.765
MagicDance	31.657	0.773	0.106	97.304
AniPortrait	31.865	0.725	0.093	59.587
X-Portrait	31.962	0.826	0.074	56.547
Ours	**32.213**	**0.831**	**0.064**	**55.346**

Table 1: Quantitative comparisons of self-reenactment.

that by leveraging a large amount of image data, the network's generalization ability is significantly enhanced, resulting in effective driving performance for faces with various appearances.

In the second stage, we transition to a multi-frame training mode, with a primary focus on temporal consistency. During this stage, we exclusively utilize video data, randomly selecting a continuous sequence of L frames from the video as a training clip, where L is set to 16 in our experiments. To support long video inference and video continuation, we draw inspiration from Open-Sora by randomly masking the initial and final frames of the clip to serve as conditional frames. For these conditional frames, no noise is added during the diffusion training process, and the diffusion time embedding is set to 0. During inference, the final frame of the previously inferred clip can be used as the conditional frame for the current clip, thereby enabling long duration inference and video continuation.

Experiments

Implementation Details

Datasets Our data comprises both image and video datasets. The former includes several publicly available face datasets, such as VGGFace2 (Cao et al. 2018), CelebA (Liu et al. 2015), and FFHQ (Karras, Laine, and Aila 2019). We filter out low-resolution data, resulting in a final count of 1M images. These datasets encompass faces of various ethnicities, multiple angles, and diverse lighting conditions, which significantly enhance the model's robustness to diverse appearances. The video datasets include several face video datasets, such as CelebV-HQ (Zhu et al. 2022), TalkingHead-1KH (Wang, Mallya, and Liu 2021), HDTF (Zhang et al. 2021), MEAD (Wang et al. 2020), and VFHQ (Xie et al. 2022). We filter out cases with low resolution, facial occlusion, excessive background motion, and extremely low lighting conditions. This results in a final set of 160K video clips, each ranging from 3 seconds to 5 minutes in length, with a cumulative duration of nearly 1K hours. For evaluation, we split 100 clips from VFHQ and CelebV-HQ to form a test set. Additionally, we collect 200 portraits with diverse appearances as the reference images, which encompass a wide range of characteristics, including different ages, skin tones, and styles ranging from realistic to anime.

Training and inference We utilize the STDiT2 structure from Open-Sora as our architecture. The backbone consists of 28 blocks, while the ControlNet comprises 13 blocks. The motion mapping network is composed of 3 convolutional layers with a stride of 2, supporting 8x downsampling. All samples are cropped to a resolution of 512x512, and we do not perform alignment on the faces. We use the VAE from SD1.5 (Rombach et al. 2022) to encode images into the latent space, where the latent features have a resolution of 64x64. We train the model in two stages using 4

Figure 4: **Qualitative comparisons of cross-reenactment.** Our method excels in facial expression and head pose transfer. Even with significant movements and exaggerated expressions, it surpasses comparative methods in both realism and accuracy.

A100 GPUs. In the first stage, the batch size is 32, and in the second stage, the batch size is 4. The first stage is trained for 20K steps, and the second stage is trained for 30K steps, with the total training time approaching one week. During the testing phase, we provide a portrait photo as the reference image and use a video of a different ID as the driving video. We employ Face Vid2vid to obtain the motion position map sequence as the driving motion. Our denoising process consists of 50 steps. We utilize the video continuation technique introduced above to support long video inference.

Comparisons and Evaluations

We compare our approach with several recent face reenactment methods, including non-diffusion methods such as Face Vid2vid (Wang, Mallya, and Liu 2021), MCNet (Hong and Xu 2023), TPSM (Zhao and Zhang 2022), and Liveportrait (Guo et al. 2024), as well as diffusion-based methods like FADM (Zeng et al. 2023), MagicDance (Chang et al. 2023a), AniPortrait (Wei, Yang, and Wang 2024) and X-Portrait (Xie et al. 2024). The outputs of all compared methods are resized to a resolution of 256x256, as most of them produce outputs at this resolution.

Self Reenactment Given a test video, we use the first frame as the reference image, while the remaining frames serve as the driving sequence and also as the ground truth. We present the qualitative results in the Figure 3. As shown, compared to non-diffusion methods, such as Face Vid2vid, MCNet, and TPSM, our results exhibit better perceptual quality, without blurriness, artifacts, or severe facial distortions. Although LivePortrait improves image quality and driving accuracy with an enhanced network and larger dataset, it cannot drive regions beyond the face, significantly reducing its realism. Compared to methods that use facial landmarks as the motion representation, such as MagicDance and AniPortrait, our method demonstrates more accurate expression reenactment, with more precise lip and head pose movements, while better preserving the subject's identity. Additionally, compared to SD-based diffusion methods, our DiT-based approach shows stronger spatial-temporal modeling capabilities. For instance, in the second row, when the subject makes large movements, our method captures the movements of the hair and shoulders in a manner more consistent with physical realism.

Method	**ID Sim** ↑	**Exp Acc** ↓	**Pose Acc** ↓
Face Vid2vid	0.592	0.785	0.058
MCNet	0.367	0.723	0.039
TPSM	0.522	0.754	0.043
LivePortrait	0.664	0.685	0.034
FADM	0.604	0.799	0.053
MagicDance	0.613	0.813	0.051
AniPortrait	0.623	0.805	0.041
X-Portrait	0.654	0.695	0.036
Ours	**0.673**	**0.664**	**0.031**

Table 2: Quantitative comparisons of cross-reenactment.

For quantitative comparison, we adopt Peak PSNR, SSIM (Assessment 2004), LPIPS (Zhang et al. 2018) and FID (Heusel et al. 2017) to measure the discrepancy between the network output and the ground truth. Table 1 presents our comparison results, demonstrating that our method achieves superior image quality and realism compared to both non-diffusion and diffusion methods.

Cross Reenactment Figure 4 presents the qualitative comparison results of cross reenactment. It can be observed that methods such as Face Vid2vid, MCNet, and TPSM exhibit noticeable facial distortion artifacts. This issue is also present in MagicDance and AniPortrait, which is attributed

Method	ID Sim ↑	Exp Acc ↓	Pose Acc ↓
1D motion	0.664	0.693	0.053
2D motion	**0.673**	**0.664**	**0.031**

Table 3: Ablation of motion position maps.

to their use of either relatively weak decoders for image rendering or sparse and inaccurate facial keypoints for driving the portrait. Consequently, their appearance preservation and identity retention capabilities are inferior to ours. LivePortrait utilizes the same dense motion to drive the portrait, resulting in better motion accuracy, such as subtle lip movements and eyelid blinks. However, thanks to the stronger spatial-temporal modeling capabilities of the DiT network, our method demonstrates greater physical realism in large movements. For instance, in the case shown in the first row, the shoulders in LivePortrait and X-Portrait remain static, whereas in our method, they exhibit consistent movements in response to head translation. This indicates that the portrait videos generated by our method possess greater vividness and realism.

For quantitative comparison, we adopt the metrics proposed by X-Portrait to measure the accuracy of cross reenactment, which include ID similarity, expression accuracy, and head pose accuracy. We employ ArcFace (Deng et al. 2019) to compute the cosine similarity between the facial features of the reference image and the generated image to characterize ID similarity. For motion accuracy, which encompasses expression and head pose accuracy, we utilize a SOTA facial expression and pose recognition method (Retsinas et al. 2024) to extract the expression and pose parameters from both the generated and driving images, and we characterize the accuracy using the L1 distance between them. As shown in Table 2, our method significantly outperforms Face Vid2vid, MCNet, TPSM, MagicDance, and AniPortrait. Compared to the current state-of-the-art methods, LivePortrait and X-Portrait, we achieve a slight improvement. This demonstrates that our implicit dense motion representation can accurately capture facial expressions and head movements, and our robust DiT network effectively assimilates the appearance of the reference image while preserving its identity.

Ablation Studies

Ablation of Motion Position Map

Here, we discuss the rationale for using the 2D motion position map $f_{r\to d}$ instead of implicit keypoints x_d as the motion guidance. The method of incorporating the motion position map using ControlNet has been detailed above. Implicit keypoints, which can be considered as a 1D vector, specifically $x_d \in \mathbb{R}^{K\times 3}$ where K is the number of keypoints, lack pixel-level spatial correspondence with the driving image. Therefore, it is not feasible to directly integrate them into the driving latents using the ControlNet approach. Therefore, we employ cross-attention to interact implicit keypoints with the driving latents, a method that has been successfully applied in text-to-image generation tasks.

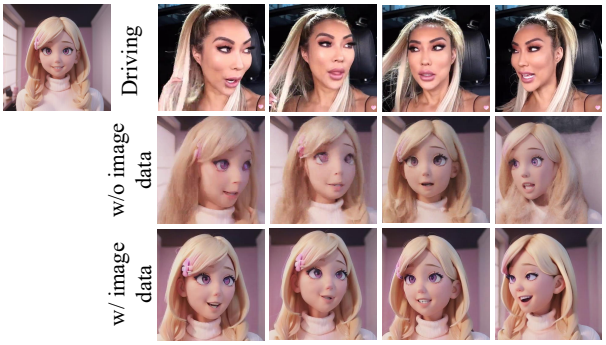

Figure 5: Ablation of image training data.

We present the comparative results in Table 3. It can be observed that both achieve good ID similarity and comparable expression accuracy. However, in terms of pose accuracy, our 2D motion position map significantly outperforms the 1D implicit keypoints. We attribute this to the fact that the 2D motion provides precise spatial guidance, which can be considered a strong guide. In contrast, the 1D motion only offers semantic-level guidance without strong spatial constraints, making it a weak guide. This is analogous to text-to-image generation tasks, where text can only provide class-level generation guidance but struggles to precisely direct the spatial positions of the generated image.

Ablation of Image Training Data

We present the results of training with only video data compared to training with both image and video data in Figure 5. It is evident that when additional image data is utilized, the model achieves better appearance preservation for non-realistic styles. This demonstrates that the diverse appearances in the image data enhance the model's robustness to unseen styles.

Conclusion and Future Work

This paper presents a robust portrait animation generation framework that integrates several effective techniques, including DiT-based rendering backbone, precise motion representation, efficient reference attention, and random augmentation techniques leveraging large image datasets. The integration of these technologies enables the framework to produce vivid and realistic portrait animations, advancing the indistinguishability between synthetic and real animations. While this paper ensures high-quality portrait generation, animation generation efficiency remains relatively low due to the large network model, making real-time applications currently impractical. Meanwhile, reliance on Face Vid2Vid's motion encoder, which fails to decouple pose and expression, may result in facial morphology inconsistencies in generated animations, particularly during pronounced head movements. Future work will employ transformer-based acceleration techniques or diffusion training distillation methods to expedite animation generation, enabling real-time human-computer interaction scenarios. Developing an advanced motion encoder to enhance motion control capabilities is also essential.

References

Assessment, I. Q. 2004. From error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 93.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 67–74. IEEE.

Chang, D.; Shi, Y.; Gao, Q.; Fu, J.; Xu, H.; Song, G.; Yan, Q.; Yang, X.; and Soleymani, M. 2023a. Magicdance: Realistic human dance video generation with motions & facial expressions transfer. *arXiv preprint arXiv:2311.12052*.

Chang, D.; Shi, Y.; Gao, Q.; Xu, H.; Fu, J.; Song, G.; Yan, Q.; Zhu, Y.; Yang, X.; and Soleymani, M. 2023b. MagicPose: Realistic Human Poses and Facial Expressions Retargeting with Identity-aware Diffusion. In *Forty-first International Conference on Machine Learning*.

Chen, J.; Wu, Y.; Luo, S.; Xie, E.; Paul, S.; Luo, P.; Zhao, H.; and Li, Z. 2024. Pixart-δ: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2023. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*.

Chu, X.; Li, Y.; Zeng, A.; Yang, T.; Lin, L.; Liu, Y.; and Harada, T. 2024. GPAvatar: Generalizable and precise head avatar from image (s). *arXiv preprint arXiv:2401.10215*.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Deng, Y.; Wang, D.; and Wang, B. 2024. Portrait4D-v2: Pseudo Multi-View Data Creates Better 4D Head Synthesizer. *arXiv preprint arXiv:2403.13570*.

Drobyshev, N.; Casademunt, A. B.; Vougioukas, K.; Landgraf, Z.; Petridis, S.; and Pantic, M. 2024. EMOPortraits: Emotion-enhanced Multimodal One-shot Head Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8498–8507.

Guo, J.; Zhang, D.; Liu, X.; Zhong, Z.; Zhang, Y.; Wan, P.; and Zhang, D. 2024. LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control. *arXiv preprint arXiv:2407.03168*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hong, F.-T.; and Xu, D. 2023. Implicit identity representation conditioned memory compensation network for talking head video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23062–23072.

Hong, F.-T.; Zhang, L.; Shen, L.; and Xu, D. 2022. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3397–3406.

Hu, L. 2024. 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*, 8153–8163.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Ren, Y.; Li, G.; Chen, Y.; Li, T. H.; and Liu, S. 2021. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13759–13768.

Retsinas, G.; Filntisis, P. P.; Danecek, R.; Abrevaya, V. F.; Roussos, A.; Bolkart, T.; and Maragos, P. 2024. 3D Facial Expressions through Analysis-by-Neural-Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2490–2501.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First order motion model for image animation. *Advances in neural information processing systems*, 32.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Tian, L.; Wang, Q.; Zhang, B.; and Bo, L. 2024. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*.

Vaswani, A. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wang, K.; Wu, Q.; Song, L.; Yang, Z.; Wu, W.; Qian, C.; He, R.; Qiao, Y.; and Loy, C. C. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, 700–717. Springer.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10039–10049.

Wei, H.; Yang, Z.; and Wang, Z. 2024. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*.

Xie, L.; Wang, X.; Zhang, H.; Dong, C.; and Shan, Y. 2022. 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*, 657–666.

Xie, Y.; Xu, H.; Song, G.; Wang, C.; Shi, Y.; and Luo, L. 2024. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Xu, Z.; Zhang, J.; Liew, J. H.; Yan, H.; Liu, J.-W.; Zhang, C.; Feng, J.; and Shou, M. Z. 2024. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1481–1490.

Ye, Z.; Jiang, Z.; Ren, Y.; Liu, J.; He, J.; and Zhao, Z. 2023. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*.

Zeng, B.; Liu, X.; Gao, S.; Liu, B.; Li, H.; Liu, J.; and Zhang, B. 2023. Face animation with an attribute-guided diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 628–637.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. 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*, 3661–3670.

Zhao, J.; and Zhang, H. 2022. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3657–3666.

Zheng, Z.; Peng, X.; Yang, T.; Shen, C.; Li, S.; Liu, H.; Zhou, Y.; Li, T.; and You, Y. 2024. Open-Sora: Democratizing Efficient Video Production for All.

Zhu, H.; Wu, W.; Zhu, W.; Jiang, L.; Tang, S.; Zhang, L.; Liu, Z.; and Loy, C. C. 2022. CelebV-HQ: A large-scale video facial attributes dataset. In *European conference on computer vision*, 650–667. Springer.