DeX-Portrait: Disentangled and Expressive Portrait Animation via Explicit and Latent Motion Representations

Yuxiang Shi[1,2,*] Zhe Li[2,*] Yanwen Wang[3,2] Hao Zhu[3] Xun Cao[3] Ligang Liu[1]
[1]University of Science and Technology of China [2]Central Media Technology Institute, Huawei [3]Nanjing University

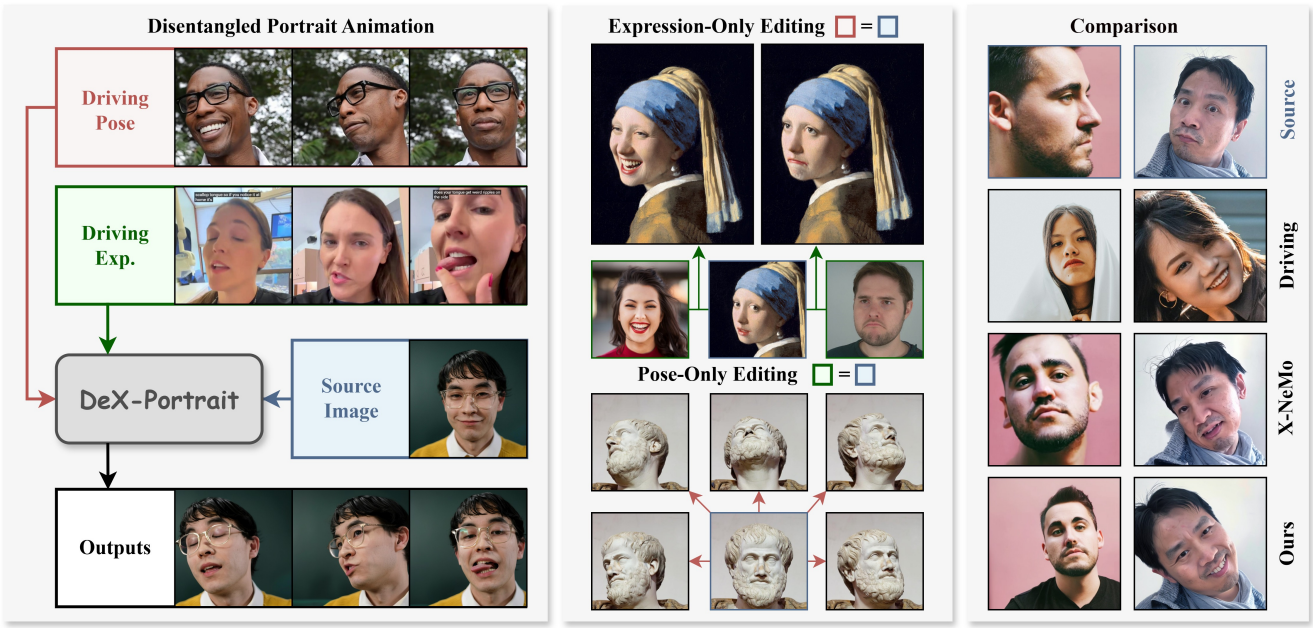


Figure 1. **Left:** Given an arbitrary source portrait image, driving expression images, and driving pose images, DeX-Portrait achieves disentangled and expressive portrait animation. **Middle:** DeX-Portrait enables expression-only and pose-only editing while keeping the other driving signal exactly the same as the source. **Right:** Compared with the state-of-the-art work, X-NeMo [69], our method offers superior fine-grained control over head pose including rotation, translation and scale.

Abstract

Portrait animation from a single source image and a driving video is a long-standing problem. Recent approaches tend to adopt diffusion-based image/video generation models for realistic and expressive animation. However, none of these diffusion models realizes high-fidelity disentangled control between the head pose and facial expression, hindering applications like expression-only or pose-only editing and animation. To address this, we propose DeX-Portrait, a novel approach capable of generating expressive portrait animation driven by disentangled pose and expression signals. Specifically, we represent the pose as an explicit global transformation and the expression as an implicit latent code. First, we design a powerful motion trainer to learn both pose and expression encoders for extracting precise and decomposed driving signals. Then we propose to inject the pose transformation into the diffusion model through a dual-branch conditioning mechanism, and the expression latent through cross attention. Finally, we design a progressive hybrid classifier-free guidance for more faithful identity consistency. Experiments show that our method outperforms state-of-the-art baselines on both animation quality and disentangled controllability.

1. Introduction

One-shot portrait animation, aiming at animating a source portrait image using a driving video, has been a popular topic due to its value in digital content creation. With the recent development of diffusion-based image/video generation models [28, 41, 50], researchers tend to modify and finetune them for expressive portrait animation [9, 69]. Although these SOTA diffusion-based approaches realize

high-quality facial animation, they still suffer from the controllability, especially individual controls on the head pose and facial expression, hindering the applications like expression-only or pose-only editing.

To achieve disentangled pose and expression controls, a plausible way [36, 53] is to represent the portrait motion with the head pose and expression blendshapes of 3D Morphable Models (3DMM) [13, 29, 52]. However, the performance of these methods is limited by the accuracy of 3DMM trackers and the representation ability of blendshapes. Consequently, they struggle to capture complex and subtle facial motions like sticking out the tongue and frowning. On the other hand, the state-of-the-art approach, X-NeMo [69], encodes the portrait motion as a 1D latent code capable of capturing expressive facial expression. Unfortunately, the latent code entangles the pose and expression and fails to precisely control the head rotation, scale and translation as shown in Fig. 1.

To this end, we propose *DeX-Portrait*, a diffusion-based framework that leverages explicit and latent motion representations for both disentangled and expressive portrait animation. Specifically, the head pose is represented as an explicit global transformation including a rotation, translation and scale (RTS), while the facial expression is represented as a latent code. The first challenge lies in disentangling the pose and expression encoders for extracting expression-agnostic pose transformation and pose-agnostic expression code. Thus we design a powerful GAN-based motion trainer (Fig. 2 (a)) by firstly applying 3D warping using a RTS-derived transformation and then modulating the generator using the expression code through Adaptive Instance Normalization (AdaIN) [20]. To prevent the pose leakage from the expression encoder, we design a series of augmentation strategies such as central cropping, random rotation and cross-view driving.

With the disentangled pose and expression encoders obtained, the second challenge lies in how to effectively inject the pose and expression signals into the diffusion model. We propose a novel dual-branch pose conditioning mechanism for precise pose control. As shown in Fig. 2, In the first branch, we map the pose RTS into a ray map and concatenate it with the noisy latent. In the other branch, we warp the intermediate features of the source portrait images through a 3D warping module and concatenate them with the corresponding features in the denoising UNet. Such a dual-branch pose injection enables the diffusion model to precisely control the head rotation, translation and scale. The expression code is injected via cross attention [49] like previous works [33, 54, 69]. Thanks to the hybrid motion representations and injection methods, our model realizes both expressive and disentangled portrait animation. In addition, inspired by FLOAT [24], we propose a progressive hybrid classifier-free guidance (CFG) in the denoising process by incorporating the pose and expression conditions successively for more stable identity consistency.

In conclusion, our core technical contributions are:

- DeX-Portrait, a diffusion-based framework that leverages explicit and latent motion representations for portrait animation, realizing both disentangled and expressive pose and expression controls.
- A powerful motion trainer that learns disentangled and precise pose and expression encoders through 3D warping and AdaIN modules.
- A dual-branch pose conditioning mechanism that injects the pose transformation into the diffusion model through the 3D warping module and the ray map.
- A progressive bybrid CFG that gradually incorporates the expression condition for more consistent identity.

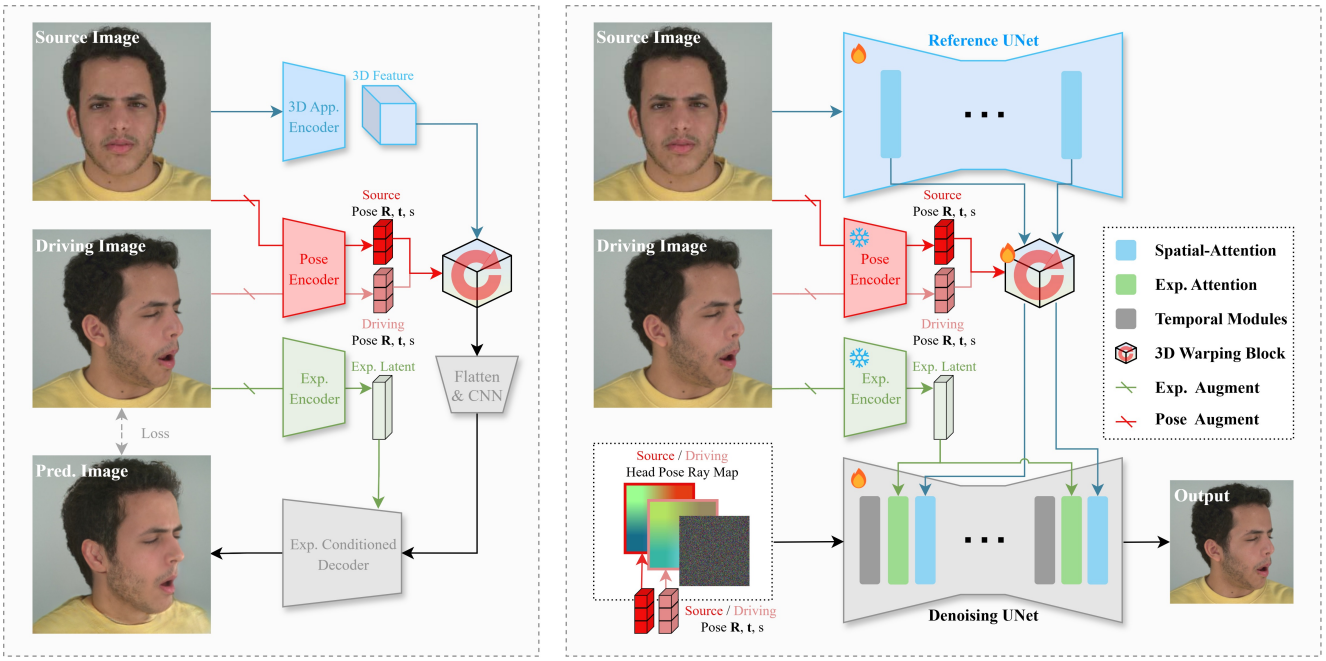
2. Related Work

2.1. Generalizable Portrait Animation

GAN or diffusion based portrait animation models first encode driving videos into motion representations, which are then used to animate arbitrary source portraits. Traditional approaches relied on explicit motion representations, such as 3D Morphable Models (3DMM) [21, 46–48, 68], facial landmarks [34, 39, 55], or dense optical flow maps [43] to disentangle the appearance and motion. While structured representations enable interpretable control, the process of explicit feature extraction inherently introduces biases, leading to limited accuracy in modeling dynamic expressions and poor generalization to large pose variations or complex scenarios. To address these issues, recent works have shifted to implicit motion representations, embedding motion information directly into latent spaces for end-to-end training [12, 14, 22, 56]. Among them, [3, 6, 51, 59] regard motion as a style and leverage StyleGAN-like architectures to produce animated results. Subsequently, diffusion-based models [9, 33, 54, 69] have also begun to incorporate such implicit control signals. However, implicit motion encoders rely on GANs for training, which often leads to entanglement between motion representations and identity features in the latent space. Different from previous approaches, DeX-Portrait incorporates an implicit expression representation and an explicit head pose representation. Via a novel motion injection method, our model can maximally mitigate the identity leakage issue while ensuring high-fidelity facial expression generation results.

2.2. Disentanglement

Recent years, several portrait animation models aim for disentangled control of facial expressions and head posesfor diverse applications. These methods can be categorized into three classes: The first class leverages manually defined facial models (e.g., 3DMM) to extract expression/pose



(a) Motion Trainer. (b) Diffusion-based Disentangled Animator.

Figure 2. Our pipeline consists of two stages: (a) Training a disentangled pose and expression encoder using a motion trainer. (b) Taming a latent diffusion model for disentangled and expressive portrait animation.

parameters of the driving face [36, 58, 67], and achieves disentangled control by replacing these parameters during inference. The second class relies on semantic extractors (e.g., action units, CLIP) for explicit disentangled control [7, 62, 66]. Both suffer from limited feature extraction accuracy and often fail to faithfully replicate the expressions of the driving subject. The third type employs latent pose and expression features [11, 12, 14, 37, 51], guiding feature extractors to extract mutually disentangled expression and pose features via physically interpretable network architectures, loss functions, or other approaches. Such methods have significantly improved generation accuracy; unfortunately, they can only be end-to-end trained based on GANs, limiting output fidelity. Our model incorporates a unique GAN-based motion trainer to train high-precision disentangled latent expression and pose features, and integrates a Diffusion-based generator to achieve high-fidelity results.

3. Method

Given a source portrait image $\mathbf{I}_\text{s}$ and a driving sequence $\{\mathbf{I}_\text{d}\}$, our objective is to generate a portrait animation sequence $\{\hat{\mathbf{I}}(\text{ID}_\text{s}, \text{pose}_\text{d}, \text{exp}_\text{d})\}$ controlled by the head pose and facial expression from the driving sequence, while preserving the identity and background consistent with the source image. Different from previous arts [54, 69], we also aim to disentangle the driving pose and expression to enable pose-only or expression-only editing as well as disentangled animation.

As illustrated in Fig. 2, our method contains two steps:

1. **Disentangled Motion Trainer.** We design a powerful GAN-based motion trainer and augmentation strategies to learn disentangled pose and expression encoders. Specifically, we first transform and augment the driving image into the pose and expression image and extract the pose transformation and expression latent, respectively. A 3D appearance feature is encoded from the source image and warped from the source pose to driving pose. Then we modulate the warped feature by the expression latent through AdaIN [20], producing an animated image and comparing it with the ground truth during training.
2. **Diffusion-based Disentangled Animation.** We employ the latent diffusion model (LDM) [41] as the backbone and injects the source identity through a reference UNet following [19, 69, 70]. Given the driving pose and expression signal, we propose a dual-branch pose injection with a cross-attention expression injection for disentangled portrait animation.

3.1. Preliminaries

Latent Diffusion Model (LDM). LDM [41] is a series of diffusion models [18, 45] that generate images in the latent space of pre-trained variational autoencoders (VAE). During training, LDM corrupts a clean latent z at time step t with a Gaussian noise ϵ_t to obtain a noisy latent z_t, follow-

ing DDPM [18]. Then a UNet-based [42] denoising network $\hat{\epsilon}_\theta$ is then trained to predict ϵ_t under the condition $\boldsymbol{c}$ using an MSE loss:

$$L_\theta = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon_t - \hat{\epsilon}_\theta(z_t, \boldsymbol{c}; t)\|_2^2\right]. \quad (1)$$

At inference time, the model begins from a pure Gaussian noise and applies a multi-step denoising process to generate meaningful latent samples according to the conditions.

Reference UNet Architecture for Animation. Owing to its powerful generative capabilities, LDM has been extensively adopted as the backbone network for motion-driven human animation synthesized from a single source image. For skeletal animation, Animate Anyone [19] pioneered to introduce a reference UNet to extract fine-grained features from the source image, which are then injected into the denoising network through spatial attention. Moreover, a temporal attention module [15] is also incorporated for temporal coherence. This paradigm was subsequently extended to portrait animation [61, 64, 69], a direction that our work also embraces.

3.2. Disentangled Motion Trainer

As shown in Fig. 2 (a), we design a powerful GAN-based motion trainer and several augmentation strategies to learn precise and disentangled pose and expression encoders. This GAN consists of three encoders (including a 3D appearance encoder, an explicit pose encoder, and a latent expression encoder) and a StyleGAN2-like [23] decoder.

3D Appearance Encoder. Following [12, 14, 56], a 3D feature containing the source identity is obtained through an appearance encoder instantiated by 2D and 3D CNNs.

Explicit Pose Encoder. We represent the head pose as an explicit global transformation $\mathbf{P} \in \mathbb{R}^{3\times4}$ including rotation $\mathbf{R}$, translation $\mathbf{t}$ and scale s (RTS) following [14, 56]:

$$\mathbf{P} = \begin{bmatrix} s\mathbf{R} & \mathbf{t} \end{bmatrix}. \quad (2)$$

From the perspective of representation, the transformation $\mathbf{P}$ with 6 $(3 + 2 + 1)$ degrees of freedom refrains from the expression leakage. Then we train a pose encoder instantiated by a ConvNeXt [31] to extract RTS from the portrait images. Obtaining the source and driving pose transformations ($\mathbf{P}_\text{s}$ and $\mathbf{P}_\text{d}$), we warp the 3D source feature from the source to driving pose under the transformation of $\mathbf{P}_\text{d}\mathbf{P}_\text{s}^{-1}$.

Latent Expression Encoder. We represent the facial expression as a 1D latent code with a dimension of 512 because of its expressive performance as demonstrated in X-NeMo [69]. We extract the latent code from the input portrait image using an expression encoder instantiated by a face alignment network (FAN) [5]. Following the information bottleneck principle from previous works [51, 69], the expression latent refrains from the identity leakage. However, it may retain the head pose information leaked from the input image.

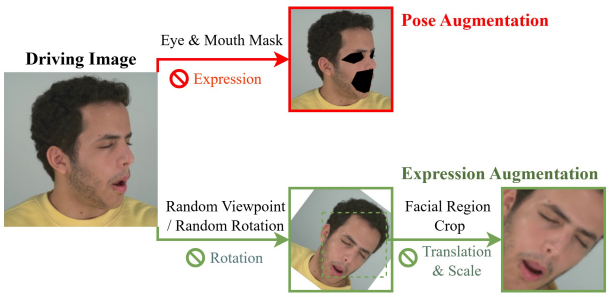


Figure 3. Illustration of the pose and expression augmentation.

Pose & Expression Augmentation. To this end, we propose a series of augmentation strategies on the pose and expression driving image to prevent expression or pose leakage, as shown in Fig. 3. For the pose input, we cover the eye and mouth regions using MediaPipe [32] landmarks to eliminate most expression information. For the expression input, we first perform random rotation or select another viewpoint (applicable only to multi-view datasets), enabling the expression encoder to be insensitive to head rotation. Then we crop the facial region and resize it to a fixed size of 224×224 via the MediaPipe bounding box to eliminate the head translation and scale. Overall, the above augmentation strategies promote the disentanglement between the pose and expression encoders.

Expression-conditioned Decoder. In the subsequent stage, we use a StyleGAN2-like generator to decode the warped appearance features into a target image while injecting the expression latent via AdaIN [20].

3.3. Diffusion-based Disentangled Animation

After obtaining disentangled expression and pose encoders from the motion trainer, we employ a latent diffusion model (LDM) and a reference UNet architecture for portrait animation, as mentioned in Sec. 3.1. Next, we elaborate on the methodology for injecting driving pose and expression signals into LDM to achieve disentangled and expressive animation. Correspondingly, the overall framework is illustrated in Fig. 2 (b).

Dual-branch Pose Injection. Regarding the injection method for pose information, existing practices typically render the pose into 2D skeleton maps [9] or spheres [33] and then inject them into the denoising network in a spatial manner. However, neither 2D skeletons nor spheres can accurately characterize the head pose. Therefore, we propose two novel methods for pose injection: ray map and reference warping.

In the first branch, inspired by Plücker ray map [38] that is widely used in camera pose controlled video generation [16, 30], we propose to transform the head pose into a ray map, and concatenate the source and driving ray maps with

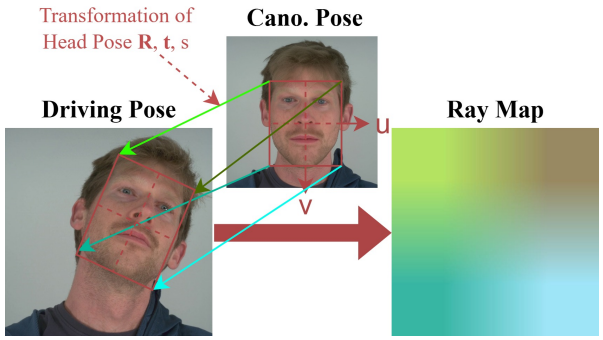


Figure 4. Illustration of the ray map of head pose.

the noisy latent for pose conditioned generation. Given a head pose $\mathbf{P}$ (Eq. 2), the ray map is formulated as

$$RayMap(u,v) = \mathbf{P}[u,v,0,0]^\top - [u,v,0,0]^\top \\ (u,v) \in [-1,1]^2, \tag{3}$$

where (u, v) is the coordinate of each pixel in the ray map. In terms of physical meaning, as illustrated in Fig. 4, each pixel on the ray map represents a vector from the canonical head pose to a target one. Given the spatial ray maps derived from both the source and driving poses, our method can achieve precise control over head rotation, translation, and scale, while guaranteeing identity consistency particularly in scenarios involving long-distance pose transitions.

However, we empirically found that relying solely on the ray map lead to edge misalignment between the synthesized result and the original image in the expression-only editing scenario, as illustrated in Fig. 8. We observe that LDM inherently possesses 3D perceptual capabilities, thereby allowing us to leverage pose signals for the direct manipulation of its intermediate latent features. Specifically, we first reshape the 2D source features in the reference UNet into 3D tensors, then warp them from the source pose to the driving pose, and finally convert them back to 2D via a flatten operation. Given that the warped source features are spatially aligned with the latent features in the denoising UNet, we first apply a convolutional projection layer to these warped features, then directly perform element-wise addition of the processed source features to the latent features in the denoising UNet. By virtue of the reference warping-based injection mechanism, our method achieves more precise pose control, particularly in the expression-only editing scenario, as illustrated in Fig. 8.

Cross-attention Expression Injection. Since the expression latent is a global feature, we perform cross attention [49] between the latent features in the denoising UNet and the expression latent following [26, 27, 69].

Figure 5. Pose-only generation preserves strong identity consistency (second from right). Compared with the original CFG, our method achieves better consistency with the source portrait (e.g., facial shapes) in scenarios involving significant pose and expression variations.

3.4. Progressive Hybrid CFG

To enhance controllability and sample fidelity, classifier-free guidance (CFG) [17] is commonly employed in the conditional diffusion process. In the CFG scheme, each denoising step computes both a conditional noise estimation $\hat{\epsilon}_\theta(z_t, \boldsymbol{c}; t)$ and an unconditional one $\hat{\epsilon}_\theta(z_t, \emptyset; t)$. The final noise estimation is computed as

$$\widetilde{\epsilon}_\theta(z_t, \boldsymbol{c}; t) \triangleq \omega\hat{\epsilon}_\theta(z_t, \boldsymbol{c}; t) + (1-\omega)\hat{\epsilon}_\theta(z_t, \emptyset; t), \tag{4}$$

where $\omega = 2.5$ denotes the CFG scale, a hyperparameter that regulates the conditioning strength.

We empirically found that the original CFG may yield unexpected results with inconsistency identity, especially when the source portrait faces sideways. We hypothesize the underlying reason is that the identity, pose, and expression conditions are entangled at each step throughout the denoising process. Drawing inspiration from [4, 24], we propose a progressive hybrid CFG that gradually incorporates pose and expression conditions. Specifically, over the 35 steps of DDIM [44], we exclude the expression condition within the initial 5 steps; subsequently, we incorporate the expression condition incrementally across the next 5 steps; finally, we employ the full conditions for the remaining 25 steps:

$$\widetilde{\epsilon}^*_\theta(z_t, \boldsymbol{c}; t) \triangleq \\ \begin{cases} \widetilde{\epsilon}_\theta(z_t, \boldsymbol{c}|_{\text{exp}}; t) & 30 < t \le 35, \\ \widetilde{\epsilon}_\theta(z_t, \boldsymbol{c}|_{\text{exp}}; t)\frac{t-25}{5} + \widetilde{\epsilon}_\theta(z_t, \boldsymbol{c}; t)\frac{30-t}{5} & 25 < t \le 30, \\ \widetilde{\epsilon}_\theta(z_t, \boldsymbol{c}; t) & t \le 25, \end{cases} \tag{5}$$

where $\boldsymbol{c}$ denotes all the conditions including identity, head pose, expression, $\boldsymbol{c}|_{\text{exp}}$ represents the conditions excluding expression, and $\widetilde{\epsilon}^*_\theta$ is our final noise estimation. As illus-

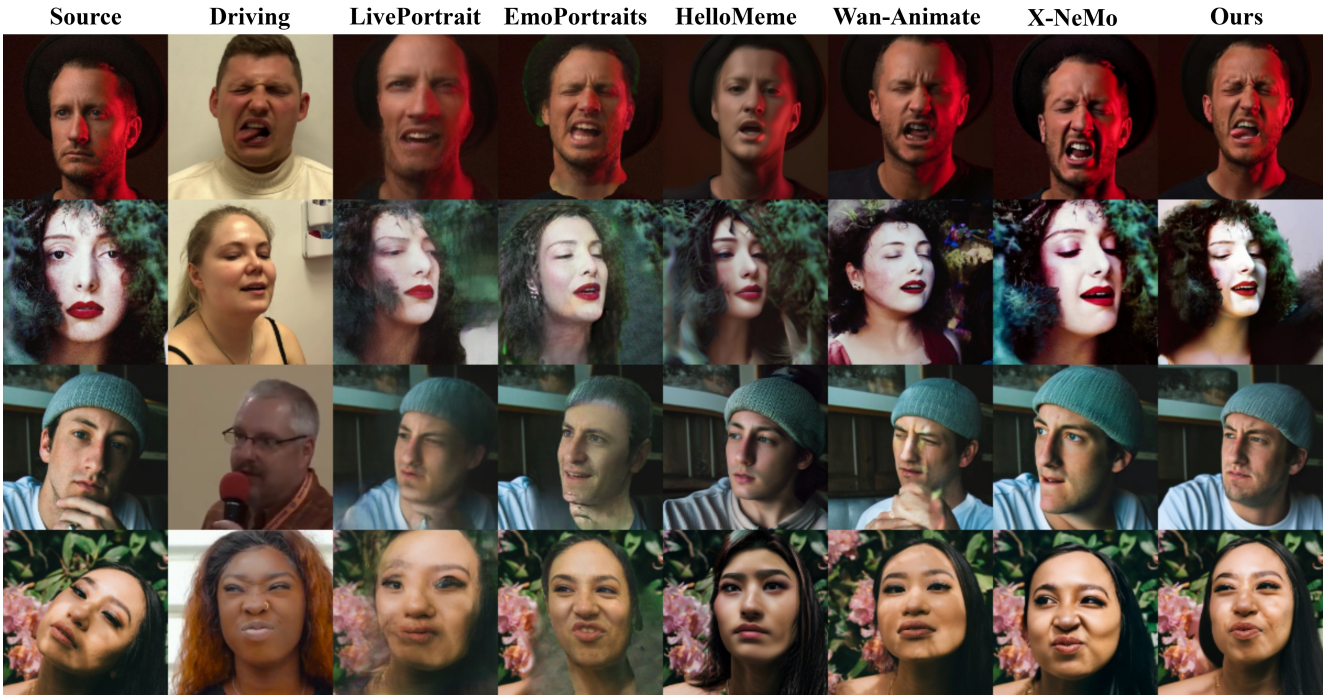


Figure 6. Qualitative comparison on cross-reenactment.

trated in Fig. 5, our progressive CFG can produce expressive animation with more consistent identity.

4. Experiment

4.1. Implenmentation details

We utilize two multi-view portrait video datasets (NerSemble [25], ava-256 [35]) and two in-the-wild monocular datasets (PFHQ [8], VFHQ [63]) for joint training. We process them to a fixed resolution of 512×512.

The whole training process involves 3 stages:

1. Motion Training. This stage is trained to obtain precise and disentangled expression and pose encoders, with a batch size of 112 and a learning rate of 1×10^{-4} for 200k iterations.
2. Diffusion Training. During this phase, we freeze the expression and pose encoders and train the reference and denoising UNets with a batch size of 48 and a learning rate of 1×10^{-5} for 120k iterations.
3. Temporal Training. Only the temporal module is trained using 24-frame video sequences, with a batch size of 8 and a learning rate of 1×10^{-5} for 80k iterations.

4.2. Benchmark, Baselines and Metrics

Benchmark. We collected a total of 150 copyright-free in-the-wild portrait photos from Life of Pix [1] and Unsplash [2], covering different ethnicities, lighting, poses and expressions. Meanwhile, we gathered 150 video clips (covering large-scale pose and expression variations) from a portrait video dataset TalkingHead1kH [57] and an extreme expression dataset FEED [12].

Baselines. We compare our method against state-of-the-art baselines, including diffusion-based methods (Wan-Animate [9], X-NeMo [69] and HelloMeme [66]) and GAN-based methods (LivePortrait [14] and EMOPortraits [12]). Among them, LivePortrait, EMOPortraits and HelloMeme can disentangle the pose and expression controls.

Metrics. We adopt PSNR, SSIM [60] and LPIPS [65] to evaluate the differences between the prediction and ground truth for self-reenactment scenarios. Since there exists no ground truth for cross-reenactment and disentangled-reenactment, we utilize CSIM [10], Average Expression Distance (AED) and Average Pose Distance (APD) computed using MediaPipe [32] for evaluating identity similarity, expression and pose accuracy.

4.3. Comparison

Self-Reenactment. In each collected video sequence, we select one frame as the reference image and use the other frames as the driving video to test the self-reenactment performance of different models. Then we reuse the driving video as the ground truth to compute PSNR, SSIM, and LPIPS. Tab. 1 demonstrate that our method outperforms other approaches on PSNR and LPIPS, with only a marginal deficit in SSIM compared to Wan-Animate.

Cross-Reenactment. We utilize portrait photos as source

Figure 7. Qualitative comparison on disentangled-reenactment.

Method	Self-Reenactment			Cross-Reenactment			Disentangled-Reenactment		
	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	AED↓	APD↓	CSIM↑	AED↓	APD↓
EMOPortraits [12]	20.010	0.758	0.248	0.255	0.0561	0.264	0.217	0.0586	0.155
LivePortrait [14]	27.760	0.857	0.098	0.459	0.0696	0.236	0.458	0.0695	0.195
HelloMeme [66]	18.870	0.660	0.292	0.313	0.0772	0.173	0.302	0.0874	0.226
X-NeMo [69]	21.060	0.741	0.207	0.492	0.0518	0.551	N/A	N/A	N/A
Wan-Animate [9]	27.970	0.865	0.098	0.551	0.0588	0.180	N/A	N/A	N/A
Ours	28.590	0.862	0.088	0.623	0.0515	0.145	0.631	0.0546	0.100

Table 1. Quantitative comparisons between our method and baselines. "N/A" means X-NeMo and Wan-Animate do not support disentangled reenactment. We highlight the best scores with orange shading, and the second best with light orange.

Figure 8. Qualitative ablation study of reference warping on the expression-only editing scenario.

images and video clips as driving videos from our benchmark for cross-reenactment comparison. We present qualitative comparison in Fig. 6. In contrast to other methods, our approach excels in the identity consistency, pose accuracy (especially the scale and translation) and expressiveness (e.g., tongues and squinting). The quantitative metrics are also reported in Tab. 1. GAN-based methods including LivePortrait [14] and EmoPortraits [12] suffer from blurriness and motion distortion due to their limited generative capability. Among the diffusion-based methods, HelloMeme [66] struggles to capture nuanced facial motions, as it utilizes a CLIP-based [40] motion encoder that is unsuitable for the specific task of portrait animation. Wan-Animate [9] leverages a pre-trained expression encoder from LIA [59] to enable expression control; however, its performance is constrained by a limited set of only 20 linear expression bases. X-NeMo [69], one of the state-of-the-art portrait animation baselines, delivers realistic and expressive animation, yet struggles to accurately reenact the driving pose, especially positional translation and scale variations, attributed to its motion representation that combines an entangled latent and a simplistic spatial triplet. Overall, thanks to the explicit pose and latent expression representations and effective injection mechanism, our method realizes expressive and precise controls over head pose and facial expression for cross-reenactment scenarios.

Disentangled-Reenactment. To verify the disentangled controllability of our model, we compare our method with related works (including LivePortrait [14], EmoPortraits

[12], and HelloMeme [66]) that support disentangled animation capabilities, using distinct driving pose and expression inputs. Given a source portrait photo $\mathbf{I}_s$, we select two different video clips from the bench mark as the driving pose $\{\mathbf{I}_d^{pose}\}$ and expression $\{\mathbf{I}_d^{exp}\}$, respectively, producing animation results $\{\hat{\mathbf{I}}\}$. We evaluate identity consistency via CSIM between $\{\hat{\mathbf{I}}\}$ and $\mathbf{I}_s$, pose control accuracy via APD between $\{\hat{\mathbf{I}}\}$ and $\{\mathbf{I}_d^{pose}\}$, and expression control accuracy via AED between $\{\hat{\mathbf{I}}\}$ and $\{\mathbf{I}_d^{exp}\}$, respectively. We present qualitative and quantitative comparisons in Fig. 7 and Tab. 1, respectively. Both comparisons demonstrate that our method achieves precise control over pose and expression, while preserving identity consistency with the source portrait. Moreover, thanks to the powerful disentangled capability, our method also enables expression-only and pose-only editing, as shown in Fig. 1 and the Supp. Mat.

4.4. Ablation Study

Head Pose Ray Map. We evaluate the head pose ray map by eliminating it, i.e., using only the reference warping for pose injection. Fig. 9 illustrates that the ray maps provide long-range correspondences between the source and driving poses, thereby enabling more stable identity consistency, especially when the source and driving rotations and scales exhibit substantial differences.

Reference Warping. We found that relying solely on the ray map for pose injection could result in inconsistent boundaries and backgrounds in expression-only editing scenarios, which leads to visible seams when pasted back onto the original full image, as shown in Fig. 8. Reference warping delivers a robust signal of an identity matrix derived by the pose transformation in this scenario, thereby enabling expression-only modification.

Pose & Expression Augmentation. As shown in Fig. 10, it can be observed that after removing the pose & expression augmentation, the model's generation consistency for expressions and poses exhibits a significant decrease. This indicates that the proposed strategy can effectively guide the pose and expression extractors to extract their corresponding features, thereby avoiding the accuracy loss caused by the mutual interference between pose information and expression information.

Quantitative Ablation Studies. All the numerical results are reported in Tab. 2. It demonstrates that these core contributions yield the most consistent identity, the most accurate pose reenactment, and high-fidelity expressiveness.

5. Discussion

Conclusion. We present DeX-Portrait, a new diffusion-based portrait animation framework that leverages explicit and latent motion representations for both disentangled and expressive animation. We propose to represent head pose and facial expression as a global transformation and a latent code, respectively, and design a dedicated motion trainer along with augmentation strategies for learning mutually disentangled pose and expression encoders. Based on the pretrained motion encoders, we propose a dual-branch pose injection mechanism coupled with cross-attention based expression injection—tailored for the diffusion model—enabling precise and independent control over portrait animation. Overall, our method outperforms other state-of-the-art approaches on both expressiveness and disentangled controllability.

Limitation. Our model is exclusively trained on real human datasets, and thus lacks generalization ability to other styles, e.g., cartoons. Moreover, our model struggles to perform reliably in scenarios involving multiple portraits and significant occlusions.

Ethics Statement. We acknowledge that our method could potentially be exploited to produce synthetic misinformation videos. Thus we emphasize the necessity of exercising responsible use of this technology, accompanied by clear synthetic content disclaimers.

Figure 9. Qualitative ablation of the head pose ray map.

Figure 10. Qualitative ablation of the augmentations.

Method	Cross-Reenactment		Disentangled-Reenactment	
	CSIM↑	AED/APD↓	CSIM↑	AED/APD↓
w/o rap map	0.609	**0.0506**/0.162	0.609	**0.0542**/0.105
w/o warping	0.619	0.0507/0.166	0.631	0.0573/0.121
w/o augmentation	0.619	0.0583/0.283	0.629	0.0634/0.168
Ours	**0.623**	0.0515/**0.145**	**0.631**	0.0546/**0.100**

Table 2. Quantitative ablation studies.

References

[1] Pix of life. https://www.lifeofpix.com/. 6

[2] Unsplash. https://unsplash.com/. 6

[3] Shaojie Bai, Te-Li Wang, Chenghui Li, Akshay Venkatesh, Tomas Simon, Chen Cao, Gabriel Schwartz, Ryan Wrench, Jason Saragih, Yaser Sheikh, et al. Universal facial encoding of codec avatars from vr headsets. *arXiv preprint arXiv:2407.13038*, 2024. 2

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 5

[5] 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 *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030, 2017. 4

[6] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13786–13795, 2020. 2

[7] Shao-Yu Chang, Jingyi Xu, Hieu Le, and Dimitris Samaras. Talking head generation via au-guided landmark prediction. *arXiv preprint arXiv:2509.19749*, 2025. 3

[8] Xiaoxu Chen, Jingfan Tan, Tao Wang, Kaihao Zhang, Wenhan Luo, and Xiaochun Cao. Toward real-world blind face restoration with generative diffusion prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9): 8494–8508, 2024. 6

[9] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025. 1, 2, 4, 6, 7

[10] 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

[11] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2663–2671, 2022. 3

[12] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8498–8507, 2024. 2, 3, 4, 6, 7, 8

[13] Xuan Gao, Haiyao Xiao, Chenglai Zhong, Shimin Hu, Yudong Guo, and Juyong Zhang. Portrait video editing empowered by multimodal generative priors. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2

[14] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 2, 3, 4, 6, 7

[15] 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. *arXiv preprint arXiv:2307.04725*, 2023. 4

[16] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 4

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4

[19] 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. 3, 4

[20] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2, 3, 4

[21] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021. 2

[22] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 2

[23] 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. 4

[24] Taekyung Ki, Dongchan Min, and Gyeongsu Chae. Float: Generative motion latent flow matching for audio-driven talking portrait. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14699–14710, 2025. 2, 5

[25] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 6

[26] Tobias Kirschstein, Simon Giebenhain, and Matthias Nießner. Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5481–5492, 2024. 5

[27] Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. *arXiv preprint arXiv:2502.20220*, 2025. 5

[28] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 1

[29] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2

[30] Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3d scenes from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 798–810, 2025. 4

[31] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 4

[32] 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. 4, 6

[33] Yuxuan Luo, Zhengkun Rong, Lizhen Wang, Longhao Zhang, and Tianshu Hu. Dreamactor-m1: Holistic, expressive and robust human image animation with hybrid guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11036–11046, 2025. 2, 4

[34] 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. 2

[35] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, et al. Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. *Advances in Neural Information Processing Systems*, 37:83008–83023, 2024. 6

[36] Lingzhou Mu, Baiji Liu, Ruonan Zhang, Guiming Mo, Jiawei Jin, Kai Zhang, and Haozhi Huang. Flap: Fully-controllable audio-driven portrait video generation through 3d head conditioned diffusion model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10437–10446, 2025. 2, 3

[37] Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong Cun, Ying Shan, and Dong-ming Yan. Dpe: Disentanglement of pose and expression for general video portrait editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2023. 3

[38] Julius Plücker. Xvii. on a new geometry of space. *Philosophical Transactions of the Royal Society of London*, (155): 725–791, 1865. 4

[39] Di Qiu, Zhengcong Fei, Rui Wang, Jialin Bai, Changqian Yu, Mingyuan Fan, Guibin Chen, and Xiang Wen. Skyreels-a1: Expressive portrait animation in video diffusion transformers. *arXiv preprint arXiv:2502.10841*, 2025. 2

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 7

[41] 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. 1, 3

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[43] 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. 2

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

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

[46] Xusen Sun, Longhao Zhang, Hao Zhu, Peng Zhang, Bang Zhang, Xinya Ji, Kangneng Zhou, Daiheng Gao, Liefeng Bo, and Xun Cao. Vividtalk: One-shot audio-driven talking head generation based on 3d hybrid prior. In *2025 International Conference on 3D Vision (3DV)*, pages 713–722. IEEE, 2025. 2

[47] Felix Taubner, Ruihang Zhang, Mathieu Tuli, Sherwin Bahmani, and David B Lindell. Mvp4d: Multi-view portrait video diffusion for animatable 4d avatars. *arXiv preprint arXiv:2510.12785*, 2025.

[48] Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B Lindell. Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5318–5330. IEEE Computer Society, 2025. 2

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 5

[50] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1

[51] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation

learning for fine-grained controllable talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17979–17989, 2023. 2, 3, 4

[52] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20333–20342, 2022. 2

[53] Lizhen Wang, Zhurong Xia, Tianshu Hu, Pengrui Wang, Pengfei Wei, Zerong Zheng, Ming Zhou, Yuan Zhang, and Mingyuan Gao. Dreamactor-h1: High-fidelity human-product demonstration video generation via motion-designed diffusion transformers. *arXiv preprint arXiv:2506.10568*, 2025. 2

[54] Qiang Wang, Mengchao Wang, Fan Jiang, Yaqi Fan, Yonggang Qi, and Mu Xu. Fantasyportrait: Enhancing multi-character portrait animation with expression-augmented diffusion transformers. *arXiv preprint arXiv:2507.12956*, 2025. 2, 3

[55] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. *arXiv preprint arXiv:1910.12713*, 2019. 2

[56] 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/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 2, 4

[57] 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/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 6

[58] Youjia Wang, Taotao Zhou, Minzhang Li, Teng Xu, Minye Wu, Lan Xu, and Jingyi Yu. Neural relighting and expression transfer on video portraits. *arXiv preprint arXiv:2107.14735*, 2021. 3

[59] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022. 2, 7

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

[61] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 4

[62] Mengting Wei, Tuomas Varanka, Xingxun Jiang, Huai-Qian Khor, and Guoying Zhao. Magicface: High-fidelity facial expression editing with action-unit control. *arXiv preprint arXiv:2501.02260*, 2025. 3

[63] 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. 6

[64] 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. 4

[65] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[66] Shengkai Zhang, Nianhong Jiao, Tian Li, Chaojie Yang, Chenhui Xue, Boya Niu, and Jun Gao. Hellomeme: Integrating spatial knitting attentions to embed high-level and fidelity-rich conditions in diffusion models. *arXiv preprint arXiv:2410.22901*, 2024. 3, 6, 7, 8

[67] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8652–8661, 2023. 3

[68] Xiaochen Zhao, Jingxiang Sun, Lizhen Wang, Jinli Suo, and Yebin Liu. Invertavatar: Incremental gan inversion for generalized head avatars. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024. 2

[69] Xiaochen Zhao, Hongyi Xu, Guoxian Song, You Xie, Chenxu Zhang, Xiu Li, Linjie Luo, Jinli Suo, and Yebin Liu. X-nemo: Expressive neural motion reenactment via disentangled latent attention. *arXiv preprint arXiv:2507.23143*, 2025. 1, 2, 3, 4, 5, 6, 7

[70] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024. 3