AniFeats: Animate 3D Feature Meshes for Character Video Generation

Beijia Lu, Zekai Gu, Zhiyang Dou, Haotian Yuan, Peng Li,
Chenyang Si, Yukang Cao, Yuming Jiang, Yuan Liu, Wenping Wang, and Ziwei Liu

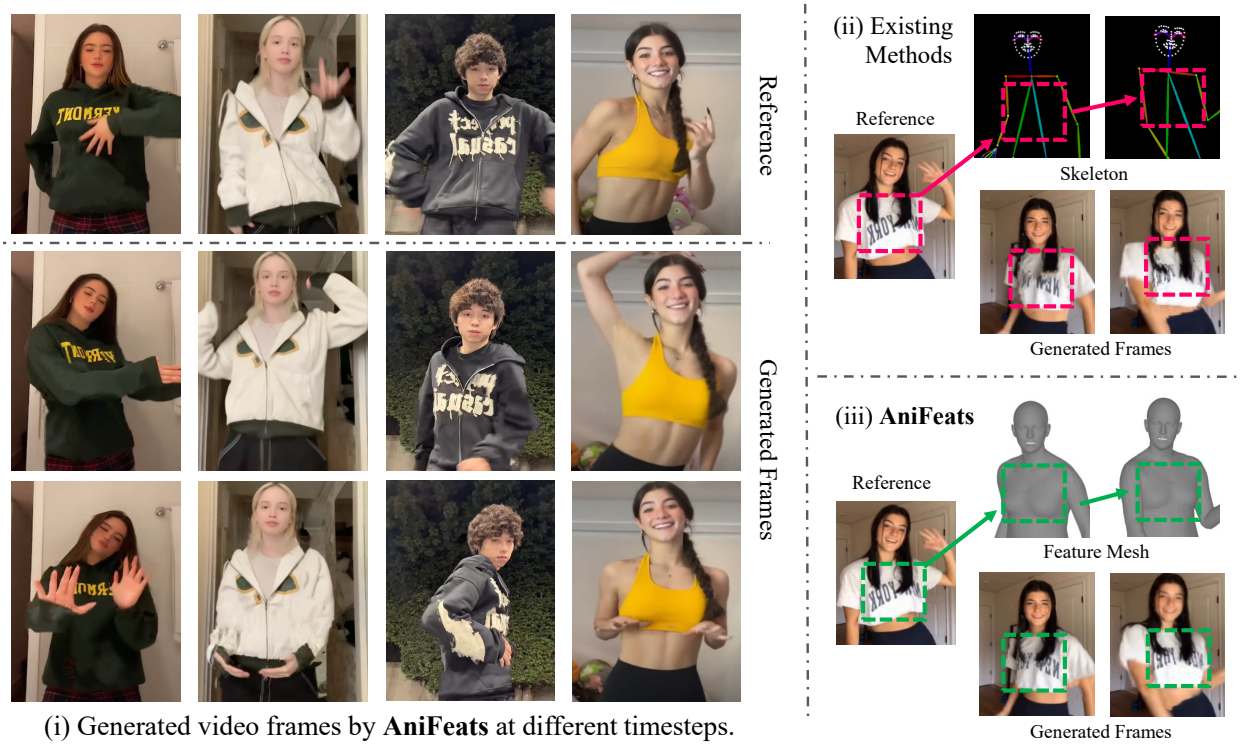

Fig. 1. (i) **AniFeats** generates consistent character videos from reference images and motion sequences. (ii) Existing methods [16], [43], [52], [61] only implicitly associate the reference image with the generated image, leading to inconsistency in the generation. (iii) AniFeats explicitly transfers the features from the reference image to the generated images via animated feature meshes, improving the consistency.

***Abstract*—Generating high-quality character animation videos is a fascinating yet challenging task. Existing methods use geometry guidance signals like skeletons, normal maps, or depth maps in a diffusion model to generate character videos from a single reference image. Although these approaches have shown encouraging results, they solely rely on cross attention layers to extract geometry guidance which inevitably leads to temporal inconsistencies and reduced quality. In this paper, we present a novel framework AniFeats to generate high-quality character animation videos. In contrast to existing methods, our key insight is to incorporate explicit features on 3D character meshes during the video generation to achieve significantly improved temporal consistency. Specifically, AniFeats extracts detailed features from the reference image, projects them onto 3D feature meshes based on SMPL-X, and utilizes rendered feature maps from the animated 3D feature meshes as guidance throughout the generation process. This approach directly links local patterns in the input image to those in the output video, effectively strengthening temporal coherence. Extensive experiments demonstrate that AniFeats generates high-quality, temporally consistent character animations with remarkably enhanced realism. Our code and models will be publicly available at https://github.com/Beijia11/AniFeats**

***Index Terms*—Video Generation, Character Animation, Human Video Generation.**

Beijia Lu is with Carnegie Mellon University. Zekai Gu, Peng Li and Yuan Liu are with Hong Kong University of Science and Technology. Zhiyang Dou is with The University of Hong Kong. Haotian Yuan is with Tsinghua University. Chenyang Si is with Nanjing University. Yuming Jiang is with Alibaba DAMO Academy. Yukang Cao and Ziwei Liu are with Nanyang Technological University. Wenping Wang is with Texas A&M University.

E-mail: beijialu@andrew.cmu.edu, skygoo2000@gmail.com, zhiyang0@connect.hku.hk, yuanht22@mails.tsinghua.edu.cn, plibp@connect.ust.hk, chenyang.si.mail@gmail.com, yukang.cao@ntu.edu.sg, yumingj80@gmail.com, yuanly@ust.hk, wenping@tamu.edu, ziwei.liu@ntu.edu.sg.

I. Introduction

CHARACTER animation videos are among the most vital forms of video content, serving an irreplaceable role in content creation for the gaming, film, and AR/VR industries.

Their ability to bring stories and characters to life enables immersive narratives that resonate deeply with audiences. Early-stage works [33], [36], [55] in this area are mainly based on GAN [11] to generate human images of novel poses but are limited by the representation capacity and stability of GANs and fall short in producing high-quality animation videos.

Recent diffusion models [14], [32] have shown a strong ability to generate high-quality videos. Thus, many works [16], [35], [52], [61] tried to utilize diffusion models to generate character animation videos. These methods typically take a reference image of a specific character and a motion sequence as input for generating an animation video that matches the given character and motion, using both as control guidance to the diffusion model for generation.

Although encouraging results are achieved, existing methods [16], [35], [52], [61] still struggle to maintain identity consistency with the input reference image and temporal consistency over time, which restricts the quality of the generated videos. This is because existing works represent target motions solely with 2D images, i.e., skeleton images, depth maps, semantic maps, or normal maps, as conditions for the diffusion model. Then, attention layers are applied to implicitly associate the input reference images with these target motions. For example, as illustrated in Fig. 1 (ii), the diffusion model has to implicitly learn the chest of the target motion corresponding to the chest in the reference image through attention layers. Such implicitly inferred correspondences between target motions and reference images may lack accuracy and can vary over time, resulting in generated videos that struggle to maintain both identity and temporal consistency. This phenomenon becomes even more severe when the input reference image has an incompatible human pose with the target motion sequence as shown in Fig. 2.

In this paper, we address the above challenges by proposing a novel method called **AniFeats** for high-quality character animation video generation. The key idea of AniFeats is to explicitly transfer the image features of the reference image to the generated frame via the animation of the 3D SMPL-X mesh, as shown in Fig. 1 (iii). Specifically, we first extract latent features from the reference image and then build 3D feature meshes by projecting these features onto the 3D SMPL-X mesh estimated from the reference image. Later, given the target motion sequences, we deform the 3D feature SMPL-X mesh accordingly and rasterize this animated feature mesh on target viewpoints to get feature maps. These feature maps contain the features extracted from the reference image but are animated to the target motion sequences and thus explicitly build correspondences between the input reference image and the target motion sequence. Finally, we utilize these rasterized feature maps as the condition to guide the diffusion generation.

Adopting the above animated feature meshes for video generation has the following two prominent advantages. First, this improves the temporal consistency. The animated 3D features associate the same 3D regions across different frames, which serves as a strong signal for the diffusion model to maintain cross-frame temporal consistency. Second, the explicit feature transferring improves the identity consistency and bridges the gap between the input reference pose and the target motion. Because even when the input reference pose deviates largely from the target motion sequence, the feature extraction on the reference image follows the estimated reference poses to select correct regions as shown in Fig. 2.

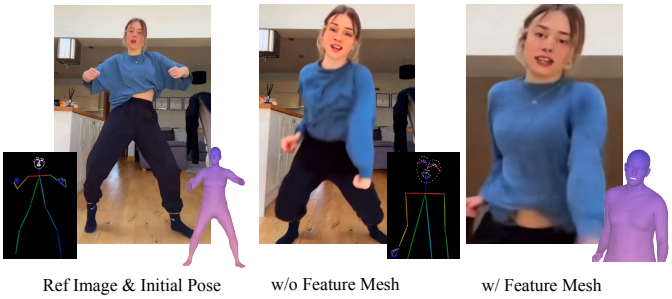

Fig. 2. **Comparison of generated results on the Unpaired Dataset**. The explicit feature mesh in AniFeats effectively reduces the discrepancy between the reference pose and the target motion.

We validate the effectiveness of AniFeats on a large number of challenging cases where our method demonstrates higher identity preservation and temporal consistency in character videos compared to baseline methods. Our method enables animating different characters with the same motion sequences and also the same character with different motions. Furthermore, we further showcase the capability of AniFeats in 4D video generation with robust temporal and spatial coherence for producing high-quality, consistent animations.

II. Related Work

A. Image-guided 2D human animation

Traditional approaches for 3D human animation [18], [20], [34], [55] typically rely on learning the human deformation field from multi-view or monocular videos, or multiple static images. However, these methods often require extensive video capture, which limits their practical application in real-world scenarios. Fortunately, recent advances in diffusion models have led to techniques capable of animating humans [5], [8]–[10], [23], [35], [38], [39], [41], [43] from a single input image. DreamPose [22] leverages the pre-trained Stable Diffusion model and incorporates both CLIP [31] and VAE [24] for efficient image encoding. DisCo [42] takes an innovative approach by using dual independent ControlNets [58] to separately control pose and background, offering more fine-grained control over the animation process. AnimateDiff [13] introduces a temporal layer to the denoising UNet to improve temporal coherence in animations. Animate Anyone [16] uses a UNet-based ReferenceNet to extract features from reference images, while MagicAnimate [52] employs a ControlNet based on DensePose [12] inputs for more accurate pose guidance compared to traditional OpenPose [7] keypoints. Champ [61] leverages SMPL [25] as their motion conditions to achieve more consistent results. However, it only considers the geometry priors derived from SMPL. Follow-Your-Pose [26], [53] leverages optical flow to separate the background and successfully achieves multi-person animation. Unfortunately, existing methods face significant challenges due to the lack of

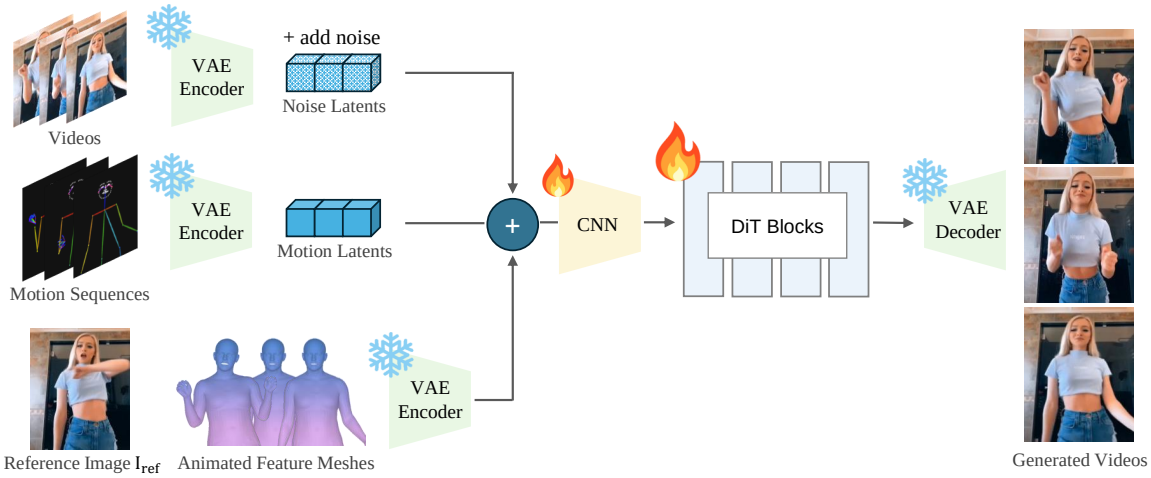

Fig. 3. **Pipeline of AniFeats.** Given an input motion sequence and a reference image, we first extract the feature map from the reference image and build the corresponding animated feature meshes. The video and the associated conditions are then encoded using a VAE encoder and concatenated before being processed by the CNN downsampler. Finally, a diffusion transformer model (DiT) leverages these conditions to denoise the noise latent and a VAE decoder generates the character video.

3D spatial information in the input. This limitation makes it difficult to maintain consistent human textures during animation, especially when dealing with complex scenarios such as large pose variations or significant rotation.

The concurrent works like Human4DiT [35], Animate-X [37], and HumanVid [44] achieve strong results by utilizing a large-scale dataset or DiT-based framework for even multi-human video generation. The contribution of these works is orthogonal to ours because we focus on utilizing 3D features as conditions and are also compatible with their framework.

B. Animatable 3D Human Reconstruction from Images

Recent advances have explored generating animatable 3D human models from monocular images or videos for applications in character animation and virtual environments. TeCH [51] employs text-guided reconstruction to produce life-like clothed humans, while SiTH [15] uses image-conditioned diffusion to recover textured 3D humans from single views. IntrinsicAvatar [40] applies physically based inverse rendering and explicit ray tracing to reconstruct dynamic humans from monocular videos, and AnimatableGaussian [47] enables fast, high-quality multi-avatar reconstruction via Gaussian splatting. While these methods yield fully animatable 3D representations, they require complete geometry reconstruction, often with high computational cost. In contrast, our approach uses 3D feature meshes derived from geometry signals such as skeletons, normal maps, and depth maps. Although stored as 2D images, these cues inherently encode 3D structure, enabling efficient animation guidance without full mesh reconstruction and avoiding the heavy optimization pipelines of prior methods.

C. Parametric 3D human model

Parametric 3D human models [25], [28] provide a compact and expressive representation of the human body. SMPL [25] established the foundation with a statistical mesh model, while SMPL-X [28] extended it with additional joints and expression parameters, enabling more detailed reconstruction [56], [57]. Recent works further address clothing and hair [3], [21], [48], [60], broadening applications to realistic scenarios. Despite limited training data, these models have become strong 3D priors for reconstruction [49], [50] and novel view synthesis [30], [46]. In our framework, we leverage the SMPL-X [28] mesh to link reference images and target motions, ensuring identity and temporal consistency in generated videos.

III. Method

Given a reference image $\mathbf{I}_{\text{ref}}$ containing a human and a motion sequence $\theta_{1:N} = \{\theta_1, \ldots, \theta_N\}$ with N denoting the number of frames, our goal is to generate a video sequence $\mathbf{I}_{1:N} = \{\mathbf{I}_1, \ldots, \mathbf{I}_N\}$ that preserves the appearance of the human of I_{ref} while following the specified motion $\theta_{1:N}$.

A. Overview

Pre-processing. On the given reference image, we apply the SMPLer-X [6] to estimate the pose θ_{ref} and the body parameters β_{ref}. Then, we combine the estimated body parameter β_{ref} with the given poses $\theta_{1:N}$ to get a set of deformed SMPL-X [28] meshes $\mathbf{M}_{1:N} := \{\mathbf{M}(\beta_{\text{ref}}, \theta_i) | i = 1, ..., N\}$, where $\mathbf{M}(\beta_{\text{ref}}, \theta_i)$ means applying the body parameters β_{ref} and the pose θ_i to generate a posed SMPL-X mesh.

Overview. Based on the given SMPL-X mesh sequence, we construct a set of condition maps to control our diffusion model for the video generation, as introduced in Sec. III-B. Our key idea is to construct a 3D feature mesh and render the feature map to explicitly associate the reference image and the generated frames, which is illustrated in Sec. III-C. Finally, we introduce our training strategy in Sec. III-D

B. Pose Conditioned Diffusion

In this section, we introduce our conditional diffusion generative model using the conditions from an SMPL-X

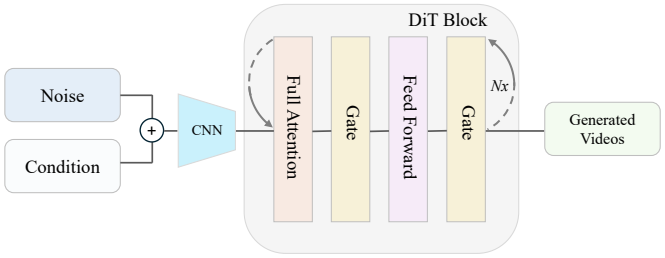

Fig. 4. **Architecture of Diffusion Transformer (DiT) blocks**. In each DiT block, the concatenated embedding sequentially passes through the Full Attention, Gate, and Feed Forward modules. After processing through the predefined blocks, the latent is decoded to reconstruct the video.

parametric model as guidance for generation. An overview of our conditional generative model is shown in Fig. 3.

a) Conditional diffusion generation: We apply a transformer-based video diffusion model [29] with additional conditions to generate our video. The video diffusion model uses a VAE encoder to compress the video and conditions into the latent space and apply an inverse Markov chain to generate data on the latent space. Starting from a pure Gaussian noise $\mathbf{z}_T$ with T denoting the maximum diffusion timestep, we iteratively apply a diffusion transformer $\epsilon(\mathbf{z}_t, t, \mathbf{c})$ to denoise $\mathbf{z}_t$ to $\mathbf{z}_0$, and the output denoised latent is processed by a VAE decoder to get the final video. $\mathbf{c}$ represents all the conditions constructed from the pose sequence $\theta_{1:N}$ and the reference image $\mathbf{I}_{ref}$.

b) Condition construction: We construct the skeleton images corresponding to the pose sequences as the conditions to the diffusion model. Then, the pretrained VAE is applied to compress both the reference image and the skeleton images into a shared latent space, where they are concatenated along the feature dimension. A CNN-based downsampling module is applied to align the latents to the same input shape as CogVideoX [54]. Subsequently, the concatenated embeddings are then processed by several diffusion transformer (DiT) blocks, whose detailed architecture is illustrated in Fig. 4. Then, the model decodes the latents using a 3D causal VAE decoder to reconstruct the video. We select skeleton images as conditional guidance rather than normal maps or depth maps because our feature map already incorporates both normal and depth information, which encodes 3D properties and texture cues. Combining these appearance-rich features with motion sequences enables the model to capture both temporal dynamics and detailed visual information, leading to improved pose control and appearance consistency.

c) Discussion on reference image condition: Previous UNet-based diffusion models, like Champ [61], process the reference image with a separated reference UNet and then apply the cross attention between the denoising UNet and the reference UNet to inject controls. In contrast, our method is different from these works in terms of directly concatenating all the conditions instead of applying cross attention layers. When we directly applied the cross attention strategy to our DiT-based framework, we observed a significant drop in performance. The model struggled to capture appearance information from the conditioning input. We hypothesize that

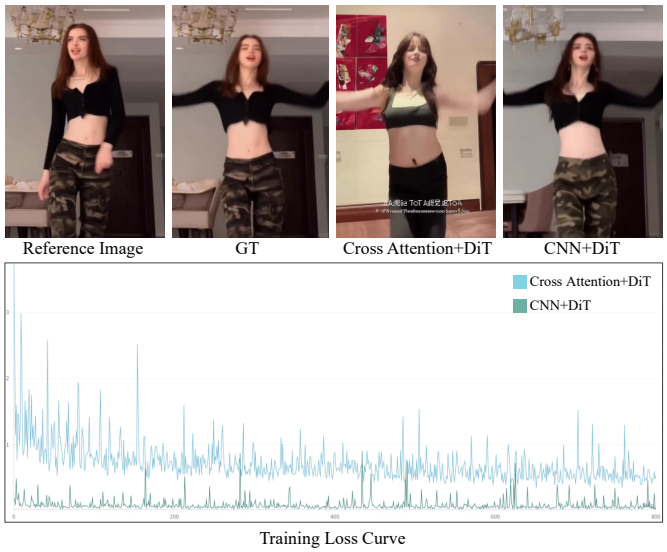

Fig. 5. **Comparison of Different Feature Injection Methods**. The upper panel compares visual results: the first column shows the reference image used as model input, the second column presents the ground truth (GT), and the third and fourth columns display results from cross-attention-based conditioning and CNN-based concatenation, respectively. The lower panel shows the corresponding loss curves, indicating that the cross-attention + DiT setting fails to converge compared in training.

this issue stems from fundamental differences in the nature of intermediate features between UNet and DiT. Unlike UNet, where hierarchical features exhibit strong spatial structures conducive to cross-attention, the DiT blocks process information in a more abstract and global manner. This structural difference likely reduces the effectiveness of cross-attention for conditioning, which makes it less suitable for direct conditional information incorporation and necessitates an alternative integration strategy. Thus, we propose a simple alternative to concatenate all conditions as inputs, which performs well on our datasets.

The comparison in Fig. 5 illustrates the effect of different conditioning strategies. The cross-attention-based approach struggles to maintain appearance consistency and fails to converge during training, whereas our CNN-based concatenation effectively preserves structural details and enhances temporal coherence.

d) Motivation of 3D feature meshes: As stated in the introduction, the skeleton images only contain information about the current target poses while being agnostic to the input reference images. Thus, the current diffusion model DiT blocks implicitly associate the reference image with the current target pose, which brings difficulty in maintaining temporal consistency and identity consistency. In the following, we introduce our idea of constructing 3D feature meshes and explicitly rendering the features of the reference image on the target pose and viewpoints as conditions for improved temporal and identity consistency.

C. Feature Mesh Extraction

In this section, we construct 3D feature meshes to render feature maps as conditions of our diffusion model, as shown

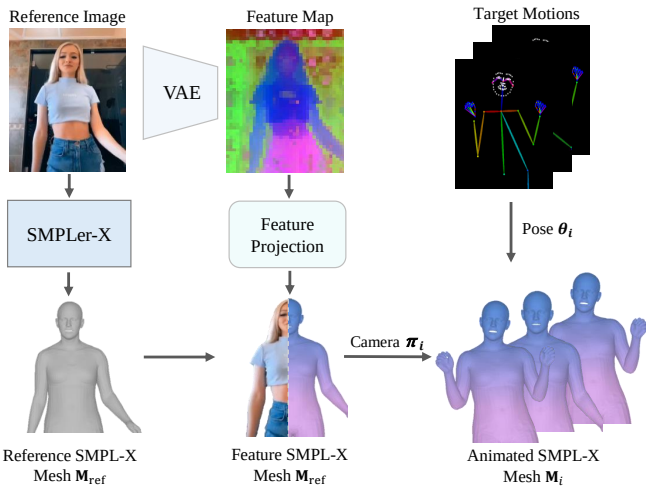

Fig. 6. **Construction of feature meshes and conditional feature maps**. To explicitly integrate 3D information, we estimate the SMPL-X mesh from the reference image and assign the latent features extracted by the pretrained VAE to the SMPL-X mesh. Then, the SMPL-X mesh is animated with a target motion and we render feature maps from animated meshes as conditions for diffusion models.

in Fig. 6.

a) 3D feature SMPL-X mesh: To construct the 3D feature mesh, we first extract the pretrained VAE latent features on the input reference image. Then, we project all the vertices of the SMPL-X mesh $\mathbf{M}_{\text{ref}} = \mathbf{M}(\beta_{\text{ref}}, \theta_{\text{ref}})$ onto the reference image to interpolate the extracted VAE features. These interpolated features are associated with the vertices of this SMPL-X mesh. When we deform this SMPL-X mesh $\mathbf{M}_{\text{ref}}$ with a target pose θ_i to get the mesh $\mathbf{M}_i = \mathbf{M}(\beta_{\text{ref}}, \theta_i)$, these associated features are also transformed to the mesh $\mathbf{M}_i$. We apply the rasterization technique to render a feature map based on these associated features. The rendered feature map is concatenated with the video and image latents processed by VAE encoder. Then, the concatenated latents is added to the denoising DiT branch of our diffusion model to generate the video.

b) Discussion: Such rendered features play two essential roles in improving both temporal and identity consistency. First, the rendered features directly come from the input reference image, which explicitly associates the reference image with the generated images and thus improves the identity consistency. Second, for temporal consistency, the features of a 3D vertex on two different video frames are exactly the same as each other, which acts like an anchor in the feature space as a temporal consistency constraint. For example, the feature on the shoulder region at the i-th video frame will be exactly the same as the feature on the shoulder at another j-th frame so the denoising diffusion transformer can easily learn that these two regions should have the same appearances, which thus improves the temporal consistency. In summary, with the help of these 3D feature meshes, we significantly improve both the temporal and identity consistency.

D. Training

We adopt the CogVideoX-Fun-V1.1-5B-Pose model as our base model. During the training process, we freeze the weights of the VAE encoder and decoder and only allow the DiT blocks and CNN downsampler to be updated. To initiate the training, a frame is randomly selected from a human video to serve as the reference image, and the entire video serves as the final target. The objective of this training strategy is to effectively train our model to learn to reference the strong guidance provided by the reference image and the feature map during the generation.

IV. Experiments

A. Implementation Details

We train AniFeats on a comprehensive and high-quality dataset comprising approximately 6000 videos. This dataset is constructed from a combination of the training set of the TikTok [19] dataset, the Champ's [61] training sample dataset, and additional in-the-wild videos created from TikTok and YouTube. We train the AniFeats model for 10k steps on 8 H800 GPUs using the AdamW optimizer with a learning rate of 1e-4. The learning rate follows a cosine schedule with restarts, with 100 warmup steps at the beginning of training.

B. Experimental Settings

a) Datasets: To ensure fair comparisons and adhere to established benchmarks in the field of character animation, we employ the test set of the TikTok dataset, which is exactly the same as used in previous works [16], [52], [61] for evaluation. Other than the TikTok dataset, we also adopted a self-collected dataset for evaluation, which consists of 100 videos mainly showing human dances. Besides the datasets for quantitative evaluation, we also include more wild examples to show the qualitative results.

b) Metrics: Our evaluation methodology follows standard metrics commonly used in previous methods [52], assessing both single-frame image quality and overall video fidelity. For single-frame quality, we use metrics like L1 error, Structural Similarity Index (SSIM) [45], Learned Perceptual Image Patch Similarity (LPIPS) [59], and Peak Signal-to-Noise Ratio (PSNR). Video fidelity is measured using Fréchet Video Distance (FVD). To measure pose accuracy, we compute the Mean Per Joint Position Error (MPJPE) between the target pose sequence and the poses extracted from the generated videos using an off-the-shelf pose estimation model.

c) Baselines: We perform a comprehensive comparison with several state-of-the-art methods for character video generation: (1) MagicAnimate [52] and Animate Anyone [16] are diffusion-based approaches that employ 2D guidance, effectively combining temporal modeling with appearance preservation to animate human 2D images given motion sequences. (2) Champ [61] uses a 3D parametric model within a latent diffusion framework, enhancing shape alignment and providing robust motion guidance for high-quality human animation. (3) UniAnimate [43] is a diffusion-based animation method that integrates disentangled condition modules for pose, human, and background into a pretrained diffusion model for realistic animation. We implement baseline methods with official codes except for Animate Anyone which we adopt the model reproduced by Moore AnimateAnyone [4].

Fig. 7. Qualitative comparison of AniFeats (**ours**) with Animate Anyone [16], MagicAnimate [52], Champ [61], and UniAnimate [43].

C. Comparisons

a) Quantitative comparisons: In Table I, we present the quantitative comparison between AniFeats and all baselines on the TikTok [19] dataset and our self-collected dataset respectively. As shown by the results, with the 3D features as additional conditions to our diffusion model, our approach outperforms previous state-of-the-art methods by showing more consistency with the groundtruth. Though our improvements in PSNR, SSIM, and LPIPS are not very significant because the generation results could be reasonable but not strictly aligned with the ground truth, we can see concretely increased consistency in the following qualitative comparison.

b) Qualitative comparisons: Fig. 7 presents qualitative comparisons between AniFeats and baseline methods. Notably, methods such as AnimateAnyone [16], and MagicAnimate [52], which depend solely on 2D poses or skeleton images in the generation, struggle to preserve consistency in the human shape, particularly when the target pose and the human orientation are largely different from those of the reference image. These limitations result in human pose distortion or a loss of detail in the generated videos. While the previous state-of-the-art method, Champ [61] and UniAnimate [43], who utilize a 3D SMPL mesh and transferred pose as guidance to establish a unified representation of body shape and pose in video generation, they still fail to maintain the

TABLE I
QUANTITATIVE COMPARISONS ON THE TIKTOK [2] AND SELF-COLLECTED DATASETS.

Methods	TikTok Dataset						Self-Collected Dataset				
	L1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	MPJPE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	MPJPE ↓
AnimateAnyone [42]	-	29.56	0.718	0.285	171.90	120.5	29.42	0.712	0.290	176.5	100.6
MagicAnimate [52]	3.13E-04	29.16	0.714	0.239	179.07	79.2	28.98	0.705	0.242	185.23	85.0
Champ [61]	3.02E-04	29.84	0.773	0.235	170.20	94.6	29.70	0.759	**0.230**	175.34	76.9
UniAnimate [43]	**2.66E-04**	30.77	**0.811**	0.231	148.08	60.8	29.75	0.762	0.240	161.17	65.2
Ours	2.89E-04	**30.82**	0.799	**0.230**	**146.53**	**60.1**	**29.77**	**0.765**	0.237	**159.20**	**62.1**

Fig. 8. **Qualitative results of ablation studies** on the effectiveness of different conditioning inputs. "w/o Mesh" refers to using only the pose map without incorporating the SMPL-X mesh. "w/o 3D Feature Map" includes the normal map to represent the mesh but excludes its feature maps as conditions. "Ours" represents the full model, which integrates both the pose map and the feature mesh to enhance appearance consistency and temporal coherence.

TABLE II
QUANTITATIVE RESULTS OF ABLATION STUDIES ON DIFFERENT CONDITIONING INPUTS IN THE DIFFUSION MODEL. "W/O MESH" USES ONLY THE POSE MAP AS GUIDANCE, EXCLUDING THE SMPL-X MESH. "W/O 3D FEATURE MAP" INCLUDES ONLY THE NORMAL MAP TO REPRESENT THE 3D MESH. "OURS" INTEGRATES BOTH POSE MAP AND FEATURE MESH.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	MPJPE ↓
w/o. Mesh	28.52	0.691	0.276	182.13	67.2
w/o. 3D Feature Map	29.06	0.744	0.251	164.04	64.5
Ours	**29.77**	**0.762**	**0.237**	**159.20**	**62.1**

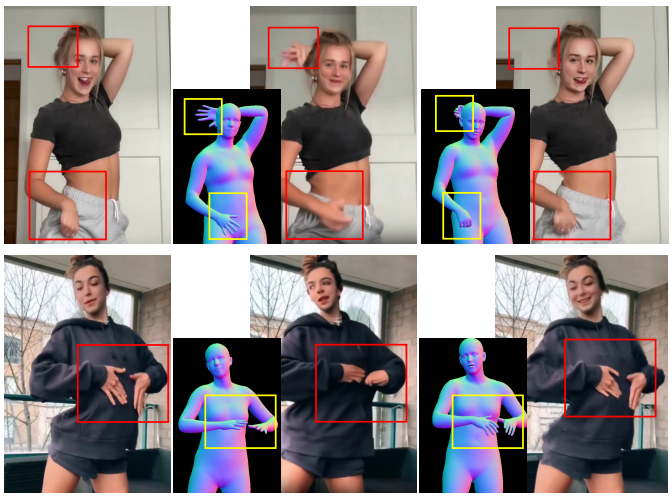

Fig. 9. **Qualitative results of ablation studies** on the usage of different parametric human models. "w/ SMPL" refers to using the SMPL model, which has 10 shape parameters and represents only body shape and pose. "w/ SMPL-X" means the model includes body, hands, and facial expressiveness with 16 shape parameters.

TABLE III
QUANTITATIVE RESULTS OF ABLATION STUDIES ON THE USAGE OF DIFFERENT PARAMETRIC HUMAN MODELS.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	MPJPE ↓
SMPL	29.34	0.772	0.228	74.3
SMPL-X	**29.63**	**0.788**	**0.219**	**62.1**

identity consistency and frame-to-frame temporal consistency in challenging cases. As a result, in the generated video, local patterns may either disappear or exhibit inconsistencies when compared to the reference image. In comparison, AniFeats extracts detailed features from the reference image and projects them onto meshes to render feature maps that serve as guidance in the generation process. This approach not only preserves human identity in the generation but also maintains temporal consistency in character animation, demonstrating a superior capability for generating realistic character animation.

D. Ablation Studies on condition maps

To verify the effectiveness of our 3D feature meshes in AniFeats, we conduct ablation studies on the self-collected dataset. Specifically, we evaluate three configurations: (1) using only the pose map as the conditioning input, (2) incorporating both the pose map and the normal map, which indicates that a 3D mesh is used as condition, and (3) employing both the pose map and the feature mesh, which is the setup of our model.

a) Quantitative results.: As summarized in Table II, we evaluate the impact of different conditioning inputs on generation quality and video fidelity metrics. Removing the SMPL-X mesh and using only the pose map as the conditioning input (w/o Mesh) results in the weakest performance. Incorporating both the pose map and the normal map (w/o 3D Feature Map), which introduces the 3D mesh, improves fidelity and spatial consistency, leading to a 9.06% decrease in LPIPS and a 9.93% reduction in FVD compared to using only the pose map. However, the absence of feature maps still limits overall performance. Our full model (Ours), which employs both the pose map and the feature mesh, further enhances

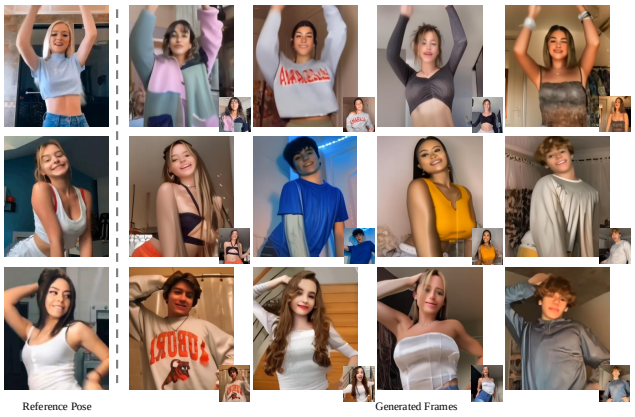

Fig. 10. **Cross-identity reenactment**. The first row shows the original image of a target pose. The following rows show the images of different characters reenacted to the same target pose (right-bottom is the reference image of the identity).

consistency, improving generation quality while achieving a 14.13% decrease in LPIPS and a 12.59% reduction in FVD. These results demonstrate that incorporating feature meshes not only improves spatial accuracy but also strengthens temporal coherence, leading to more realistic and stable video generation.

b) Qualitative Results: As shown in Figure 8, using only the pose map fails to maintain appearance consistency, leading to missing patterns on clothing. Similarly, relying solely on the 3D mesh and pose without feature maps results in incomplete texture reconstruction, as the model lacks sufficient appearance guidance. When incorporating feature maps from the SMPL-X mesh as conditioning inputs, the model accurately captures and transfers patterns from the reference image to the target, preserving fine details and ensuring temporal consistency in the generated video.

E. Ablation Studies on parametric Human Models

a) Quantitative results.: To analyze the effect of different parametric human models, we compare SMPL and SMPL-X as the underlying 3D representations in our framework. As shown in Table III, using SMPL-X consistently improves generation quality in multiple evaluation metrics. Specifically, using SMPL-X instead of SMPL results in a 2.07% improvement in SSIM and a 3.95% reduction in LPIPS.

b) Qualitative results: From the qualitative results shown in Figure 9, using SMPL as the parametric human model leads to inconsistencies in motion transfer, particularly in articulated regions such as hands and facial expressions. This results in noticeable blurring and loss of detail, especially when the character undergoes complex movements. Additionally, the lack of expressive facial modeling in SMPL causes discrepancies between the reference image and the generated frames, reducing overall visual fidelity.

This result can be attributed to SMPL-X's richer expressiveness, which incorporates articulated hands, fingers, and facial expressions in addition to body shape and pose. These additional degrees of freedom enable more precise motion transfer and a better alignment between the generated frames and the reference character.

F. More Results

a) Cross-identity reenactment: To comprehensively evaluate the robustness of AniFeats, we test its cross-identity reenactment by sampling three motion sequences from the TikTok dataset to animate different characters. This setup examines the model's ability to generalize across identities and motions while maintaining high-fidelity synthesis. As shown in Figure 10, our method realistically transfers detailed appearance attributes from reference images and naturally adapts to diverse motion dynamics, achieving stable generation that preserves both temporal consistency and identity fidelity.

b) Temporal Consistency: AniFeats is explicitly designed to ensure temporal consistency in character animation, preserving coherent appearance details even under complex and rapid motions. As illustrated in Figure 11, our method maintains structural integrity and fine-grained details of the reference identity by leveraging 3D feature meshes as explicit guidance. This mitigates common issues in 2D-based generation, such as flickering, drift, or unnatural warping, enabling smooth and stable animations with high visual fidelity.

c) 4D Character Video Generation: By leveraging the 3D feature mesh as explicit guidance, AniFeats enables the generation of dynamic character videos that maintain both spatial and temporal consistency across frames. As illustrated in Figure 12, our approach enables synthesizing character motions while allowing for controlled camera movement. By adjusting the camera pose relative to the 3D feature mesh, our method introduces a level of multi-view consistency into the animation process. However, changing viewpoints on a dynamically moving human remains a challenging problem, as motion and appearance must remain coherent across varying perspectives. Our model takes a step in this direction by incorporating 3D-aware representations, offering a more structured way to handle motion consistency. While traditional methods primarily rely on 2D cues, our approach suggests the potential for improving stability in motion reenactment, with possible applications in virtual avatars, gaming, and immersive content creation.

d) Robustness to Challenging Cases: We further evaluate the robustness of our method under scenarios involving viewpoints without rendered features, detailed facial expressions, and loose or non-rigid clothing, as illustrated in Figure 13. In (a), we examine unrendered-view feature maps, where the 3D feature mesh does not provide explicit texture guidance for the current viewpoint. In such cases, the corresponding region in the feature map is set to zero; nevertheless, our model, aided by auxiliary cues such as DWPose, synthesizes plausible appearances without introducing temporal inconsistency. Subfigure (b) highlights our ability to preserve high-quality facial details, benefiting from the use of SMPL-X, which provides 3,502 dedicated facial vertices and 86 blendshapes, significantly more than the standard SMPL, thus ensuring accurate facial representation even in close-up or highly expressive sequences. Subfigures (b) and (c) further

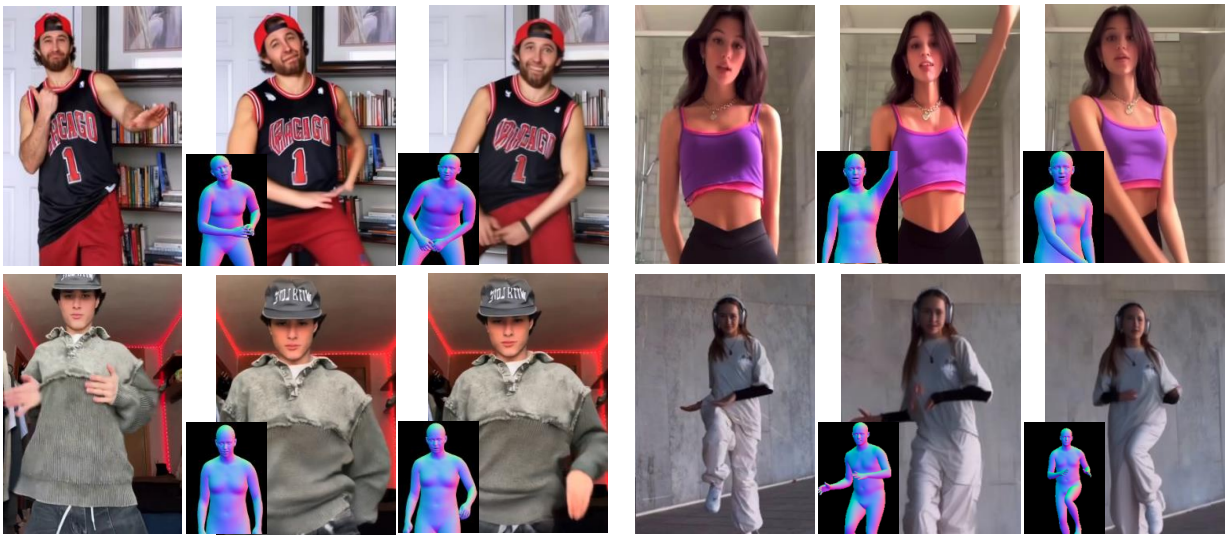

Fig. 11. Qualitative results of our method to demonstrate the temporal consistency. We show two images of the same identity reenacted in two different poses. The consistency of their cloth patterns and identity demonstrates the effectiveness of AniFeats.

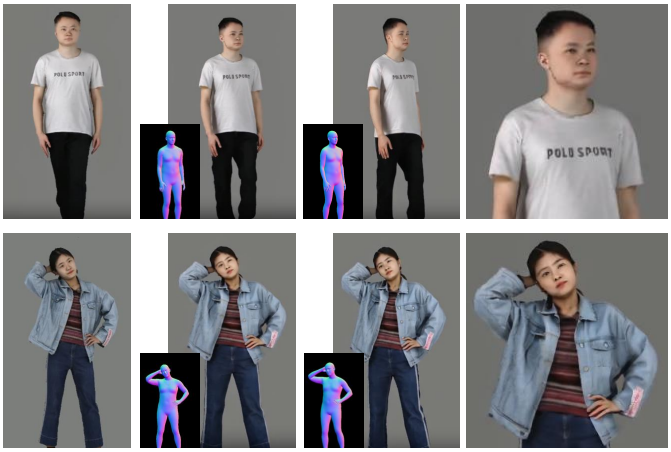

Fig. 12. **4D generation**. AniFeats generates 4D character videos from different viewpoints with a fixed pose.

demonstrate that our approach can faithfully reconstruct loose or non-rigid clothing and out-of-mesh regions (e.g., hair), maintaining structural coherence and temporal stability despite these regions not being explicitly represented in the SMPL-X mesh. These results collectively indicate that our method generalizes effectively to challenging conditions, delivering consistent, high-fidelity animations across diverse viewpoints, clothing types and appearance details.

e) Limitations: Although AniFeats produces visually plausible and consistent character videos from motion sequences and reference images, it may still encounter challenges in generating physically accurate animations. In cases where the estimated mesh is misaligned with the input image, unprojected features can be mapped to incorrect mesh locations, resulting in local artifacts such as distorted geometry or misplaced texture details as shown in Figure 14. A potential future direction is to incorporate the explicit avatar reconstruction like recent Gaussian splatting-based avatar reconstruction methods [1], [17], [27] in the diffusion model for better generation quality.

V. Conclusion

In this paper, we propose AniFeats, a novel framework that explicitly integrates a 3D body model to enhance temporal consistency in character animation. AniFeats extracts visual features from a reference image, projects them onto 3D Feature Meshes constructed based on SMPL-X, and renders these feature maps as structured conditions for video generation. By leveraging the 3D spatial priors, our method establishes a strong correspondence between the reference appearance and the generated frames, ensuring coherence in both identity and motion representation. Extensive experiments demonstrate that AniFeats achieves superior temporal stability and visual fidelity, outperforming existing approaches in generating high-quality, realistic character animations with consistent motion dynamics.

References

[1] Rameen Abdal, Wang Yifan, Zifan Shi, Yinghao Xu, Ryan Po, Zhengfei Kuang, Qifeng Chen, Dit-Yan Yeung, and Gordon Wetzstein. Gaussian shell maps for efficient 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9441–9451, 2024.

[2] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[4] Moore Animate Anyone. Moore animate anyone, 2024.

[5] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Fig. 13. **Robustness to Challenging Cases**. (a) Unrendered viewpoints with missing 3D feature map regions still yield plausible results guided by auxiliary cues. (b) High-quality facial details are preserved even under close-up and expressive poses. (b)(c) Loose clothing and out-of-mesh regions are reconstructed coherently and consistently, maintaining structural and temporal stability.

Fig. 14. **SMPL estimation errors.** Example of mesh–image misalignment leading to incorrect feature projection, resulting in local artifacts such as distorted geometry or misplaced texture details in the generated video.

[6] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. In *NeurIPS*, 2023.

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[8] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *Forty-first International Conference on Machine Learning*, 2023.

[9] Abdelrahman Eldesokey and Peter Wonka. Latentman: Generating consistent animated characters using image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7510–7519, 2024.

[10] Haipeng Fang, Zhihao Sun, Ziyao Huang, Fan Tang, Juan Cao, and Sheng Tang. Dance your latents: Consistent dance generation through spatial-temporal subspace attention guided by motion flow. *arXiv preprint arXiv:2310.14780*, 2023.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[12] RA Guler, Natalia Neverova, and IK DensePose. Dense human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[15] SiT Ho, Chuan Han, Xingzhe Wu, Yueqi Qian, Dahua Lin, and Bo Dai. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 58–68, 2024.

[16] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023.

[17] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 634–644, 2024.

[18] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9352–9364, 2023.

[19] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021.

[20] Suyi Jiang, Haoran Jiang, Ziyu Wang, Haimin Luo, Wenzheng Chen, and Lan Xu. Humangen: Generating human radiance fields with explicit priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12543–12554, 2023.

[21] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *IEEE conference on computer vision and pattern recognition*, 2018.

[22] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion video synthesis with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[23] Jeongho Kim, Min-Jung Kim, Junsoo Lee, and Jaegul Choo. Tcan: Animating human images with temporally consistent pose guidance using diffusion models. *arXiv preprint arXiv:2407.09012*, 2024.

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[26] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4117–4125, 2024.

[27] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. *arXiv preprint arXiv:2407.21686*, 2024.

[28] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[29] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[30] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021.

[31] 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*. PMLR, 2021.
[32] 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*, 2022.
[33] Kripasindhu Sarkar, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Humangan: A generative model of human images. In *2021 International Conference on 3D Vision (3DV)*, pages 258–267. IEEE, 2021.
[34] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image, 2021.
[35] Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4dit: Free-view human video generation with 4d diffusion transformer. *arXiv preprint arXiv:2405.17405*, 2024.
[36] 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.
[37] Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. *arXiv preprint arXiv:2410.10306*, 2024.
[38] Zhengyan Tong, Chao Li, Zhaokang Chen, Bin Wu, and Wenjiang Zhou. Musepose: a pose-driven image-to-video framework for virtual human generation. *arxiv*, 2024.
[39] Chenyang Wang, Zerong Zheng, Tao Yu, Xiaoqian Lv, Bineng Zhong, Shengping Zhang, and Liqiang Nie. Diffperformer: Iterative learning of consistent latent guidance for diffusion-based human video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6169–6179, 2024.
[40] Haimin Wang, Yang Jiang, Linjie Xu, Yue Wang, Xintao Yu, Menglei Zhao, Chen Change Loy, and Bo Dai. Intrinsicavatar: Physically based inverse rendering of dynamic humans from monocular videos via explicit ray tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 184–194, 2024.
[41] Qilin Wang, Zhengkai Jiang, Chengming Xu, Jiangning Zhang, Yabiao Wang, Xinyi Zhang, Yun Cao, Weijian Cao, Chengjie Wang, and Yanwei Fu. Vividpose: Advancing stable video diffusion for realistic human image animation. *arXiv preprint arXiv:2405.18156*, 2024.
[42] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2023.
[43] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *arXiv preprint arXiv:2406.01188*, 2024.
[44] Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, et al. Humanvid: Demystifying training data for camera-controllable human image animation. *arXiv preprint arXiv:2407.17438*, 2024.
[45] 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.
[46] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 16210–16220, 2022.
[47] Jiawei Wu, Yang Jiang, Min Ye, Song Zhang, Xintao Yu, Menglei Zhao, Chen Qian, Chen Change Loy, and Bo Dai. Animatable 3d gaussian: Fast and high-quality reconstruction of multiple human avatars. In *ACM SIGGRAPH 2024 Conference Proceedings*, pages 1–12. ACM, 2024.
[48] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
[49] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 512–523, 2023.
[50] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022.
[51] Zhiqiang Xu, Xinyu Li, Jian Wang, Zhong Zheng, Lizhen Ma, Xu Chen, Chen Change Loy, and Xiaokang Yang. Tech: Text-guided reconstruction of lifelike clothed humans. In *2024 International Conference on 3D Vision (3DV)*, pages 152–162. IEEE, 2024.
[52] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498*, 2023.
[53] Jingyun Xue, Hongfa Wang, Qi Tian, Yue Ma, Andong Wang, Zhiyuan Zhao, Shaobo Min, Wenzhe Zhao, Kaihao Zhang, Heung-Yeung Shum, et al. Follow-your-pose v2: Multiple-condition guided character image animation for stable pose control. *arXiv preprint arXiv:2406.03035*, 2024.
[54] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.
[55] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided human animation from a single image in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
[56] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *arXiv preprint arXiv:2207.06400*, 2022.
[57] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *IEEE International Conference on Computer Vision*, 2021.
[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
[59] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
[60] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
[61] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Qingkun Su, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. *arXiv preprint arXiv:2403.14781*, 2024.

Beijia Lu is currently a master student at Carnegie Mellon University, supervised by Prof. Jun-Yan Zhu. She received her B.S. degree in Mathematics from City University of Hong Kong in 2024. Her research interests lie in computer graphics and computer vision.

Zekai Gu is a research assistant at Hong Kong University of Science and Technology advised by Prof. Yuan Liu. He received a master's degree from the National University of Singapore advised by Prof. Marcelo H Ang Jr. His research interest includes Generative AI and 3D computer vision.

Zhiyang Dou is an MPhil student in Computer Graphics Group at The University of Hong Kong, supervised by Prof. Wenping Wang and Prof. Taku Komura. He received the B. Eng. degree with honors at Shandong University, advised by Prof. Shiqing Xin. His research interest lies in Character Animation, Geometric Modeling and Processing, Simulation, Computer Graphics.

Haotian Yuan is currently pursuing a Bachelor of Engineering degree in Computer Science at the Institute for Interdisciplinary Information Sciences, Tsinghua University. His research interests include 3d computer vision and robotics.

Peng Li is a Ph.D student at HKUST advised by Yike Guo and Wenhan Lou. He received a M.S degree from Tsinghua University in 2023, and a B.S. degree from Xidian University, China, in 2020. His research interest includes depth estimation, 3D reconstruction and generation.

Chenyang Si is an Assistant Professor in School of Intelligence Science and Technology at Nanjing University. He was a research fellow at Nanyang Technological University, Singapore, working with Prof. Ziwei Liu. Prior to this, he worked as a Research Scientist at the Sea AI Lab of Sea Group. He completed Ph.D. degree in 2021 at CASIA, supervised by Prof. Tieniu Tan, co-supervised by Prof. Liang Wang and Prof. Wei Wang.

Yukang Cao is a Postdoctoral Research Fellow at Nanyang Technological University, Singapore, working with Prof. Ziwei Liu. Prior to this, he obtained his Ph.D. degree in the Department of Computer Science, The University of Hong Kong advised by Prof. Kwan-Yee K. Wong, and received his B.Eng from Zhejiang University in 2020.

Yuming Jiang is currently a Research Scientist at Alibaba DAMO Academy. He obtained the Ph.D. degree from Nanyang Technological University, Singapore, supervised by Prof. Ziwei Liu and Prof. Chen Change Loy. He got the bachelor degree in computer science from Yingcai Honors College, University of Electronic Science and Technology of China.

Yuan Liu is an Assistant Professor in the Division of Integrated Systems Design at the Hong Kong University of Science and Technology. He did his PostDoc at Nanyang Technological University, Singapore under the supervision of Prof. Ziwei Liu. He obtained his PhD degree at the University of Hong Kong advised by Prof. Wenping Wang. Prior to that, he obtained both his Master's and Bachelor's degrees from Wuhan University.

Wenping Wang is a Professor of Computer Science & Engineering at Texas A&M University. His research interests include computer graphics, computer visualization, computer vision, robotics, medical image processing, and geometric computing. He has published over 300 technical papers in these fields. He is journal associate editor of Computer Aided Geometric Design (CAGD) and IEEE Transactions on Visualization and Computer Graphics, and has chaired a number of international conferences, including Pacific Graphics 2012, ACM Symposium on Physical and Solid Modeling (SPM) 2013, SIGGRAPH Asia 2013, and Geometry Summit 2019. He received the John Gregory Memorial Award for his contributions in geometric modeling. He is an ACM Fellow and IEEE Fellow.

Ziwei Liu is an Associate Professor at College of Computing and Data Science in Nanyang Technological University, Singapore. Previously, he was a research fellow in Chinese University of Hong Kong with Prof. Dahua Lin and a post-doc researcher in University of California, Berkeley with Prof. Stella Yu. He is the recipient of PAMI Mark Everingham Prize, MIT TR Innovators under 35 Asia Pacific, ICBS Frontiers of Science Award, CVPR Best Paper Award Candidate and Asian Young Scientist Fellowship.