Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance

Shenhao Zhu*[1], Junming Leo Chen*[2], Zuozhuo Dai[3], Zilong Dong[3], Yinghui Xu[2], Xun Cao[1], Yao Yao[1], Hao Zhu†[1], and Siyu Zhu†[2]

[1] Nanjing University, Nanjing, China
[2] Fudan University, Shanghai, China
[3] Alibaba Group, Hangzhou, China
shenhaozhu@smail.nju.edu.cn, leochenjm@gmail.com,
{caoxun, yaoyao, zh}@nju.edu.cn, {zuozhuo.dzz, list.dzl}@alibaba-inc.com,
{xuyinghui, siyuzhu}@fudan.edu.cn

Abstract. In this study, we introduce a methodology for human image animation by leveraging a 3D human parametric model within a latent diffusion framework to enhance shape alignment and motion guidance in current human generative techniques. The methodology utilizes the SMPL(Skinned Multi-Person Linear) model as the 3D human parametric model to establish a unified representation of body shape and pose. This facilitates the accurate capture of intricate human geometry and motion characteristics from source videos. Specifically, we incorporate rendered depth images, normal maps, and semantic maps obtained from SMPL sequences, alongside skeleton-based motion guidance, to enrich the conditions of the latent diffusion model with comprehensive 3D shape and detailed pose attributes. A multi-layer motion fusion module, integrating self-attention mechanisms, is employed to fuse the shape and motion latent representations in the spatial domain. By representing the 3D human parametric model as the motion guidance, we can perform parametric shape alignment of the human body between the reference image and the source video motion. Experimental evaluations on benchmark datasets demonstrate the methodology's superior ability to generate high-quality human animations that accurately capture pose and shape variations. Furthermore, our approach also exhibits superior generalization capabilities on the proposed in-the-wild dataset.

Keywords: Latent Diffusion Model · Human Image Animation · 3D human parametric model · Motion Guidance

1 Introduction

Recent advancements in generative diffusion models, particularly latent diffusion models, have significantly propelled the field of image animation forward [25,

* These authors contributed equally to this work.
† Corresponding Author

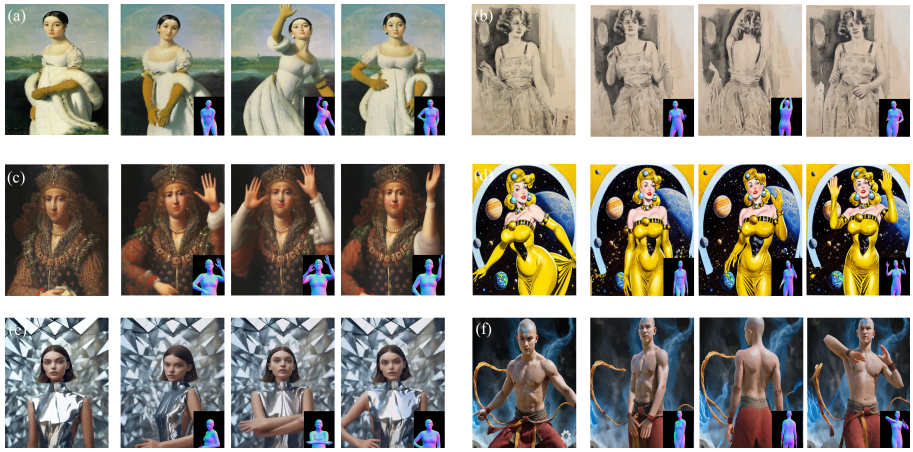


Fig. 1: The proposed methodology showcases a novel ability to produce temporally consistent and visually authentic human image animations by leveraging a reference image and a prescribed motion sequence articulated through 3D human parametric models. Furthermore, it demonstrates an enhanced capacity to refine shape alignment and motion guidance within the resulting videos. This approach facilitates the animation of a wide range of characters, encompassing portraits exhibiting substantial domain variations, such as: (a) a neoclassical oil painting portraying a woman adorned in a white dress and fur coat; (b) a watercolor portrait of a woman; (c) an oil panel painting titled "The Queen of Armenia", as well as characters derived from Text-to-Image diffusion models with the following prompts: (d) a painting of a woman in a yellow dress, heavy metal comic cover art, space theme; (e) a woman in a silver dress posing for a picture, trending on cg society, futurism, with bright blue eyes; (f) a realistic depiction of Aang, the last airbender, showcasing his mastery of all bending elements while in the powerful Avatar State.

61, 63, 66]. These advancements have found broad application in virtual reality experiences, interactive storytelling, and digital content creation, producing a plethora of sophisticated dynamic visual content. Within the realm of human image animation, techniques typically rely on a reference image and motion guidance specific to humans, such as skeletons [19,52], semantic maps [45,56], and dense motion flows [23, 66], to generate controllable human animation videos. In this domain, two predominant approaches prevail: those based on GANs [12, 30] and diffusion models [17, 48].

GAN-based methods [42, 44, 49, 53, 54, 59] commonly employ warping functions to spatially transform the reference image according to input motion for generating sequential video frames. By leveraging the inherent generative visual capabilities of GANs, these methods aim to fill in missing regions and improve visually implausible areas within the generated content. Despite yielding promising results in dynamic visual content generation, GAN-based approaches often encounter challenges in effectively transferring motion, particularly in scenarios involving substantial variations in human identity and scene dynamics between the reference image and the source video motion. This limitation manifests as

unrealistic visual artifacts and temporal inconsistencies in the synthesized content.

Simultaneously, diffusion-based methodologies [14, 23, 23, 52, 62] incorporate the reference image and various dynamics as conditions at both the appearance and motion levels. By harnessing the generative capabilities of latent diffusion models in conjunction with condition guidance, these techniques facilitate the direct generation of human animation videos. Recent diffusion models [23, 52], grounded in data-driven strategies, notably those leveraging CLIP-encoded visual features [35] extracted from reference images pretrained on a vast collection of image-text pairs, in conjunction with diffusion models and temporal alignment modules, have demonstrated effectiveness in addressing the generalization challenges inherent in GAN-based approaches. Therefore, inspired by advanced methodologies such as Animate Anyone [19] and MagicAnimate [56], the aim of this paper is to further optimize shape alignment and pose guidance mechanisms.

In the present study, we propose that the use of a reference image in conjunction with pose guidance, provided through sequential skeleton or dense pose data, may present certain limitations concerning both pose alignment and motion guidance. In a progressive step forward, we advocate for adopting a 3D parametric human model, such as SMPL [26], to encode the 3D geometry of the reference image and extract human motion from the source videos. Firstly, diverging from approaches that segregate the representation of body shape and pose (e.g. dense pose and skeleton methods that primarily emphasize pose), the SMPL model offers a unified representation that encompasses both shape and pose variations using a low-dimensional parameter space. Consequently, in addition to pose information, SMPL model also provides guidance on human geometry-related surface deformations, spatial relationships (e.g. occlusions), contours, and other shape-related features. Secondly, owing to the parametric nature of the SMPL model, we can establish geometric correspondence between the reconstructed SMPL from the reference image and the SMPL-based motion sequences extracted from the source video. This enables us to adjust the parametric SMPL-based motion sequences, thereby enhancing the motion and geometric shape conditions within latent diffusion models. Thanks to the generalization of SMPL to different body shapes, we can effectively handle the substantial variations in body shapes between the reference image and source video.

Incorporating the SMPL model as a guiding framework for both shape and pose within a latent diffusion model, our methodology is structured around three fundamental components: 1) The sequences derived from the SMPL model corresponding to the source video are projected onto the image space, resulting in the generation of depth images, normal maps, and semantic maps that encapsulate essential 3D information. The depth image plays a critical role in capturing the 3D structure of the human form, while the normal map enables the representation of the orientation of the human body. The semantic map aids in accurately managing interactions between different components of the human body during the process of animation generation. 2) Our analysis demonstrates that incorporating skeleton-based motion guidance enhances the precision of guid-

ance information, particularly for intricate movements such as facial expressions and finger movements. As a result, the skeleton is maintained as an auxiliary input to complement the aforementioned maps. 3) In the process of integrating depth, normal, semantic, and skeleton maps through feature encoding, the employment of self-attention mechanisms facilitates the feature maps, processed via self-attention, in learning the representative saliency regions within their respective layers. Such multi-layer semantic fusion enhances the model's capability to comprehend and generate human postures and shapes. Finally, the inclusion of these multi-layer feature embeddings conditioned on a latent video diffusion model leads to precise image animation both in pose and shape.

Our proposed methodology has been evaluated through thorough experiments using the popular TikTok and UBC fashion video datasets, showcasing its effectiveness in improving the quality of human image animation. Furthermore, we have conducted a comparative analysis of our methodology against state-of-the-art approaches on a novel video dataset gathered from diverse real-world scenarios, demonstrating the robust generalization capabilities of our proposed approach.

2 Related Work

Diffusion Models for Image Generation. Diffusion-based models [4, 21, 33, 36, 39, 41] have rapidly emerged as a fundamental component in the domain of text-to-image generation, renowned for their capacity to yield highly promising generative outcomes. To address the considerable computational requirements inherent in diffusion models, the Latent Diffusion Model, as proposed in [39], introduces a technique for denoising within the latent space. This method not only enhances the computational efficiency of these models but also preserves their ability to generate high-fidelity images. Moreover, in the endeavor to enhance control over the visual generation, recent studies such as ControlNet [63], T2I-Adapter [31], and IP-Adapter [58] have delved into the incorporation of supplementary encoder layers. These layers facilitate the assimilation of control signals encompassing aspects such as pose, depth, and edge information and even permit the utilization of images in conjunction with textual prompts. This progression signifies a significant advancement towards more controlled and precise image generation, facilitating the creation of images characterized by not only superior quality but also enriched contextual accuracy and detail.

Diffusion Models for Human Image Animation. The task of animating human images, a significant endeavor within the domain of video generation, aims to seamlessly create videos from one or multiple static images [1, 6, 8, 10, 20, 22, 34, 37, 43–45, 60, 61, 64, 66]. The recent advancements of diffusion models in the text-to-image domain have sparked interest in exploring their utility for animating human images. PIDM [5] introduces a texture diffusion module that is specifically crafted to align the texture patterns of the source and target images closely, thereby enhancing the realism of the resultant animated output. DreamPose [23] capitalizes on the capabilities of the pre-trained Stable Diffusion model

by incorporating both CLIP [35] and VAE [24] for image encoding. It integrates these embeddings with an adapter. Similarly, DisCo [52] innovatively segregates the control of pose and background using dual independent ControlNets [63], providing finer control over the animation process. Animate Anyone [19] utilizes a UNet-based ReferenceNet to extract features from reference images. It includes pose information via a lightweight pose guider. Expanding on the principles introduced by AnimateDiff [14], Animate Anyone integrates a temporal layer into the denoising UNet to enhance temporal coherence. MagicAnimate [56] follows a similar approach but employs a ControlNet tailored for DensePose [13] inputs instead of the more commonly used OpenPose [7] keypoints to provide more precise pose guidance. This paper primarily builds upon esteemed diffusion-based methodologies and advances the optimization of appearance alignment and motion guidance mechanisms. This is achieved by introducing a 3D parametric model for geometric reconstruction of the reference image and motion modeling of the source video sequence.

Pose Guidance in Human Image Animation. DWpose [57] stands out as an enhanced alternative to OpenPose [7], offering more accurate and expressive skeletons. This improvement has proven beneficial for diffusion models in generating higher quality images, with its adoption as a condition signal in various works [9, 19]. The work presented in DensePose [38] aims to establish dense correspondences between an RGB image and a surface-based representation. The SMPL [26] model is a 3D model renowned for its realistic depiction of human bodies through skinning and blend shapes. Its widespread adoption spans fields like human reconstruction [2, 16] and interaction with environments [15, 29]. It also serves as essential ground truth for neural networks in pose and shape analysis [28, 32]. In this paper, we consider SMPL, the 3D parametric model, to reconstruct the poses as well as the shapes from the source video, and obtain more complete condition for appearance alignment and pose guidance.

3 Method

Figure 2 illustrates the overview of our proposed approach. Given an input human image and a reference video depicting a motion sequence, the objective is to synthesize a video where the person in the image replicates the actions observed in the reference video, thereby creating a controllable and temporally coherent visual output. In Section 3.1, we present an overview of the latent diffusion model and the SMPL model to establish the foundational concepts necessary for subsequent discussions. Section 3.2 elaborates on the application of the SMPL model to extract pose and shape information from the source video, enabling the generation of multiple outputs containing pose and shape details. In Section 3.3, these outputs are then utilized to provide multi-layer pose and shape guidance for the human image animation within the latent diffusion model framework. Lastly, Section 3.4 provides a comprehensive exposition of the network architecture along with a detailed description of the training and inference procedures employed in the proposed methodology.

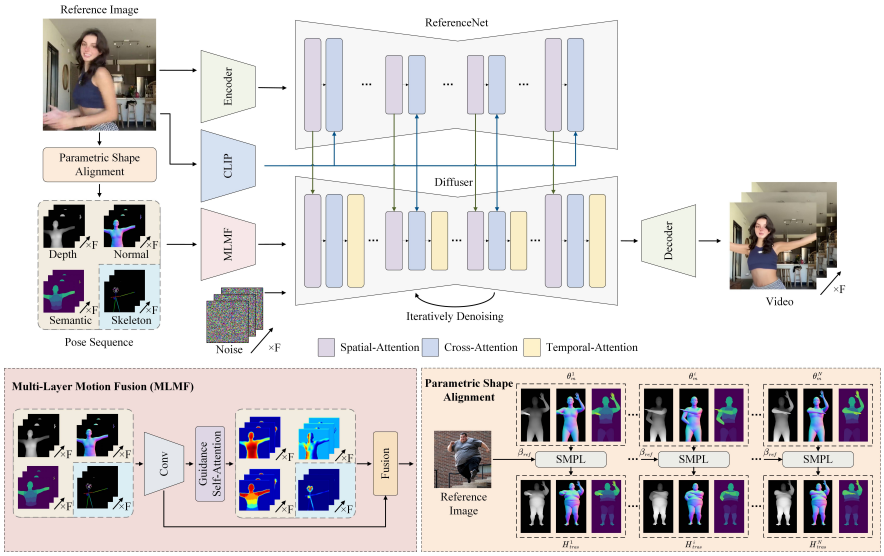


Fig. 2: The overview of our proposed approach. Given an input human image and a reference video depicting a motion sequence, we obtain the pose sequence corresponding to the reference image through Parametric Shape Alignment as 3D motion guidance. MLMF is employed to encode multi-layer 3D-related motion information. Referencenet and Temporal-attention ensure identity consistency and temporal coherence, respectively.

3.1 Preliminary

Latent Diffusion Models. The Latent Diffusion Model (LDM) proposed by Rombach et al. [39] presents a novel approach in the domain of Diffusion Models by incorporating two distinct stochastic processes, namely diffusion and denoising, into the latent space. Initially, a Variational Autoencoder (VAE) [24] is trained to encode the input image into a low-dimensional feature space. Subsequently, the input image I is transformed into a latent representation $\boldsymbol{z}_0 = \mathcal{E}(I)$ using a frozen encoder $\mathcal{E}(\cdot)$. The diffusion process involves applying a variance-preserving Markov process [17, 46, 48] to $\boldsymbol{z}_0$, where noise levels increase monotonically to generate diverse noisy latent representations:

$$\boldsymbol{z}_t = \sqrt{\overline{\alpha}_t}\boldsymbol{z}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}, \quad \epsilon \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{1}$$

Here, $t = 1, ..., T$ signifies the time steps within the Markov process, where T is commonly configured to 1000, and $\overline{\alpha}_t$ represents the noise intensity at each time step. Subsequent to the ultimate diffusion iteration, $q(\boldsymbol{z}_T \mid \boldsymbol{z}_0)$ approximates a Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

The denoising process involves the prediction of noise $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c})$ for each timestep t from $\boldsymbol{z}_t$ to $\boldsymbol{z}_{t-1}$. Here, $\boldsymbol{\epsilon}_\theta$ denotes the noise prediction neural networks, exemplified by architectures like the U-Net model [40], while c_{text} signifies the

text embedding derived from the CLIP mechanism. The loss function quantifies the expected mean squared error (MSE) between the actual noise $\boldsymbol{\epsilon}$ and the predicted noise $\boldsymbol{\epsilon}_\theta$ conditioned on timestep t and noise $\boldsymbol{\epsilon}$:

$$L = \mathbb{E}_{\mathcal{E}(I), c_{\text{text}}, \epsilon \sim \mathcal{N}(0,1), t} \left[\omega(t) \| \epsilon - \epsilon_\theta(z_t, t, c_{\text{text}}) \|_2^2 \right], t = 1, ..., T \tag{2}$$

Here, $\omega(t)$ represents a hyperparameter that governs the weighting of the loss at timestep t. Following training, the model is capable of progressive denoising from an initial state $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ to $\boldsymbol{z}_0$ using a fast diffusion sampler [27, 47]. Subsequently, the denoised $\boldsymbol{z}_0$ is decoded back into the image space I utilizing a frozen decoder $\mathcal{D}(\cdot)$

SMPL model. The SMPL model, as introduced in the work by Loper et al. [26], stands as a prevalent methodology within the domains of computer graphics and computer vision, particularly in the realm of realistic human body modeling and animation. This model is structured around a parametric shape space that effectively captures the nuanced variations in body shape exhibited among individuals alongside a pose space that intricately encodes the articulation of the human body. Through the amalgamation of these two spaces, the SMPL model exhibits the capability to produce anatomically plausible and visually realistic human body deformations spanning a diverse spectrum of shapes and poses. The SMPL model operates on low-dimensional parameters for the pose, denoted as $\theta \in \mathbb{R}^{24\times3\times3}$, and shape, denoted as $\beta \in \mathbb{R}^{10}$. By inputting these parameters, the model generates a 3D mesh representation denoted as $M \in \mathbb{R}^{3\times N}$ with $N = 6890$ vertices. A vertex-wise weight $W \in \mathbb{R}^{N\times k}$ is applied to evaluate the relations between the vertex and the body joints $J \in \mathbb{R}^{3\times k}$, which could then be used for human part segmentation.

3.2 Multi-Layer Motion Condition

SMPL to Guidance Conditions. Given a reference human image I_{ref} and a sequence of driving motion video frames $I^{1:N}$, where N denotes the total number of frames, we obtain the 3D human parametric SMPL model, H_{ref} and $H_{\text{m}}^{1:N}$, respectively, utilizing an existing framework known as 4D-Humans [11]. In order to extract comprehensive visual information from the pixel space, we render the SMPL mesh to obtain 2D representations. This includes encoding depth maps, which contain distance information from each pixel to the camera, which is crucial for reconstructing the 3D structure of the scene. Additionally, we encode normal maps, which depict the surface orientation at each point in the image and can capture geometric information related to the orientation of the human body surface. Furthermore, semantic segmentation maps provide class information for each pixel in the image, enabling accurate handling of interactions between different components of the human body.

Parametric Shape Alignment. As a key of human video generation, animating the reference human image by driving motion sequence while keeping the reference appearance and shape remains challenging. Previous skeleton-based methods use only sparse keypoints to guide the animation and thus ignore the

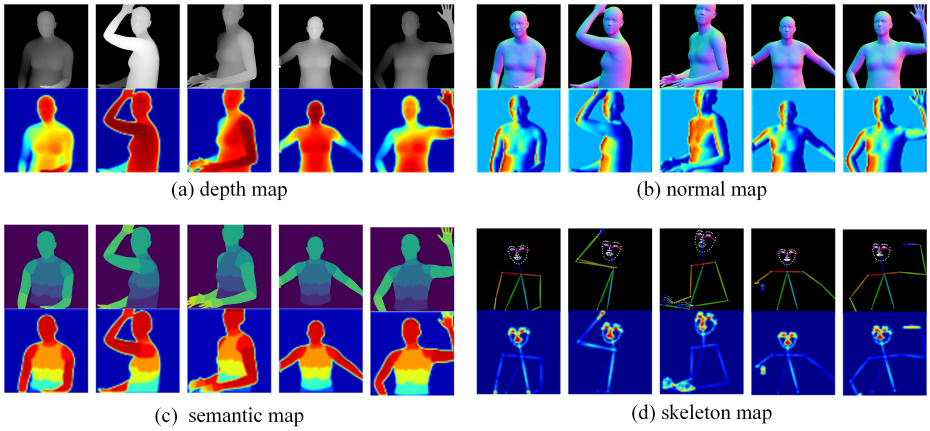


Fig. 3: Multi-layer motion condition and corresponding cross attention maps. Each set of images (above) comprises representations of a depth map, normal map, semantic map, and DWpose skeleton rendered from the corresponding SMPL sequences. The subsequent images (below) illustrate the output of the guidance self-attention.

shape variety of humans. With the parametric human model, our work can easily align both shape and pose between reference human and motion sequence. Given a SMPL model H_{ref} fitted on reference image I_{ref} and a SMPL sequence $H_{\text{m}}^{1:N}$ from N frames driving video $I^{1:N}$, we aim to align the shape β_{ref} of H_{ref} to the pose sequence $\theta_{\text{m}}^{1:N}$ of $H_{\text{m}}^{1:N}$. The aligned SMPL model of each frame $i \in [1, N]$ is then formulated as:

$$H_{\text{trans}}^{i} = \text{SMPL}(\beta_{\text{ref}}, \theta_{\text{m}}^{i}) \tag{3}$$

We then take corresponding conditions rendered from the $H_{\text{trans}}^{1:N}$ to guide video generation on image I_{ref}, which produces pixel-level aligned human shape and enhances the human appearance mapping process in generated animation video.

3.3 Multi-Layer Motion Guidance

Now we have completed the shape-level alignment between the parametric SMPL model reconstructed based on the reference image and the SMPL model sequence of the source video using parametric shape alignment. Subsequently, depth maps, normal maps, and semantic maps are rendered from the aligned SMPL model sequence. Additionally, a skeleton was introduced as an auxiliary input to enhance the representation of intricate movements, such as facial expressions and finger movements. As shown in Figure 3, leveraging latent feature embedding and the self-attention mechanism to be introduced below, we can spatially weight the multi-layer embeddings of human shapes and poses, resulting in the generation of multi-layer semantic fusion as the motion guidance.

Guidance Self-Attention. ControlNet [63] is frequently used in human animation tasks to control generated actions considering additional guidance.

However, introducing multiple guidance condition to ControlNet would result in a computational burden that is unaffordable. In light of this, we are inspired by the advanced work [19] and propose a guidance encoder designed to encode our multilevel guidance. Through this approach, we achieve the simultaneous extraction of information from the guidance while fine-tuning a pre-trained denoising U-Net. The encoder consists of a series of lightweight networks. We assign a guidance network to each guidance condition to encode its features. For each guidance network, we first extract features of the guidance condition through a set of convolutional layers. Considering the presence of multilevel guidance conditions, which involve different characteristics of the human body, a self-attention module is appended after the convolutional layers. This module facilitates the precise capture of corresponding semantic information for each of the multi-layer guidance condition. In particular, Figure 3 illustrates the self-attention map of depth, normal, semantic, and skeleton feature embeddings post-training. The analysis reveals distinct patterns: the depth condition predominantly focuses on the geometric contours of the human figure; the normal condition emphasizes the orientation of the human body; the semantic condition prioritizes the semantic information of different body parts; and the skeleton attention provides detailed constraints on the face and hands.

Multi-Layer Motion Fusion. In order to preserve the integrity of the pre-trained denoising U-Net model, we opt to use a convolutional layer with zero initialization as the output layer to extract the features of each guidance condition. The guidance encoder consolidates the feature embeddings from all the guidance conditions by aggregating them through summation, yielding the ultimate guidance feature denoted as y. This operation can be expressed mathematically as:

$$y = \sum_{i=1}^{N} \mathcal{F}^i(\cdot, \theta^i), \tag{4}$$

where N signifies the total count of guidance conditions incorporated, i is the index of the pose guidance, and θ is the input pose image. Subsequently, the guidance feature is combined with the noisy latent representation before being fed into the denoising fusion module.

3.4 Network

Network Structure. In this section, we present the comprehensive pipeline of our proposed method illustrated in Figure 2. Our approach introduces a video diffusion model that incorporates motion guidance derived from 3D human parametric models. Specifically, we employ the SMPL model to extract a continuous sequence of SMPL poses from the motion data. This conversion results in multilevel guidance that encapsulates both 2D and 3D characteristics, thereby enhancing the model's comprehension of human shape and pose attributes. To integrate this guidance effectively, we introduce a motion embedding module that incorporates the multilayer guidance into the model. The multiple latent embeddings of motion guidance are individually refined through self-attention

mechanisms and subsequently fused together using a multi-layer motion fusion module. Furthermore, we encode the reference image using a VAE encoder and a CLIP image encoder. To ensure video consistency, we utilize two key modules: the ReferenceNet and the temporal alignment module. The VAE embeddings are fed into the ReferenceNet, which is responsible for maintaining consistency between the characters and background in the generated video and the reference image. Additionally, we employ a motion alignment strategy that utilizes a series of motion modules to apply temporal attention across frames. This process aims to mitigate any discrepancies between the reference image and the motion guidance, thus enhancing the overall coherence of the generated video content.

Training. The training process consists of two distinct stages. During the initial stage, training is conducted solely on images, with the exclusion of motion modules within the model. We freeze the weights of the VAE encoder and decoder, as well as the CLIP image encoder, in a frozen state while allowing the Guidance Encoder, Denoising U-Net, and reference encoder to be updated during training. To initiate this stage, a frame is randomly selected from a human video to serve as a reference, and another image from the same video is chosen as the target image. The multi-layer guidance extracted from the target image is then input into the Guidance network. The primary objective of this stage is to generate a high-quality animated image utilizing the multilevel guidance derived from the specific target image.

In the second training phase, the incorporation of the motion module serves to augment the temporal coherence and fluidity of the model. This module is initialized with the pre-existing weights obtained from AnimateDiff. A video segment comprising 24 frames is extracted and employed as the input data. During the training of the motion module, the Guidance Encoder, Denoising U-Net, and reference encoder, which were previously trained in the initial stage, are held constant.

Inference. During the inference process, animation is performed on a specific reference image by aligning the motion sequences extracted from in-the-wild videos or synthesized ones. Parametric shape alignment is utilized to align the motion sequence with the reconstructed SMPL model derived from the reference image at the pixel level, providing a basis for animation. To accommodate the input of a video clip comprising 24 frames, a temporal aggregation technique [50] is employed to concatenate multiple clips. This aggregation method aims to produce a long-duration video output.

4 Experiments

4.1 Implementations

Dataset. We have curated an in-the-wild dataset comprising approximately 5,000 high-fidelity authentic human videos sourced from reputable online repositories, encompassing a total of 1 million frames. The dataset is segmented as follows: Bilibili (2540 videos), Kuaishou (920 videos), Tiktok & Youtube (1438

Method	L1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID-VID ↓	FVD ↓
MRAA	3.21E-04	29.39	0.672	0.296	54.47	284.82
DisCo	3.78E-04	29.03	0.668	0.292	59.90	292.80
MagicAnimate	3.13E-04	29.16	0.714	0.239	21.75	179.07
Animate Anyone	-	29.56	0.718	0.285	-	171.9
Ours	3.02E-04	29.84	0.773	0.235	26.14	170.20
Ours*	**2.94E-04**	**29.91**	**0.802**	**0.234**	**21.07**	**160.82**

Table 1: Quantitative comparisons on Tiktok dataset. * indicates that the proposed approach is fine-tuned on the Tiktok training dataset.

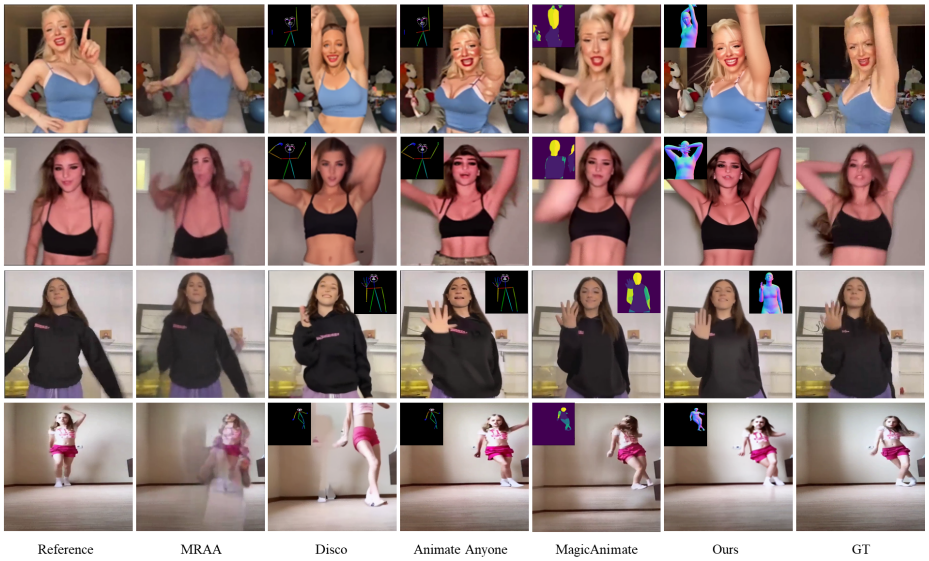


Fig. 4: Qualitative comparisons between our and the state-of-the-art approaches on TikTok and proposed unseen dataset.

videos), and Xiaohongshu (430 videos). These videos feature individuals of varying ages, ethnicities, and genders, depicted in full-body, half-body, and close-up shots set against diverse indoor and outdoor backdrops. In order to enhance our model's capacity to analyze a wide range of human movements and attire, we have included footage of dancers showcasing various dance forms in diverse clothing styles. In contrast to existing datasets characterized by pristine backgrounds, our dataset capitalizes on the diversity and complexity of backgrounds to aid our model in effectively distinguishing the foreground human subjects from their surroundings. To maintain fairness and align with established benchmarks in the field of image animation, the identical test set as utilized in MagicAnimate [56] has been employed for TikTok evaluation.

Implementation. Our experiments were facilitated by the computational power of 8 NVIDIA A100 GPUs. The training regimen is structured in two



Fig. 5: Qualitative comparisons between our and the state-of-the-art approaches on UBC fashion video datasets.

phases: initially, we processed individual video frames—sampling, resizing, and center-cropping them to a uniform resolution of 768x768 pixels. This stage spanned 60,000 steps with a batch size of 32. Subsequently, we dedicated attention to the temporal layer, training it over 20,000 steps with sequences of 24 frames and a reduced batch size of 8 to enhance temporal coherence. Both stages adhered to a learning rate of 1e-5. During inference, to achieve continuity over extended sequences, we employed a temporal aggregation method, which facilitated the seamless integration of results from distinct batches, thereby generating longer video outputs.

4.2 Comparisons

Baselines. We perform a comprehensive comparison with several state-of-the-art methods for human image animation: (1) MRAA [45] is state-of-the-art GAN-based animation approaches, which estimate optical flow from driving sequences to warp the source image and then inpaint the occluded regions using GAN models. (2) DisCo [52] is the state-of-the-art diffusion-based animation method that integrates disentangled condition modules for pose, human, and background into a pretrained diffusion model to implement human image animation. (3) MagicAnimate [56] and Animate Anyone [19] are newer diffusion-based animation methods that employ more complex model structure and train on more general data which makes them perform quite well on TikTok dataset. In all the qualitative (video and visual) comparative experiments, we employed the open-source implementation of Animate Anyone from MooreThreads and MagicAnimate from the original authors. For all quantitative experimental comparisons, we directly referenced the relevant statistics from the original literature.

Evaluation Metrics. Our evaluation methodology adheres to the established metrics utilized in existing research literature. Specifically, we assess both single-frame image quality and video fidelity. The evaluation of single-frame

Animate Anyone: https://github.com/MooreThreads/Moore-AnimateAnyone

MagicAnimate: https://github.com/magic-research/magic-animate

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
MRAA	-	0.663	0.311	321.5
DisCo	28.32	0.694	0.286	295.4
MagicAnimate	29.14	0.713	0.235	193.6
Animate Anyone	28.86	0.727	0.242	199.2
Ours	**29.87**	**0.806**	**0.211**	**173.5**

Table 2: Quantitative comparisons on proposed in-the-wild dataset.

quality incorporates metrics such as the L1 error, Structural Similarity Index (SSIM) [55], Learned Perceptual Image Patch Similarity (LPIPS) [65], and Peak Signal-to-Noise Ratio (PSNR) [18]. Video fidelity, on the other hand, is evaluated through the Frechet Inception Distance with Fréchet Video Distance (FID-FVD) [3] and Fréchet Video Distance (FVD) [51].

Evaluation on Benchmark Dataset. Table 1 presents a concise quantitative analysis of various methods evaluated on the TikTok dataset, focusing on key metrics such as L1 loss, PSNR, SSIM, LPIPS, FID-VID, and FVD. The proposed method, both in its original and fine-tuned (* indicated) forms, demonstrates superior performance across most metrics, particularly highlighting its effectiveness in achieving lower L1 loss, higher PSNR and SSIM values, and reduced LPIPS, FID-VID, and FVD scores. Notably, the fine-tuned version of our approach shows the best overall results, indicating the benefits of dataset-specific optimization. Figure 4 and Figure 5 provide additional qualitative comparisons on such benchmarks.

Evaluation on Proposed In-the-wild Dataset. In order to further compare the robustness of various methods, distinct from datasets such as TikTok and UBC fashion that exhibit domain proximity, we have constructed a testing dataset comprising 100 high-fidelity authentic human videos sourced from online repositories. These videos exhibit significant variations in the shape, pose, and appearance of the individuals depicted. The statistical comparison presented in Table 2 collectively illustrates the efficacy of the proposed approach in generalizing to various domains.

Cross ID Animation. As shown in Figure 6, a comparative analysis is conducted between our approach and state-of-the-art baseline methods for cross-identity animation, specifically focusing on the task of animating reference images through motion sequences derived from disparate videos.

Multi-view Animation. Although our method may not match the direct rendering of 3D human representations for consistent novel views, we employ a sequence of SMPLs for multi-view consistent motion guidance and utilize generative models to achieve satisfactory multi-view results. More details are provided in the supplementary material.

4.3 Limitations and Future Works

Despite the enhanced shape alignment capabilities and motion guidance offered by the human parametric SMPL model, its modeling capacity for faces and

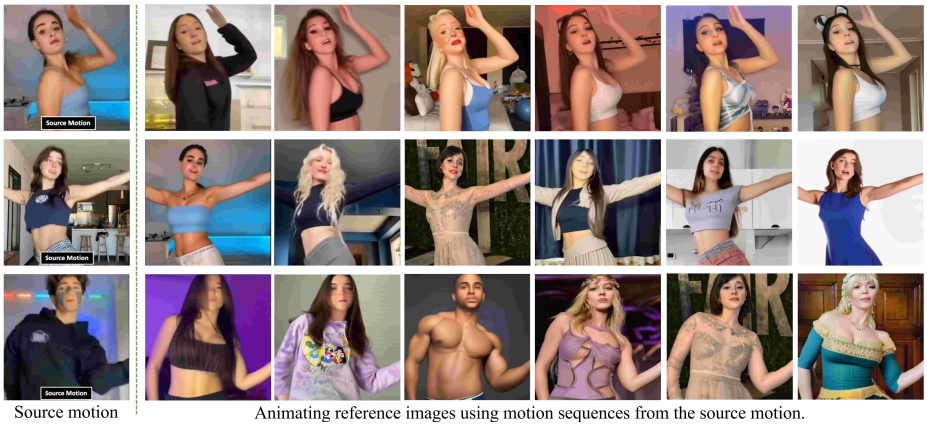


Fig. 6: The demonstration of cross ID animation from the proposed approach.

hands is limited. Consequently, the guidance effect for faces and hands does not match the efficacy of feature-based methods. This limitation underpins the incorporation of DWpose as an additional constraint for facial and hand modeling. As illustrated in Figure 3, the self-attention mechanism further amplifies the saliency of faces and hands within the skeleton map. However, it is important to note that since SMPL and DWpose are solved independently, a potential discrepancy in consistency between them exists. Although this discrepancy did not manifest significantly in our experiments, it nonetheless introduces a potential source of error.

5 Conclusion

This paper introduces a novel approach to human image animation that integrates the SMPL 3D parametric human model with latent diffusion models, aiming to enhance pose alignment and motion guidance. By leveraging the unified representation of shape and pose variations offered by the SMPL model, along with depth, normal, and semantic maps, this method further improves the ability to capture realistic human movements and shapes of previous techniques. The inclusion of skeleton-based motion guidance and self-attention mechanisms for feature map integration further refines the animation process, enabling the creation of dynamic visual content that more accurately reflects human anatomy and movement. Experimental validation across various datasets confirms the effectiveness of this approach in producing high-quality human animations, showcasing its potential to advance digital content creation in fields requiring detailed and realistic human representations.

Acknowledgement. This work was supported by the NSFC grant 62441204.

References

1. AlBahar, B., Saito, S., Tseng, H.Y., Kim, C., Kopf, J., Huang, J.B.: Single-image 3d human digitization with shape-guided diffusion. In: SIGGRAPH Asia (2023)
2. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
3. Balaji, Y., Min, M.R., Bai, B., Chellappa, R., Graf, H.P.: Conditional gan with discriminative filter generation for text-to-video synthesis. In: Proceedings of theInternational Joint Conference on Artificial Intelligence (2019)
4. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
5. Bhunia, A.K., Khan, S., Cholakkal, H., Anwer, R.M., Laaksonen, J., Shah, M., Khan, F.S.: Person image synthesis via denoising diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
6. Cao, Y., Cao, Y.P., Han, K., Shan, Y., Wong, K.Y.K.: Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. arXiv preprint arXiv:2304.00916 (2023)
7. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: 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. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
9. Feng, M., Liu, J., Yu, K., Yao, Y., Hui, Z., Guo, X., Lin, X., Xue, H., Shi, C., Li, X., et al.: Dreamoving: A human dance video generation framework based on diffusion models. arXiv preprint arXiv:2312.05107 (2023)
10. Fu, J., Li, S., Jiang, Y., Lin, K.Y., Qian, C., Loy, C.C., Wu, W., Liu, Z.: Stylegan-human: A data-centric odyssey of human generation. arXiv preprint **arXiv:2204.11823** (2022)
11. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. arXiv preprint arXiv:2305.20091 (2023)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in Neural Information Processing Systems (2014)
13. Guler, R., Neverova, N., DensePose, I.: Dense human pose estimation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
14. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
15. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3d scenes by learning human-scene interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
16. He, T., Xu, Y., Saito, S., Soatto, S., Tung, T.: Arch++: Animation-ready clothed human reconstruction revisited. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems (2020)

18. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: International Conference on Pattern Recognition. IEEE (2010)
19. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117 (2023)
20. Hu, S., Hong, F., Pan, L., Mei, H., Yang, L., Liu, Z.: Sherf: Generalizable human nerf from a single image. arXiv preprint arXiv:2303.12791 (2023)
21. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023)
22. Jiang, S., Jiang, H., Wang, Z., Luo, H., Chen, W., Xu, L.: Humangen: Generating human radiance fields with explicit priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12543–12554 (2023)
23. Karras, J., Holynski, A., Wang, T.C., Kemelmacher-Shlizerman, I.: Dreampose: Fashion video synthesis with stable diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
24. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
25. Li, B., Rajasegaran, J., Gandelsman, Y., Efros, A.A., Malik, J.: Synthesizing moving people with 3d control. arXiv preprint arXiv:2401.10889 (2024)
26. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
27. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in Neural Information Processing Systems (2022)
28. Lu, J., Lin, J., Dou, H., Zhang, Y., Deng, Y., Wang, H.: Dposer: Diffusion model as robust 3d human pose prior. arXiv preprint arXiv:2312.05541 (2023)
29. Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.J.: Learning to dress 3d people in generative clothing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
30. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
31. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
32. Mu, J., Sang, S., Vasconcelos, N., Wang, X.: Actorsnerf: Animatable few-shot human rendering with generalizable nerfs. arXiv preprint arXiv:2304.14401 (2023)
33. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
34. Prokudin, S., Black, M.J., Romero, J.: Smplpix: Neural avatars from 3d human models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1810–1819 (2021)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR (2021)
36. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)

37. Ren, Y., Li, G., Liu, S., Li, T.H.: Deep spatial transformation for pose-guided person image generation and animation. IEEE Transactions on Image Processing **29**, 8622–8635 (2020)
38. Riza Alp Guler, Natalia Neverova, I.K.: Densepose: Dense human pose estimation in the wild. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
40. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. Springer (2015)
41. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems (2022)
42. Sarkar, K., Liu, L., Golyanik, V., Theobalt, C.: Humangan: A generative model of humans images (2021)
43. Sarkar, K., Mehta, D., Xu, W., Golyanik, V., Theobalt, C.: Neural re-rendering of humans from a single image (2021)
44. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in Neural Information Processing Systems **32** (2019)
45. Siarohin, A., Woodford, O.J., Ren, J., Chai, M., Tulyakov, S.: Motion representations for articulated animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
46. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. PMLR (2015)
47. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2020)
48. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2020)
49. Tian, Y., Ren, J., Chai, M., Olszewski, K., Peng, X., Metaxas, D.N., Tulyakov, S.: A good image generator is what you need for high-resolution video synthesis. arXiv preprint arXiv:2104.15069 (2021)
50. Tseng, J., Castellon, R., Liu, K.: Edge: Editable dance generation from music. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
51. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
52. Wang, T., Li, L., Lin, K., Zhai, Y., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for referring human dance generation in real world. arXiv preprint arXiv:2307.00040 (2023)
53. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)

54. Wang, Y., Bilinski, P., Bremond, F., Dantcheva, A.: G3an: Disentangling appearance and motion for video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
55. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
56. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. arXiv preprint arXiv:2311.16498 (2023)
57. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
58. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
59. Yoon, J.S., Liu, L., Golyanik, V., Sarkar, K., Park, H.S., Theobalt, C.: Pose-guided human animation from a single image in the wild (2021)
60. Yoon, J.S., Liu, L., Golyanik, V., Sarkar, K., Park, H.S., Theobalt, C.: 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)
61. Yu, W.Y., Po, L.M., Cheung, R.C., Zhao, Y., Xue, Y., Li, K.: Bidirectionally deformable motion modulation for video-based human pose transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
62. Zhang, J., Yan, H., Xu, Z., Feng, J., Liew, J.H.: Magicavatar: Multimodal avatar generation and animation. arXiv preprint arXiv:2308.14748 (2023)
63. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
64. Zhang, P., Yang, L., Lai, J.H., Xie, X.: Exploring dual-task correlation for pose guided person image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
65. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
66. Zhao, J., Zhang, H.: Thin-plate spline motion model for image animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)