High-Fidelity Character Animation: Generating Coherent and Controllable Motion Videos from Static Images

Yongming Huang[a,b] , Zhuojun Xia[a,b] , Can Bu[a,b]

[a]School of Automation, Southeast University, Nanjing 210096, China
[b] Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, 210096, China
{huang_ym, 213212931, bucan}@seu.edu.cn

Abstract

This paper focuses on video-driven character animation, which aims to generate temporally continuous and appearance-consistent role animation videos from static images using pose signals. Although diffusion models have become the mainstream solution, existing methods still face significant challenges in cross-frame detail consistency, motion controllability, and long-term temporal modeling. To address these limitations, we propose High-Fidelity Character Animation—a innovative diffusion-based video generation framework. First, we design an Identity Preservation Module to enhance the consistency of identity features in the process of video generation. Second, we propose Domain Specific Adaptation Optimization to enhance generalization across diverse scenarios and complex poses. Fianlly, we develop a Spatio-temporal Optimization with Detail Enhancement to improve the representation of key region details and ensure temporal consistency across long video sequences. Experiments show that the performance of our model on two benchmark datasets (TikTok dataset and TED-talks dataset) achieves the state-of-the-art, exceeding the strongest baselines by 22.8% and 30.2% respectively in terms of video quality metrics.

1. Introduction

Video-driven character animation aims to generate temporally consistent and visually consistent character animation videos from static images by analyzing pose information in input videos [32, 41]. At present, it has important application prospects in virtual anchor, game NPC generation and other fields [3, 19]. The early research of this technology comes from the development of GAN, such as using latent space interpolation to realize simple character action generation [10] and using pix2pix framework to complete pose transfer [18]. However, the images or videos generated by GAN-based methods have the problems of poor temporal coherence, local distortion, and blurred details in occluded areas [33].

In recent years, some studies based on diffusion model framework to solve character animation video tasks have achieved some results [15, 28]. However, current diffusion-based approaches [2, 12, 22] struggle to capture fine-grained details, leading to distortions in appearance features during temporal sequences—particularly with subtle textures. These methods also fail to maintain stability during long-term motion sequences due to inherent frame-wise processing that neglects temporal consistency. In addition, these works often rely on the CLIP encoder to encode the reference appearance [27], which is an encoder trained to match the semantic features of the text in the diffusion model, leading to the loss of fine features in the feature encoding [29].

In this paper, we propose High-Fidelity Character Animation, a video framework for character animation, to ensure high-fidelity character appearance features and long-term temporal modeling consistency in the generated video. First, we inherit the network design and pre-trained weights from Stable Diffusion (SD) [15, 28], and then model the temporal consistency between different video frames by inserting a temporal attention layer into the diffusion network [12]. Secondly, we design an Identity Preservation Module into our diffusion network to maintain high-fidelity character detail textures. Then, a strategy of Domain Specific Adaptation Optimization is designed to enhance the generalization ability of the model in various scenes and complex poses. Finally, based on the above methods, we introduce a Spatio-temporal Optimization with detail enhancement to further improve the high fidelity of character animation.

We summarize our contributions as follows: (1) In this paper we propose a High-Fidelity Character Animation framework, which can maintain the spatial and temporal consistency of the character's appearance details in the generated video. At the same time, the application of temporal consistency modeling can generate high-definition character animation video with arbitrary duration without temporal jitter and flicker. (2) High-Fidelity Character Animation achieves state-of-the-art performance on multiple metrics across the TikTok [19] and TED-Talks[33] benchmark datasets.

2. Related Work

2.1. Diffusion Models for Image Generation

The remarkable success of diffusion models has greatly promoted various research advances in text-to-image generation, and it has become the mainstream method for text-to-image generation tasks [15, 29]. The breakthrough of diffusion model in the field of image generation began with the proposal of DDPM [15], which achieved high-quality image synthesis through the reverse denoising process. Key advances include Latent Diffusion Models [28], which introduce latent space dimensionality reduction techniques to significantly reduce computational complexity. Its core architecture adopts a U-Net variant that fuses text conditions through cross-attention layers, enabling efficient image generation on GPUs [6]. In addition, ControlNet [47] realizes spatial alignment condition injection through zero convolution layer, which lays the foundation for pose control. The GLIDE model [25] implements the text-image semantic alignment mechanism, which can generate images with semantic consistency. However, it is difficult to retain high-frequency texture details in semantic level feature extraction, which becomes the bottleneck of character animation. These works provide technical reserves for video generation, but have limitations in fine-grained appearance preservation [2, 42].

2.2. Diffusion Models for Animation

With the maturity of text-to-image technology, researchers begin to expand into the temporal dimension and study the diffusion model of character animation video generation [6, 35]. AnimateDiff [12] achieves the video extension capability of pre-trained models with a pluggable motion module. In terms of condition control, VideoComposer [42] incorporates image conditions into the diffusion process, and Gen-2 [6] shows the possibility of multi-condition fusion generation. Recent works such as Text2Video-Zero [21] explore zero-shot generation without video data fine-tuning, and FollowYourPose [24] achieves accurate control of motion trajectory through pose temporal encoding. MimicMotion model [22] further expands the application of diffusion model in character animation generation. In

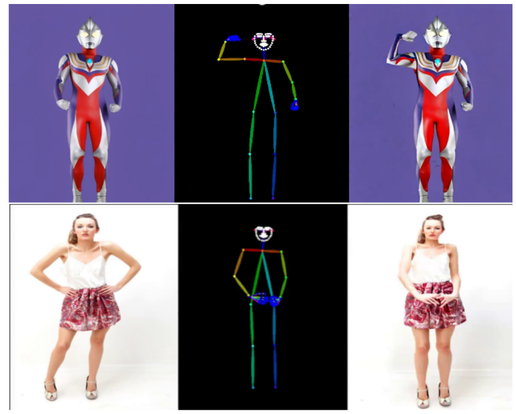


Figure 1. Given a reference image (the leftmost image in each group) and a motion pose sequence (the middle pose image in each group), High-Fidelity Character Animation can generate a controllable animation video results for any character that are consistent with the appearance characteristics of the reference character and stable in time series.

this model, a confidence-aware posture guidance strategy is introduced to realize the fine control of human motion posture. Momask model [4] uses the generative mask modeling technology to intelligently fill and repair the occluded part, making the generated 3D human motion more complete and realistic. These research advances provide important technical reserves for image-to-video generation tasks. However, although the introduction of these models and methods has greatly improved the quality of character animation video generation, these methods [12, 40, 42] still face the problems of inconsistent character identity characteristics across frames, inconsistent and stable generation of long videos, and excessive computational complexity and high computing power requirements.

3. Method

This paper aims to achieve pose-guided image-video character animation generation. Given a reference image $I_{ref} \in \mathbb{R}^{H \times W \times 3}$ and a target pose sequence $P = \{p_t\}_{t=1}^{T}$, the model must generate a character animation video that complies with physical laws and has consistent identity characteristics $V = \{v_t\}_{t=1}^{T}$, as shown in Figure 1. However, current diffusion models generally use CLIP encoders [27, 29] or other appearance encoders to encode reference images. The extracted semantic features are sparse, and it is difficult to effectively capture high-frequency detail information such as clothing texture and hair, resulting in shortcomings in the generated video detail representation and inter-frame consistency [6, 29, 44]. Therefore, we propose an improved Reference ProcessingNet based on the existing research results [2, 17]. Then we introduce an Identity Preservation Module to enhance the consistency of identity features in the process of video generation. In addition, we also design a Domain-specific Adaptation Optimization to improve the generalization ability of the model in different application scenarios by fine-tuning . To further solve the shortcomings of the original diffusion framework in modeling the temporal dependence between frames, we introduce the temporal layer into the diffusion network structure to better model the temporal consistency of character animation videos [7, 12, 42].

The overall pipeline of the High-Fidelity Character Animation proposed in this paper is shown in Figure 2. Initially, the reference image undergoes preprocessing, which includes color space transformation and geometric enhancement. Subsequently, the processed image is fed into the VAE encoder to extract latent space features denoted as $Z_{ref} \in \mathbb{R}^{h\times w\times c}$, and into the CLIP encoder to obtain high-level semantic features. These two feature representations are then jointly input into a modified Reference ProcessingNet in order to generate fine-grained conditional embeddings.Next, the pose sequence undergoes feature encoding via a lightweight PoseNet model , producing conditional guidance signals $f_{pose}$for action generation. Unlike prior ControlNet-based approaches, this design minimizes computational overhead while maintaining efficiency .Within the Denoising UNet, reference image features and pose features fuse through cross-attention: spatial attention layer ensures alignment of facial features while temporal attention layer maintains temporal coherence across frames. The resulting video clips are then reconstructed by the VAE decoder. To resolve temporal discontinuities in generated video segments, Gaussian-weighted fusion is applied to overlapping regions , ensuring smooth inter-frame transitions and coherent animation output.

3.1. Identity Preservation Module

Within the diffusion model framework, single-layer spatial attention [17, 35, 44] in diffusion models causes progressive distortion of facial and clothing features during cross-frame propagation, as CLIP encoder or other image encoders are limited in their ability to capture high-frequency details. Inspired by the ControlNet hierarchical condition injection mechanism [47], we propose a feature decoupling framework that separates identity into character feature and clothing attributes for independent constraint.

Character feature alignment constraint. Inspired by the ability of Arcface model to enhance discrimination between classes [5, 30], the pre-trained ArcFace network is used to extract character embedding feature vectors to ensure angular cosine space consistency between generated and reference frames. The constraint process is defined as follows:

$$\mathcal{L}_{face} = 1 - \frac{f_{ref}\cdot f_{gen}}{||f_{ref}||f_{gen}|} + \lambda||f_{ref} - f_{gen}||_2^2 \ , \quad (1)$$

where $||.||_2$ is the L2 loss, f_{ref} and f_{gen} respectively represent the embedding vectors of the characters in the reference image and the generated image.λ is the weight coefficient used to balance the influence between the cosine similarity and the Euclidean distance.

Clothing Attribute Constraint. We adapt the Gram matrix method from Neural Style Transfer [8, 20], utilizing the VGG19 network [23, 34, 46] to extract the style statistical information in feature maps. Specifically, shallow features capture high-frequency details (e.g., silk luster), while mid-level features model structural elements (e.g., collar contours). For the **l**-th layer feature map $F^l \in \mathbb{R}^{C\times H\times W}$, the Gram matrix is defined as:

$$G_{ij}^l = \frac{1}{CHW}\sum_{h=1}^{H}\sum_{w=1}^{W} F_{i,h,w}^l F_{j,h,w}^l \ . \quad (2)$$

Thus, the clothing style constraint term is obtained as:

$$\mathcal{L}_{style} = \sum_{l\in \text{conv1\_2,conv2\_2}} ||G_{ref}^l - G_{gen}^l||_F \ , \quad (3)$$

where G_{ref}^l , G_{ref}^l represents the Gram matrix of the original character clothing attributes and the target character clothing attributes, $||.||_F$ represents the Frobenius norm, which is used to quantify the style difference. VGG19's conv1_2 captures texture features, while conv2_2 models fold morphology.

Through the above design, we can preserve high-frequency texture priors and avoids CLIP's semantic sparsity limitation through detail-enhanced representation learning.

3.2. Domain Specific Adaptation Optimization

Due to the domain limitations of pretrained diffusion models,the diffusion framework shows limited adaptability to rapid action changes and complex clothing structures . We jointly optimize clothing types and action-posture types, introduce a Domain Specific Adaptation Optimization, and further fine-tune it using the UCF101 dataset [36].

Adaptation method for specific clothing types. To enhance specific clothing type adaptation, we extract feature vectors f_{cloth} for 12 clothing categories from 3,200 UCF101 dataset samples using Reference ProcessingNet, then fuse them with backbone features via weighted integration:

$$\begin{aligned} \alpha &= \sigma(v^T[f_{main}; f_{cloth}), \\ f_{fusion} &= f_{main} + \alpha\cdot(W_a f_{cloth}), \end{aligned} \quad (4)$$

where the weight coefficient α is used to dynamically adjust the influence intensity of different clothing features on the generation process; σ represents the Sigmoid function; v is a learnable parameter vector; W_a is a learnable projection matrix; f_{main} represents the clothing feature vector.

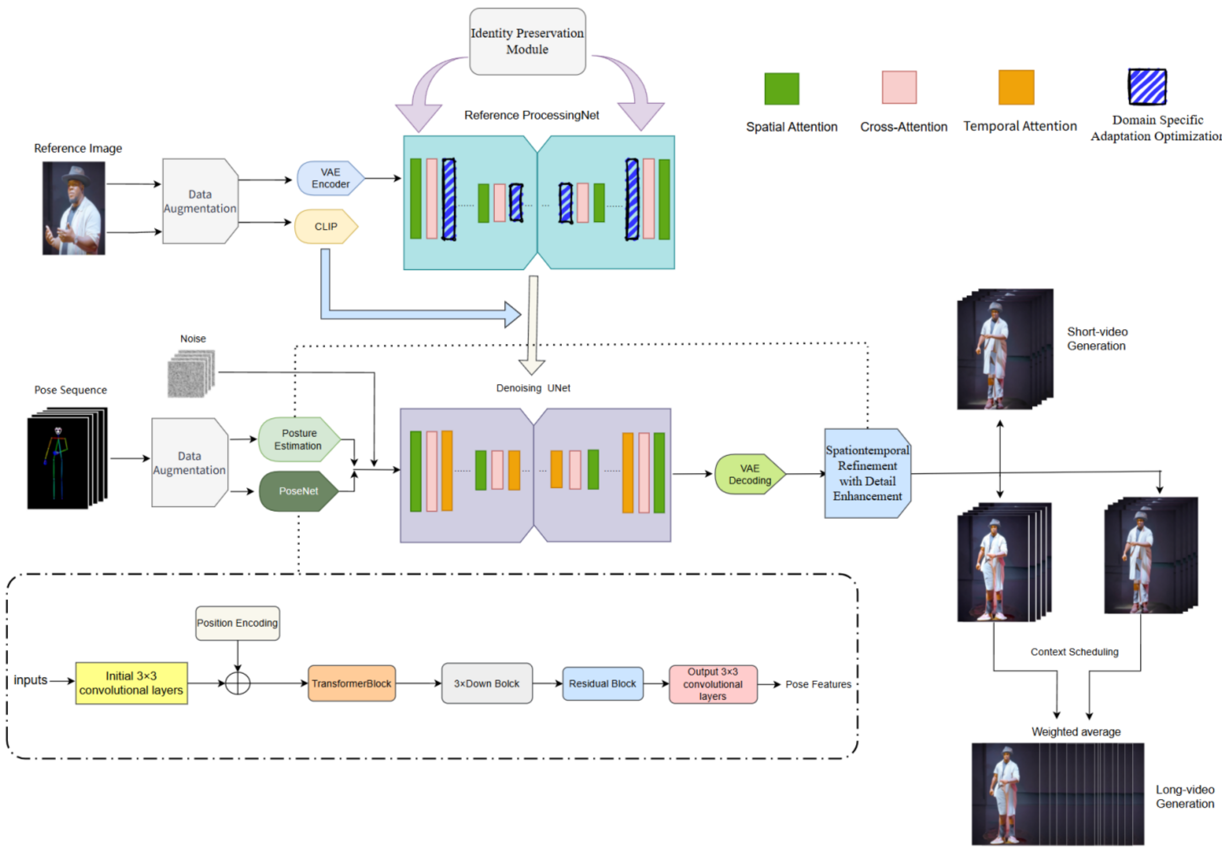


Figure 2. The pipeline of our **High-Fidelity Character Animation**.Our pipeline preprocesses the reference image and employs dual-path encoding: VAE captures latent appearance features , CLIP extracts semantic context , and a lightweight PoseNet replaces ControlNet for pose encoding. These features undergo denoising in UNet before VAE decoding outputs video segments. Gaussian-weighted fusion blends overlapping regions to generate long and coherent character animations.

Adaptation method for specific action-posture types. To improve the model's adaptability to specific action types, we extract features using the spatio-temporal convolutional network [7, 37] on UCF101 dataset action classes, then align motion and pose features via cross-attention to inject high-weight action contexts for pose correction. The relevant definitions are as follows:

$$Q = W_q f_{pose}, K = W_k f_{motion}, V = W_v f_{motion},$$
$$f_{fusion} = \sum_{j=1}^{256} \alpha_{ij} V_j, \tag{5}$$
$$\alpha_{ij} = \frac{Q_i \cdot K_j}{\sqrt{d}},$$

where f_{motion} is the feature vector output by the spatio-temporal convolutional network, f_{pose} is the posture feature output by PoseNet, α_{ij} is the corresponding attention weight, Q_i is the Query vector at the i-th posture feature position, and K_j is the Key vector at the j-th action feature position. Our model has better adaptability to different field scenarios and related complex action postures through the joint optimization design.

3.3. Spatio-temporal Optimization with Detail Enhancement

This paper introduces a spatio-temporal optimization with detail enhancement strategy to improve the temporal coherence of the model in long video sequences and the ability to represent details in key areas such as hands.

Temporal Consistency Modeling. To extend from a static diffusion model to video generation and effectively capture the temporal dependence between video frames for temporal consistency modeling, we insert a learnable time layer into the denoising Unet structure of the diffusion network [2, 12, 42]. Specifically, for a given batch of feature maps $x \in \mathbb{R}^{b \times t \times h \times w \times c}$ (b represents the batch size and t represents the number of frames) , we first reshape it to $x' \in \mathbb{R}^{(b \times h \times w) \times t \times c}$. Then, we add a sinusoidal positional encoding [26, 39] to the time series $x' = \left\{ x'_1, x'_2 \dots x'_f \right\}$. Subsequently,we perform a self-attention mechanism in the time dimension: $Attention(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d}})V$, where $Q = W_q x', K = W_k x', V = W_v x'$. And the weight matrices W_q, W_k, W_v are all randomly initialized. The attention output is fused through a residual connection: $\widehat{x} = x + W_o \cdot Attention(Q,K,V)$ .To maintain the stability of the pre-trained model, we adopt a zero-initialization strategy and initialize the output projection matrixas W_o a zero matrix [9, 47], so that the module behaves as an identity mapping in the initial state. In addition, we insert the time layer only after the downsampling blocks of the UNet, avoiding the cross-attention layer to prevent prevent the text condition information from being disturbed.

Enhancement of Hand Details. In the process of generating character animation using the generation method based on the diffusion model, there are often problems with distorted hand details. In particular, there are phenomena such as adhesion in the finger gap area and blurred finger contours. We designed a three-stage hand detail enhancement algorithm, whose core idea is derived from the morphological operations in image processing [13, 31].

1. **Dynamic Kernel Design:** Dynamically adjust the size and shape of the morphological kernel according to the finger spacing to adapt to the changes in different gestures and finger spacings.The dynamic kernel is defined as follows:

$$K(p_i,p_j) = \begin{cases} \text{disk}(r_{ij}), & |p_i - p_j| < \tau_d \\ \text{line}(p_i,p_j,w_{ij}), & \text{otherwise} \end{cases} \tag{6}$$

where r_{ij} is the kernel radius determined by the distance between adjacent fingertips, p_i, and p_j represent the coordinate of the key points of the fingertip, τ_d is the distance threshold selected empirically, w_{ij} is the width of the linear kernel.

2. **Directional Dilation:** Use a dynamically adjusted structural element in the finger gap area to effectively separate the finger gaps. Strengthen the boundaries of the finger gaps through the dilation operation.

3. **Morphological Reconstruction and Detail Restoration:** The unnatural edges that may be introduced by dilation are removed through the erosion operation, while maintaining the natural shape of the finger gaps.

Context Scheduling. To address the issues of video memory limitations and global consistency faced during the generation of long videos, this design introduces a context scheduling strategy. Specifically, this strategy divides a long video into multiple "context windows" . Each window contains a certain number of consecutive frames. Then, these windows are processed in batches, and an overlapping area is set between the windows to ensure the overall consistency of the video. For example, when processing a 120 video, divide it into multiple 24-frame context Windows, with an overlap of 4 frames between adjacent windows. Mathematically, it can be expressed as: $W_k = \{F_i, F_{i+1}, \ldots, F_{i+23}\}$, $i = k \times (24-4)$,where W_k represents the k-th window, and F_i represents the i-th frame.

Considering that different regions in the video have different impacts on the final result, the design adopts a weighted fusion strategy for the overlapping area to improve the video quality.For the overlapping 4 frames, different Gaussian weights are assigned according to their positions within the window. The central frame has the highest weight, and the weights of the edge frames gradually decrease.The final frame is generated by the weighted sum, and the formula is as follows: $I^t_{final} = \sum_k w_k(t) \cdot I^t_{start}(k)$, where w_k is the Gaussian weight, and $I^t_{start}(k)$ is the result corresponding to the *k*-th window within the *t* frames.

3.4. Training Strategy

Our model method is an extension of the SD model [15, 28], the core training objective of the Stable Diffusion model is: $\mathcal{L}_{SD} = E_{z_t,c,\varepsilon,t}||\epsilon - \epsilon_\theta(z_t,c,t)||^2_2$,where the VAE encoder maps the image x to the latent space $z_t = \varepsilon(x)$,c is the conditional vector such as the CLIP text embedding , $\epsilon \sim N(0,I)$ is the standard Gaussian noise, and $t \in [1,T]$ represents the diffusion time step. .Since we have introduced a time layer in our model, we adopt a two-stage training strategy. In the first stage, we temporarily ignore the time layer and train the main body of the model. At the same time, we train the Reference ProcessingNet and the PoseNet.The loss term of this stage is calculated as

$$\mathcal{L}_1 = \mathbb{E}_{z_t, I_{\text{ref}}, p, \epsilon \sim \mathcal{N}(0,1), t} \left\| \epsilon - \epsilon_\theta(z_t, y_{\text{ref}}, y_p, t) \right\|_2^2, \tag{7}$$

where p represents the target pose (the coordinate tensor of bone key points),I_{ref} represents the reference image (RGB three channels), y_{ref} represents the multi - scale features of the reference image extracted by the Reference ProcessingNet, and y_p represents the pose features extracted by the PoseNet. In the second stage, we only train the time layer while keeping the parameters of the rest of the network fixed.The mathematical expression of the loss function calculation at this stage is as follows:

$$\mathcal{L}_2 = \mathbb{E}_{z_t^{1:K}, I_{\text{ref}}, p^{1:K}, \epsilon^{1:K} \sim \mathcal{N}(0,1), t} \left\| \epsilon^{1:K} - \epsilon_\theta \left(z_t^{1:K}, c^{1:K} \right) \right\|_2^2, \tag{8}$$

where $z_t^{1:K}$ represents a sequence of K consecutive noisy late variables, $\epsilon^{1:K}$ represents a sequence of K real Gaussian noise, $p^{1:K}$ represents a sequence of K target poses, and $c^{1:K}$ represents a conditional feature sequence, which consists of two parts: the reference image features and the pose feature sequence.

4. Experiments

To evaluate our model, we train on more than 3,000 character animation videos collected online.Using DensePose

Method	L1↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FID-VID↓	FVD↓
MRAA[33]	6.43E-04	28.79	0.673	0.245	85.42	103.52	658.12
TPS[49]	8.06E-04	29.62	0.715	0.239	118.91	122.69	743.81
Disco[40]	3.49E-04	29.58	**0.784**	0.272	64.79	84.13	365.53
IPC	5.73E-04	29.21	0.735	0.258	89.84	77.20	412.73
MagicAnimate	3.18E-04	30.15	0.692	0.215	55.47	69.68	217.98
HFCA(ours)	**2.56E-04**	**30.27**	0.763	**0.187**	**43.12**	**39.57**	**168.27**

Table 1. Quantitative results on TikTok dataset [19] with best results in **bold**; (↑: higher better, ↓: lower better).

Method	L1↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FID-VID↓	FVD↓
MRAA[33]	6.12E-04	32.48	0.725	0.278	65.11	41.75	215.64
TPS[49]	6.58E-04	30.89	0.682	0.295	97.45	84.56	552.03
Disco[40]	3.75E-04	32.16	0.779	0.221	51.54	64.21	462.73
IPC	4.31E-04	31.52	0.751	0.238	68.29	56.27	381.28
MagicAnimate	3.29E-04	32.74	**0.827**	0.192	43.77	35.93	176.32
HFCA(ours)	**2.87E-04**	**32.68**	0.793	**0.167**	**29.84**	**27.85**	**122.94**

Table 2. Quantitative comparisons on TED-talks dataset [33]

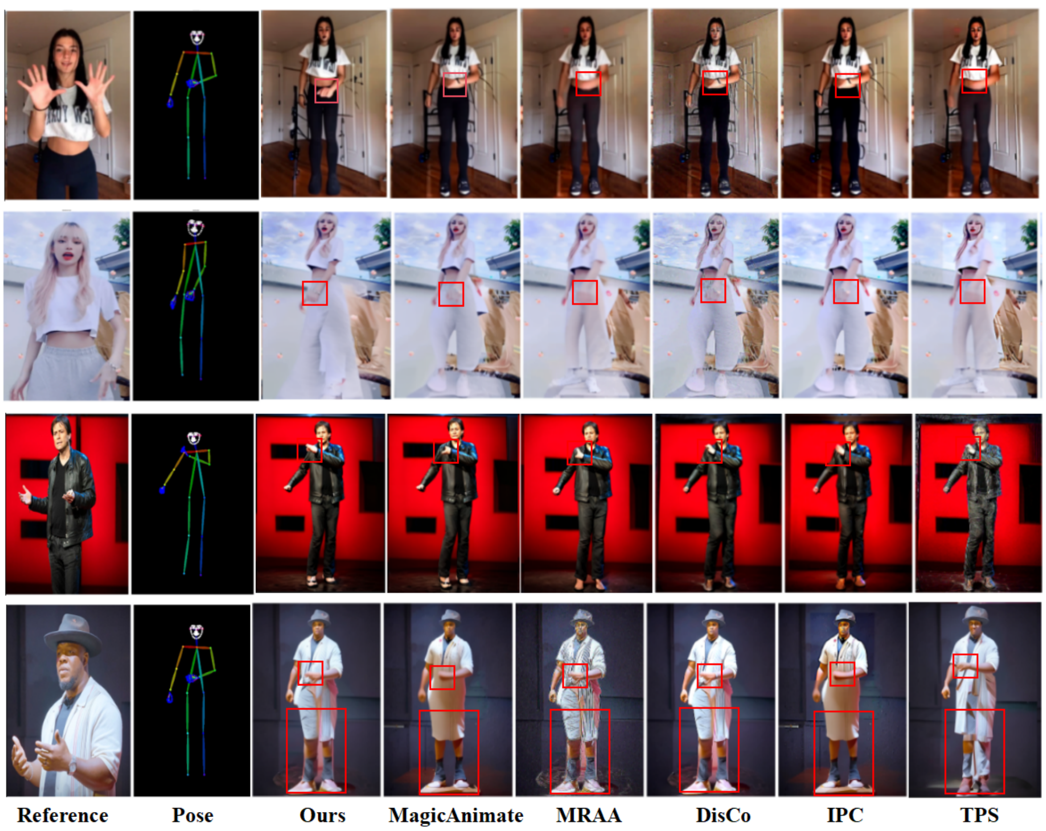


Figure 3. Qualitative comparisons on TikTok[19] and TED-Talks[33] datasets: Reference image and target pose overlaid atop, generated frames below. Red boxes highlight detail differences, demonstrating our method's superior feature preservation.

[11], we extracted per-frame UV coordinates and skeletal keypoints to generate aligned pose feature maps. Each reference image and its pose map formed Image-Pose training pairs for appearance-pose mapping learning.All images are resized to 640×640 with center cropping for resolution uniformity.We carried out training and testing on two specific benchmark datasets, namely the TikTok dataset [19] and the TED-talks dataset [33]. For fair comparison, we adopted MagicAnimate's train-test splits [44] on TikTok and TED-talks benchmark datasets.

4.1. Comparisons

Baseline Test. This paper comprehensively compares High-Fidelity Character Animation with several current advanced human image animation methods.(1) GAN-based approaches (MRAA [33], TPS [49]) employ spatial transformation networks with occlusion masks to map driving poses to reference images. (2)Diffusion-based methods include DisCo [40], which utilizes branched ControlNet modules for pose,human body and background decomposition integrated via cross-attention, and MagicAnimate [44] combining appearance encoders with temporal diffusion models and video fusion techniques. (3) We also construct a baseline model that integrates the temporal layer structure in Stable Diffusion, ControlNet [47], IP-Adapter [45], and Magicvideo [50], named IPC. Benchmark evaluations cover dance generation on TikTok dataset [19] and speech gesture synthesis on TED-Talks dataset ensuring task-specific performance validation.

Evaluation metrics. In this study, we use consistent quantitative metrics across two benchmark datasets: image-level quality (L1 error, SSIM [43], LPIPS [48], PSNR[16], FID [14]) and video-level fidelity (FID-FVD [1], FVD [38]). Despite varied metrics in prior work (e.g., MRAA/MagicAnimate's AKD/MKR/AED for TED-Talks), we ultimately decide to uniformly use the metrics of L1 error, SSIM , LPIPS , PSNR, FID , FID-FVD and FVD for the two benchmark datasets to avoid the impact caused by differences in metrics.

Quantitative comparisons. Table 1 and Table 2 shows the quantitative comparison results of our model and various baseline methods on two benchmark datasets. In the human dance generation task,the test results on the TikTok dataset are shown in Table 1.HFCA achieves state-of-the-art image-level metrics (L1/PSNR/LPIPS/FID).In addition, HFCA has achieved significant performance improvements of 43.2% and 22.8% in the two key video quality evaluation metrics of FID-VID and FVD respectively Compared with the strongest baseline MagicAnimate.The results in the Ted-talk dataset are shown in Table 2. HFCA has achieved the best FID-VID and FVD metrics on this dataset. Compared with the strongest baseline method MagicAnimate, the FID-VID is improved by 22.4% and the FVD is improved by 30.2%.

Qualitative comparison. We similarly employed a qualitative comparison approach to evaluate the performance of the output animation in terms of visual quality. As shown in Figure 3, we show a qualitative comparison between HFCA and various classes of baseline methods on the TikTok dataset. GAN-based methods (MRAA/TPS) fail to generate ideal output under extreme pose variations due to spatial transformation limitations.Baseline methods based on diffusion models (IPC, DisCo, and MagicAnimate) perform better in terms of single-frame image quality. However, DisCo's frame-independent generation causes temporal fragmentation, yielding inconsistent backgrounds and blurred edges.While IPC and MagicAnimate improve background consistency, they exhibit identity erosion with blurred facial features and unrealistic details due to inadequate identity preservation. In contrast, our method achieves superior fidelity through dual-path optimization of temporal coherence and identity feature preservation module.For TED-Talks' speech scenarios emphasizing facial expressions and gestures, diffusion methods (IPC/MagicAnimate/DisCo) surpass GAN-based approaches in pose control but exhibit edge blurring and posture distortions. HFCA demonstrates superior identity preservation and background stability through domain adaptation.

IPM	L1↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FID-VID↓	FVD↓
w/o	4.39	26.84	**0.785**	0.260	64.08	72.75	384.65
w/	**2.56**	**30.27**	0.763	**0.187**	**43.12**	**39.57**	**168.27**

Table 3. Ablation study for Identity Preservation Module,with best results in **bold**.Note: Values in columns L1 are scaled by 10^{-4}(Be consistent with Table 4 and Table 5 below).

DSAO	L1↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FID-VID↓	FVD↓
w/o	6.83	25.91	0.612	0.319	69.75	83.01	529.62
w/	**2.56**	**30.27**	**0.763**	**0.187**	**43.12**	**39.57**	**168.27**

Table 4. Ablation study for Domain Specific Adaptation Optimization.

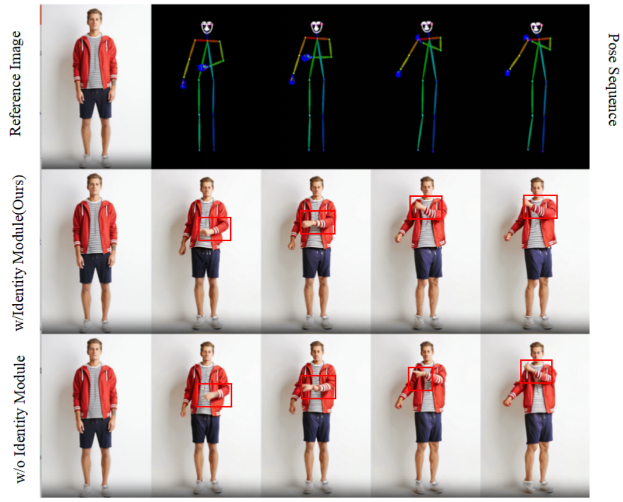
STODE	L1↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FID-VID↓	FVD↓
w/o	8.12	23.89	0.541	0.432	85.36	105.21	648.92
o	3.88	27.40	**0.795**	0.284	57.93	44.30	215.06
w/	**2.56**	**30.27**	0.763	**0.187**	**43.12**	**39.57**	**168.27**

Table 5. Ablation study for Spatio-temporal optimization with Detail Enhancement."o" represents the application of only the time consistency modeling part.

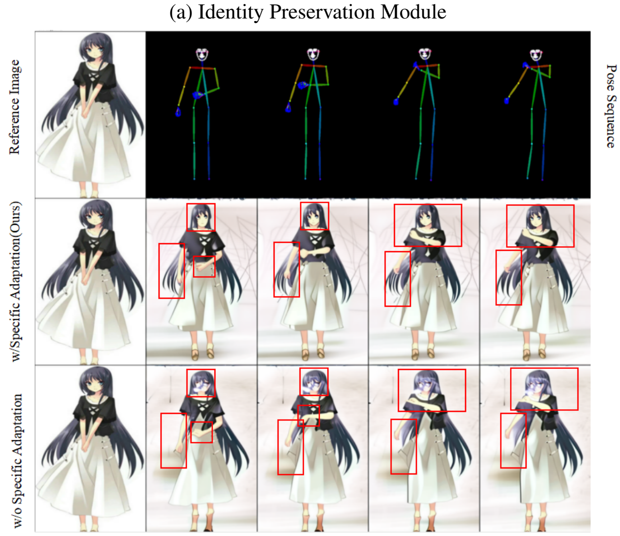
4.2. Ablation Study

To verify the effectiveness of the three major innovations in High-Fidelity Character Animation: the Identity Preservation Module, Domain Specific Adaptation Optimization and Spatio-temporal Optimization with Detail Enhancement, we conducted corresponding ablation experiments on the TikTok dataset.

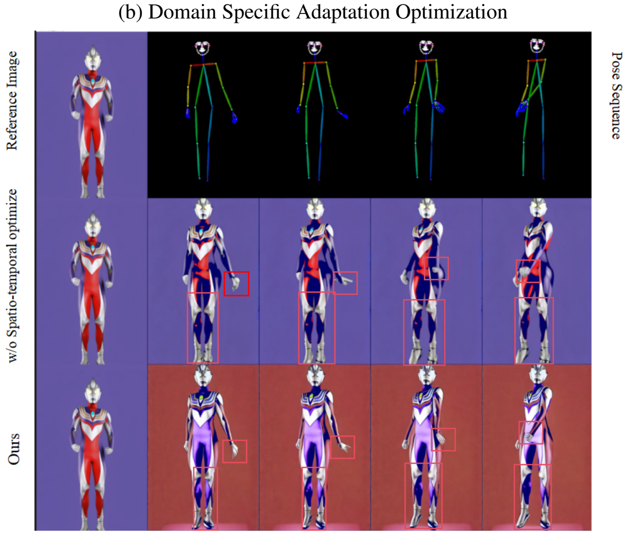
Identity Preservation Module. To evaluate the role of the proposed Identity Preservation Module, we trained a version of High-Fidelity Character Animation that did not include this module and conducted corresponding compara-



(a) Identity Preservation Module



(b) Domain Specific Adaptation Optimization



(c) Spatio-temporal Optimization with Detail Enhancement

Figure 4. Visual ablation study: (a) Identity Preservation Module; (b) Domain Specific Adaptation Optimization; (c) Spatio-temporal Optimization with Detail Enhancement. Red boxes highlight key differences.

tive experiments.The results shown in Table 3 indicate that when the identity feature preservation module is removed, the evaluation metrics of single-frame quality and video fidelity both decline. In addition, the qualitative ablation experiment shown in Figure 4a further validates the role of this module. As shown in the figure, the model with the identity feature preservation module performs better in terms of identity consistency: it helps preserve texture details.

Domain Specific Adaptation Optimization. Ablation studies confirm Domain-Specific Adaptation Optimization(DSAO) significantly boosts performance. we compare the output performance of the model with and without applying this optimization strategy.Quantitative results Table 4 demonstrate DSAO's superiority across metrics, while qualitative analysis Figure 4b shows enhanced its advantages—preserving hairstyles and facial features without distortion. Without DSAO, models exhibit detail degradation.

Spatio-temporal Optimization with Detail Enhancement. Ablation studies for Spatio-temporal Optimization with Detail Enhancement (STODE) employ three progressive configurations:1) The model directly splices images in chronological to generate character animation videos.2) The model only adopts the temporal consistency modeling part of STODE (insert a temporal layer in the network).3) The model fully applies Spatio-temporal Optimization with Detail Enhancement.Quantitative results Table 5 confirm STODE's critical necessity—its absence causes significant fidelity degradation, while full implementation outperforms temporal-only modeling across metrics. Qualitative analysis is shown in Figure 4c.We adjust the background color of the character animation to highlight the visual differences.It demonstrates STODE's superior hand detail enhancement and spatiotemporal consistency, maintaining feature stability across consecutive frames.This comprehensive validation establishes STODE's efficacy in detail processing and maintaining spatio-temporal consistency.

5. Conclusion

In this paper, we propose **High-Fidelity Character Animation**, a diffusion-based framework transforming static images into pose-driven animated videos. Key innovations—Identity Preservation,Domain Specific Adaptation and Spatio-temporal Optimization enhance detail fidelity and temporal coherence. Experiments demonstrate state-of-the-art performance in frame quality and video generation, outperforming existing methods.

Acknowledgments

This research is financially supported by the Jiangsu Provincial Key R and D Programme 261 (BE2023030,BE20221-54).

References

[1] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, page 2, 2019. 7

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 1, 2, 4

[3] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh. Ted-talks gesture dataset for nonverbal communication analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 6789–6799, 2023. 1

[4] Liang Chen, Jiayi Zhang, Jing Lin, Guoxian Song, Jun Liu, Qiong Liu, and Otmar Hilliges. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2023. 2

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 3

[6] Patrick Esser, Johanna Chiu, Patrick Atighehchian, Jonathan Granskog, and Alexandros Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*, pages 1–10, 2023. 2

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019. 3, 4

[8] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 3

[9] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010. 5

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27:2672–2680, 2014. 1

[11] Riza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 6

[12] Yuwei Guo, Chen 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. 1, 2, 3, 4

[13] Robert M Haralick, Stanley R Sternberg, and Xinhua Zhuang. Image analysis using mathematical morphology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (4):532–550, 1987. 5

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. 7

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2, 5

[16] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 7

[17] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 2, 3

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 1

[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. 1, 2, 6, 7

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711, 2016. 3

[21] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023. 2

[22] Chen Li, Kang Liu, Rui Wang, Chen Qian, and Dahua Lin. Mimicmotion: Motion-based 3d character animation with confidence-aware pose guidance. *arXiv preprint arXiv:2308.04399*, 2023. 1, 2

[23] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016. 3

[24] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, 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*, pages 1–9, 2023. 2

[25] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2022. 2

[26] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI preprint*, 2018. 4

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 1, 2

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 5

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lijuan Li, Jay Whang, Emily L Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2

[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 3

[31] Jean Serra. *Image analysis and mathematical morphology*. Academic Press, 1982. 5

[32] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, pages 7137–7147, 2019. 1

[33] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. 1, 2, 6, 7

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[35] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Junting An, Saining Zhang, Qiyuan Hu, Oran Yang, Or Litany, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14307–14317, 2023. 2, 3

[36] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3

[37] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015. 4

[38] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. *arXiv preprint arXiv:1912.01881*, 2019. 7

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 4

[40] J. Wang, X. Chen, Z. Xu, and L. Zhang. Disco: Disentangled control for realistic dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13653–13662, 2023. 2, 6, 7

[41] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems*, pages 1144–1156, 2018. 1

[42] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023. 2, 3, 4

[43] 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. 7

[44] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jingfeng Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498*, 2023. 2, 3, 6, 7

[45] Hu Ye, Jun Zhang, Sifei Liu, and Xiaogang Wang. Ip-adapter: Learning compatible image prompts for text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11256–11265, 2023. 7

[46] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833, 2014. 3

[47] 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*, pages 3836–3847, 2023. 2, 3, 5, 7

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 7

[49] Jian Zhao and Hao Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 6, 7

[50] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 7