SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

October 2025, Vol. 68, Iss. 10, 200103:1–200103:14
https://doi.org/10.1007/s11432-024-4592-3

Special Topic: Large Multimodal Models

UniAnimate: taming unified video diffusion models for consistent human image animation

Xiang WANG[1], Shiwei ZHANG[2], Changxin GAO[1], Jiayu WANG[2], Xiaoqiang ZHOU[3], Yingya ZHANG[2], Luxin YAN[1] & Nong SANG[1*]

[1]*Key Laboratory of Ministry of Education for Image Processing and Intelligent Control,*
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[2]*Alibaba Group, Hangzhou 310052, China*
[3]*School of Automation, University of Science and Technology of China, Hefei 230026, China*

Received 15 May 2024/Revised 28 November 2024/Accepted 4 March 2025/Published online 28 September 2025

Abstract Recent diffusion-based human image animation techniques have demonstrated impressive success in synthesizing videos that faithfully follow a given reference identity and a sequence of desired movement poses. Despite this, there are still two limitations: (i) an extra reference model is required to align the identity image with the main video branch, which significantly increases the optimization burden and model parameters; (ii) the generated video is usually short in time (e.g., 24 frames), hampering practical applications. To address these shortcomings, we present a `UniAnimate` framework to enable efficient and long-term human video generation. First, to reduce the optimization difficulty and ensure temporal coherence, we map the reference image along with the posture guidance and noise video into a common feature space by incorporating a unified video diffusion model. Second, we propose a unified noise input that supports random noised input as well as first frame conditioned input, which enhances the ability to generate long-term video. Finally, to further efficiently handle long sequences, we explore an alternative temporal modeling architecture based on a state space model to replace the original computation-consuming temporal Transformer. Extensive experimental results indicate that `UniAnimate` achieves superior synthesis results over existing state-of-the-art counterparts in both quantitative and qualitative evaluations. Notably, `UniAnimate` can even generate highly consistent one-minute videos by iteratively employing the first frame conditioning strategy. Code and models are publicly available at https://unianimate.github.io/.

Keywords video generation, human image animation, diffusion model, large multi-modal models, temporal modeling

Citation Wang X, Zhang S W, Gao C X, et al. UniAnimate: taming unified video diffusion models for consistent human image animation. Sci China Inf Sci, 2025, 68(10): 200103, https://doi.org/10.1007/s11432-024-4592-3

1 Introduction

Human image animation [1, 2] is an attractive and challenging task that aims to generate lifelike and high-quality videos in accordance with the input reference image and target pose sequence. This task has made unprecedented progress and showcased the potential for broad applications [3–6] with the rapid advancement of video generation methods [4, 7–13], especially the iterative evolution of generative models [14–22].

Existing methods can be broadly categorized into two groups. The first group [1, 2, 23, 24] usually leverages an intermediate pose-guided representation to warp the reference appearance and subsequently utilizes a generative adversarial network (GAN) [25] for plausible frame prediction conditioning on previously warped subjects. However, GAN-based approaches generally suffer from training instability and poor generalization issues [3, 26], resulting in non-negligible artifacts and inter-frame jitters. The second group [3–5, 26–29] employs diffusion models to synthesize photo-realistic videos. For instance, Disco [26] disentangles the control signals into three conditions, i.e., subjects, backgrounds, and dance moves, and applies a ControlNet-like architecture [30] for holistic background modeling and human pose transfer. Animate Anyone [3] and MagicAnimate [4] utilize a 3D-UNet model [31] for video denoising and exploit an additional reference network mirrors from the main 3D-UNet branch, excluding temporal Transformer modules, to extract reference image features for appearance alignment. To encode target pose information, a lightweight pose encoder is also utilized to capture desired motion characterizations. These

* Corresponding author (email: nsang@hust.edu.cn)

methods inherit the advantages of stable training and strong transferable capabilities of diffusion models, demonstrating superior performance to GAN-based approaches [3, 4, 26].

Despite these advancements, the existing diffusion-based methods still have two limitations: (i) they require an extra reference network to encode reference image features and align them with the main branch of the 3D-UNet, resulting in increased training difficulty and model parameter count; (ii) they usually employ temporal Transformers to model the temporal information, but Transformers require quadratic computations in the temporal dimension, which limits the generated video length. Typical methods [3,27] can only generate 24 frames, restricting practical deployment. Although the slide window strategy [4] that employs temporally overlapped local windows to synthesize videos and average the intersection parts is able to generate longer videos, we empirically observed that there are usually non-smooth transitions and appearance inconsistencies at the segment connections with the reference image.

To address the aforementioned limitations, we propose the `UniAnimate` framework for consistent human image animation. Specifically, we leverage a unified video diffusion model to simultaneously handle the reference image and noised video, facilitating feature alignment and ensuring temporally coherent video generation. Additionally, to generate smooth and continuous video sequences, we design a unified noised input that allows random noised video or first frame conditioned video as input for video synthesis. The first frame conditioning strategy can generate subsequent frames based on the final frame of the previously generated video, ensuring smooth transitions. Moreover, to alleviate the constraints of generating long videos at once, we utilize temporal Mamba [32–34] to replace the original temporal Transformer, significantly improving the efficiency. By this means, `UniAnimate` can enable highly consistent human image animation and is able to synthesize long-term videos with smooth transitions, as displayed in Figure 1. We conduct a comprehensive quantitative analysis and qualitative evaluation to verify the effectiveness of our `UniAnimate`, highlighting its superior performance compared to existing state-of-the-art methods.

2 Related work

This work is highly relevant to the fields of video generation, human image animation, and temporal coherence modeling. We will give a brief discussion of them below.

Video generation. Recent success in the diffusion models has remarkably boosted the progress of text-to-image generation [14, 30, 35–42]. However, generating videos from input conditions is a considerably more challenging task than its image counterpart due to the higher dimensional properties of video [8, 9]. Different from static images, video exhibits an additional temporal dimension, comprising a sequence of frames, which is crucial for understanding dynamic visual content and textual input. In order to model spatio-temporal dependencies, Make-A-Video [8] and ModelscopeT2V [43] adopt the 3D-UNet framework, which is a temporal extension of 2D-UNet [44] by integrating temporal layers such as temporal Transformers. This paradigm has also been widely followed in subsequent work [15,18,45–57]. SEINE [57] leverages random mask strategy to interpolate videos and can be extended to image-to-video generation. However, SEINE does not support the coordination of the first frame with the reference frame and poses during image-to-video generation. In pursuit of higher controllability both spatially and temporally, some techniques such as Gen-1 [58] and VideoComposer [16] attempt to introduce additional guided conditions, e.g., depth maps and motion vectors, for controllable general video synthesis [20, 21, 59–61]. In this work, we concentrate on the human-centered image animation task, which requires precise control of both human-related appearance attributes and the desired target pose motion to create plausible videos.

Human image animation. Animating human images along with the driving pose sequence is a challenging yet useful video creation task. With the rapid development of generative networks, various approaches have been proposed. Previous studies [62–65] mainly focused on exploring GAN-based generation, typically leveraging a motion network to predict dense appearance flows and perform feature warping on reference inputs to synthesize realistic images that follow the target poses. However, these techniques often suffer from instability training and mode collapse issues [3], struggling to precisely control the generated human motions and yield sub-optimal synthesis quality. As a result, extensive efforts [3–5, 26, 27, 29, 66, 67] start to establish image animation architectures on diffusion models [14, 40] due to their superior training stability and impressive high-fidelity results through an iterative refining process. For instance, Disco [26] develops a hybrid diffusion architecture based on ControlNet [30] with

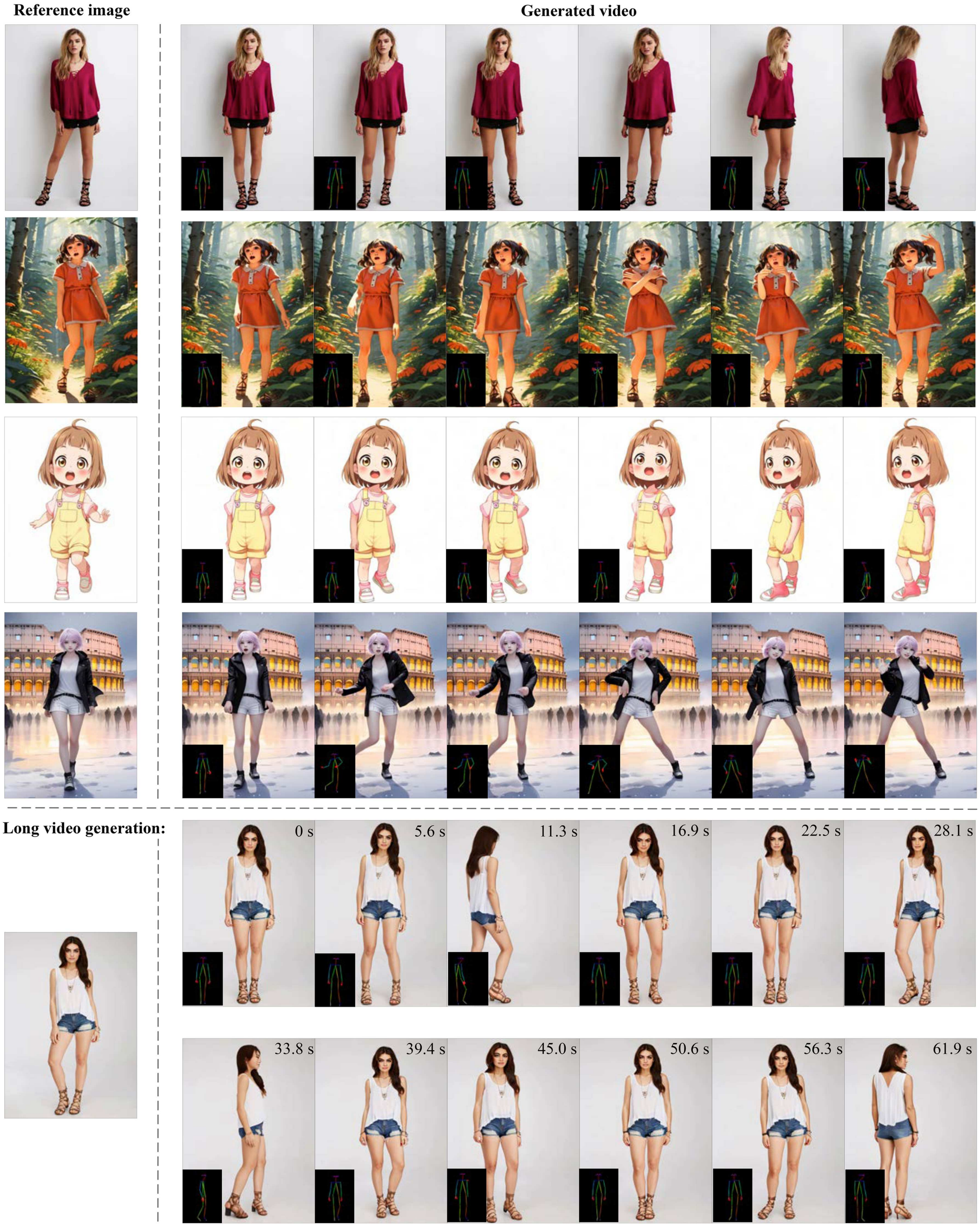

Figure 1 (Color online) Example videos synthesized by the proposed `UniAnimate`. Given a reference image and a target pose sequence, `UniAnimate` can generate temporally consistent and high-quality character videos that seamlessly adhere to the input conditional guidance. Note that our method is not trained on any cartoon character dancing videos, displaying excellent cross-domain transfer capability. In addition, by iteratively employing the first frame conditioning strategy, `UniAnimate` can generate high-fidelity one-minute videos.

disentangled control of human foreground, background, and poses to allow composable human dance generation. MagicAnimate [4] and Animate Anyone [3] take advantage of video diffusion models for enhanced temporal consistency and introduce a two-stage learning strategy to decouple appearance alignment and motion guidance respectively. Commonly, existing diffusion-based frameworks usually build on a ControlNet-like 3D-UNet model [3, 15, 30] to maintain temporal coherence and introduce a reference encoder, which is a replica of 3D-UNet excluding the temporal Transformer layers, to preserve the intricate appearance of the reference image. Although promising, these advances often require multiple separate networks with non-negligible parameters, leading to increasing optimization difficulties, and face challenges in long-term video generation due to the quadratic complexity of the temporal Transformer.

Temporal coherence modeling. Temporal dynamics analysis plays a pivotal role in many video understanding tasks [68–70]. Previous work on temporal modeling broadly falls into two categories: convolution-based [68, 71, 72] and RNN-based [12, 73, 74]. However, convolution-based approaches suffer from limited receptive fields [75], leading to difficulty in modeling long-range temporal dependencies. Although RNN-based methods can perceive long-term relations, they encounter the dilemma of unparalleled calculation, causing computational inefficiency. To capture non-local associations and enable parallel calculation, many recent researches [8, 16, 45, 69, 70, 75–83] seek to employ Transformers for sequential modeling, displaying remarkable performance in various downstream applications, e.g., action recognition [70, 75], action detection [69], and video generation [15, 16, 43]. Nevertheless, temporal Transformers still face huge computational costs due to the quadratic complexity, especially when dealing with long sequences. Mamba [32], a type of fundamental state space models (SSMs) [84], which conceptually merges the merits of parallelism and non-locality, has demonstrated convincing potential in a wide range of downstream natural language processing [32] and computer vision fields [33, 85]. Inspired by the excellent performance and linear time efficiency of Mamba in long sequence processing [33, 34, 85–87], this paper attempts to introduce Mamba into the human image animation task as a strong and promising alternative for temporal coherence modeling.

3 Method

Human image animation aims to generate a high-quality and temporally consistent video based on the input reference image and target pose sequence. The challenges in this task involve maintaining temporal consistency and natural appearance throughout the generated video. To this end, we present our proposed `UniAnimate`, which addresses the limitations of existing diffusion-based methods for consistent and long-term human image animation. We will first briefly introduce the basic concepts of the latent diffusion model. Subsequently, the detailed pipeline of `UniAnimate` will be described.

3.1 Preliminaries of latent diffusion model

The optimization and inference of traditional pixel-level diffusion models [9, 37, 40] require prohibitive calculations in the high-dimensional RGB image space. To reduce the computational cost, latent diffusion models [11, 13, 35, 49, 53] propose employing denoising procedures in the latent space of a pre-trained variational autoencoder (VAE). In particular, a VAE encoder is first employed to embed the input sample to the down-sampled latent data $\mathbf{z}_0$. Subsequently, a Markov chain of forward diffusion process q is defined to progressively add stochastic Gaussian noise of T steps to the clean latent data $\mathbf{z}_0$. The forward diffusion step can be formulated as

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}), \quad t = 1, 2, \ldots, T, \tag{1}$$

where $\beta_t \in (0, 1)$ denotes the noise schedule. As t gradually increases, the total noise imposed on the original $\mathbf{z}_0$ becomes more intense, and eventually $\mathbf{z}_t$ tends to be a random Gaussian noise. The objective of the diffusion model $\boldsymbol{\epsilon}_\theta$ is to learn a reversed denoising process p that aims to recover the desired clean sample $\mathbf{z}_0$ from the noised data $\mathbf{z}_t$. The denoising process $p(\mathbf{z}_{t-1}|\mathbf{z}_t)$ can be estimated by $\boldsymbol{\epsilon}_\theta$ as the following form:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{z}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{z}_t, t)), \tag{2}$$

where $\boldsymbol{\mu}_\theta(\mathbf{z}_t, t)$ is the approximated objective of the reverse diffusion process, and θ means the parameters of the denoising model $\boldsymbol{\epsilon}_\theta$. In many video generation techniques [3, 16, 31], the denoising model is a 3D-UNet model [31]. In the optimization stage, a simplified L2 loss is usually applied to minimize the discrepancies between predicted noise and real ground-truth noise:

$$\mathcal{L} = \mathbb{E}_\theta\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, c)\|^2\right] \tag{3}$$

in which c is the input conditional guidance. After the reversed denoising stage, the predicted clean latent is fed into the VAE decoder to reconstruct the predicted video in the pixel space.

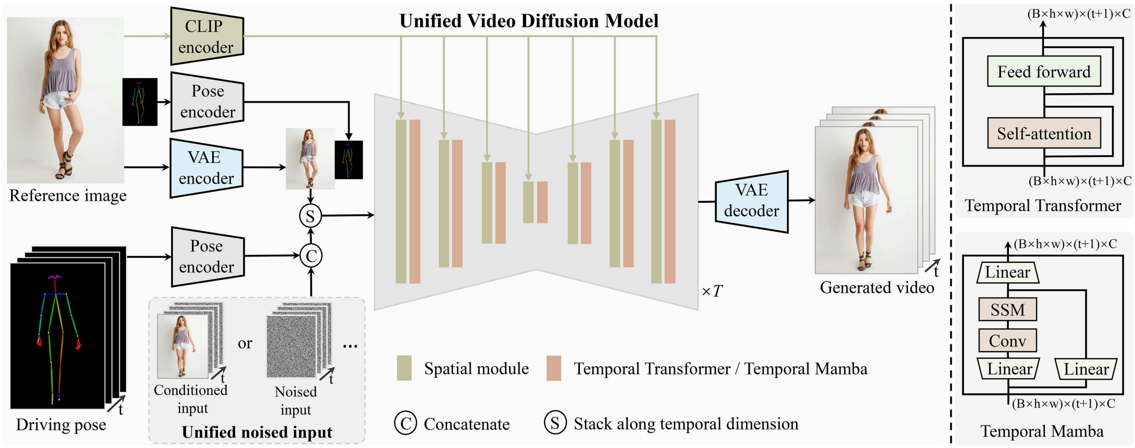

Figure 2 (Color online) The overall architecture of the proposed `UniAnimate`. Firstly, we utilize the CLIP encoder and VAE encoder to extract latent features of the given reference image. To facilitate the learning of the human body structure in the reference image, we also incorporate the representation of the reference pose into the final reference guidance. Subsequently, we employ a pose encoder to encode the target driven pose sequence and concatenate it with the noised input along the channel dimension. The noised input is derived from the first frame conditioned video or a noised video. Then, the concatenated noised input is stacked with the reference guidance along the temporal dimension and fed into the unified video diffusion model to remove noise. The temporal module in the unified video diffusion model can be the temporal Transformer or temporal Mamba. Finally, a VAE decoder is adopted to map the generated latent video to the pixel space.

3.2 UniAnimate

`UniAnimate` aims to create visually appealing and temporally coherent videos that correspond to the given reference image and pose sequence. To align the appearance between the given image and the generated video, we design a unified video diffusion model to embed the reference information and estimated video content in the shared feature space. In addition to the driving pose sequence, the source pose of the reference image is also incorporated to provide corresponding spatial position and layout information of the human body. To ensure long-term video generation, a first frame conditioning strategy is introduced, and we explore an alternative based on Mamba [32] for temporal coherence modeling. The overall framework of the proposed `UniAnimate` is displayed in Figure 2.

Unified video diffusion model. To tackle the problem of temporally consistent human image animation, we leverage the widely used 3D-UNet structure [16, 31, 43] for video creation. Unlike previous human image animation methods that employ two separate networks, namely a referenceNet for encoding the appearance of the reference image and a main 3D-UNet branch for synthesizing human motion videos, `UniAnimate` proposes to take advantage of a unified video diffusion model. This unified structure is able to jointly encode the appearance of the reference image and synthesize the motion of the generated video. The advantages of this strategy are twofold: (1) the feature representations of the reference image and the generated video exist in the same feature space, facilitating appearance alignment, and (2) the parameters of the framework are reduced, making optimization more feasible. Additionally, different from previous methods [3, 4], which need to learn character structure information implicitly from the reference image, we propose to extract the reference skeletal pose from the reference image, explicitly incorporating the position and layout information of the reference human. Specifically, the reference image is first encoded into latent space using a VAE encoder, resulting in a feature representation of size $C_1 \times h \times w$, where C_1, h, and w represent the channel, width, and height, respectively. The reference pose is also processed through a pose encoder, extracting layout information. The reference image and pose features are then fused to obtain the final reference representation f_{ref} with a shape of $C \times h \times w$. To incorporate target pose information, we use a pose encoder to encode the driving pose sequence and concatenate the resulting driving pose features and the input noised latent to obtain the fused features $f_v \in \mathbb{R}^{t\times C\times h\times w}$, where t means the temporal length. Subsequently, the reference representation f_{ref} and the fused features f_v are stacked along the temporal dimension, resulting in combined features $f_{merge} \in \mathbb{R}^{(t+1)\times C\times h\times w}$. Finally, the combined features are then fed into the unified video diffusion model for jointly appearance alignment and motion modeling.

Unified noised input. Due to memory limitations, it is not possible to generate a long video in a single pass. Instead, multiple short video segments need to be synthesized separately and eventually merged into one long video. Typically, existing methods [3, 4] utilize the slide window strategy that employs temporally overlapped local windows to synthesize short videos and average the intersection parts to generate longer videos. However, in our experiments (Section 5), we empirically find that this slide window strategy may suffer from discontinuities between segments and usually cannot preserve appearance consistency with the reference image. To address this issue, we propose a unified noised input that allows random noised video or first frame conditioned video as input for video synthesis. The first frame conditioning manner takes the beginning frame of a video as the condition for generating videos starting from the frame. By leveraging this strategy, the last frame of the previous short video segment can be used as the first frame of the next segment, enabling seamless and visually coherent long-term animation. The first frame conditioning strategy offers two advantages: (1) it supports user-defined input images as the starting frame, combined with the target pose sequence for human image animation, and (2) it is able to generate consistent long videos with smooth transitions by iteratively employing the first frame conditioning strategy.

Temporal modeling manners. Previous methods [3,4,16,31] usually employ temporal Transformers to model the motion patterns in the video. While these methods have shown impressive progress, the quadratic complexity relationship between temporal Transformers and input video length limits the video length that can be generated in a single segment. In this paper, we explore a new temporal modeling approach called temporal Mamba [32–34] for the human image animation task. Mamba [32] is commonly treated as a type of linear time-invariant system that can map a sequential input $x(s) \in \mathbb{R}^L$ to a response state $y(s) \in \mathbb{R}^L$ and can be typically formulated as

$$\begin{aligned} h'(s) &= \mathbf{A}h(s) + \mathbf{B}x(s), \\ y(s) &= \mathbf{C}h(s), \end{aligned} \tag{4}$$

where $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are parameter matrices, and $h(s)$ is a hidden state. In the application, a discretized version of Mamba that adopts a bidirectional scanning mechanism [33] is leveraged by us to handle temporal dependencies. Temporal Mamba exhibits a linear complexity relationship with the generated video length. As will be demonstrated in the experimental section, performance of temporal Mamba surpasses or matches that of temporal Transformers, while requiring less memory consumption.

Training and inference. During training, we follow the conventional video generation paradigm [16, 31] and train the model to generate clean videos by estimating the imposed noise. To facilitate the multi-condition generation, we introduce random dropout to the input conditions (e.g., the first frame and reference image) at a certain ratio (e.g., 0.5). At the inference stage, our `UniAnimate` supports human video animation using only a reference image and a target pose sequence, as well as the input of a first frame. To generate long videos composed of multiple segments, we utilize the reference image for the first segment. For subsequent segments, we use the reference image along with the first frame of the previous segment to initiate the next generation.

4 Experiments

In this section, we first describe the experimental setups of `UniAnimate`. Afterward, a comprehensive qualitative and quantitative evaluation with existing state-of-the-art techniques will be implemented to validate the effectiveness of the proposed method in generating temporally smooth videos for the human image animation task.

4.1 Experimental setups

Datasets. Following previous studies [2, 3, 88], the comparative experiments are conducted on two standard and widely-used datasets, namely TikTok [88] and Fashion [2]. The TikTok dataset consists of 340 training videos and 100 testing videos. Each video has a duration of 10–15 s. To ensure a fair comparison, we follow the settings of prior methods [3,4,26], where 10 videos from the test set are selected for both qualitative and quantitative comparisons. Fashion is a dataset with simple and clean backgrounds, containing 500 training videos and 100 testing videos, with each video covering approximately 350 frames. To enhance the robustness and generalization of our model, similar to [3, 27], we additionally collect around 10k TikTok-like internal videos. It is worth noting that, to enable fair comparisons with existing

Table 1 Quantitative comparison with existing methods on the TikTok dataset. "PSNR*" indicates that the modified metric is applied to avoid numerical overflow. The best results are in bold. ↑ means the higher the better and ↓ means the opposite. The second best results are underlined. The modified metric is at https://github.com/Wangt-CN/DisCo/issues/86.

Method	L1 ↓	PSNR ↑	PSNR* ↑	SSIM ↑	LPIPS ↓	FVD ↓
FOMM [62] (NeurIPS19)	3.61E−04	–	17.26	0.648	0.335	405.22
MRAA [64] (CVPR21)	3.21E−04	–	18.14	0.672	0.296	284.82
TPS [65] (CVPR22)	3.23E−04	–	18.32	0.673	0.299	306.17
DreamPose [29] (ICCV23)	6.88E−04	28.11	12.82	0.511	0.442	551.02
DisCo [26] (CVPR24)	3.78E−04	29.03	16.55	0.668	0.292	292.80
MagicAnimate [4] (CVPR24)	3.13E-04	29.16	–	0.714	0.239	179.07
Animate Anyone [3] (CVPR24)	–	29.56	–	0.718	0.285	171.90
Champ [27] (ArXiv24)	2.94E-04	29.91	–	0.802	0.234	160.82
`UniAnimate`	**2.66E−04**	**30.77**	**20.58**	**0.811**	**0.231**	**148.06**

methods, we train our `UniAnimate` solely on the TikTok and Fashion benchmarks without incorporating extra videos and report experimental results in Subsections 4.2 and 4.3.

Detailed implementation. In the experiments, we use DWpose [89] to extract pose sequences for model optimization. The visual encoder of the multi-modal CLIP-Huge model [90] in Stable Diffusion v2.1 [35] is used to encode the CLIP embedding of the reference image. The pose encoder is composed of several convolution layers and has a similar structure as STC-encoder in VideoComposer [16]. Like previous approaches [3,4,27], we employ a pre-trained video generation model [18] for model initialization. In our experiments, the temporal Mamba architecture is initialized with pre-trained weights derived from the foundational video model [18] trained on the WebVid10M dataset. The experiments are conducted on 8–16 NVIDIA A100 GPUs. During the training phase, videos are resized to a spatial resolution of 768×512. We randomly input video segments of uniformly sampled 16 or 32 frames into the model to learn temporal consistency. We utilize the AdamW optimizer [91] with a learning rate of 5E−5 to optimize the network. For noise sampling, DDPM [14] with 1000 steps is performed during training. In the inference stage, we warp the length of the driving pose to roughly align with the reference pose and adopt the DDIM sampler [40] with 50 steps for accelerated sampling.

Evaluation metrics. We quantitatively evaluate our method using various metrics. In particular, four widely-used image metrics, namely L1, PSNR [92], SSIM [93], and LPIPS [94], are applied to measure the visual quality of the generated results. Besides these image metrics, we also leverage the Fréchet video distance (FVD) [95] as a video evaluation metric, which quantifies the discrepancy between the generated video distribution and the real video distribution.

4.2 Comparisons with state-of-the-art methods

For a comprehensive evaluation, we compare our proposed method with existing approaches in terms of both quantitative and qualitative measures. Additionally, a human evaluation is further conducted to verify the efficacy.

Quantitative comparisons. To validate the effectiveness of our proposed method, we compare it with existing state-of-the-art approaches, including Disco [26], MagicAnimate [4], Animate Anyone [3], and Champ [27]. These methods adopt ControlNet-like structures to achieve appearance alignment. As shown in Table 1, our `UniAnimate` outperforms existing state-of-the-art competitors across all the evaluation metrics on the TikTok dataset. For example, `UniAnimate` reaches a FVD of 148.06, achieving the best video fidelity among recent studies. The quantitative results of both image and video metrics demonstrate our model's excellent ability to learn and generate realistic content, highlighting the capability of `UniAnimate` to effectively capture and reproduce the underlying data distribution of training samples. The experiment on the Fashion dataset is also conducted, as illustrated in Table 2 [3,23,24,29,64,65,96,97]. From the comparison, we can observe that `UniAnimate` exhibits superior structural preservation capacity, obtaining the best SSIM of 0.940. `UniAnimate` also achieves impressive performance on other metrics, and these results collectively reaffirm the ability of `UniAnimate` to synthesize visually fidelity animations in the fashion video domain. In addition, unlike other methods that require an extra reference model, such as Animate Anyone, which requires about 2.1 B parameters, our method only requires about 1.4 B model parameters, which greatly reduces the model complexity and optimization burden.

Qualitative comparisons. In addition to quantitative measures, we also provide a qualitative comparison in Figure 3. We showcase the comparison of `UniAnimate` with other competitive methods on

Table 2 Quantitative comparison with existing methods on the Fashion dataset. "*w/o finetune*" represents the method without additional finetuning on the fashion dataset. "PSNR*" indicates that the modified metric[1] is applied to avoid numerical overflow. The best results are in bold. The second best results are underlined. ↑ means the higher the better and ↓ means the opposite.

Method	PSNR ↑	PSNR* ↑	SSIM ↑	LPIPS ↓	FVD ↓
MRAA [64] (CVPR21)	–	–	0.749	0.212	253.6
TPS [65] (CVPR22)	–	–	0.746	0.213	247.5
DPTN [24] (CVPR22)	–	24.00	0.907	0.060	215.1
NTED [96] (CVPR22)	–	22.03	0.890	0.073	278.9
PIDM [97] (CVPR23)	–	–	0.713	0.288	1197.4
DBMM [23] (ICCV23)	–	24.07	0.918	0.048	168.3
DreamPose [29] (ICCV23)	–	–	0.885	0.068	238.7
DreamPose *w/o finetune* [29] (ICCV23)	34.75	–	0.879	0.111	279.6
Animate Anyone [3] (CVPR24)	**38.49**	–	0.931	0.044	81.6
UniAnimate	37.92	**27.56**	**0.940**	**0.031**	**68.1**

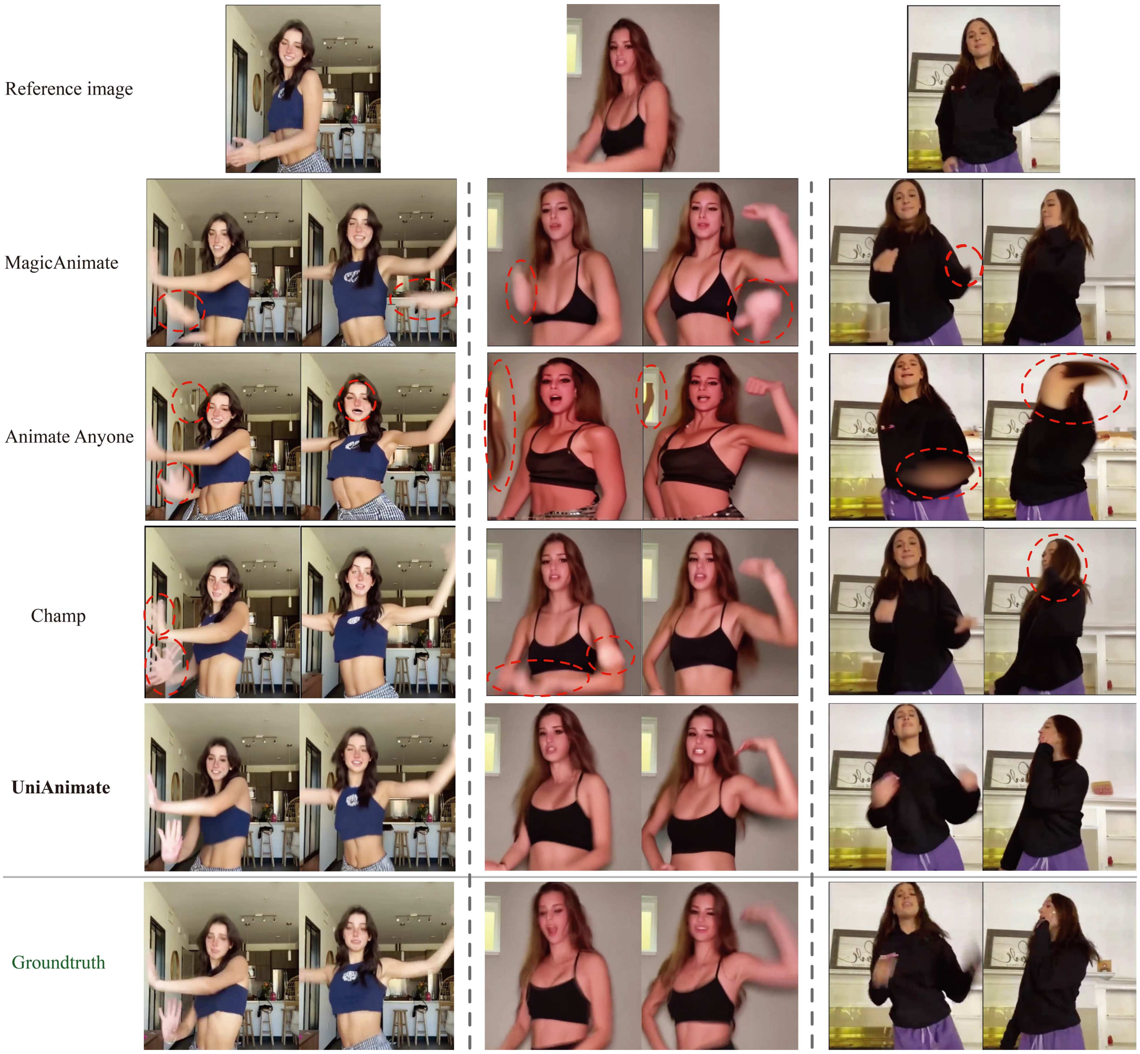

Figure 3 (Color online) Qualitative comparison with existing state-of-the-art methods on the TikTok dataset. Three typical state-of-the-art methods namely MagicAnimate [4], Anymate Anyone [3], and Champ [27] are compared.

the TikTok test set. The results of Animate Anyone [3] are obtained by leveraging the publicly available reproduced code[1]. From the visualizations, we can observe that MagicAnimate [4] exhibits instances of limb generation and appearance misalignment, while Animate Anyone introduces undesirable artifacts. Champ [27] may produce some artifacts that are discordant, such as unreasonable hand counts. These

[1] https://github.com/MooreThreads/Moore-AnimateAnyone.

Table 3 User study. We ask users to rate the generated video results on the TikTok dataset in terms of visual quality, identity preservation, and temporal consistency. The best results are in bold. ↑ means the higher the better.

Method	Visual quality (%) ↑	Identity preservation (%) ↑	Temporal consistency (%) ↑
MagicAnimate [4]	76	81	82
Animate Anyone [3]	67	84	71
Champ [27]	74	77	85
UniAnimate	**85**	**89**	**91**

Table 4 Ablation study on the TikTok dataset. "PSNR*" indicates that the modified metric is applied to avoid numerical overflow. "UniAnimate *w/o* unified VDM" implies a ControlNet-like structure, i.e., two separate networks to encode the appearance and temporal coherence, respectively. The best results are in bold. ↑ means the higher the better and ↓ means the opposite.

Method	L1 ↓	PSNR* ↑	SSIM ↑	LPIPS ↓	FVD ↓
UniAnimate *w/o* reference pose	3.07E−04	18.45	0.735	0.276	182.41
UniAnimate *w/o* unified VDM	3.12E−04	18.09	0.712	0.291	205.28
UniAnimate	**2.66E−04**	**20.58**	**0.811**	**0.231**	**148.06**

Table 5 Quantitative comparison of different temporal modeling manners on the TikTok dataset. The best results are in bold. ↑ means the higher the better and ↓ means the opposite.

Method	L1 ↓	PSNR* ↑	SSIM ↑	LPIPS ↓	FVD ↓
Temporal Mamba	**2.47E−04**	**20.81**	0.804	**0.222**	156.26
Temporal Transformer (default)	2.66E−04	20.58	**0.811**	0.231	**148.06**

methods fail to produce satisfactory results. In contrast, the proposed UniAnimate consistently generates high-quality and coherent pose transfer results that adhere to the input conditions, demonstrating remarkable controllability. We attribute our advanced performance to the use of a unified video diffusion model to handle both reference image and noised video simultaneously, resulting in a common feature space for appearance alignment and motion modeling, facilitating model optimization.

Human evaluation. In order to further assess the performance of our method, we incorporate an additional human evaluation. We randomly sampled 50 images and pose sequences and used them to generate videos. The videos are evaluated by 4 different human raters. Each rater is asked to score based on visual quality, identity preservation, and temporal consistency, with scores ranging from 0.2, 0.4, 0.6, 0.8, and 1.0, with 1.0 (i.e., 100%) being best and 0.2 (i.e., 20%) being very poor. We averaged the obtained scores to get the final results. The evaluation results are shown in Table 3. The human evaluation results indicate that our method displays favorable visual aesthetics, reliable controllability, and enhanced temporal consistency.

4.3 Ablation study

Analysis of network components. To generate temporally consistent videos that are visually aligned with the given reference image, we introduce a unified video diffusion model. This architecture employs a shared 3D-UNet to handle both appearance alignment and motion modeling. To further improve the appearance alignment, we incorporate a reference pose to facilitate understanding of the reference image's layout and human structure. We conduct an ablation study on the proposed unified video diffusion model architecture and the effect of reference pose, as shown in Table 4. The results indicate that each module contributes significantly to the overall performance improvement. Example cases in Figure 4 further demonstrate the crucial roles of each module. For instance, removing the reference pose may result in undesired artifacts such as "disconnected limbs." On the other hand, the synthesized results may display appearance inconsistencies (e.g., mismatched backgrounds) with the reference image without the unified video diffusion model. This can be attributed to the fact that aligning features into the same space becomes a challenging task with two separate networks. In contrast, our method exhibits remarkable results in appearance alignment.

Varying temporal modeling manners. In our UniAnimate, we introduce the temporal Mamba as an alternative component for temporal modeling in the human image animation model. As illustrated in Table 5, we find that temporal Mamba achieves comparable performance to temporal Transformer, both providing effective temporal modeling. From the comparative results in Figure 5, we notice that both temporal Mamba and temporal Transformer exhibit excellent visually appealing results under our UniAnimate framework, and the generated results are basically close to real videos without any obvious

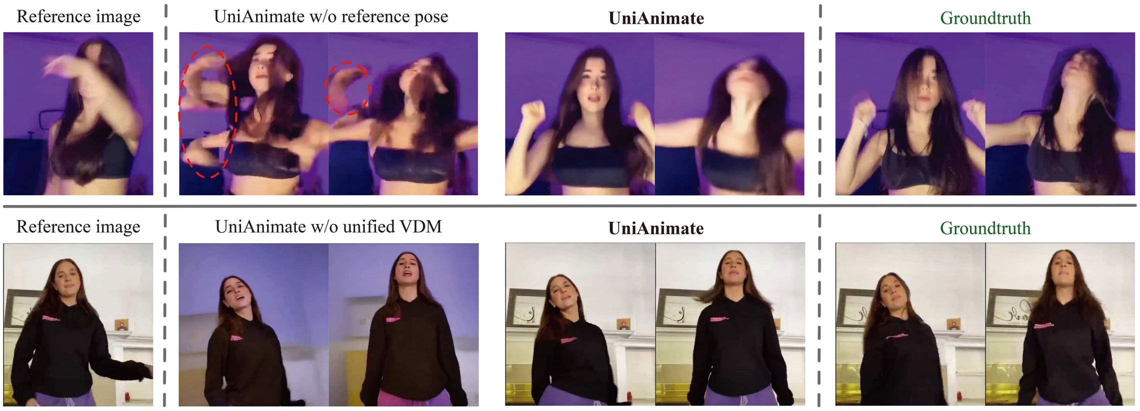

Figure 4 (Color online) Ablation study. To ensure a fair comparison, the same random noise is imposed on the baseline methods and `UniAnimate`.

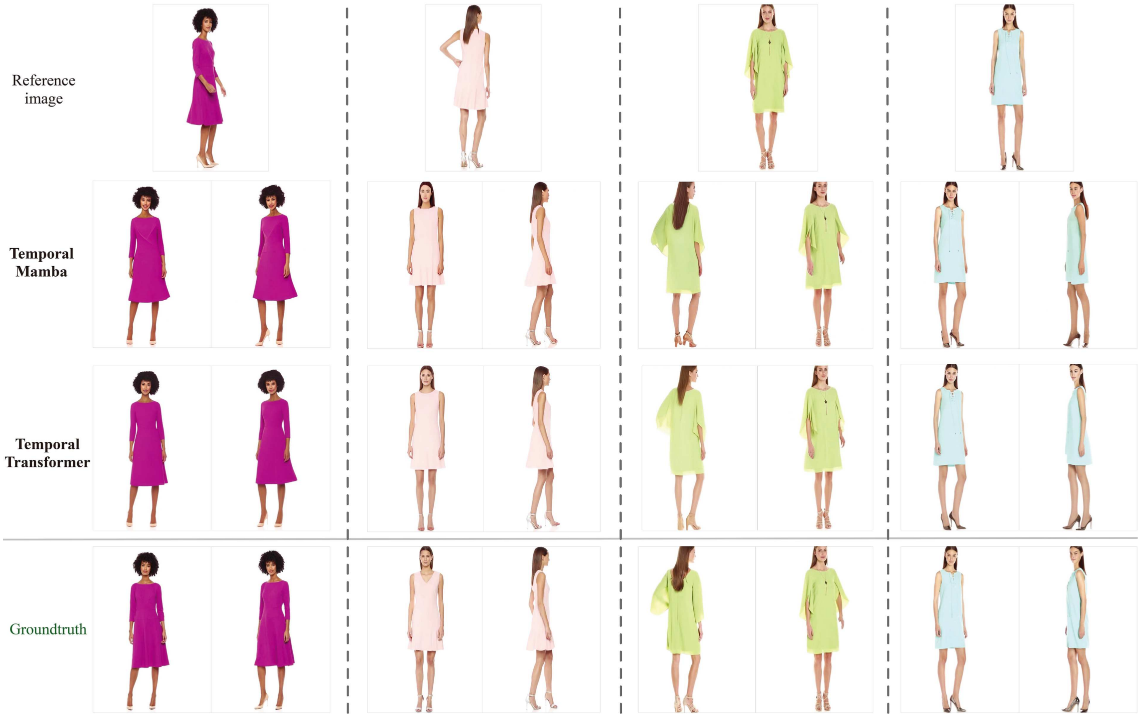

Figure 5 (Color online) Generated video examples on the Fashion dataset. Results are generated by `UniAnimate` with temporal Mamba and temporal Transformer.

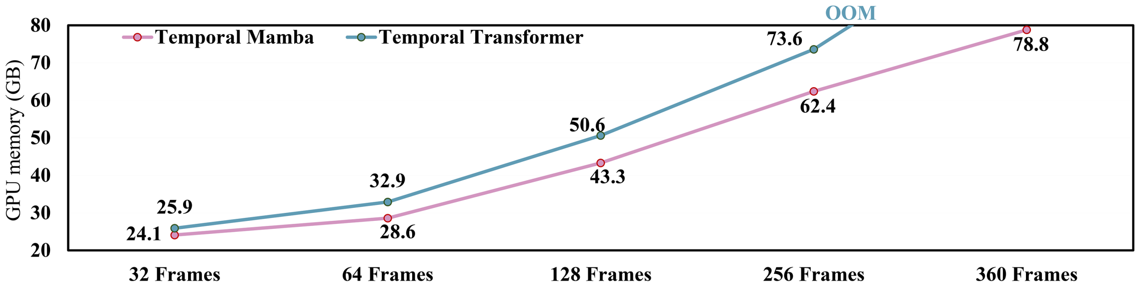

Figure 6 (Color online) Quantitative comparison of different temporal modeling manners about inference memory cost (GB). "OOM" is short for out of memory. Experiments are conducted on NVIDIA 80G A100 GPUs. Note that inference memory and training memory are not the same, and training memory will be much larger since extra gradient calculations are involved. The inference overhead is for the entire framework, including the CLIP encoder and pose encoder.

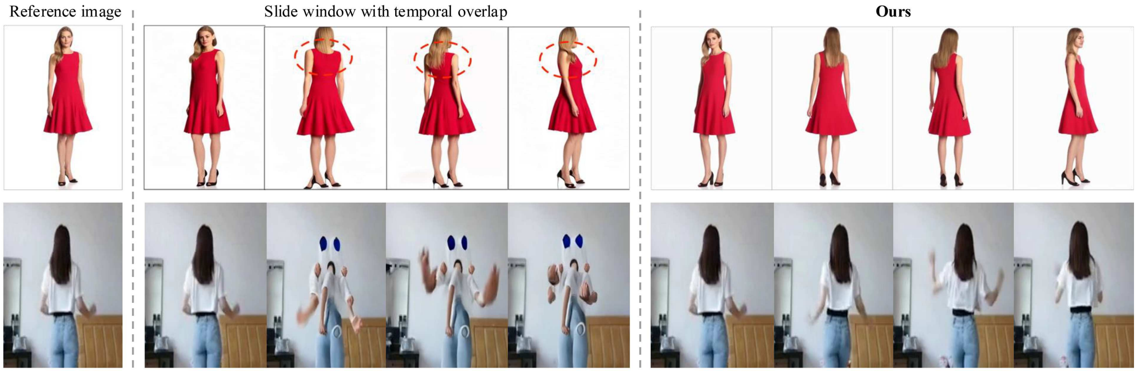

Figure 7 (Color online) Qualitative comparison of different long video generation strategies. Existing methods using the slide window strategy may straggle to synthesize smooth transitions, resulting in discontinuous appearance and inconsistent background.

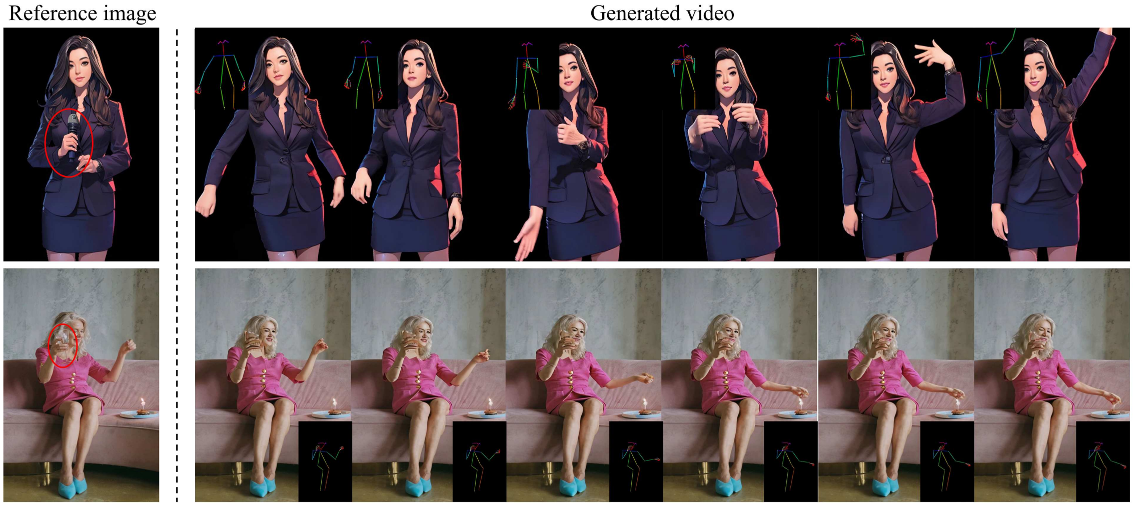

Figure 8 (Color online) Failure cases. If the person in the reference image is holding objects, the generated video may contain artifacts or cause objects to be missing. We attribute this to (1) the scarcity of training videos in which people hold objects in their hands and (2) the lack of explicit conditional guidance to control human-object interactions.

artifacts. Notably, we observe experimentally that temporal Mamba is particularly suitable for handling long sequences in terms of memory cost, as the computational resource overhead grows linearly with time, as shown in Figure 6. We hope that this proposed temporal modeling mechanism can lay the foundation for future research in this domain, especially long-range temporal modeling.

5 Long video generation with smooth transitions

To enable the generation of long videos, we incorporate the unified noised input that supports first frame conditioning, which allows us to continue generating subsequent video frames by leveraging the final frame of the previously generated segment and the reference image. As illustrated in Figure 7, we compare our first frame conditioning solution with the slide window strategy used in [4] and observe that the slide window strategy may suffer from unsatisfactory transition results, such as discontinuous appearance and inconsistent background. We attribute this to the fact that the input pose sequence and the difficulty of denoising are different between two adjacent windows, so directly averaging the intersecting parts may damage the generated results and bring in artifacts. In contrast, the first frame conditioning technique used in our `UniAnimate` can keep the last frame of the previous segment the same as the beginning frame of the following segment, thus achieving a smooth transition.

6 Limitations

Although UniAnimate achieves superior results compared to existing state-of-the-art approaches, there are still some limitations. (i) Generating realistic and fine-grained details in facial and hand regions (see the second line of Figure 8) remains challenging. (ii) During the long video generation process, the completion of invisible parts by different video segments may be inconsistent. This inconsistency can occasionally lead to temporal artifacts, disrupting the overall continuity of the generated videos. (iii) If the character in the reference image is holding objects, the generated video may contain artifacts or cause objects to be missing, as displayed in Figure 8. We attribute this to the scarcity of training videos in which people hold objects in their hands and the lack of explicit conditional guidance to control human-object interactions. In the future, we will focus on collecting high-quality HD videos and designing cross-segment interaction strategies to achieve more consistent human image animation results.

7 Conclusion

In this paper, we presented UniAnimate, a novel approach for generating high-fidelity, temporally smooth videos for human image animation. By introducing the unified video diffusion model, the unified noised input, and temporal Mamba, we address the appearance misalignment limitation of existing methods and achieve improved video generation quality and efficiency. Extensive experimental results quantitatively and qualitatively validate the effectiveness of the proposed UniAnimate and highlight its potential for practical application deployment.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. U22B2053, 623B2039) and Alibaba Group through Alibaba Research Intern Program.

References

1 Yang C, Wang Z, Zhu X, et al. Pose guided human video generation. In: Proceedings of European Conference on Computer Vision, 2018. 201–216

2 Zablotskaia P, Siarohin A, Zhao B, et al. Dwnet: dense warp-based network for pose-guided human video generation. 2019. ArXiv:1910.09139

3 Hu L. 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, 2024. 8153–8163

4 Xu Z, Zhang J, Liew J H, et al. MagicAnimate: temporally consistent human image animation using diffusion model. 2023. ArXiv:2311.16498

5 Chang D, Shi Y, Gao Q, et al. Magicdance: realistic human dance video generation with motions & facial expressions transfer. 2023. ArXiv:2311.12052

6 Jiang Y, Yang S, Qiu H, et al. Text2human: text-driven controllable human image generation. ACM Trans Graphics, 2022, 41: 1–11

7 Hong W, Ding M, Zheng W, et al. Cogvideo: large-scale pretraining for text-to-video generation via Transformers. In: Proceedings of International Conference on Learning Representations, 2023

8 Singer U, Polyak A, Hayes T, et al. Make-A-Video: text-to-video generation without text-video data. In: Proceedings of International Conference on Learning Representations, 2023

9 Ho J, Chan W, Saharia C, et al. Imagen video: high definition video generation with diffusion models. 2022. ArXiv:2210.02303

10 Luo Z, Chen D, Zhang Y, et al. Videofusion: decomposed diffusion models for high-quality video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 10209–10218

11 Zhang S, Wang J, Zhang Y, et al. I2VGen-XL: high-quality image-to-video synthesis via cascaded diffusion models. 2023. ArXiv:2311.04145

12 Tulyakov S, Liu M Y, Yang X, et al. MocoGAN: decomposing motion and content for video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. 1526–1535

13 Ma Y, Zhang S, Wang J, et al. Dreamtalk: when expressive talking head generation meets diffusion probabilistic models. 2023. ArXiv:2312.09767

14 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 33: 6840–6851

15 Guo Y, Yang C, Rao A, et al. Animatediff: animate your personalized text-to-image diffusion models without specific tuning. In: Proceedings of International Conference on Learning Representations, 2024

16 Wang X, Yuan H, Zhang S, et al. Videocomposer: compositional video synthesis with motion controllability. In: Proceedings of Advances in Neural Information Processing Systems, 2023

17 Chen H, Xia M, He Y, et al. Videocrafter1: open diffusion models for high-quality video generation. 2023. ArXiv:2310.19512

18 Wang X, Zhang S, Yuan H, et al. A recipe for scaling up text-to-video generation with text-free videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024

19 Wang X, Zhang S, Zhang H, et al. Videolcm: video latent consistency model. 2023. ArXiv:2312.09109

20 Zhang Y, Wei Y, Jiang D, et al. Controlvideo: training-free controllable text-to-video generation. 2023. ArXiv:2305.13077

21 Zhao M, Wang R, Bao F, et al. Controlvideo: adding conditional control for one shot text-to-video editing. 2023. ArXiv:2305.17098

22 Wang Y, Bilinski P, Bremond F, et al. G3an: disentangling appearance and motion for video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 5264–5273

23 Yu W Y, Po L M, Cheung R C, et al. Bidirectionally deformable motion modulation for video-based human pose transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 7502–7512

24 Zhang P, Yang L, Lai J H, et al. Exploring dual-task correlation for pose guided person image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 7713–7722

25 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems, 2014

26 Wang T, Li L, Lin K, et al. Disco: disentangled control for referring human dance generation in real world. In: Proceedings of International Conference on Learning Representations, 2024

27 Zhu S, Chen J L, Dai Z, et al. Champ: controllable and consistent human image animation with 3D parametric guidance. 2024. ArXiv:2403.14781

28 Ma Y, He Y, Cun X, et al. Follow your pose: pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 4117–4125

29 Karras J, Holynski A, Wang T C, et al. Dreampose: fashion video synthesis with stable diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 22680–22690

30 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. 3836–3847

31 Blattmann A, Rombach R, Ling H, et al. Align your latents: high-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 22563–22575

32 Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. 2023. ArXiv:2312.00752

33 Zhu L, Liao B, Zhang Q, et al. Vision mamba: efficient visual representation learning with bidirectional state space model. 2024. ArXiv:2401.09417

34 Li K, Li X, Wang Y, et al. Videomamba: state space model for efficient video understanding. 2024. ArXiv:2403.06977

35 Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 10684–10695

36 Nichol A Q, Dhariwal P, Ramesh A, et al. Glide: towards photorealistic image generation and editing with text-guided diffusion models. In: Proceedings of International Conference on Machine Learning, 2022. 16784–16804

37 Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with clip latents. 2022. ArXiv:2204.06125

38 Mou C, Wang X, Xie L, et al. T2I-adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models. 2023. ArXiv:2302.08453

39 Huang L, Chen D, Liu Y, et al. Composer: creative and controllable image synthesis with composable conditions. In: Proceedings of International Conference on Machine Learning, 2023

40 Song J, Meng C, Ermon S. Denoising diffusion implicit models. In: Proceedings of International Conference on Learning Representations, 2021

41 Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 35: 36479–36494

42 Liu M, Wei Y X, Wu X H, et al. Survey on leveraging pre-trained generative adversarial networks for image editing and restoration. Sci China Inf Sci, 2023, 66: 151101

43 Wang J, Yuan H, Chen D, et al. Modelscope text-to-video technical report. 2023. ArXiv:2308.06571

44 Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015. 234–241

45 Wu J Z, Ge Y, Wang X, et al. Tune-a-video: one-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 7623–7633

46 Chai W, Guo X, Wang G, et al. Stablevideo: text-driven consistency-aware diffusion video editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 23040–23050

47 Ceylan D, Huang C H P, Mitra N J. Pix2video: video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 23206–23217

48 Zhou D, Wang W, Yan H, et al. Magicvideo: efficient video generation with latent diffusion models. 2022. ArXiv:2211.11018

49 An J, Zhang S, Yang H, et al. Latent-shift: latent diffusion with temporal shift for efficient text-to-video generation. 2023. ArXiv:2304.08477

50 Xing Z, Dai Q, Hu H, et al. Simda: simple diffusion adapter for efficient video generation. 2023. ArXiv:2308.09710

51 Qing Z, Zhang S, Wang J, et al. Hierarchical spatio-temporal decoupling for text-to-video generation. 2023. ArXiv:2312.04483

52 Yuan H, Zhang S, Wang X, et al. Instructvideo: instructing video diffusion models with human feedback. 2023. ArXiv:2312.12490

53 Wei Y, Zhang S, Qing Z, et al. Dreamvideo: composing your dream videos with customized subject and motion. 2023. ArXiv:2312.04433

54 Chen H, Zhang Y, Cun X, et al. Videocrafter2: overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 7310–7320

55 Wang Y, Chen X, Ma X, et al. Lavie: high-quality video generation with cascaded latent diffusion models. 2023. ArXiv:2309.15103

56 Zhang D J, Wu J Z, Liu J W, et al. Show-1: marrying pixel and latent diffusion models for text-to-video generation. Int J Comput Vis, 2025, 133: 1879–1893

57 Chen X, Wang Y, Zhang L, et al. Seine: short-to-long video diffusion model for generative transition and prediction. In: Proceedings of International Conference on Learning Representations, 2024

58 Esser P, Chiu J, Atighehchian P, et al. Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 7346–7356

59 Xing J, Xia M, Liu Y, et al. Make-your-video: customized video generation using textual and structural guidance. 2023. ArXiv:2306.00943

60 Yin S, Wu C, Liang J, et al. Dragnuwa: fine-grained control in video generation by integrating text, image, and trajectory. 2023. ArXiv:2308.08089

61 Chen T S, Lin C H, Tseng H Y, et al. Motion-conditioned diffusion model for controllable video synthesis. 2023. ArXiv:2304.14404

62 Siarohin A, Lathuilière S, Tulyakov S, et al. First order motion model for image animation. In: Proceedings of Advances in Neural Information Processing Systems, 2019. 32

63 Li Y, Huang C, Loy C C. Dense intrinsic appearance flow for human pose transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 3693–3702

64 Siarohin A, Woodford O J, Ren J, et al. Motion representations for articulated animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 13653–13662

65 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. 3657–3666

66 Xu Z, Wei K, Yang X, et al. Do you guys want to dance: zero-shot compositional human dance generation with multiple persons. 2024. ArXiv:2401.13363

67 Zhu B, Wang F, Lu T, et al. Poseanimate: zero-shot high fidelity pose controllable character animation. 2024. ArXiv:2404.13680

68 Lea C, Flynn M D, Vidal R, et al. Temporal convolutional networks for action segmentation and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017. 156–165

69 Wang X, Zhang S, Qing Z, et al. Oadtr: online action detection with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 7565–7575

70 Arnab A, Dehghani M, Heigold G, et al. Vivit: a video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 6836–6846

71 Wang X, Zhang S, Qing Z, et al. Self-supervised learning for semi-supervised temporal action proposal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 1905–1914

72 Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3D residual networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017. 5533–5541

73 Gupta S, Keshari A, Das S. RV-GAN: recurrent gan for unconditional video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 2024–2033

74 Li Y, Min M, Shen D, et al. Video generation from text. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2018

75 Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? In: Proceedings of International Conference on Machine Learning, 2021. 4

76 Liang D K, Chen X W, Xu W, et al. TransCrowd: weakly-supervised crowd counting with transformers. Sci China Inf Sci, 2022, 65: 160104

77 Li K, Guo D, Wang M. ViGT: proposal-free video grounding with a learnable token in the transformer. Sci China Inf Sci, 2023, 66: 202102

78 Shao Y F, Geng Z C, Liu Y T, et al. CPT: a pre-trained unbalanced transformer for both Chinese language understanding and generation. Sci China Inf Sci, 2024, 67: 152102

79 Chen H X, Li H X, Li Y H, et al. Sparse spatial transformers for few-shot learning. Sci China Inf Sci, 2023, 66: 210102

80 Li Z, Yang B, Liu Q, et al. Monkey: image resolution and text label are important things for large multi-modal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024

81 Liu Y, Yang B, Liu Q, et al. Textmonkey: an OCR-free large multimodal model for understanding document. 2024. ArXiv:2403.04473

82 Wang X, Zhang S, Yuan H, et al. Few-shot action recognition with captioning foundation models. 2023. ArXiv:2310.10125

83 Wang X, Zhang S, Cen J, et al. CLIP-guided prototype modulating for few-shot action recognition. Int J Comput Vis, 2024, 132: 1899–1912

84 Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces. 2021. ArXiv:2111.00396

85 Liu Y, Tian Y, Zhao Y, et al. Vmamba: visual state space model. 2024. ArXiv:2401.10166

86 Yang C, Chen Z, Espinosa M, et al. Plainmamba: improving non-hierarchical mamba in visual recognition. 2024. ArXiv:2403.17695

87 Chen G, Huang Y, Xu J, et al. Video mamba suite: state space model as a versatile alternative for video understanding. 2024. ArXiv:2403.09626

88 Jafarian Y, Park H S. 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, 2021. 12753–12762

89 Yang Z, Zeng A, Yuan C, et al. Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 4210–4220

90 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning, 2021. 8748–8763

91 Loshchilov I, Hutter F. Decoupled weight decay regularization. 2017. ArXiv:1711.05101

92 Hore A, Ziou D. Image quality metrics: PSNR vs. SSIM. In: Proceedings of International Conference on Pattern Recognition, 2010. 2366–2369

93 Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process, 2004, 13: 600–612

94 Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. 586–595

95 Unterthiner T, van Steenkiste S, Kurach K, et al. Towards accurate generative models of video: a new metric & challenges. 2018. ArXiv:1812.01717

96 Ren Y, Fan X, Li G, et al. Neural texture extraction and distribution for controllable person image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 13535–13544

97 Bhunia A K, Khan S, Cholakkal H, et al. Person image synthesis via denoising diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 5968–5976