DiffPerformer: Iterative Learning of Consistent Latent Guidance for Diffusion-based Human Video Generation

Chenyang Wang[†,1], Zerong Zheng[2], Tao Yu[3], Xiaoqian Lv[1], Bineng Zhong[4], Shengping Zhang[*,1,5], Liqiang Nie[1]
[1]Harbin Institute of Technology [2]NNKosmos Technology [3]Tsinghua University
[4]Guangxi Normal University [5]Peng Cheng Laboratory
{c.wang, xiaoqian.lv}@stu.hit.edu.cn, zrzheng1995@foxmail.com, ytrock@tsinghua.edu.cn, bnzhong@gxnu.edu.cn, s.zhang@hit.edu.cn, nieliqiang@gmail.com

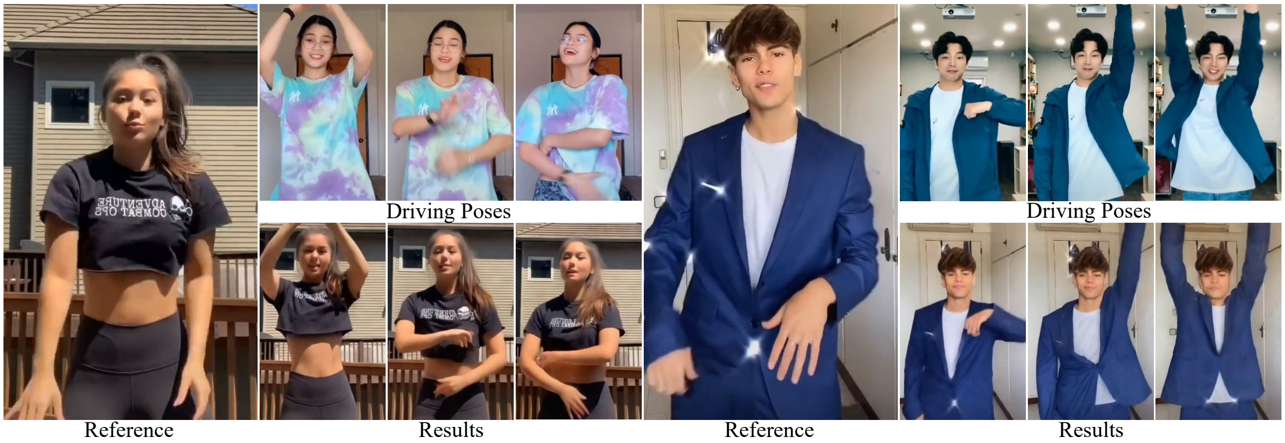

Figure 1. **Human performance videos generated by DiffPerformer** given the reference appearance and driving poses. The generated videos contain realistic dynamic details and coherent appearances across the long-range sequence especially under challenging poses.

Abstract

Existing diffusion models for pose-guided human video generation mostly suffer from temporal inconsistency in the generated appearance and poses due to the inherent randomization nature of the generation process. In this paper, we propose a novel framework, DiffPerformer, to synthesize high-fidelity and temporally consistent human video. Without complex architecture modification or costly training, DiffPerformer finetunes a pre-trained diffusion model on a single video of the target character and introduces an implicit video representation as a proxy to learn temporally consistent guidance for the diffusion model. The guidance is encoded into VAE latent space and an iterative optimization loop is constructed between the implicit video representation and the diffusion model, allowing to harness the smooth property of the implicit video representation and the generative capabilities of the diffusion model in a mutually beneficial way. Moreover, we propose 3D-aware human flow as a temporal constraint during the optimization to explicitly model the correspondence between driving poses and human appearance. This alleviates the misalignment between driving poses and target performer and therefore maintains the appearance coherence under various motions. Extensive experiments demonstrate that our method outperforms the state-of-the-art methods. The code is available at https://github.com/aipixel/DiffPerformer.

1. Introduction

Pose-guided human video generation aims to produce a video of a specific character performing the given poses, which has a wide range of applications such as human-

[†] Work done during an internship at Tsinghua University.
[*] Corresponding author.

computer interaction, motion analysis and VR. Although significant efforts have been devoted to developing effective human video generation, it is still a challenging task due to the variety of body poses and intricate human details.

In early stage, researchers adopt GANs to tackle the task [8, 23, 50, 51]. Unfortunately, GANs suffer from mode collapse and unstable training, failing to generate pleasant results for diverse visual contents in the real world such as human performance. Recently, diffusion models [13, 44] show promising results in generating photo-realistic and diverse images. More importantly, diffusion models are capable of generating images under conditions [18, 37, 47, 61], allowing users to control the generated content over various attributes. Motivated by this progress, several methods propose to apply diffusion models for controllable generation of human images or videos [19, 26, 52, 56], and demonstrate great potential in generating high-quality human appearances for the given poses.

Unfortunately, challenges remain when it comes to generating temporally consistent videos of human performance using diffusion models. Some existing works use only text prompts to control the human identity [26, 56], making them unable to produce authentic and consistent appearance. Others like DisCo [52] and DreamPose [19] condition diffusion models on a reference image to control the human identity and train the networks on large-scale real human motion datasets. However, temporal jittering and flickering are still ubiquitous in their results even after adopting spatio-temporal attention [19] or finetuning on person-specific videos [52]. We speculate that the temporal inconsistency is primarily caused by the inherent randomization nature in the generation process of diffusion models, and consequently it cannot be easily resolved through model tuning or architecture modification.

Therefore, in this paper, we address the issue of temporal consistency from a different perspective. Unlike existing methods that apply complex modification to diffusion models and large-scale training on human motion datasets, we simply finetune a video diffusion model on a single video and introduce a temporal proxy to iteratively learn a consistent latent guidance, which enables the diffusion model to achieve coherent human performance representation and synthesis. The core idea is inspired by the recent success in text-to-3D generation, where 2D image diffusion models are adapted to generate view-consistent images through the optimization of 3D proxy representation [28]. We extend this observation to the temporal domain and introduce an implicit video representation as the temporal proxy. It maps pixel positions into color values with coordinate-based MLPs [20, 24] and factorizes the human performance video into a canonical space and a temporal deformation field [32]. Such a representation enforces a prior on video smoothness and appearance coherence. With this implicit video representation, we present DiffPerformer, a novel human performance synthesis framework where the implicit video representation and diffusion models are mutually beneficial, allowing high-fidelity human video synthesis without temporal inconsistency as shown in Fig. 1.

Our framework starts by finetuning a personalized pose-guided diffusion model on a single video to yield frames corresponding to given pose sequence. These frames are then distilled into the video representation to leverage inherent smoothness of the representation for effectively eliminating temporal inconsistencies, yet it may result in a loss of appearance details. Hence we encode the smoothed frames into latent space and employ the diffusion model to enhance the details from the consistent latent guidance in return, ideally closing a feedback loop. Furthermore, we utilize the implicit video representation as the denoising initialization for the diffusion model, eventually leading to an iterative joint optimization algorithm. In this way, we can fully harness the smooth property of our video representation and the generative capabilities of the diffusion model. Besides, we propose 3D-aware human flow as a temporal constraint during the iterative joint optimization to improve the alignment between poses and appearance. It explicitly models the correspondence between pose and human appearance by mapping the driven signals into the shape of a specific character, which alleviates the shape misalignment between guided poses and target characters and maintains the appearance coherence under various motion.

The contributions are summarized as follows:

- We propose a human performance synthesis framework, DiffPerformer, that for the first time introduces an implicit video representation as consistent latent guidance to enforced the temporal consistency of the diffusion model, enabling high-fidelity, coherent and pose-aligned human video synthesis.
- We present an iterative joint optimization algorithm to integrate the diffusion model and the implicit video representation in a mutually beneficial way, which fully harnesses the smooth property of the video representation and the generative capabilities of diffusion models.
- We present 3D-aware human flow as a temporal constraint to explicitly build the correlation between motion and specific character, leading to content consistency under various motion.
- Extensive evaluations and applications demonstrate that our method outperforms the state-of-the-art methods on pose-guided human video generation.

2. Related Work

Text-to-video Diffusion Model. Despite the remarkable progress in text-to-image (T2I) diffusion models [13, 36, 37, 44, 54, 59], expanding this progress to the video domain is still challenging as these models fail to maintain temporal consistency across frames. To address these issues, there have been significant efforts in video diffu-

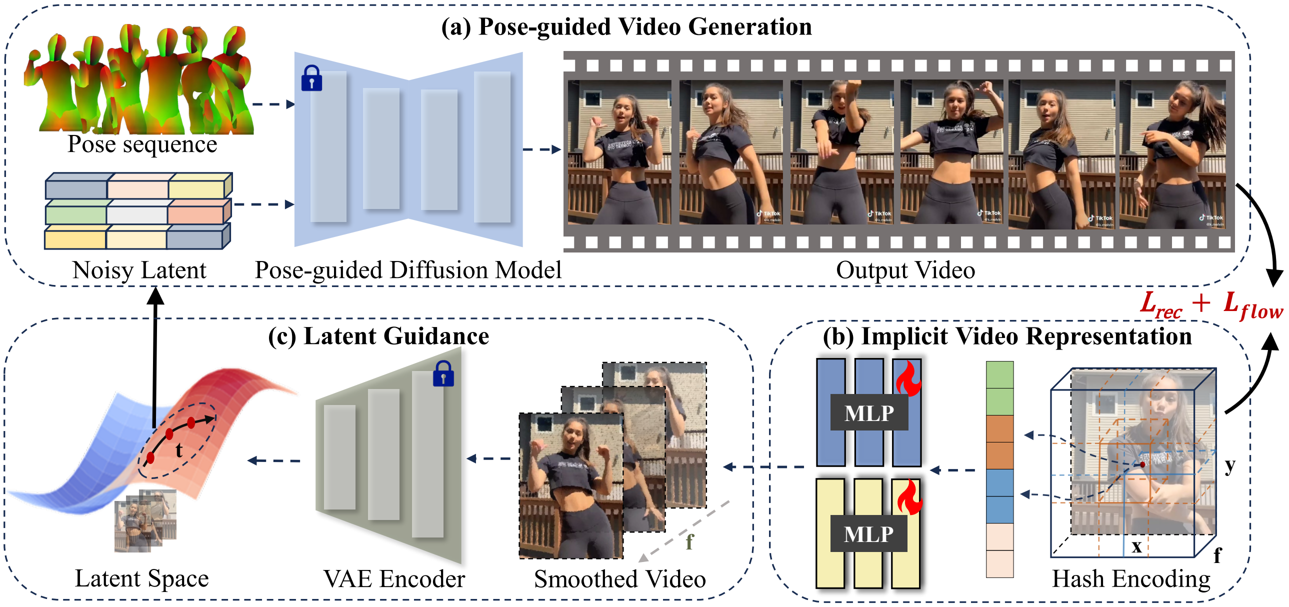

Figure 2. **Overview of DiffPerformer** comprised of (a) Pose-guided Video Generation, (b) Implicit Video Representation and (c) Latent Guidance. These components close a feedback loop that embeds implicit video representation as a temporal proxy to enforce the pose-guided diffusion model to generate a pose-aligned and coherent human video without temporal inconsistency.

sion models [25, 27, 58], aiming to learn the temporal distribution from a large-scale video dataset to generate videos. However, these methods require large-scale video datasets to train, which makes them computationally expensive. Recently, many researches explore to make use of the learned image prior from T2I models for video generation [1, 3, 21, 65]. For instance, Make-A-Video [41] adds spatio-temporal convolution and attention layers to pretrained T2I models for generating consistent video from given texts. Also based on the T2I model, Latent-Shift [1] designs a feature offset strategy to splice multi-frame features for temporal alignment, while Catanzaro et al. [9] separate the noise into shared and independent parts to generate consistent results. Unfortunately, these methods can only generate short video clips randomly or based on a given text and do not allow for finer control.

Controllable Diffusion Model. Adding controls over various attributes into image and video synthesis is important for real-world content creation. ControlNet [61] provides a flexible and effective way to add condition for image generation without modifying the structure of pre-trained diffusion models. Composer [15] decomposes an image into multiple factors to train a diffusion model with all these factors, which allows various levels of conditions to control the generation. Motivated by controlled image generation methods, VideoComposer [53] proposes a spatial-temporal condition encoder to inject various control signals for customized video generation. Tune-a-video [56] extends spatial-temporal attention in T2I models from one image to multiple images to produce consistent videos. Besides, many methods [5, 10, 16, 29, 31, 35, 57] attempt to achieve zero-shot controllable video editing through fine-tuning T2I models on a single test video. Based on these approaches, recent research has been conducted for generating videos with consistent character and realistic movements. Ma et al. [26] design inter-frame attention and finetune the pre-trained ControlNet [61] with many background videos to generate pose-guided video. Unfortunately, this method is limited to producing the character with a given text (e.g., Iron Man) and cannot be adapted to a specific real-world person. DreamPose [19] designs an adapter to fuse the encoding results of the given reference image to ensure the authenticity and consistency of a specific person, but it fails to process challenging motions. DisCo [52] presents an architecture with disentangled control to improve the faithfulness of human video synthesis. However, it still suffers from inconsistency due to the lack of temporal information.

Implicit Video Neural Representations. Implicit neural representations are powerful to represent images [6, 42] and 3D scenes [28, 34, 48] in many applications, such as image process [7, 20] and free-view rendering [22, 33, 39, 40, 63, 64]. The capability of implicit representation also benefits the performance of video processing. OmniMotion [49] proposes a globally consistent motion representation to estimate the motion of every pixel in a video. LNA [20] and Lu *et al.* [24] leverage a layer-based implicit video neural representation to enable video edit. CoDef [32] design a temporal deformation field as a new type of video representation to achieve in video editing. Motivated by these methods, we also employ an implicit video representation to enforce the synthesized video to be temporally consistent in terms of both semantics and appearance.

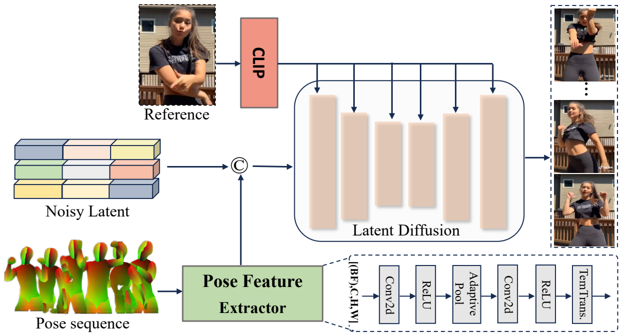

Figure 3. **Illustration of the pose-guided diffusion model.** Given the reference image and pose sequence, the model generates a high-fidelity and pose-aligned human video.

3. Overview

3.1. Preliminary: Latent Diffusion Models

Diffusion models (DMs) [13, 44] are generative models designed to learn a data distribution $p(\boldsymbol{x})$ from the reverse prediction of a Markovian diffusion process. The latent diffusion model (LDM) [37] is a variant that operates in the latent space to improve the computational efficiency when generating high-fidelity images. It leverages an autoencoder to achieve the transformation between images and latent spaces and applies a denoising U-Net [38] for noise estimation. During training, LDM uses an encoder $\mathcal{E}_{\text{enc}}$ to compress an image $\mathbf{I}$ into lower-dimensional latent space $\boldsymbol{z} = \mathcal{E}_{\text{enc}}(\mathbf{I})$. Then, a forward deterministic Gaussian process in T time steps is applied to the image latent $\boldsymbol{z}$ to produce the noisy latent $\boldsymbol{z}_T \sim \mathcal{N}(0, 1)$. A denoising U-Net is trained to predict the noise at each step $t \in \{1, \ldots, T\}$ in order to perform the reverse process. The optimization objective is formulated as

$$\mathcal{L}_{\text{LDM}} = \left\| \epsilon - \epsilon_\theta \left(\boldsymbol{z}_t, \boldsymbol{c}, t \right) \right\|_2 , \tag{1}$$

where ϵ_θ is the denoising U-Net , ϵ is additive Gaussian noise and $\boldsymbol{c}$ represents the condition embedding. After training, LDM applies a sampling process to generate $\boldsymbol{z}$, which is decoded into a high-resolution image by a decoder $\mathcal{E}_{\text{dec}}$.

3.2. Our Framework

A key challenge in generating realistic human videos under given poses (e.g., keypoints [4], DensePose [11]) is to ensure the alignment between the character's movements and the driving signals while maintaining temporally consistent appearance. As shown in Fig. 2, our proposed DiffPerformer achieves this via constructing an iterative optimization loop between implicit video representation and a pose-guided diffusion model. Its core idea is to regard the temporally consistent representation as a proxy to enforce a prior on video smoothness and appearance coherence. Specifically, DiffPerformer first constructs a pose-guided diffusion model and finetunes it on a single video to embed the character appearance, which alleviates the requirement of training on large-scale datasets (Sec. 4.1). Next, a hashing-based implicit video representation is introduced to provide a consistent latent guidance for diffusion models to ensure the temporal consistency of the generated video (Sec. 4.2). We design an iterative optimization loop to connect the latent guidance with the diffusion network, which fully harnesses the smooth property of the video representation and the generative capabilities of the diffusion model (Sec. 4.3).

4. Method

4.1. Pose-guided Diffusion Model

Pretrained image diffusion models [15, 52, 61] have demonstrated their capability to produce diverse and high-quality images but are incapable of generating consistent multi-frame content. To overcome this limitation, we adopt VideoComposer [53], a pretrained video diffusion model to construct a pose-guided diffusion model, achieving controllable realistic human video generation. The main body of the pose-guided diffusion model extends the 2D UNet to a 3D UNet by introducing temporal layers to handle the time dimension. Meanwhile, we extract the features of poses and concatenate them with the input as shown in Fig. 3. To further maintain the character identity, we inject the CLIP embeddings of the reference image I_{ref} into the diffusion model to jointly guide the denoising process.

To better leverage the generative motion priors, we initialize the U-Net of our pose-guided diffusion model using the pertained weights provided by VideoComposer [53]. Nonetheless, the absence of pose-aligned data hampers it from keeping consistent human identities in the synthesized videos. Therefore, we first use Eq. (1) to finetune the pose-guided diffusion model on the video of the specific character in order to further embed the pose-aligned authentic content into the network weights. Additionally, we also finetune the decoder of VAE $\mathcal{E}_{\text{dec}}$ on the given video using

$$\mathcal{L}_{\text{enc}} = \sum_{n=1}^{N} \left\| I_n - \mathcal{E}_{\text{dec}} \left(\mathcal{E}_{\text{enc}} \left(I_n \right) \right) \right\|_2 , \tag{2}$$

where $n \in \{1, \ldots, N\}$ represents the video frame index. After finetuning the diffusion model and the VAE decoder, we acquire a person-specific pose-guided diffusion model, from which we synthesize the initial results V_p for the target character given the pose sequence P. Note that this synthesis process can start from either random noises or other specific initialization.

4.2. Implicit Video Representation

Although we redesign the diffusion model to adapt to pose-guided video generation, the stochasticity of the sampling process leads to undesirable flickering artifacts in the generated video V_p. Therefore, we improve its temporal smoothness by introducing an implicit video representation as a proxy to "distill" the generated video, as discussed below.

4.2.1 Constructing Implicit Video Representation

Following CoDef [32], the implicit video representation first constructs a canonical space to models the texture and details of human appearance. It is achieved by leveraging coordinate-based MLPs $\mathcal{C}$ to map the pixel positions into color values

$$(r, g, b) = \mathcal{C}\left(x, y\right), \tag{3}$$

where (x, y) represents the pixel coordinates and (r, g, b) is the corresponding color values. Then we define a temporal deformation field $\mathcal{D}$ with MLPs to predict the observation-to-canonical deformation of each frame. Therefore, given a pixel position in frame n, its color values are obtained from the implicit video representation via

$$(r, g, b) = \mathcal{C}\left(\mathcal{D}\left(x, y, n\right)\right). \tag{4}$$

However, dynamic human performance videos exhibit complex textures and dynamic details, posing challenges for the networks to represent high-frequency dynamic appearance details. To overcome this obstacle, we adopt a 3D multi-resolution hashing encoding [30] to encode the pixel position of each frame (x, y, n) into high-dimensional features, which allows the representation to capture high-frequency details. Following InstantNGP [30], the multi-resolution hash encoding divides the 3D grid in L levels. Each level operates independently and stores feature vectors at the vertices of a grid. The feature of (x, y, n) in each level is looked up via tri-linear interpolated from its 8-neighboring vertices. After applying the 3D multi-resolution hash encoding, the Eq. (4) can be updated as

$$(r, g, b) = \mathcal{C}\left(\mathcal{D}\left(\mathcal{H}\left(x, y, n\right)\right)\right), \tag{5}$$

where $\mathcal{H}$ denotes the hash encoding. In this way, the implicit video representation predicts the color values of each frame from the pixel position efficiently and effectively.

4.2.2 Training Implicit Video Representation

To train our implicit video representation, we penalize the pixel-wise difference between the video from the diffusion model and the output of the implicit video representation

$$\mathcal{L}_{\text{rec}} = \sum_{n=1}^{N} \left\| I_n - \mathcal{C}\left(\mathcal{D}\left(\mathcal{H}\left(x, y, n\right)\right)\right) \right\|_1, \tag{6}$$

where I_n is the frame n of V_p.

3D-aware Human Flow. With the aforementioned design, our video representation can enhance the stability of generated video in the temporal domain. However, it fails to correct mistakes like misalignment between poses and appearance due to the lack of pose guidance in the optimization process. We mitigate it by proposing a 3D-aware human flow and regarding it as a temporal constraint to improve the coherence between appearance and poses. The proposed flow is calculated from the body mesh estimated from driving videos, which explicitly builds the correspondence between given poses and the reference appearance. Due to the constant mesh topology, it is easy to obtain more accurate motion directions of the human body compared to learning-based flow estimation methods, such as RAFT [45].

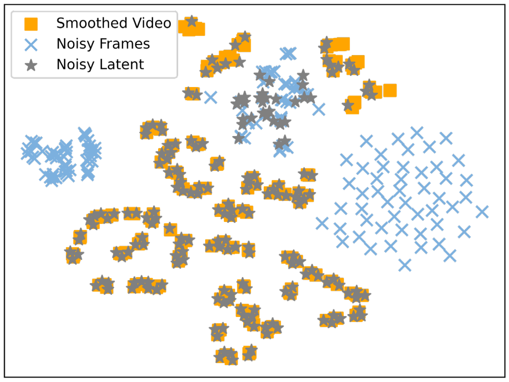

Figure 4. **Visualization of latent features from 200 frames using T-SNE**. Compared with the latent features of noisy images (blue), the noisy latent (grey) exhibits a greater conformity to the latent of the smoothed video (orange), which allows the refined video to preserve the temporal consistency and style of the smoothed video.

Specifically, we first utilize an off-the-shelf human motion capture method [60] to estimate the 3D human mesh $\Theta_n = \{\boldsymbol{\theta}_n, \boldsymbol{\beta}_n, \boldsymbol{\pi}_n\}$ of frame n in driving videos, where $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ represent the pose, shape, and camera parameters, respectively. However, directly employing the Θ_n to compute temporal constraint is inaccessible due to the shape difference between the reference person and driven signals. Therefore, DiffPerformer also extracts the mesh $\Theta_r = \{\boldsymbol{\theta}_r, \boldsymbol{\beta}_r, \boldsymbol{\pi}_r\}$ of the reference image and transforms the shape of the driving pose to the reference character while maintaining the pose parameters. The transformed mesh of frame n in driving video is revised as $\Theta_t = \{\boldsymbol{\theta}_n, \boldsymbol{\beta}_r, \boldsymbol{\pi}_n\}$. Then, we render the 2D depth $D = \{d_1, \ldots, d_N\}$ for all frames from $\{\Theta_1, \ldots, \Theta_N\}$ and combine the depths and meshes to determine the 3D visible points in the estimated meshes. By adopting the nearest neighbor algorithm to attach each pixel in depth map an index from the visible points, our strategy takes into account as dense points as possible for flow calculation. The illustration is shown in Supp.Mat.. Then, the 3D-aware human flow $\mathcal{F}_{n\rightarrow n+1}$ is obtained by calculating the position offset of each point, and the flow loss is formalized as

$$\mathcal{L}_{\text{flow}} = \sum_{n=1}^{N-1} \left\| \mathcal{D}\left(\mathcal{H}\left(x, y, n\right)\right) - \mathcal{D}\left(\mathcal{H}\left(\left(x, y\right) + \mathcal{F}^{x,y}_{n\rightarrow n+1}\right), n+1\right) - \mathcal{F}^{x,y}_{n\rightarrow n+1} \right\|_1. \tag{7}$$

The flow loss efficiently regularizes the consistency of the appearance and poses. The total loss for the implicit

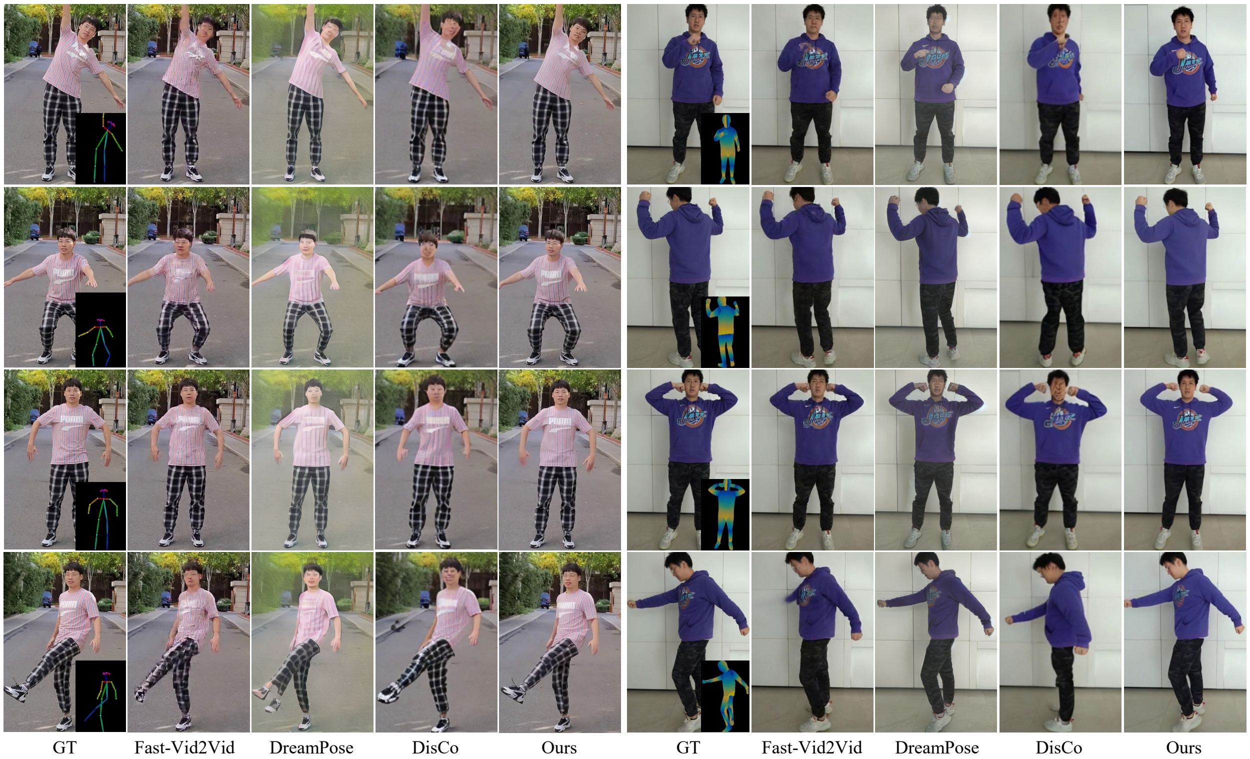

Figure 5. Qualitative comparisons on daily videos. Note that target poses are not available during training. Zoom in for the best view.

video representation is finally defined as

$$\mathcal{L}_{\text{ivr}} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{flow}}, \tag{8}$$

where λ is a hyper-parameter. After optimization, the implicit video representation reconstructs a smoothed video V_s, which alleviates the inconsistency of V_p while maintaining the intricate performer appearance.

4.3. Iterative Joint Optimization

After obtaining the generated pose-aligned video V_p from the pose-guided diffusion model (Sec. 4.1), DiffPerformer leverages Eq. (8) to reconstruct a smoothed result V_s using implicit video representation (Sec. 4.2). As mentioned before, the generated pose-aligned video V_p suffers from an incoherent appearance because of the randomization nature of the denoising diffusion process. Improved by our implicit video representation, the smoothed results V_s become coherent in the temporal domain but scarifies the fine-grained details. To combine the best of both worlds, we propose an iterative joint optimization strategy to borrow the consistency of the smoothed video while refining its details using the pose-guided diffusion model, as shown in Fig. 2.

Specifically, we adopt the outputs of implicit video representation as the latent guidance of diffusion model, and regard the outputs of denoising diffusion process as the optimization goal of implicit video representation. This loop optimization strategy harnesses the smooth property of our implicit video representation and the generative capabilities of the pose-guided diffusion model, leading high-fidelity human video synthesis while maintaining the appearance coherence according to the given poses.

Latent Guidance. To refine the smoothed video using the diffusion model, an intuitive idea is directly adding random noise to the smoothed video and re-denoise it. However, since the denoising process is operated on the latent space from VAE, adding noise on the image level destroys the original consistency of the video and perturbs the direction of generation. Considering this shortcoming, we adopt to add noise to the smoothed video in latent space and take the noisy latent features as the denoising initialization for the diffusion model. Compared with noisy frames, noised latent features closely resemble the original video distribution to maintain temporal consistency, which is illustrated in Fig. 4. This enables the sampling of the diffusion model along a consistent direction from video guidance. Meanwhile, the added noise allows the diffusion model to refine the details of the smoothed video. Then, V_p is updated with the generated video guided by the consistent latent features.

Loop Optimization. To stabilize and accelerate the optimization process, we design a loop optimization strategy. Specifically, we let V_p be obtained through refining V_s using the diffusion model, and in return take V_p as the optimization goal for V_s as in Eq. (6). As the optimization proceeds, V_p is regularly updated every 2000 optimization steps of V_s.

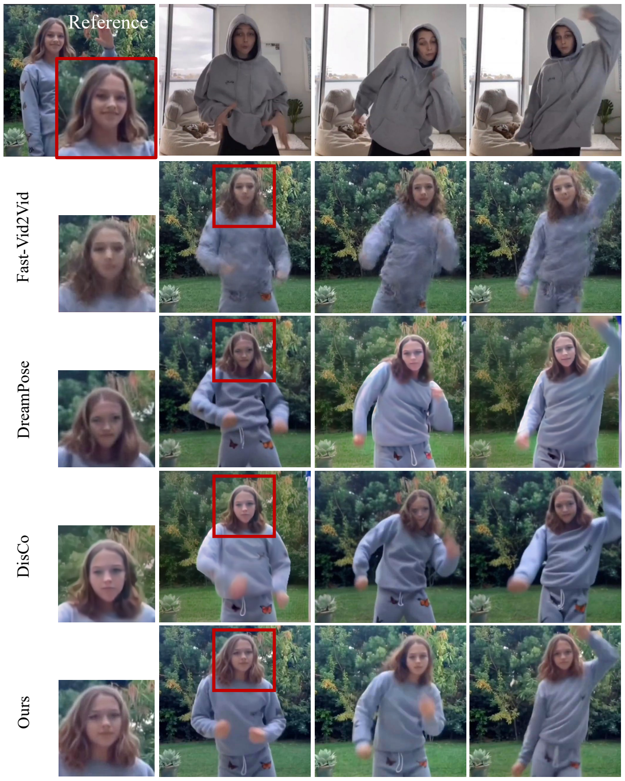

Figure 6. Qualitative comparisons on the Tiktok dataset. Zoom in for the best view.

To warm up the optimization loop, we initialize the implicit video representation using the reference image I_{ref} at the beginning of optimization to facilitate faster convergence. The detailed procedure is shown in Supp.Mat.. In addition, as the details of the V_s gradually converge, the intensity of the noise added to the latent decreases in accordance, which reduces the impact of the generated randomness on the video content. With such iterative joint optimization, DiffPerformer can synthesize realistic human videos with a temporally consistent, pose-aligned human appearance.

5. Experiments

5.1. Experimental Setup

Training Details. We implement DiffPerformer in the Pytorch framework and all experiments are conducted on a single NVidia RTX 3090 GPU with resolution 512×512. For the pose-guided diffusion model, we finetune it on each instance for 10 epochs with a learning rate of 5×10^{-6}. The VAE decoder is finetuned for 2000 steps with a learning rate of 5×10^{-5}. We use a DDIM sampler [43] for 50 steps when performing the denoising diffusion process, and the implicit video representation is optimized in 10000 steps with a learning rate of 1×10^{-3}.

Evaluation Metrics. We evaluate our method and the baselines with the metrics of both the per-frame quality and the temporal quality. To evaluate the image frame quality, we use three widely-adopted metrics, i.e., PSNR [14], SSIM [55], LPIPS [62], FID [12] and L1. For video quality evaluation, we report FID-VID [2] and FVD [46].

Table 1. Quantitative evaluation on test poses.

Method	Fast-Vid2Vid [66]	DreamPose [19]	DisCo [52]	Ours
PSNR ↑	30.4	28.00	29.73	**30.72**
SSIM ↑	0.64	0.42	0.42	**0.69**
LPIPS ↓	0.29	0.41	0.49	**0.22**
FID ↓	38.42	69.84	55.96	**36.00**
L1 ↓	5.96E-5	1.01E-4	6.29E-5	**4.33E-5**
FID-VID ↓	27.06	32.98	23.67	**22.32**
FVD ↓	298.79	529.13	326.12	**254.39**

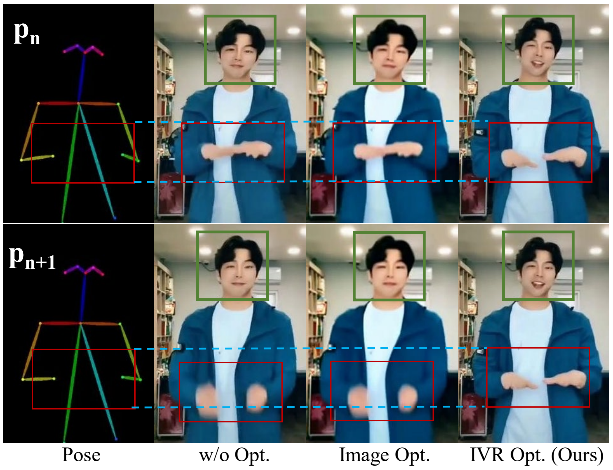

Figure 7. Ablation study of the implicit video representation.

5.2. Comparison

To prove the superiority of our method, we compare with two state-of-the-art methods for diffusion-based human video generation, namely Disco [52] and DreamPose [19]. For a fair comparison, we use the provided checkpoint of these methods and finetune them on the target video. In addition, we also compare with Fast-Vid2Vid [66], a representative GAN-based method for human video synthesis. The experiments are conducted on causally captured videos and the TikTok dataset [17]. Details in Supp.Mat..

Qualitative Comparison. We evaluate the qualitative performance of the proposed DiffPerformer as shown in Figs. 5 and 6. Fig. 5 shows the generation from two kinds of pose signals (keypoints and densepose) on daily videos. Fast-vid2vid and DisCo suffer from artifacts and are unrealistic, especially in the area of the face. On the contrary, our method generates facial results very close to the reference frame and exhibits more detail. Besides, DreamPose fails to restore the accurate color and obtains pose misaligned and inconsistent frames. In Fig. 6, we further compare these methods on the TikTok dataset to validate the performance in various poses (e.g., dancing). It can be seen that the face identity is lost in the compared methods. Meanwhile, the details like limbs blur, if not disappear, in their results.

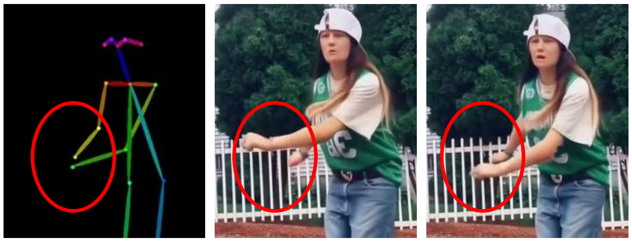

Figure 8. Ablation study of the 3D-aware human flow. The flow helps improve the accuracy of pose alignment.

Besides, DreamPose produces an unreasonable body shape compared with the reference frame. These results indicate that our method realizes higher-fidelity and temporally more consistent results, significantly outperforming others on identity preservation, photo-realism and pose alignment. More results and analysis are provided in Supp.Mat..

Quantitative Comparison. Table 1 reports the quantitative comparison among all methods. As we can see, DiffPerformer outperforms state-of-the-art methods in terms of all metrics, indicating that our results are perceptually best. Image-level metrics (PSNR, SSIM, LPIPS, FID, L1) demonstrate that our method can synthesize the most realistic and compelling results than current state-of-the-arts. Moreover, FID-VID and FVD verify that DiffPerformer is able to generate temporally consistent videos with high-fidelity content, notably surpassing existing approaches.

5.3. Ablation Study

Implicit Video Representation. Implicit video representation (denoted as IVR) is the core component that improves the temporal consistency of the generated videos from the pose-guided diffusion. To verify its effectiveness, we conduct experiments with three different settings: (1) output of the pose-guided diffusion model (w/o Opt.), (2) regarding the coarse video from pose-guided diffusion as smoothed video in Fig. 2 (Image Opt.), (3) the whole framework of DiffPerformer (IVR Opt.). We take two adjacent frames for visualisation, and the results are shown in Fig. 7. We can see that the result of the diffusion process suffers from temporal discontinuities in pose and facial identity. Besides, refining the original output using the diffusion model cannot improve the temporal consistency and even has the problem of color corruption. On the contrary, IVR ensures the temporal consistency and enhance the texture details. Thanks to the flow loss in IVR, the pose misalignment is also corrected in the refined results. More experiment settings and analysis about IVR are provided in Supp.Mat..

3D-aware Human Flow. 3D-aware human flow (denoted as 3DHF) is presented to model the correspondence between appearance and poses during optimization. To evaluate its effectiveness, we conduct various experiments, and present the results in Fig. 8. We can see that when removing the constraint of 3D-aware human flow in the optimization, the pose misalignment occurs especially in the body parts of large-magnitude motion (e.g., limbs). In other words, our 3D-aware human flow is beneficial for keeping the content consistent with challenging poses.

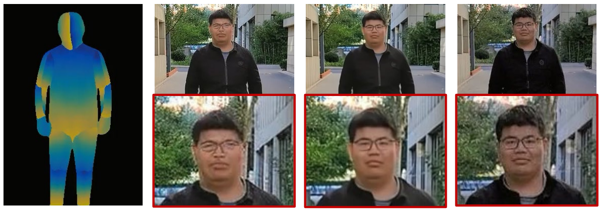

Figure 9. Ablation of VAE Finetuning. Finetuning the VAE decoder can yield more photorealistic details.

VAE Decoder Finetuned. To generate more photorealistic results, we finetune the decoder of VAE in the pose-guided diffusion model. We conduct the experiments to study the effects of finetuning. The comparative results are shown in Fig. 9, which illustrates that the finetuned VAE recovers more details especially in the facial area. Meanwhile, the results also get rid of blurry appearance and become more clear and consistent with character appearance.

6. Conclusion and Discussion

This paper proposes a novel framework, DiffPerfomer, to generate high-fidelity and temporally consistent human video according to driving poses. DiffPerformer introduces an implicit video representation as guidance for the finetuned diffusion model to enforce a prior on video smoothness and appearance coherence, which offers a different perspective to address the problem of temporal inconsistency in the generation process. The proposed iterative joint optimization algorithm with 3D-aware human flow makes the representation and diffusion model mutually beneficial in latent space, improving temporal consistency and enhancing the details under various poses. Thanks to the design of iterative learning of latent guidance, DiffPerformer exhibits superior effectiveness in generating human video.

Limitations. Due to the lack of background content guidance, our method does not work as well in videos under a moving camera. Besides, due to the use of diffusion model, it is more time-consuming compared to methods that do not rely on diffusion models.

Potential Social Impact. DiffPerformer possesses the potential to revolutionize the content creation industry. However, since our method can make realistic personalized appearances, it should be careful about the possibility of misuse, like creating deceptive “deepfakes”.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (Nos. 62272134, 62236003, 62072141 and 62125107), Shenzhen College Stability Support Plan (Grant No. GXWD20220817144428005) and The Major Key Project of PCL (PCL2023A10-2).

References

[1] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. 3

[2] 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*, 2019. 7

[3] 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 *CVPR*, 2023. 3

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE TPAMI*, 2021. 4

[5] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *CVPR*, 2023. 3

[6] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, 2021. 3

[7] Zeyuan Chen, Yinbo Chen, Jingwen Liu, Xingqian Xu, Vidit Goel, Zhangyang Wang, Humphrey Shi, and Xiaolong Wang. Videoinr: Learning video implicit neural representation for continuous space-time super-resolution. In *CVPR*, 2022. 3

[8] Oran Gafni, Lior Wolf, and Yaniv Taigman. Vid2game: Controllable characters extracted from real-world videos. In *ICLR*, 2020. 2

[9] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *ICCV*, 2023. 3

[10] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 3

[11] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 4

[12] 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 *NeurIPS*, 2017. 7

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 4

[14] Alain Horé and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *ICPR*, 2010. 7

[15] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 3, 4

[16] Ziqi Huang, Kelvin C.K. Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *CVPR*, 2023. 3

[17] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *CVPR*, 2021. 7

[18] Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. *arXiv preprint arXiv:2304.08483*, 2023. 2

[19] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025*, 2023. 2, 3, 7

[20] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM TOG*, 2021. 2, 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. *arXiv preprint arXiv:2303.13439*, 2023. 3

[22] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *CVPR*, 2023. 3

[23] Wen Liu, Wenhan Luo Lin Ma Zhixin Piao, Min Jie, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *ICCV*, 2019. 2

[24] Erika Lu, Forrester Cole, Tali Dekel, Weidi Xie, Andrew Zisserman, David Salesin, William T Freeman, and Michael Rubinstein. Layered neural rendering for retiming people in video. *ACM TOG*, 2020. 2, 3

[25] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 2023. 3

[26] 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. *arXiv preprint arXiv:2304.01186*, 2023. 2, 3

[27] Kangfu Mei and Vishal M. Patel. VIDM: video implicit diffusion models. In *AAAI*, 2023. 3

[28] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3

[29] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[30] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 2022. 5

[31] Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, and Vishal M Patel. Unite and conquer: Plug & play multi-modal synthesis using diffusion models. In *CVPR*, 2023. 3

[32] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. *arXiv preprint arXiv:2308.07926*, 2023. 2, 3, 5

[33] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021. 3

[34] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. *ACM TOG*, 2021. 3

[35] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. 3

[36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4

[39] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *CVPR*, 2023. 3

[40] Ruizhi Shao, Jingxiang Sun, Cheng Peng, Zerong Zheng, Boyao Zhou, Hongwen Zhang, and Yebin Liu. Control4d: Efficient 4d portrait editing with text. In *CVPR*, 2024. 3

[41] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023. 3

[42] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020. 3

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 7

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2, 4

[45] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 5

[46] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7

[47] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *SIGGRAPH*, 2023. 2

[48] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, 2021. 3

[49] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *ICCV*, 2023. 3

[50] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Nikolai Yakovenko, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018. 2

[51] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Bryan Catanzaro, and Jan Kautz. Few-shot video-to-video synthesis. In *NeurIPS*, 2019. 2

[52] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2023. 2, 3, 4, 7

[53] 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. In *CVPR*, 2023. 3, 4

[54] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He,

Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 2

[55] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 7

[56] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023. 2, 3

[57] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 3

[58] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *CVPR*, 2023. 3

[59] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 2

[60] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE TPAMI*, 2023. 5

[61] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 4

[62] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[63] Xiaochen Zhao, Lizhen Wang, Jingxiang Sun, Hongwen Zhang, Jinli Suo, and Yebin Liu. Havatar: High-fidelity head avatar via facial model conditioned neural radiance field. *ACM TOG*, 2023. 3

[64] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *ACM TOG*, 2023. 3

[65] 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. 3

[66] Long Zhuo, Guangcong Wang, Shikai Li, Wanye Wu, and Ziwei Liu. Fast-vid2vid: Spatial-temporal compression for video-to-video synthesis. In *ECCV*, 2022. 7