MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model

Zhongcong Xu[1] Jianfeng Zhang[2] Jun Hao Liew[2] Hanshu Yan[2] Jia-Wei Liu[1]
Chenxu Zhang[2] Jiashi Feng[2] Mike Zheng Shou[1]*
[1]Show Lab, National University of Singapore [2]ByteDance
zhongcongxu@u.nus.edu

Abstract

This paper studies the human image animation task, which aims to generate a video of a certain reference identity following a particular motion sequence. Existing animation works typically employ the frame-warping technique to animate the reference image towards the target motion. Despite achieving reasonable results, these approaches face challenges in maintaining temporal consistency throughout the animation due to the lack of temporal modeling and poor preservation of reference identity. In this work, we introduce MagicAnimate, a diffusion-based framework that aims at enhancing temporal consistency, preserving reference image faithfully, and improving animation fidelity. To achieve this, we first develop a video diffusion model to encode temporal information. Second, to maintain the appearance coherence across frames, we introduce a novel appearance encoder to retain the intricate details of the reference image. Leveraging these two innovations, we further employ a simple video fusion technique to encourage smooth transitions for long video animation. Empirical results demonstrate the superiority of our method over baseline approaches on two benchmarks. Notably, our approach outperforms the strongest baseline by over 38% in terms of video fidelity on the challenging TikTok dancing dataset. Code and model will be made available.

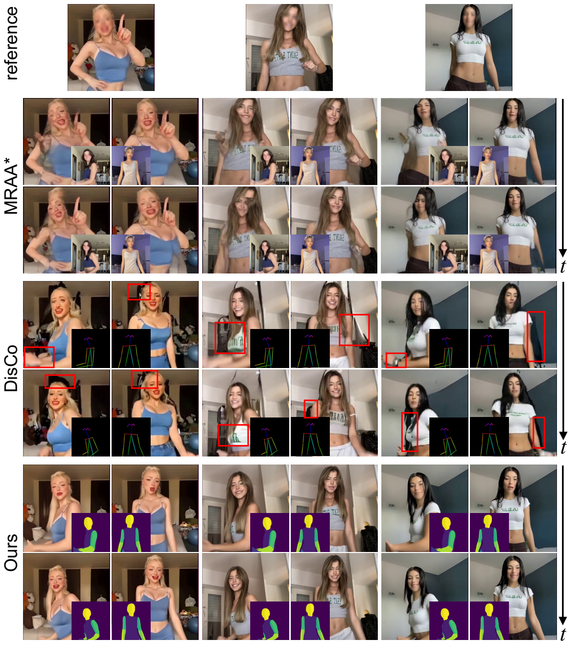


Figure 1. Given a sequence of motion signals, MagicAnimate produces temporally consistent animation for reference identity images, whereas state-of-the-art methods fail to generalize or preserve the reference appearance, as highlighted in red boxes. The motion sequence is overlaid at the corner. *Note that MRAA directly uses video frames as the driving signal. The complete video results can be found on our Project Page.

1. Introduction

Given a sequence of motion signals such as video, depth, or pose, the image animation task aims to bring static images to life. The animation of humans, animals, cartoons, or other general objects, has attracted much attention in research [27, 28, 51]. Among these, human image animation [15, 34, 47] has been the most extensively explored, given its potential applications across various domains, including social media, movie industry, and entertainment, *etc*. In contrast to traditional graphic approaches [9, 40], the abundance of data enables the development of low-cost data-driven animation frameworks [6, 7, 12, 35, 42, 46].

Existing data-driven methods for human image animation can be categorized into two primary groups based on the generative backbone models used, namely GAN-based and diffusion-based frameworks. The former [27, 35] typically employs a warping function to deform the reference image into the target pose and utilize GAN models to extrapolate the missing or occluded body parts. In contrast, the latter [15, 34] harness appearance [21] and pose condi-

[*] Corresponding author

tions [49] to generate the target image based on pretrained diffusion models [23]. Despite generating visually plausible animations, these methods typically exhibit several limitations: 1) GAN-based methods possess restricted motion transfer capability, resulting in unrealistic details in occluded regions and limited generalization ability for cross-identity scenarios, as depicted in Figure 1. 2) Diffusion-based methods, on the other hand, process a lengthy video in a frame-by-frame manner and then stack results along the temporal dimension. Such approaches neglect temporal consistency, resulting in flickering results. In addition, these works typically rely on CLIP [21] to encode reference appearance, which is known to be less effective in preserving details, as highlighted in the red boxes in Figure 1.

In this work, to address the aforementioned limitations, we develop a human image animation framework called MagicAnimate that offers long-range temporal consistency, robust appearance encoding, and high per-frame quality. To achieve this, we first develop a video diffusion model that encodes temporal information by incorporating temporal attention blocks into the diffusion network. Secondly, we introduce an innovative appearance encoder to preserve the human identity and background information derived from the reference image. Unlike existing works that employ CLIP-encoded visual features, our appearance encoder is capable of extracting dense visual features to guide the animation, which leads to better preservation of identity, background, clothes, *etc*. To further improve per-frame fidelity, we additionally devise an image-video joint training strategy to leverage diverse single-frame image data for augmentation, which provides richer visual cues to improve the modeling capability of our framework for details. Lastly, we leverage a surprisingly simple video fusion technique to enable long video animation with smooth transitions.

In summary, our contributions are three-fold: (1) We propose MagicAnimate, a novel diffusion-based human image animation approach that integrates temporal consistency modeling, precise appearance encoding, and temporal video fusion, for synthesizing temporally consistent human animation of arbitrary length. (2) Our method achieves state-of-the-art performance on two benchmarks. Notably, it surpasses the strongest baseline by more than 38% in terms of video quality on the challenging TikTok dancing dataset. (3) MagicAnimate showcases robust generalization ability, supporting cross-identity animation and various downstream applications, including unseen domain animation and multi-person animation.

2. Related Work

2.1. Data-driven Animation

Prior efforts in image animation have predominantly concentrated on the human body or face, leveraging the abundance of diverse training data and domain-specific knowledge, such as keypoints [3, 20, 44], semantic parsing [18], and statistical parametric models [31, 41, 42, 46]. Building upon these motion signals, a long line of work [26, 29, 31, 35, 36, 45] has emerged. These approaches can be classified into two categories based on their animation pipeline, *i.e*., implicit and explicit animation. Implicit animation methods transform the source image to the target motion signal by deforming the reference image in sub-expression space [31] or manipulating the latent space of a generative model [19, 20, 32, 36]. The generative backbone conditions on target motion signal to synthesize animations. Conversely, explicit methods warp the source image to the target by either 2D optical flow [22, 26–28, 45, 51], 3D deformation field [3, 17, 35], or directly sawpping the face of target image [18]. In addition to deforming the source image or 3D mesh, recent research efforts [29, 41, 42, 46] explore explicitly deforming points in 3D neural representations for human body and face synthesis, showcasing improved temporal and multi-view consistency.

2.2. Diffusion Models for Animation

The remarkable progress in diffusion models [23, 24, 30] has propelled text-to-image generation to unprecedented success, spawning numerous subsequent works, such as controllable image generation [49] and video generation [39], *etc*. Recent works have embraced diffusion models for human-centric video generation [16] and animation [34]. Among these works, a common approach [37] develops a diffusion model for generating 2D optical flow and then animates the reference image using frame-warping technique [27]. Moreover, many diffusion-based animation frameworks [15, 34, 47] employ Stable Diffusion [23] as their image generation backbone and leverage ControlNet [49] to condition the animation process on OpenPose [5] keypoint sequences. For the reference image condition, they usually adopt a pretrained image-language model, CLIP [21], to encode the image into a semantic-level text token space and guide the image generation process through cross-attention. While these works yield visually plausible results, most of them process each video frame independently and neglect the temporal information in animation videos, which inevitably leads to flickering animation results.

3. Method

Given a reference image I_{ref} and a motion sequence $\boldsymbol{p}^{1:N} = [\boldsymbol{p}_1, \cdots, \boldsymbol{p}_N]$, where N is the number of frames, our objective is to synthesize a continuous video $I^{1:N} = [I_1, \cdots, I_N]$ with the appearance of I_{ref} while adhering to the provided motion $\boldsymbol{p}^{1:N}$.

Existing diffusion-based frameworks [15, 34] process each frame independently, neglecting the temporal con-

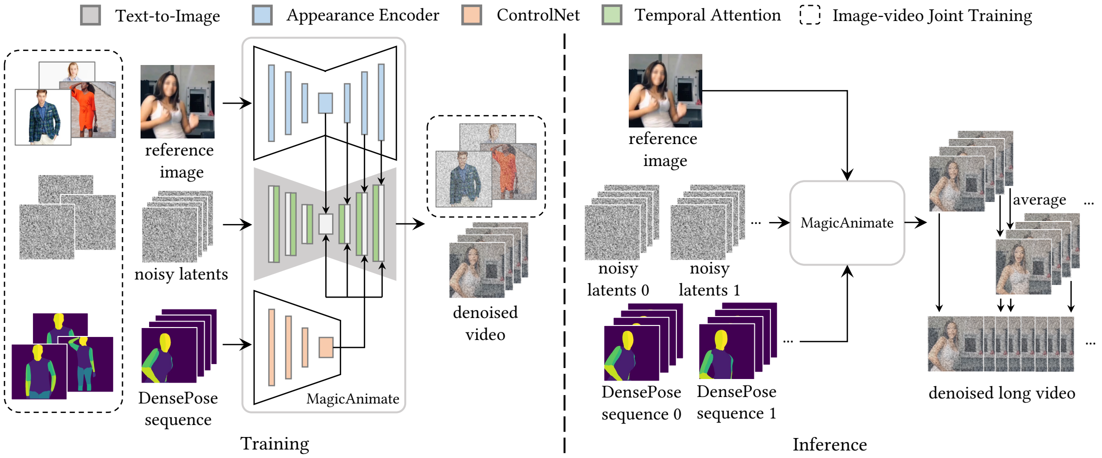


Figure 2. **MagicAnimate pipeline**. Given a reference image and the target DensePose motion sequence, MagicAnimate employs a video diffusion model and an appearance encoder for temporal modeling and identity preserving, respectively (**left panel**). To support long video animation, we devise a simple video fusion strategy that produces smooth video transition during inference (**right panel**).

sistency among different frames, which consequently results in flickering animations. To address this, we build a video diffusion model $\mathcal{F}^{\mathrm{T}}$ for temporal modeling by incorporating temporal attention blocks into the diffusion backbone (Sec. 3.1). In addition, existing works [15, 34] use CLIP [21] encoder to encode the reference image. We argue that these semantic-level features are too sparse and compact to capture intricate details. Therefore, we introduce a novel appearance encoder $\mathcal{F}_{\mathrm{a}}$ (Sec. 3.2) to encode I_{ref} into appearance embedding $\boldsymbol{y}_{\mathrm{a}}$ and condition our model on it for identity- and background-preserving animation.

The overall pipeline of our MagicAnimate (Sec. 3.3) is depicted in Figure 2. We first embed the reference image into appearance embedding $\boldsymbol{y}_{\mathrm{a}}$ using our appearance encoder. We then pass the target pose sequence, *i.e*., DensePose [8], into a pose ControlNet [49] $\mathcal{F}_{\mathrm{p}}$ to extract motion condition $\boldsymbol{y}_{\mathrm{p}}^{1:K}$. Conditioning on these two signals, our video diffusion model is trained to animate the reference human identity to follow the given motions. In practice, due to memory constraints, we process the entire video in a segment-by-segment manner. Thanks to the temporal modeling and robust appearance encoding, MagicAnimate can largely maintain temporal and appearance consistency across segments. Nevertheless, there still exists minor discontinuities between segments. To mitigate this, we leverage a simple video fusion approach to improve the transition smoothness. Specifically, as depicted in Figure 2, we decompose the entire video into overlapping segments and simply average the predictions for overlapping frames. Lastly, we also introduce an image-video joint training strategy to further enhance the reference-preserving capability and single-frame fidelity (Sec. 3.4).

3.1. Temporal Consistency Modeling

To ensure temporal consistency across video frames, we extend the image diffusion model to the video domain. Specifically, we inflate the original 2D UNet to 3D temporal UNet by inserting temporal attention layers [10, 39, 52]. The temporal UNet is denoted as $\mathcal{F}^{\mathrm{T}}(\cdot;\theta^{\mathrm{T}})$ with trainable parameters θ^{T}. The architecture of the inflated UNet blocks is illustrated in Figure 2. We begin with randomly initialized latent noise $\boldsymbol{z}_{\boldsymbol{t}}^{1:K}$ where K is the length of the video frames. We then stack K consecutive poses into a DensePose sequence $\boldsymbol{p}^{1:K}$ for motion guidance. We input $\boldsymbol{z}_{\boldsymbol{t}}^{1:K}$ to our video diffusion backbone $\mathcal{F}^{\mathrm{T}}$ by reshaping the input features from $\mathbb{R}^{N\times C\times K\times H\times W}$ into $\mathbb{R}^{(NK)\times C\times H\times W}$. Within temporal modules, we reshape the features into $\mathbb{R}^{(NHW)\times K\times C}$ to compute cross-frame information along the temporal dimension. Following prior works [10], we add sinusoidal positional encoding to make the model aware of the position for each frame within the video. As such, we compute temporal attention using the standard attention operation, which is formulated as $\mathrm{Attention}(Q,K,V)=\mathrm{Softmax}(\frac{QK^T}{\sqrt{d}})V$, where $Q=W^Q\boldsymbol{z}_{\boldsymbol{t}}$, $K=W^K\boldsymbol{z}_{\boldsymbol{t}}$, $V=W^V\boldsymbol{z}_{\boldsymbol{t}}$. Through this attention mechanism, MagicAnimate aggregates temporal information from neighboring frames and synthesizes K frames with improved temporal consistency.

3.2. Appearance Encoder

The goal of human image animation is to generate results under the guidance of a reference image I_{ref}. The core objective of our appearance encoder is representing I_{ref} with detailed identity- and background-related features that can be injected into our video diffusion model for retargeting under the motion signal guidance. Inspired by recent works on dense reference image conditioning, such as MasaCtrl [4] and Reference-only ControlNet [48], we propose a novel appearance encoder with improved identity and background preservation to enhance single-frame fidelity and temporal coherence. Specifically, our appearance encoder creates another trainable copy of the base UNet $\mathcal{F}_{\mathrm{a}}(\cdot;\theta^{\mathrm{a}})$

and compute the condition features for the reference image I_{ref} for each denoising step $\boldsymbol{t}$. This process is mathematically formulated as

$$\boldsymbol{y}_{\text{a}} = \mathcal{F}_{\text{a}}(\boldsymbol{z_t}|I_{\text{ref}}, \boldsymbol{t}, \theta^{\text{a}}), \tag{1}$$

where $\boldsymbol{y}_{\text{a}}$ is a set of normalized attention hidden states for the middle and upsampling blocks. Different from ControlNet which adds conditions in a residual manner, *we pass these features to the spatial self-attention layers in the UNet blocks* by concatenating each feature in $\boldsymbol{y}_{\text{a}}$ with the original UNet self-attention hidden states to inject the appearance information. Our appearance condition process is mathematically formulated as:

$$\begin{aligned} \text{Attention}(Q, K, V, \boldsymbol{y}_{\text{a}}) = \text{Softmax}(\frac{QK'^T}{\sqrt{d}})V', \\ Q = W^Q\boldsymbol{z_t}, K' = W^K[\boldsymbol{z_t}, \boldsymbol{y}_{\text{a}}], V' = W^V[\boldsymbol{z_t}, \boldsymbol{y}_{\text{a}}], \end{aligned} \tag{2}$$

where $[\cdot]$ denotes concatenation operation. Through this operation, we can adapt the spatial self-attention mechanism in our video diffusion model into a hybrid one. This hybrid attention mechanism can not only maintain the semantic layout of the synthesized image, such as the pose and position of the human in the image, but also query the contents from the reference image in the denoising process to preserve the details, including identity, clothes, accessories, and background. This improved preservation capability benefits our framework in two aspects: (1) our method can transfer the reference image faithfully to the target motion; (2) the strong appearance condition contributes to temporal consistency by retaining the same identity, background, and other details throughout the entire video.

3.3. Animation Pipeline

With the incorporation of temporal consistency modeling and the appearance encoder, we combine these elements with pose conditioning, *i.e.*, ControlNet [49], to transform the reference image to the target poses.

Motion transfer. ControlNet for OpenPose [5] keypoints is commonly employed for animating reference human images. Although it produces reasonable results, we argue that the major body keypoints are sparse and not robust to certain motions, such as rotation. Consequently, we choose DensePose [8] as the motion signal $\boldsymbol{p}_i$ for dense and robust pose conditions. We employ a pose ControlNet $\mathcal{F}_{\text{p}}(\cdot, \theta^{\text{p}})$, the pose condition for frame i is computed as

$$\boldsymbol{y}_{\text{p},i} = \mathcal{F}_{\text{p}}(\boldsymbol{z_t}|\boldsymbol{p}_i, \boldsymbol{t}, \theta^{\text{p}}), \tag{3}$$

where $\boldsymbol{y}_{\text{p},i}$ is a set of condition residuals added to the residuals for the middle and upsampling blocks in the diffusion model. In our pipeline, we concatenate the motion feature of each pose in a DensePose sequence into $\boldsymbol{y}_{\text{p}}^{1:K}$.

Denoising process. Building upon the appearance condition $\boldsymbol{y}_{\text{a}}$ and motion condition $\boldsymbol{y}_{\text{p}}^{1:K}$, MagicAnimate animates the reference image following the DensePose sequence. The noise estimation function $\epsilon_\theta^{1:K}(\cdot)$ in the denoising process is mathematically formulated as:

$$\epsilon_\theta^{1:K}\left(\boldsymbol{z}_t^{1:K}, \boldsymbol{t}, I_{\text{ref}}, \boldsymbol{p}^{1:K}\right) = \mathcal{F}^{\text{T}}(\boldsymbol{z}_{\boldsymbol{t}}^{1:K}|\boldsymbol{t}, \boldsymbol{y}_{\text{a}}, \boldsymbol{y}_p^{1:K}), \tag{4}$$

where θ is the collection of all the trainable parameters, namely θ^{T}, θ^{a}, and θ^{p}.

Long video animation. With temporal consistency modeling and appearance encoder, we can generate temporally consistent human image animation results for arbitrary length via segment-by-segment processing. However, unnatural transitions and inconsistent details across segments may occur because temporal attention blocks cannot model long-range consistency across different segments.

To address this challenge, we employ a sliding window method to improve transition smoothness in the inference stage. As shown in Figure 2, we divide the long motion sequence into multiple segments with temporal overlap, where each segment has a length of K. We first sample noise $\boldsymbol{z}^{1:N}$ for the entire video with N frames, and also partition it into noise segments with overlap $\{\boldsymbol{z}^{1:K}, \boldsymbol{z}^{K-s+1:2K-s}, ..., \boldsymbol{z}^{n(K-s)+1:n(K-s)+K}\}$, where $n = \lceil (N-K)/(K-s) \rceil$ and s is the overlap stride, with $s < K$. If $(N-K) \mod (K-s) \neq 0$, *i.e.*, the last segment size is less than K, for simplicity, we simply pad it with the first few frames to construct a K-frame segment. Besides, we empirically find that sharing the same initial noise $\boldsymbol{z}^{1:K}$ for all the segments improves video quality. For each denoising timestep $\boldsymbol{t}$, we predict noise and obtain $\epsilon_\theta^{1:K}$ for each segment, and then merge them into $\epsilon_\theta^{1:N}$ by averaging overlap frames. When $\boldsymbol{t} = 0$, we obtain the final animation video $I^{1:N}$.

3.4. Training

Learning objectives. We employ a multi-stage training strategy for our MagicAnimate. In the first stage, we omit the temporal attention layers temporarily and train the appearance encoder together with pose ControlNet. The loss term of this stage is computed as

$$\mathcal{L}_1 = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{t}, I_{\text{ref}}, \boldsymbol{p}_i, \epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta\|_2^2\right], \tag{5}$$

where $\boldsymbol{p}_i$ is the DensePose of target image I_i. The learnable modules are $\mathcal{F}_{\text{p}}(\cdot, \theta^{\text{p}})$ and $\mathcal{F}_{\text{a}}(\cdot, \theta^{\text{a}})$. In the second stage, we optimize only the temporal attention layers in $\mathcal{F}^{\text{T}}(\cdot, \theta^{\text{T}})$, and the learning objective is formulated as

$$\mathcal{L}_2 = \mathbb{E}_{\boldsymbol{z}_0^{1:K}, \boldsymbol{t}, I_{\text{ref}}, \boldsymbol{p}^{1:K}, \epsilon^{1:K}\sim\mathcal{N}(0,1)}\left[\|\epsilon^{1:K} - \epsilon_\theta^{1:K}\|_2^2\right]. \tag{6}$$

Image-video joint training. Human video datasets [14, 28], compared with image datasets, have a much smaller scale and are less diverse in terms of identities, backgrounds, and poses. This restricts the effective learning of reference condition capability of our animation framework. To alleviate this issue, we employ an image-video joint training strategy.

In the first stage when we pretrain the appearance encoder and pose ControlNet, we set a probability threshold τ_0 for sampling the human images from a large-scale image dataset [25]. We draw a random number $r \sim U(0,1)$, where $U(\cdot,\cdot)$ denotes uniform distribution. If $r \leq \tau_0$, we use the sampled image for training. In this case, the conditioning pose $\boldsymbol{p}_i$ is estimated from I_{ref}, and the learning objective of our framework becomes reconstruction.

Although the introduction of temporal attention in the second stage helps improve temporal modeling, we also notice that this leads to degraded per-frame quality. To simultaneously improve temporal coherence and maintain single-frame image fidelity, we also employ joint training in this stage. Specifically, we select two probability thresholds τ_1 and τ_2 empirically, and compare $r \sim U(0,1)$ with these thresholds. When $r \leq \tau_1$, we sample the training data from the image dataset, and we sample data from the video dataset otherwise. Based on the different training data, our denoising process in the training stage is formulated as

$$\epsilon_\theta^{1:K} = \begin{cases} \epsilon_\theta^{1:K}\left(\boldsymbol{z}_t, \boldsymbol{t}, I_{\text{ref}}, \boldsymbol{p}_i\right), \text{with } i = \text{ref}, & \text{if } r \leq \tau_1, \\ \epsilon_\theta^{1:K}\left(\boldsymbol{z}_t, \boldsymbol{t}, I_{\text{ref}}, \boldsymbol{p}_i\right), \text{with } i \neq \text{ref}, & \text{if } \tau_1 \leq r \leq \tau_2. \\ \epsilon_\theta^{1:K}\left(\boldsymbol{z}_t^{1:K}, \boldsymbol{t}, I_{\text{ref}}, \boldsymbol{p}^{1:K}\right), & \text{if } r \geq \tau_2 \end{cases} \tag{7}$$

4. Experiments

We evaluate the performance of MagicAnimate using two datasets, namely TikTok [14] and TED-talks [28]. The TikTok dataset comprises 350 dancing videos, while TED-talks includes 1,203 video clips extracted from TED-talk videos on YouTube. To ensure a fair comparison with state-of-the-art methods, we utilize the identical test set as DisCo [34] for TikTok evaluation and adhere to the official train/test split for TED-talks. All datasets undergo the same preprocessing pipeline. Please refer to *Sup. Mat.* for our dataset preprocessing and implementation details.

4.1. Comparisons

Baselines. We perform a comprehensive comparison of MagicAnimate with several state-of-the-art methods for human image animation: (1) **MRAA** [28] and **TPS** [51] are state-of-the-art GAN-based animation approaches, which estimate optical flow from driving sequences to warp the source image and then inpaint the occluded regions using GAN models. (2) **DisCo** [34] is the state-of-the-art diffusion-based animation method that integrates disentangled condition modules for pose, human, and background into a pretrained diffusion model to implement human image animation. (3) We construct additional baseline by combining the state-of-the-art image condition method, *i.e.*, IP-Adapter [43], with pose ControlNet [49], which is labeled as **IPA+CtrlN**. To make a fair comparison, we further add temporal attention blocks [10] into this framework and construct a video version baseline labeled as **IPA+CtrlN-V**. In addition, MRAA and TPS methods utilize ground-truth videos as driving signals. To ensure fair comparisons, we train alternative versions for MRAA and TPS using the same driving signal (DensePose) as MagicAnimate.

Evaluation metrics. We adhere to established evaluation metrics employed in prior research. For the TikTok dataset, we evaluate both single-frame image quality and video fidelity. The metrics used for single-frame quality include L1 error, SSIM [38], LPIPS [50], PSNR [13], and FID [11]. Video fidelity is assessed through FID-FVD [1] and FVD [33]. On the TED-talks dataset, we follow MRAA and report L1 error, average keypoint distance (AKD), missing keypoint rate (MKR), and average Euclidean distance (AED). However, these evaluation metrics are designed for single-frame evaluation and lack perceptual measurement of the animation results. Consequently, we also compute FID, FID-VID, and FVD on the TED-talks dataset to measure the image and video perceptual quality.

Quantitative comparisons. Table 1 provides the quantitative comparison results between MagicAnimate and baselines on two benchmark datasets. Our method surpasses all baselines in terms of reconstruction metrics, *i.e.*, L1, PSNR, SSIM, and LPIPS, on TikTok (Table 1a). Notably, MagicAnimate improves against the strongest baseline (DisCo) by 6.9% and 18.2% for SSIM and LPIPS, respectively. Additionally, MagicAnimate achieves state-of-the-art video fidelity, demonstrating significant performance improvements of 63.7% for FID-VID and 38.8% for FVD compared to DisCo.

Our method also exhibits superior video fidelity on the TED-talks dataset (Table 1b), achieving the best FID-VID of 19.00 and FVD of 131.51. This performance is particularly notable against the second-best method (MRAA), with an improvement of 28.1% for FVD. Additionally, MagicAnimate demonstrates state-of-the-art single-frame fidelity, securing the best FID score of 22.78. Compared with DisCo, a diffusion-based baseline method, MagicAnimate showcases a significant improvement of 17.2%. However, it is important to note that MagicAnimate has a higher L1 error compared to baselines. This is likely caused by the lack of background information in the DensePose control signals. Consequently, MagicAnimate is unable to learn a consistent dynamic background as presented in the TED-talks dataset, leading to an increased L1 error. Nevertheless, MagicAnimate achieves a comparable L1 error with the strongest baseline (MRAA) in foreground human regions,

Method	Image					Video	
	L1↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FID-VID↓	FVD↓
TPS* [51]	3.23E-04	29.18	0.673	0.299	53.78	72.55	306.17
MRAA* [28]	3.21E-04	29.39	0.672	0.296	54.47	66.36	284.82
TPS [51]	6.17E-04	28.17	0.560	0.449	140.37	142.52	800.77
MRAA [28]	4.61E-04	28.39	0.646	0.337	85.49	71.97	468.66
IPA [43]+CtrlN [49]	7.38E-04	28.03	0.459	0.481	69.83	113.31	802.44
IPA [43]+CtrlN [49]-V	6.99E-04	28.00	0.479	0.461	66.81	86.33	666.27
DisCo [34]	3.78E-04	29.03	0.668	0.292	**30.75**	59.90	292.80
MagicAnimate (Ours)	**3.13E-04**	**29.16**	**0.714**	**0.239**	32.09	**21.75**	**179.07**

(a) Quantitative comparisons on TikTok [14] dataset. We cite results directly from [34] for DisCo, TPS*, and MRAA*.

Method	Image					Video	
	AKD↓	MKR↓	AED↓	(L1, $L1_{fg}$)↓	FID↓	FID-VID↓	FVD↓
TPS* [51]	2.16	0.008	0.167	(1.04E-04, 7.24E-05)	22.35	6.88	81.41
MRAA* [28]	2.50	0.007	0.154	(1.06E-04, 7.12E-05)	21.13	6.03	78.29
TPS [51]	11.00	0.063	0.331	(2.22E-04, 1.43E-04)	86.65	72.49	457.02
MRAA [28]	4.37	0.024	0.246	(**1.61E-04**, **1.07E-04**)	35.75	22.97	182.78
IPA [43]+CtrlN [49]	5.14	0.022	0.375	(3.58E-04, 1.71E-04)	43.23	49.13	434.00
IPA [43]+CtrlN [49]-V	4.24	0.019	0.369	(4.43E-04, 1.66E-04)	49.21	38.48	281.42
DisCo [34]	2.96	0.019	0.253	(2.07E-04, 1.31E-04)	27.51	19.02	195.00
MagicAnimate (Ours)	**2.65**	**0.013**	**0.204**	(2.92E-04, 1.11E-04)	**22.78**	**19.00**	**131.51**

(b) Quantitative comparisons on TED-talks [28] dataset. We also report L1 error for the foreground (human regions).

Table 1. Quantitative comparisons with baselines, with best results in **bold** and second best results underlined. *The original TPS and MRAA directly use ground-truth video frames for animation, we report their results (marked in gray) only for reference.

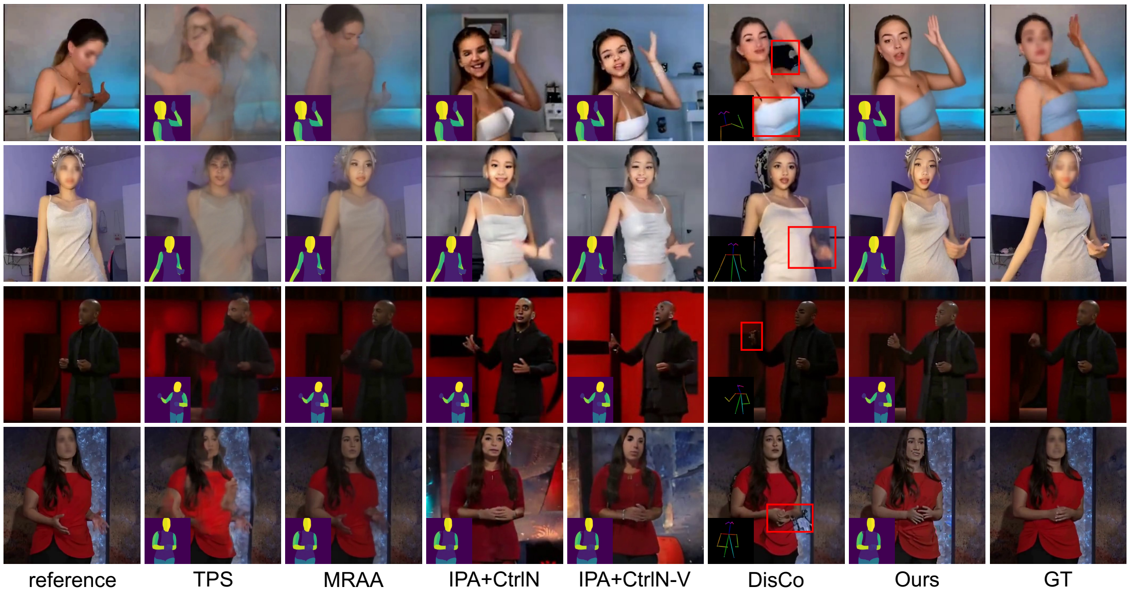


Figure 3. Qualitative comparisons between MagicAnimate and baselines on TikTok and TED-talks datasets. We overlay the target pose on the bottom left corner of the synthesized frames and highlight the artifacts generated by the strongest baseline (DisCo) in red boxes. For comprehensive video comparisons, please refer to our Project Page.

demonstrating its effectiveness for human animations. Furthermore, our method achieves the best performance for AKD, MKR, and AED, providing evidence of its superior identity-preserving ability and animation precision.

Qualitative comparisons. In Figure 3, we present qualitative comparisons between MagicAnimate and baselines. Notably, the dancing videos from the TikTok dataset exhibit significant pose variations, posing a challenge for GAN-based methods such as MRAA and TPS, as they struggle to produce reasonable results when there is a substantial pose difference between the reference image and the driving signal. In contrast, the diffusion-based baselines, IPA+CtrlN, IPA+CtrlN-V, and DisCo, show better single-frame quality. However, as IPA+CtrlN and DisCo generate each frame independently, their temporal consistency is unsatisfactory, as evidenced by the color change of the clothes and inconsistent backgrounds in the occluded regions. The video diffusion baseline, IPA+CtrlN-V, displays more consistent content, yet its single-frame quality is inferior due to weak reference conditioning. Conversely, MagicAnimate produces temporally consistent animations and high-fidelity details for the background, clothes, face, and hands.

Unlike the TikTok dataset, TED-talks dataset comprises speech videos recorded under dim lighting conditions. The motions in the TED-talks dataset primarily involve gestures, which are less challenging than dancing videos. Thus, MRAA and TPS produce more visually plausible results, albeit with inaccurate motion. In contrast, IPA+CtrlN, IPA+CtrlN-V, DisCo, and MagicAnimate demonstrate a more precise body pose control ability because these methods extract appearance conditions from the reference image to guide the animation instead of directly warping the source image. Among all these methods, MagicAnimate exhibits superior identity- and background-preserving ability, as shown in Figure 3, thanks to our appearance encoder, which extracts detailed information from reference image.

Cross-identity animation. Beyond animating each identity with its corresponding motion sequence, we further investigate the cross-identity animation capability of MagicAnimate and the state-of-the-art baselines, *i.e.*, DisCo, and MRAA. Specifically, we sample two DensePose motion sequences from the TikTok test set and use these sequences to animate reference images from other videos. Figure 1 illustrates that MRAA fails to generalize for driving videos that contain substantial pose differences, while DisCo struggles to preserve the details in the reference images, resulting in artifacts in the background and clothing. In contrast, our method faithfully animates the reference images given the target motion, demonstrating its robustness.

4.2. Ablation Studies

To verify the effectiveness of the design choices in MagicAnimate, we conduct ablative experiments on the TikTok dataset, which features significant pose variations, a wide range of identities, and diverse backgrounds.

Temporal modeling. To assess the impact of the proposed temporal attention layer, we train a version of MagicAnimate *without* it for comparison. The results, presented in Table 2a, show a decrease in both single-frame quality and video fidelity evaluation metrics when the temporal attention layers are discarded, highlighting the effectiveness of our temporal modeling. This is further supported by the qualitative ablation results presented in Figure 4a, where the model without explicitly temporal modeling fails to maintain temporal coherence for both humans and backgrounds.

Temp Attn	L1↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FID-VID↓	FVD↓
w/o	3.98	28.90	0.652	0.263	**27.54**	42.21	247.30
w/	**3.13**	**29.16**	**0.714**	**0.239**	32.09	**21.75**	**179.07**

(a) The effect of modeling temporal information.

App Enc	L1↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FID-VID↓	FVD↓
CLIP	8.00	27.94	0.461	0.481	78.35	82.50	724.96
IP-Adapter	7.89	27.98	0.481	0.442	64.17	67.65	590.99
Ours	**3.13**	**29.16**	**0.714**	**0.239**	**32.09**	**21.75**	**179.07**

(b) The effect of appearance encoder.

Spat	*Temp*	L1↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FID-VID↓	FVD↓
✗	✗	3.20	29.09	0.706	0.248	37.15	24.45	158.16
✗	✓	3.19	29.12	0.705	0.246	38.41	23.08	**156.32**
✓	✓	**3.13**	**29.16**	**0.714**	**0.239**	**32.09**	**21.75**	179.07

(c) The effect of image-video joint training.

Avg	L1↓	FID↓	FID-FVD↓
w/o	3.21	32.99	22.50
w/	**3.13**	**32.08**	**21.75**

(d) The effect of the inference-stage temporal video fusion.

Noise	L1↓	FID↓	FID-FVD↓
diff	**3.03**	32.74	22.50
same	3.13	**32.08**	**21.75**

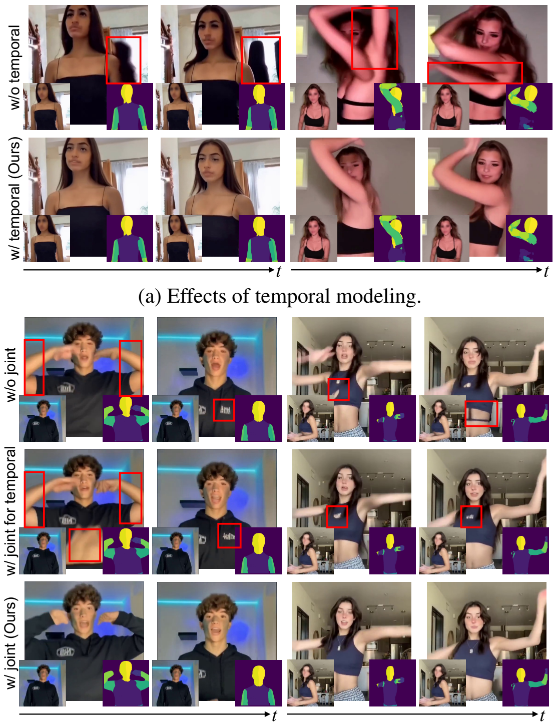
(e) The effect of sharing the same initial noises for all the video segments.

Table 2. Ablations of MagicAnimate on TikTok dataset, with best results in **bold**. We vary our architectural designs and training strategies to investigate their effectiveness. We report L1$\times 10^{-4}$ for numerical simplicity.

Appearance encoder. To evaluate the enhancement brought by the proposed appearance encoding strategy, we replace the appearance encoder in MagicAnimate with CLIP [21] and IP-Adapter [43] to establish baselines. Table 2b summarizes the ablative results. It is evident that our method significantly outperforms these two baselines in reference image preserving, resulting in a substantial improvement for both single-frame and video fidelity.

Inference-stage video fusion. MagicAnimate utilizes a video fusion technique to enhance the transition smoothness of long-term animation. Table 2d and Table 2e demonstrate the effectiveness of our design choices. In general, skipping the video fusion or using different initial random noises for different video segments diminishes animation performance, as evidenced by the performance drop for both appearance and video quality.

Image-video joint training. We introduce an image-video joint training strategy to enhance the animation quality. As shown in Table 2c, applying image-video joint training at both the appearance encoding and temporal modeling stages consistently increases the animation quality. Such improvements can also be observed in Figure 4b. Without the joint training strategy, the model struggles to model intricate details accurately, tending to produce incorrect



(a) Effects of temporal modeling.

(b) Effects of image-video joint training strategy.

Figure 4. Visualization of ablation studies, with errors highlighted in red boxes. For each frame, we overlay the reference image at the bottom left corner, and the target pose at the bottom right corner.

clothes and accessories as shown in Figure 4b.

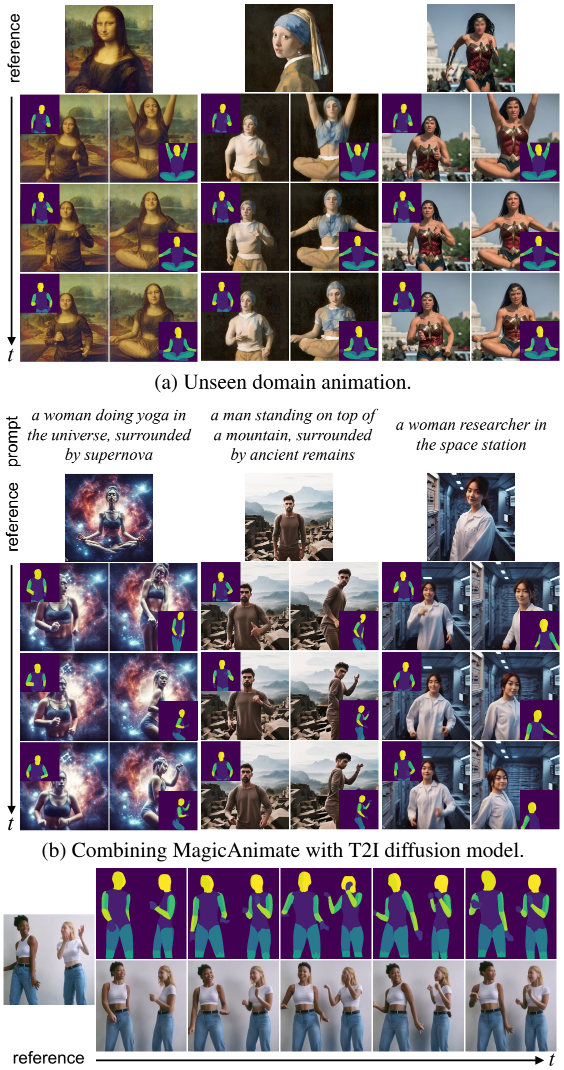
4.3. Applications

Despite being trained only on realistic human data, MagicAnimate demonstrates the ability to generalize to various application scenarios, including animating unseen domain data, integration with a text-to-image diffusion model, and multi-person animation.

Unseen domain animation. MagicAnimate showcases generalization ability for unseen image styles and motion sequences. As shown in Figure 5a, it can animate oil paintings and movie images to perform actions such as running and Yoga, maintaining a stable background and inpainting the occluded regions with temporally consistent results.

Combining with text-to-image generation. Due to its strong generalization ability, MagicAnimate can be used to animate images generated by text-to-image (T2I) models, *e.g*., DALL·E3 [2]. As shown in Figure 5b, we first employ DALL·E3 to synthesize the reference image using various prompts. These reference images can then be animated by our model to perform various actions.

Multi-person animation. MagicAnimate also exhibits strong generalization for multi-person animation. As illustrated in Figure 5c, we can generate animations for multiple individuals given the reference frame and a motion sequence, which includes two dancing individuals.


reference
t
prompt
a woman doing yoga in the universe, surrounded by supernova
a man standing on top of a mountain, surrounded by ancient remains
a woman researcher in the space station
reference
t
reference
t

(a) Unseen domain animation.

(b) Combining MagicAnimate with T2I diffusion model.

(c) Multi-person animation.

Figure 5. (a) Animation results for the unseen domain. (b) Combining MagicAnimate with DALL·E3 [2], and (c) Multi-person animation. We overlay the motion signal at the corner of each frame in (a) and (b). Video results can be found on Project Page.

5. Conclusion

This work introduces MagicAnimate, a novel diffusion-based framework designed for human avatar animation with an emphasis on temporal consistency. By effectively modeling temporal information, we enhance the overall temporal coherence of the animation results. The proposed appearance encoder not only elevates single-frame quality but also contributes to improved temporal consistency. Additionally, the integration of a video frame fusion technique enables seamless transitions across the animation video. MagicAnimate demonstrates state-of-the-art performance in terms of both single-frame and video quality. Moreover, its robust generalization capabilities make it applicable to unseen domains and multi-person animation scenarios.

Acknowledgement

This project is supported by the National Research Foundation, Singapore under its NRFF Award NRF-NRFF13-2021-0008, and the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2022).

References

[1] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, 2019. 5

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, and Yunxin Jiao. Improving image generation with better captions. `https://cdn.openai.com/papers/dall-e-3.pdf`, 2023. 8

[3] Chen Cao, Qiming Hou, and Kun Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM TOG*, 2014. 2

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023. 3

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2, 4

[6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *CVPR*, 2019. 1

[7] Zhenglin Geng, Chen Cao, and Sergey Tulyakov. 3d guided fine-grained face manipulation. In *CVPR*, 2019. 1

[8] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 3, 4

[9] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM TOG*, 2019. 1

[10] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv*, 2023. 3, 5

[11] 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. 5

[12] Fangzhou Hong, Zhaoxi Chen, Yushi LAN, Liang Pan, and Ziwei Liu. EVA3d: Compositional 3d human generation from 2d image collections. In *ICLR*, 2023. 1

[13] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *ICPR*, 2010. 5

[14] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *CVPR*, 2021. 4, 5, 6

[15] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv*, 2023. 1, 2, 3

[16] 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*, 2023. 2

[17] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit warping for animation with image sets. In *NeurIPS*, 2022. 2

[18] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*, 2019. 2

[19] Trevine Oorloff and Yaser Yacoob. Robust one-shot face video re-enactment using hybrid latent spaces of stylegan2. In *ICCV*, 2023. 2

[20] Shengju Qian, Kwan-Yee Lin, Wayne Wu, Yangxiaokang Liu, Quan Wang, Fumin Shen, Chen Qian, and Ran He. Make a face: Towards arbitrary high fidelity face manipulation. In *ICCV*, 2019. 2

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 7

[22] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *ICCV*, 2021. 2

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

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[25] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv*, 2021. 5

[26] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *CVPR*, 2019. 2

[27] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. 2019. 1, 2

[28] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 1, 2, 4, 5, 6

[29] Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Kyle Olszewski, Jian Ren, Hsin-Ying Lee, Menglei Chai, and Sergey Tulyakov. Unsupervised volumetric animation. In *CVPR*, 2023. 2

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. 2021. 2

[31] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time

face capture and reenactment of rgb videos. In *CVPR*, 2016. 2

[32] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 2018. 2

[33] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv*, 2018. 5

[34] 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*, 2023. 1, 2, 3, 5, 6

[35] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021. 1, 2

[36] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *ICLR*, 2022. 2

[37] Yaohui Wang, Xin Ma, Xinyuan Chen, Antitza Dantcheva, Bo Dai, and Yu Qiao. Leo: Generative latent image animator for human video synthesis. *arXiv*, 2023. 2

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 5

[39] 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

[40] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM TOG*, 2021. 1

[41] Hongyi Xu, Guoxian Song, Zihang Jiang, Jianfeng Zhang, Yichun Shi, Jing Liu, Wanchun Ma, Jiashi Feng, and Linjie Luo. Omniavatar: Geometry-guided controllable 3d head synthesis. In *CVPR*, 2023. 2

[42] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Jiashi Feng, and Mike Zheng Shou. Xagen: 3d expressive human avatars generation. In *NeurIPS*, 2023. 1, 2

[43] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv*, 2023. 5, 6, 7

[44] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *CVPR*, 2019. 2

[45] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *ECCV*, 2020. 2

[46] Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. Avatargen: a 3d generative model for animatable human avatars. In *ECCV Workshop*, 2023. 1, 2

[47] Jianfeng Zhang, Hanshu Yan, Zhongcong Xu, Jiashi Feng, and Jun Hao Liew. Magicavatar: Multimodal avatar generation and animation. *arXiv*, 2023. 1, 2

[48] Lvmin Zhang. Reference-only controlnet. https://github.com/Mikubill/sd-webui-controlnet/discussions/1236, 2023. 3

[49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 4, 5, 6

[50] 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. 5

[51] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *CVPR*, 2022. 1, 2, 5, 6

[52] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv*, 2022. 3