TalkingPose: Efficient Face and Gesture Animation with Feedback-guided Diffusion Model

Alireza Javanmardi[1] Pragati Jaiswal[1,2] Tewodros Amberbir Habtegebrial[1,2] Christen Millerdurai[1]
Shaoxiang Wang[1,2] Alain Pagani[1] Didier Stricker[1,2]
[1]German Research Center for Artificial Intelligence (DFKI) [2]RPTU

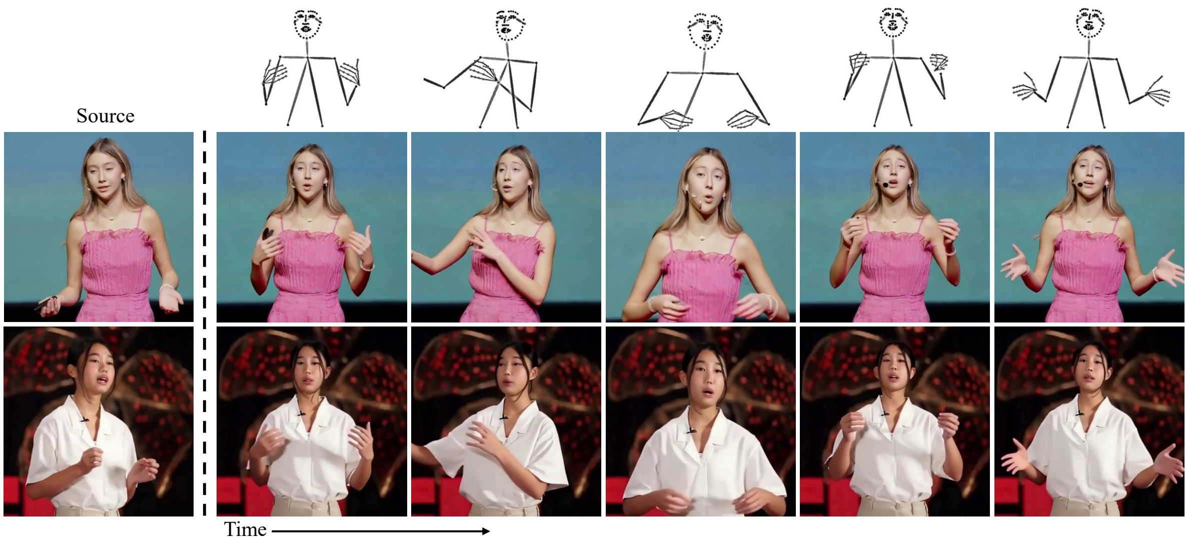

Figure 1. **TalkingPose** provides limitless face and gesture animation, while achieving the highest efficiency among current video diffusion approaches. Here, we animate the source characters (single frame input) based on the driving motions.

Abstract

Recent advancements in diffusion models have significantly improved the realism and generalizability of character-driven animation, enabling the synthesis of high-quality motion from just a single RGB image and a set of driving poses. Nevertheless, generating temporally coherent long-form content remains challenging. Existing approaches are constrained by computational and memory limitations, as they are typically trained on short video segments, thus performing effectively only over limited frame lengths and hindering their potential for extended coherent generation. To address these constraints, we propose TalkingPose, a novel diffusion-based framework specifically designed for producing long-form, temporally consistent human upper-body animations. TalkingPose leverages driving frames to precisely capture expressive facial and hand movements, transferring these seamlessly to a target actor through a stable diffusion backbone. To ensure continuous motion and enhance temporal coherence, we introduce a feedback-driven mechanism built upon image-based diffusion models. Notably, this mechanism does not incur additional computational costs or require secondary training stages, enabling the generation of animations with unlimited duration. Additionally, we introduce a comprehensive, large-scale dataset to serve as a new benchmark for human upper-body animation. Project page: https://dfki-av.github.io/TalkingPose

1. Introduction

Animating humans from a single image stems from the increasing demand for authentic, engaging digital interactions across various fields such as virtual reality, gaming, remote communication, and content creation. This technology has promising applications in virtual communication and the medical industry, especially where hand gestures are essential, such as in sign language understanding or hand-intensive fields like surgery, where precise gestures and facial expressions are crucial. The core of this challenge involves extracting the appearance of a ”target” human actor from a single source image and animating it by transferring motion from a sequence of driving frames. Consequently, recent research has focused on developing methods that either animate only the target actor’s facial expressions [24, 30, 52] or generate holistic animations of the entire body [42, 54]. Despite significant progress in this domain, two primary obstacles persist. First, the limited availability of large-scale datasets focusing on human upper-body motion constrains the robustness and generalizability of current methods. Second, both training and deploying video generative models remain computationally demanding, thereby restricting their outputs to short frame sequences and inhibiting the seamless synthesis of continuous, long-form animations.

Early methods primarily employed computer graphics techniques targeting the facial region, utilizing parametric face models to animate reconstructed heads [45, 46]. These approaches have seen notable advancements with the introduction of differentiable rendering techniques [27], which blend machine learning with traditional graphics pipelines [34, 35, 66]. However, these methods typically relied on multi-view data or RGB-D sensors for accurate face reconstruction. With the advancement of generative models, particularly Generative Adversarial Networks (GANs) [9], researchers have applied them to *one-shot* talking head synthesis. They typically use keypoints and feature warping techniques, which show impressive results when trained on facial data [41, 52]. However, extending these models to upper-body or full-body videos poses challenges in capturing detailed facial expressions and precise hand gestures.

More recently, Diffusion models [11] have demonstrated remarkable capabilities in face and body retargeting, leveraging advanced conditioning information to enable high-resolution character animation with enhanced photorealism [15, 30, 59]. Nevertheless, they face two primary challenges: ensuring consistency across animation sequences (on the order of hundreds of frames) and managing the substantial computational and memory demands of video synthesis. To address these limitations, we introduce *TalkingPose*, a novel framework that integrates hand gestures into facial animation models via a feedback-guided diffusion backbone. Unlike previous works that rely on temporal layers for consistency [3, 15, 59, 65]—requiring either training on stacks of frames [48, 64] or, training single-frame models followed by fine-tuning only the temporal layers [15, 65]—both approaches demand substantial computational resources and large-scale, temporally consistent video input. In contrast, we propose a closed-loop control (CLC) mechanism applied solely during inference, which significantly reduce frame-to-frame inconsistencies without the need for additional training or parameters. This design notably reduces computational overhead while enabling continuous, long-form animation. Furthermore, we introduce a large-scale video dataset encompassing upper-body motions—including diverse facial expressions, hand gestures, appearances, and backgrounds

In summary, our key contributions are as follows:

- We propose a closed-loop control mechanism that stabilizes diffusion models during inference, ensuring robust temporal consistency for extended video generation while significantly improving computational efficiency.
- We present *TalkingPose*, a large-scale video dataset covering diverse ages, genders, facial expressions, hand gestures, appearances, and backgrounds.
- We evaluate our method on the TED-talk [42], TikTok [18] datasets and our newly collected *TalkingPose*, showing its effectiveness through both qualitative and quantitative results.

2. Related Works

2.1. Diffusion Models for Video Generation

Diffusion models gained significant attention with Denoising Diffusion Probabilistic Models (DDPMs) [11] as powerful generative models for image generation tasks [6, 12]. Latent diffusion models [38] later improved efficiency, paving the way for applications in video generation [2, 7, 10]. These models typically use U-Net [39] or transformer architectures [50] to perform diffusion on stacks of latent embeddings. For conditional video generation, methods often employ CLIP [36] for text/image conditioning, with ControlNet [63] and T2I-Adapter [33] enhancing control over attributes such as pose, mask, and edge.

Early works attempted to generate consistent videos by training on sequences of frames using 3D U-Net models [13]. Later works introduced an additional temporal attention layer [10], which performs attention along the temporal dimension. While these strategies improve temporal coherence, they also demand extensive temporal data training, which can be computationally and memory-intensive [3, 15, 59]. Moreover, at inference, they often struggle to preserve consistency across successive frames, especially for longer videos. To address this, several works generate longer videos in a coarse-to-fine manner [53, 62] or progressively merge latent features of overlapping frames of video chunks [64], thereby enhancing transitions between frames. Nevertheless, these methods are predominantly designed to smooth transitional segments rather than comprehensively handle extended video generation. To overcome these limitations, we introduce a feedback-driven mechanism that enables consistent, continuous character video generation without extra computational overhead.

2.2. Face and Body Animation Generation

The field of face and body animation began with early approaches that primarily relied on graphics-based techniques [45, 46], such as parametric face models like FLAME [25] for facial animation and SMPL [28] for full-body reconstruction. Although these models enabled animatable 3D avatars, they often struggled with finer details, such as hair and accessories (e.g., glasses). Recent approaches based on implicit neural representations [32, 43] and 3D Gaussian Splatting (3DGS) [21] have greatly enhanced photorealism and rendering speed [34, 35, 58, 66]. However, these methods typically require multi-view training data and substantial training time for each individual avatar, limiting their scalability in broader applications.

In parallel, researchers have explored generative models that learn the underlying distribution of avatars from large-scale video datasets, enabling single-shot, instantaneous inference [41, 52, 60].

Early efforts in this domain focused on feature-warping techniques using facial keypoints in an adversarial manner [41, 52] While these self-supervised models perform well in talking head synthesis, they show limitations when trained on upper-body or full-body video datasets [41]. Some works have improved the framework by incorporating body regions [42] or inverse GAN techniques [54], yet they show limitations in capturing intricate details, particularly in facial expressions and hand movements simultaneously.

Recently, diffusion-based models have become promising alternatives for animating human faces. While they rely on driving frames, these methods also integrate other modalities such as audio [19, 26, 47, 57], and have shown strong results in generating dancing avatars guided by music or motion sequences [22, 59, 64, 65]. A separate line uses diffusion to map a driving actor's appearance onto a source video, as in MIMO and AnimateAnyone2 [16, 31]. For appearance extraction, many recent methods adopt an "Appearance/Reference Net" [3, 15, 65], which replicates the Stable Diffusion U-Net and processes the source image in parallel. They then either copy or concatenate the spatial attention layers from the Appearance Net to the main U-Net—sometimes by themselves [3, 59] or in combination with CLIP and cross-attention modules [15]. Other approaches refine specific aspects of the avatar by employing SMPL [28] to improve shape and pose, or by incorporating a face mask region and ArcFace features [48] to enhance facial realism.

3. Method

In our TalkingPose approach, we animate a target actor from a given source image using a sequence of driving frames that capture the desired motion. We begin with an overview of Stable Diffusion in Sec.3.1, followed by a description of our framework in Sec.3.2, an explanation of our feedback mechanism during inference in Sec.3.3, and an introduction to our dataset in Sec.3.4.

3.1. Preliminaries

Stable Diffusion [38] operates on a learned latent space, substantially reducing computational costs versus pixel-space diffusion. An input image $I \in \mathbb{R}^{H\times W\times C}$ is encoded into a latent representation $z = E(I) \in \mathbb{R}^{h\times w\times c}$ by a VAE encoder E [23], with a downsampling factor $f = \frac{H}{h} = \frac{W}{w} = 2^m, m \in \mathbb{N}$. The forward diffusion process incrementally adds Gaussian noise, $\epsilon \sim \mathcal{N}(0, I)$, at each step t (drawn uniformly from $\{1, \ldots, T\}$), producing a noisy latent z_t. A U-Net-based denoising network ϵ_θ then estimates ϵ from $\big(z_t, t, \tau_\theta(y)\big)$ or uses *v-prediction* to estimate a weighted combination of the noise and the clean latent x_0. Here, τ_θ is a domain-specific encoder (e.g., CLIP [36]) that maps the image y to a conditioning vector. During training, the objective

$$L_{\text{denoise}} = \mathbb{E}_{z,\,\epsilon,\,t,\,y}\Big[\|\epsilon \,-\, \epsilon_\theta\big(z_t,\,t,\,\tau_\theta(y)\big)\|^2\Big]$$

enforces accurate noise prediction, enabling effective inversion of the forward process. During inference, denoising steps (e.g., DDIM [44]) generate $\tilde{z}_0 \approx z$, which is decoded by D to reconstruct the final image $\tilde{x}_0 = D(\tilde{z}_0)$. This approach enables high-fidelity image generation with reduced computational overhead.

3.2. TalkingPose Framework

As Illustrated in Fig. 2, our method takes as input (1) a source image showing the target actor's appearance, and (2) a sequence of driving frames that convey the desired motion. The objective is to generate new frames of the target actor, preserving the subject's appearance while adopting the motion from the driving frames. Building upon the AnimateAnyone pipeline [15], we remove the temporal attention layers to enable training on individual frames rather than complete video sequences. In each training iteration, two frames are randomly selected from a video: the first captures the actor's appearance as the source frame, while the second provides the motion cues as the driving frame. Both frames are converted into latent representations via a VAE encoder. The source latent is input to the ReferenceNet [15]. Additionally, the source frame is processed by a CLIP ViT-L/14 encoder [36] to extract appearance features, which are injected into both the Denoising U-Net and the ReferenceNet via cross-attention mechanism. We refer to this integrated module as the *Appearance Encoder*. Simultaneously, the motion encoder extracts motion features from the driving frame. To

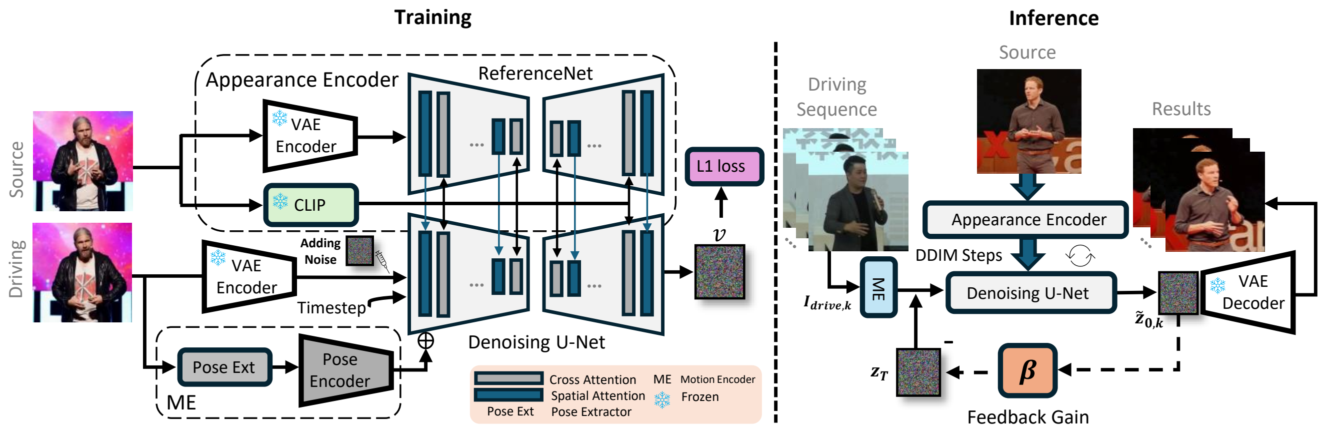

Figure 2. **TalkingPose Pipeline. Training:** Following AnimateAnyone [15], the Appearance Encoder (CLIP + ReferenceNet) obtains source features, while the Motion Encoder (driving pose extraction using method of Yang et al. [61] + Pose Encoder) prepares motion cues for the U-net. **Inference:** A single RGB source and a driving pose condition drive DDIM steps to predict a latent, which is refined via a feedback loop with proportional gain (β).

simulate the forward diffusion process, noise is added to the latent representation of the driving frame based on the selected time-step. This noisy latent, along with the motion features, is then input to the Denoising U-Net. The network is then trained to predict the added noise, with the discrepancy supervised by an L_1 loss. By accurately recovering the noise, the U-Net effectively fuses the source appearance with the driving motion. For more architectural details, readers are referred to AnimateAnyone [15].

3.3. Inference with Closed-loop Control

Generating long-form content with diffusion models poses significant computational challenges, particularly in terms of GPU resource utilization during both the training and inference phases. The memory requirements increase proportionally with the number of frames being processed. Additionally, segmenting videos into smaller parts can lead to cumulative error propagation, since using only the initial or final frame of a segment as the "source" may amplify discrepancies over time. Furthermore, the batch-wise approach tends to introduce temporal artifacts during transitions between batches.

To address these limitations, we introduce *closed-loop control* (CLC), a novel feedback-driven inference strategy that can be seamlessly integrated with any latent diffusion-based model. As illustrated in Fig. 2, we begin by extracting the appearance from a source image—potentially distinct from the identity depicted in the driving pose—and merging it with the target pose from the first driving frame. Specifically, we generate a motion encoding for this target pose and add it to Gaussian noise to create the input for our U-Net. Using DDIM steps, we obtain an output latent encoding that fuses the source image's appearance with the motion from the first driving frame. This latent is then passed through the Stable Diffusion VAE decoder to generate the initial frame of the animation.

For subsequent frames, we reintroduce the generated output latent encoding into the sampled noise, regulated by a feedback gain to maintain a consistent appearance of the avatar across frames. This feedback gain is modeled from a control theory perspective, treating it as a closed-loop control system [20, 56], Specifically, we use negative feedback to reduce unexpected deviations, counteracting small errors, minimizing oscillations, and guiding the output toward a stable, consistent appearance. The process works as follows: for each frame k, the diffusion model requires a Gaussian noise z_T as the initial input for inference. The diffusion process runs over several timesteps to produce the output latent encoding $\hat{z}_{0,k}$ for the frame k. To generate the next frame, $\hat{z}_{0,k+1}$, we apply the following state update equation:

$$x_{k+1} = z_T + \beta(\hat{z}_{0,k} - z_T) \tag{1}$$

where x_{k+1} is the U-net input before adding to the motion encoding to produce $\hat{z}_{0,k+1}$, and β is the feedback gain parameter. This feedback loop reduces errors and disturbances, ensuring consistency across generated frames over time. This approach is significantly more efficient than methods relying on temporal attention [3, 15, 59], which can also suffer from inconsistencies over extended periods. Additionally, this inference-based method eliminates the need for extra training and additional parameters while enabling the generation of an unlimited number of stable, consistent frames. Further details of our method is provided in Algorithm 1.

3.4. TalkingPose Dataset

We curate a large-scale dataset of human upper-body videos from diverse YouTube sources, aiming to capture not only facial geometry and expressions but also hand articulation, body movements, clothing diversity, and varying backgrounds. This scope goes beyond face-focused datasets, where only facial attributes are emphasized. As

Algorithm 1 Inference with Closed-loop Control (CLC)

Require:
I_{source}: source image
$\{I_{\text{drive},k}\}_{k=1}^{\infty}$: driving images
T: number of diffusion steps
β: feedback gain
$a_{\text{src}} \leftarrow$ Appearance Encoder(I_{source})
Sample $z_T \sim \mathcal{N}(0, I)$
$x_1 \leftarrow z_T$
for $k = 1$ to ∞ **do**
  $m_k \leftarrow$ MotionEncoder$(I_{\text{drive},k})$
  for $t = T$ down to 1 **do**
    $\hat{z}_{t,k} \leftarrow DM_\theta(x_k, a_{\text{src}}, m_k; t)$
  end for
  $x_{k+1} \leftarrow z_T + \beta\left(\hat{z}_{0,k} - z_T\right)$
  $\hat{I}_k \leftarrow$ Decoder$(\hat{z}_{0,k})$
end for

a result, learning generative models from upper-body data is more challenging, requiring robust representation of a broader set of visual and motion cues. Our initial collection included 21K raw videos, which, after a multi-stage preprocessing pipeline, yielded *18K videos with unique identities*. Each video typically features a single presenter centered in the frame. From these videos, we detect and crop the upper-body region at a resolution of 512×512. Samples failing to meet this criterion (e.g., insufficient resolution, missing face or hands) are discarded. The final dataset spans approximately 1250 hours of video footage at 20 FPS, encompassing diverse locations, lighting conditions, ages, and a balanced gender distribution. A detailed breakdown of the preprocessing steps and demographic statistics is provided in the Appendix.

4. Experiments

4.1. Datasets

We conduct our experiments on three benchmark datasets:
TED-talk. The TED-talk dataset [42] consists of 369 training and 42 test videos. Since some videos were no longer accessible, we replaced them with comparable recordings from the same YouTube channel; we refer to this version as TED-talk$^{\#}$.
TikTok. The TikTok dataset [18] consists of 350 single-person dance clips, each lasting 10–15 seconds. We follow the original split, using 340 videos for training and 10 for testing.
TalkingPose. We introduced the *TalkingPose* dataset with about 18K videos of people giving talks with hand gestures, using 90% for training and 10% from non-overlapping identities for testing.

4.2. Implementation Details

We trained our model for 70K steps on four NVIDIA H200 GPUs using a batch size of 32 and a learning rate of 1×10^{-5} in v-prediction setting [40]. The CLIP and VAE encoder-decoder weights remained frozen, while we fine-tuned only the ReferenceNet and the Denoising U-net, both initialized from Stable Diffusion V1.5. The Pose Guider was initialized with Gaussian weights, except for the final projection layer, which employed zero convolution. During inference, we used DDIM sampling with 30 steps and evaluated the model on 50-frame videos using an NVIDIA RTX3090 GPU. To avoid contrast saturation, we set the CFG value to 3.5. To tune the feedback gain (β) hyperparameter, we performed a grid search on 100 validation videos from the TalkingPose dataset.

4.3. Qualitative and Quantitative Comparison

Baseline. For our comparative analysis, we selected several state-of-the-art models that address both body animation and facial reenactment. Specifically, we evaluated AnimateAnyone [15], MagicPose [3], MagicAnimate [59], MimicMotion [64], Champ [65] and StableAnimator [48]. Since training scripts were unavailable for some methods [59, 64], we relied on official repositories and released pre-trained checkpoints trained on large-scale datasets with an emphasis on generalization for all baselines except AnimateAnyone. For AnimateAnyone, we used the widely recognized unofficial MooreAnimate repository, training both stages.
Metrics. We assess image quality using two variants of Peak Signal-to-Noise Ratio (PSNR) [14]: $PSNR_{int}$, which computes integer-based values as in [3, 15, 65], and $PSNR_{float}$, which avoids numerical overflow by using floating-point operations as in [48, 59, 64]. For our experiments, we mainly report $\text{PSNR}_{\text{float}}$ since it yields more accurate results. We further report Structural Similarity Index Measure (SSIM) [55] and Learned Perceptual Image Patch Similarity (LPIPS). To evaluate temporal consistency, we use FID-VID [1] and Fréchet Video Distance (FVD) [49]. All metrics are computed using the Disco evaluation toolkit [51]. For pose accuracy, we adopt Average Keypoint Distance (AKD) [8], using MediaPipe landmarks [29] for face, hands, and torso separately. For facial identity preservation, we compute cosine similarity (CSIM) based on ArcFace embeddings [5, 17, 37]. Finally, we measure lip synchronization with SyncNet [4].
Evaluation. For evaluation, we used the first frame of each video as the source and the full video as the driving sequence, allowing for frame-by-frame comparisons across all models. Methods that rely on temporal layers alleviate short-term flickering, but we observed frequent contrast shifts and frame-specific artifacts (bottom two rows of

Figure 3. **Qualitative Comparison**. This figure shows our model's ability to accurately capture facial expressions, hand gestures, and poses across diverse postures, while preserving the appearance and background of the reference frame, compared to state-of-the-art methods AnimateAnyone [15], MagicAnimate [59], Champ [65], MimicMotion [64], and StableAnimator [48] on TED-talk# [42] and TalkingPose.

Method	TED-talk#					TalkingPose				
	SSIM ↑	$\text{PSNR}_{\text{float}}$ ↑	LPIPS ↓	FID-VID ↓	FVD ↓	SSIM ↑	$\text{PSNR}_{\text{float}}$ ↑	LPIPS ↓	FID-VID ↓	FVD ↓
AnimateAnyone (Moore impl.)† [15]	0.768	**22.22**	**0.226**	<u>5.99</u>	**89.90**	<u>0.736</u>	<u>21.87</u>	<u>0.197</u>	<u>9.56</u>	**192.78**
AnimateAnyone (TED-talk)	0.832	–	0.159	–	80.50					
MagicPose [3]	0.638	16.25	0.307	15.16	329.44	0.509	14.27	0.305	23.57	348.05
MagicAnimate [59]	0.574	15.02	0.326	19.05	401.60	0.445	12.59	0.339	31.70	425.87
MimicMotion [64]	0.643	16.21	0.349	12.94	297.28	0.640	17.53	0.275	12.46	274.16
Champ [65]	0.685	18.63	0.290	12.46	217.87	0.654	18.66	0.294	15.14	304.79
StableAnimator [48]	0.703	19.56	0.279	12.22	203.90	0.669	18.63	0.248	16.52	276.77
TalkingPose (Ours)	**0.773**	<u>21.92</u>	<u>0.230</u>	**5.32**	<u>105.62</u>	**0.749**	**21.94**	**0.186**	**5.07**	<u>203.11</u>

Table 1. **Quantitative results on TED-talk# [42] and TalkingPose, evaluated with Disco [51].** Best results are in **bold**, second-best are <u>underlined</u>. For AnimateAnyone, we report both results from the original publication (gray, TED-talk) and our reproduction using the unofficial Moore implementation on TED-talk#, a variant of TED-talk where missing videos were supplemented.

Method	SSIM ↑	PSNR_{int} ↑[14]	$\text{PSNR}_{\text{float}}$ ↑[51]	LPIPS ↓	FID-VID ↓	FVD ↓
AnimateAnyone [15]	0.718	29.56	-	0.285	-	171.90
MagicPose [3]	0.752	29.53	-	0.292	46.30	-
MagicAnimate [59]	0.714	-	18.22	0.239	21.75	179.07
MimicMotion [64]	0.795	-	20.10	-	**9.30**	594.00
Champ [65]	<u>0.802</u>	29.91	-	0.234	21.07	<u>160.82</u>
StableAnimator [48]	0.801	**30.81**	<u>20.66</u>	0.232	-	**140.62**
TalkingPose (Ours)	**0.822**	<u>30.50</u>	**21.36**	**0.222**	<u>15.04</u>	226.63

Table 2. **Quantitative results on the TikTok dataset [18].** Reported values are taken from the corresponding original publications. The PSNR variants are described in Sec. 4.3.

Fig. 3, top row of Fig. 4 and the supplementary video). We analyse this behaviour in detail in the ablation studies (Sec. 4.4). As shown in Fig. 3, *MagicAnimate* and *MimicMotion* often fail to consistently preserve the character's facial identity. While other methods reliably preserve facial appearance when generating the video with various expressions. In terms of pose accuracy, *MagicAnimate* sometimes fails to replicate hand gestures precisely, especially in the TikTok dataset, whereas other methods maintain better pose consistency (see Fig. 4). Similar issues arise in the TED-talk# and TalkingPose datasets, where *TalkingPose* again outperforms the competing approaches. Quantitatively, our *TalkingPose* model achieves state-of-the-art performance in temporal consistency and appearance preservation. This is confirmed by superior image-based metrics (SSIM, PSNR, LPIPS) reported in Table 1, with consistent performance across the TED-talk#, TalkingPose, and TikTok datasets on Table 2, where it also ranks second in PSNR and LPIPS for the

†AnimateAnyone (Moore impl.): https://github.com/MooreThreads/Moore-AnimateAnyone

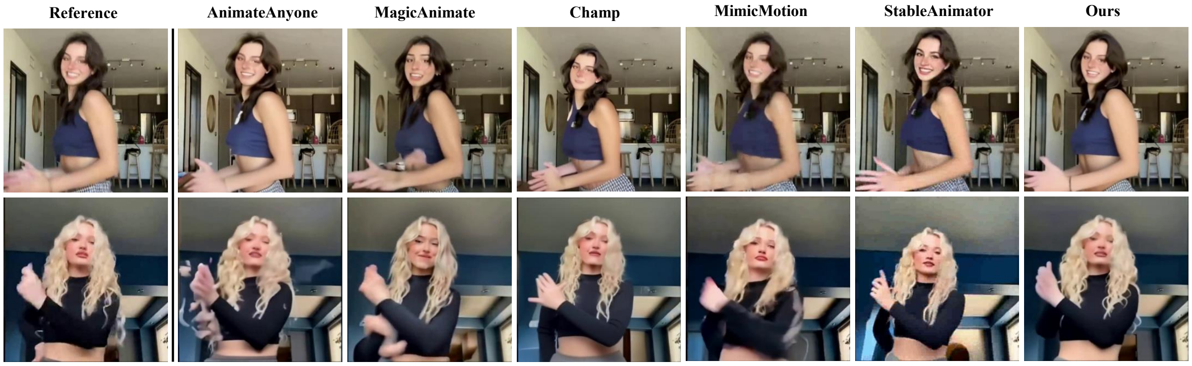

Figure 4. **Qualitative Comparison**. This figure demonstrates our model's performance compared to state-of-the-art methods AnimateAnyone [15], MagicAnimate [59], Champ [65], MimicMotion [64] and StableAnimator [48] on TikTok dataset [18].

TED-talk# dataset. For video-based evaluation, *TalkingPose* excels with FID-VID scores of 5.32 for TED-talk# and 5.07 for TalkingPose, and it obtains comparable FVD on the TikTok dataset while ranking second on the TED-talk# and TalkingPose datasets. Moreover, our approach effectively narrows the quality–efficiency gap relative to diffusion-based methods, handling diverse poses, gestures, and backgrounds with ease. Fig. 6 shows two random frames from the TalkingPose set with their difference map, highlighting its stable background, consistent appearance, and accurate pose. Although *AnimateAnyone* keeps appearance well via its motion module, TalkingPose yields slightly better quality and greater efficiency without specialised temporal layers or stacked-frame training. Additional quantitative analyses are provided in the Appendix.

To assess the effectiveness of our talking avatar animation, we evaluate lip synchronization using SyncNet [4] and $\mathrm{AKD}_{\mathrm{lip}}$, which measures alignment of lip keypoints. We conduct the evaluation on a validation set of 50 audio–video clips, each 10 seconds long, and report the results in Table Tab. 3. As shown, our method achieves comparable SyncNet scores while attaining the best performance in $\mathrm{AKD}_{\mathrm{lip}}$.

4.4. Ablation Analysis

To assess the effectiveness of our proposed configurations, we conducted a comprehensive ablation study on the TalkingPose dataset, which includes diverse human appearances under varying lighting conditions, backgrounds, and visual attributes.

Temporal Jittering Error (TJE). We introduce the *Temporal Jittering Error* (TJE) metric to quantify longer-term temporal consistency. In addition to examining consecutive frames ($\Delta = 1$), we evaluate longer-term temporal consistency by comparing pairs of frames separated by $\Delta > 1$. Concretely, let $I_t^{(\mathrm{real})}$ and $I_{t+\Delta}^{(\mathrm{real})}$ be two frames from the real video at time t and $t + \Delta$, respectively. Similarly, $I_t^{(\mathrm{gen})}$ and $I_{t+\Delta}^{(\mathrm{gen})}$ are the corresponding frames from our generated video.

We define:

$$D_t^{(\mathrm{real})} = I_{t+\Delta}^{(\mathrm{real})} - I_t^{(\mathrm{real})}, \quad D_t^{(\mathrm{gen})} = I_{t+\Delta}^{(\mathrm{gen})} - I_t^{(\mathrm{gen})},$$

and measure

$$\mathrm{error}_t = \mathrm{mean}\Big(\big|D_t^{(\mathrm{real})} - D_t^{(\mathrm{gen})}\big|\Big).$$

By choosing $\Delta > 1$, we capture more noticeable motion changes over a small window (e.g., 5 frames $\approx$ 200 ms at 25 fps), rather than relying only on single-frame intervals where slow motion might yield negligible differences. This per-frame error is aggregated across $t = 0, \ldots, N - \Delta$, providing a robust measure of how closely the generated video matches the real one over slightly longer temporal spans, thus highlighting any jitter or flicker.

As shown in Fig. 5, our baseline model is capable of faithfully synthesizing the character and background. However, some temporal jittering artifacts remain (highlighted with red rectangles). These artifacts are more pronounced in Fig. 7, where our proposed CLC strategy significantly enhances temporal coherence. Quantitatively, Tab. 4 shows that the base model already delivers good frame image quality, while incorporating a motion module reduces FVD from 552.82 to 285.65, it slightly degrades certain image-based metrics. This is also highlighted in Fig. 4 in shift contrast and artifacts in videos in supplementary. In contrast, adding our CLC method not only further reduces FVD (improving consistency beyond what the motion module achieves alone) but also yields marginal gains in image-based metrics.

To demonstrate CLC's generality, we integrated it into Champ, which shares AnimateAnyone's two-stage training. At the stage-1 checkpoint, Champ's FVD was 580.76; with CLC ($\beta = 0.05$), it fell to 276.18 on TED-Talk, confirming cross-model gains. We then evaluated noise sampling. Switching from a fixed seed to independent z_T per frame degraded FVD from 203.11 to 251.93 and LPIPS from 0.245 to 0.308, validating the fixed z_T strategy in Algorithm 1.

Finally, a grid search on the validation set (Table 5) showed gains above 0.1 introduce artifacts, so we select $\beta = 0.05$ as the optimal stability–quality trade-off.

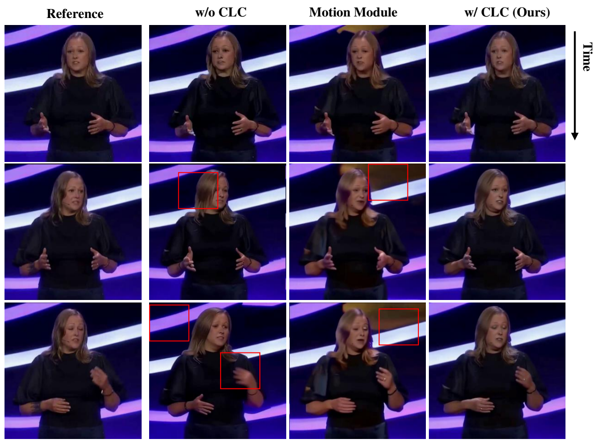

Figure 5. **Ablation Study on Temporal Analysis**. Three sample frames from generated videos under (1) baseline without CLC, (2) with motion module, and (3) our CLC method. Red boxes mark artifacts or temporal errors vs. the reference.

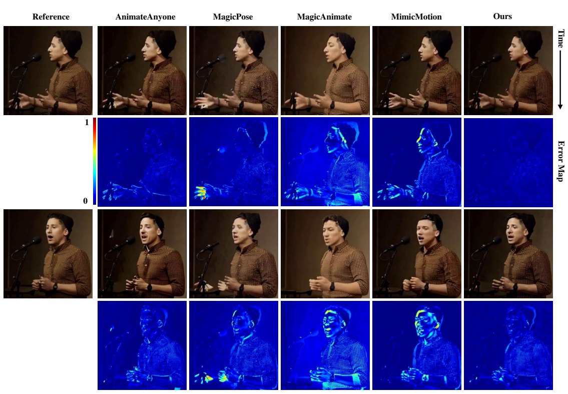

Figure 6. **Temporal Qualitative Analysis**. This figure demonstrates temporally consistent character animation and its resemblance to the provided source, along with two randomly selected pose frames from the video. The differences with the reference are highlighted using error map.

Method	SyncScore$_{\text{Audio}}$	AKD$_{\text{lip}}$
AnimateAnyone	9.11	6.80
MagicPose	9.89	7.84
MagicAnimate	10.47	7.95
MimicMotion	9.71	8.11
Champ	11.28	9.76
StableAnimator	**8.62**	7.51
TalkingPose (Ours)	9.15	**6.66**
GT	6.31	–

Table 3. Lip-sync comparison: quantitative evaluation using SyncScore and lip-only average keypoint distance (AKD$_{\text{lip}}$).

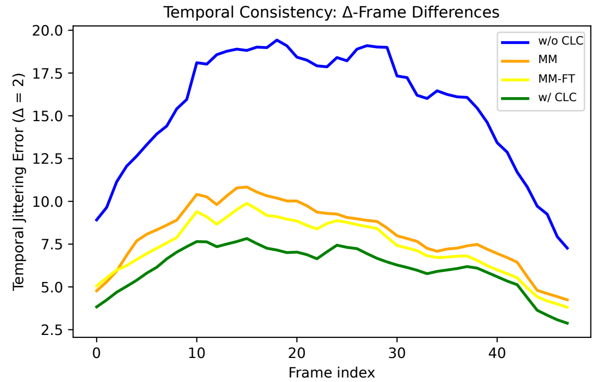

Figure 7. **Temporal Jittering Error (TJE).** "w/o CLC" refers to the model without a motion module, "MM" indicates the model with a pretrained motion module, "MM-FT" denotes the motion module fine-tuned, and "Ours" represents the Base model fitted with our CLC mechanism. Lower is better.

Model	SSIM ↑	PSNR$_{\text{float}}$ ↑	LPIPS ↓	FID-VID ↓	FVD ↓
w/o CLC	0.747	**21.96**	0.190	26.98	552.82
MM	0.696	20.31	0.219	8.46	285.65
MM FT	0.736	21.87	0.197	9.56	**192.78**
w/ CLC (Ours)	**0.749**	21.94	**0.186**	**5.07**	203.11

Table 4. **Ablation on CLC.** Results on TalkingPose with different settings: "w/o CLC" is AnimateAnyone without a motion module (1st stage only), "MM" adds a pre-trained motion module, "MM-FT" fine-tunes the motion module (i.e., AnimateAnyone trained for both stages), and "w/ CLC" is AnimateAnyone with its motion module replaced by our proposed CLC.

β Variant	SSIM ↑	PSNR$_{\text{float}}$ ↑	LPIPS ↓	FID-VID ↓	FVD ↓
$\beta : 0$	0.747	21.96	0.190	26.98	552.82
$\beta : 0.01$	0.748	**21.98**	0.188	8.17	208.50
$\beta : 0.05$	**0.749**	21.94	**0.186**	**5.07**	**203.11**
$\beta : 0.1$	0.745	21.84	0.187	9.52	206.36
$\beta : 0.2$	0.625	16.36	0.269	23.49	371.84

Table 5. **Feedback gain (β) ablation study.**

5. Conclusion

We introduced *TalkingPose*, a diffusion-based framework for animating human upper-body, including face and hand gestures, from a source appearance and driving pose sequence. Our closed-loop control mechanism ensures high temporal consistency for long video sequences. We also contribute a large-scale dataset encompassing diverse appearances, backgrounds, and gestural variations. *TalkingPose* outperforms state-of-the-art methods while offering higher efficiency, paving the way for more robust and accessible human animation techniques.

6. Acknowledgments

This work has been partially funded by the EU projects CORTEX[2] (GA: Nr 101070192) and LUMINOUS (GA: Nr 101135724).

References

[1] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, page 2, 2019. 5

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[3] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *Forty-first International Conference on Machine Learning*, 2023. 2, 3, 4, 5, 6

[4] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 5, 7

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 5

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[7] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 2

[8] Michael Gashler. Waffles: A machine learning toolkit. *Journal of Machine Learning Research*, 12(69):2383–2387, 2011. 5

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. 2

[10] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024. 2

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

[12] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 2

[13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[14] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369. IEEE, 2010. 5, 6

[15] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 2, 3, 4, 5, 6, 7

[16] Li Hu, Guangyuan Wang, Zhen Shen, Xin Gao, Dechao Meng, Lian Zhuo, Peng Zhang, Bang Zhang, and Liefeng Bo. Animate anyone 2: High-fidelity character image animation with environment affordance. *arXiv preprint arXiv:2502.06145*, 2025. 3

[17] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020. 5

[18] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021. 2, 5, 6, 7

[19] Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. *arXiv preprint arXiv:2409.02634*, 2024. 3

[20] T. Kailath. *Linear Systems*. Prentice-Hall, 1980. 4

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3

[22] Jeongho Kim, Min-Jung Kim, Junsoo Lee, and Jaegul Choo. Tcan: Animating human images with temporally consistent pose guidance using diffusion models. *arXiv preprint arXiv:2407.09012*, 2024. 3

[23] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2014. 3

[24] Tobias Kirschstein, Simon Giebenhain, and Matthias Nießner. Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5481–5492, 2024. 2

[25] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 3

[26] Gaojie Lin, Jianwen Jiang, Chao Liang, Tianyun Zhong, Jiaqi Yang, and Yanbo Zheng. Cyberhost: Taming audio-driven avatar diffusion model with region codebook attention. *arXiv preprint arXiv:2409.01876*, 2024. 3

[27] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. 2

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned

multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 3

[29] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Proceedings of the Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[30] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. *arXiv preprint arXiv:2406.01900*, 2024. 2

[31] Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Mimo: Controllable character video synthesis with spatial decomposed modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21181–21191, 2025. 3

[32] 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. *Communications of the ACM*, 65(1):99–106, 2021. 3

[33] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2

[34] Malte Prinzler, Otmar Hilliges, and Justus Thies. Diner: Depth-aware image-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12449–12459, 2023. 2, 3

[35] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 2, 3

[36] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[37] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021. 5

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[40] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022. 5

[41] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3

[42] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. 2, 3, 5, 6

[43] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 3

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[45] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6):183–1, 2015. 2, 3

[46] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2, 3

[47] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024. 3

[48] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. *arXiv preprint arXiv:2411.17697*, 2024. 2, 3, 5, 6, 7

[49] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. In *International Conference on Learning Representations (ICLR)*, 2019. 5

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 2

[51] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9326–9336, 2024. 5, 6

[52] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu.

One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2021. 2, 3

[53] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, et al. Magicvideo-v2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468*, 2024. 2

[54] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations*, 2022. 2, 3

[55] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[56] Kun Xu, Chongxuan Li, Jun Zhu, and Bo Zhang. Understanding and stabilizing gans' training dynamics using control theory. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*. JMLR.org, 2020. 4

[57] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*, 2024. 3

[58] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2024. 3

[59] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 2, 3, 4, 5, 6, 7

[60] Kewei Yang, Kang Chen, Daoliang Guo, Song-Hai Zhang, Yuan-Chen Guo, and Weidong Zhang. Face2face ρ: Real-time high-resolution one-shot face reenactment. In *European Conference on Computer Vision*, pages 55–71. Springer, 2022. 3

[61] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 4

[62] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023. 2

[63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[64] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. In *International Conference on Machine Learning*, 2025. 2, 3, 5, 6, 7

[65] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. *arXiv preprint arXiv:2403.14781*, 2024. 2, 3, 5, 6, 7

[66] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4574–4584, 2023. 2, 3