SUPPLEMENTARY MATERIAL FOR "FREEANIMATE: TRAINING-FREE HUMAN IMAGE ANIMATION WITH PREVIEW-GUIDED DENOISING"

Yuan Zeng[1] *Yujia Shi*[2,3] *Zongqing Lu*[1] *QingMin Liao*[1†]

[1]Shenzhen International Graduate School, Tsinghua University, China
[2]Harbin Institute of Technology, China
[3]Pengcheng Laboratory, China

In this Supplementary File, we first provide the fundamental concepts and background in Section A, followed by an introduction to the diffusion models for video generation in Section C. Section D delves deeper into the Preview Generation Strategy. Section E outlines the evaluation metrics and benchmark details used for assessing FreeAnimate's performance. More visualization results are presented in Section F, showcasing the performance across various datasets. Finally, Section G addresses the limitations of FreeAnimate and discusses potential future work. Detailed results and comparisons can be found on our project page: https://freeani.github.io/.

A. PRELIMINARIES

Latent Diffusion Models. Latent diffusion models (LDMs) present a novel approach by performing denoising in the latent space, rather than the image pixel space, thereby reducing computational costs without sacrificing image quality [1, 2]. A prominent example of the LDM paradigm is the Stable Diffusion (SD) model [3], which combines a Vector Quantized-Variational AutoEncoder (VQ-VAE) [4] with a time-step conditioned U-Net [5]. In this framework, the VQ-VAE encoder transforms an input image $\boldsymbol{x}$ into a latent space representation $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$, while the decoder reconstructs the image $\boldsymbol{x}'$ from a generated latent code $\boldsymbol{z}'$ as $\boldsymbol{x}' = \mathcal{D}(\boldsymbol{z}')$. The denoising process consists of T sequential steps, gradually denoising the initial Gaussian noise $\boldsymbol{z_T} \in \mathcal{N}(0, I)$ into the desired latent features $\boldsymbol{z_0}$, which can then be mapped back to the RGB image space using the VQ-VAE decoder. Each iteration of the denoising process aims to estimate the residual noise between the current latent $\boldsymbol{z_t}$ and the target latent $\boldsymbol{z_0}$. This is typically achieved by employing a denoising U-Net, with the associated loss function defined as:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(\boldsymbol{x}),c,\epsilon\sim\mathcal{N}(0,I),t}\left[\|\epsilon - \epsilon_\theta(\boldsymbol{z_t}, t, c)\|_2^2\right],$$

where $\boldsymbol{z_t}$ represents the intermediate latent at time-step t during the denoising process. The term $\epsilon_\theta(\cdot)$ refers to the U-Net used for noise prediction, and c represents the conditioning input.

DDIM Sampling and Inversion. In the denoising stage, we adopt deterministic sampling based on DDIM to generate a clean latent representation $\boldsymbol{z_0}$ from an initial standard Gaussian noise $\boldsymbol{z_t}$. Formally, latent $\boldsymbol{z_{t-1}}$ is computed from the previous step $\boldsymbol{z_t}$ as follows:

$$\boldsymbol{z_{t-1}} = \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{\boldsymbol{z_t} - \sqrt{1-\alpha_t}\,\epsilon_\theta(\boldsymbol{z_t}, t)}{\sqrt{\alpha_t}}\right)}_{\text{predicted } \boldsymbol{z_0}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta(\boldsymbol{z_t}, t)}_{\text{direction toward } \boldsymbol{z_t}} + \underbrace{\sigma_t\,\epsilon_t}_{\text{random noise}}. \tag{1}$$

Here, α_t and σ_t specify the noise schedule at time step t, and $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ denotes standard Gaussian noise. The function ϵ_θ refers to a U-Net model with learnable parameters θ. Due to the deterministic nature of DDIM, the denoising trajectory is inherently reversible. This property enables DDIM inversion, which reconstructs a noisy latent code $\boldsymbol{z_T}$ from a clean latent $\boldsymbol{z_0}$. The inversion process is defined as:

$$\hat{\boldsymbol{z}}_{\boldsymbol{t}} = \sqrt{\alpha_t}\left(\frac{\hat{\boldsymbol{z}}_{\boldsymbol{t-1}} - \sqrt{1-\alpha_{t-1}}\,\epsilon_\theta(\hat{\boldsymbol{z}}_{\boldsymbol{t-1}}, t-1)}{\sqrt{\alpha_{t-1}}}\right) + \sqrt{1-\alpha_t-\sigma_{t-1}^2}\cdot\epsilon_\theta(\hat{\boldsymbol{z}}_{\boldsymbol{t-1}}, t-1) + \sigma_{t-1}\,\epsilon_{t-1}. \tag{2}$$

The resulting inverted latent $\hat{\boldsymbol{z}}_{\boldsymbol{t}}$ can subsequently undergo standard DDIM denoising to recover a clean latent $\hat{\boldsymbol{z}}_{\boldsymbol{0}}$, which closely approximates the original $\boldsymbol{z_0}$.

ControlNet. ControlNet [6] is a flexible extension of the Stable Diffusion (SD) model, designed to enable explicit control over the geometric, structural, and semantic attributes of the generated images. It achieves this by incorporating trainable copies of the U-Net's down and middle blocks, which are used to extract features from conditioning inputs. To facilitate integration with the original SD model, ControlNet introduces additional "zero convolution" layers. These layers not only help suppress undesirable noise during training but also contribute to stabilizing the overall training process. With the rapid advancement of diffusion model research, ControlNet has been extended to support a diverse set of conditioning modalities, including depth maps, segmentation masks, skeleton maps, canny edges, and sketches. Furthermore, combining multiple ControlNet modules is simple, typically achieved by adding their individual outputs together before passing them into the SD model.

B. RELATED WORK

B.1. Training Based Human Image Animation

There has been a surge of training-based HIA methods over the past few years, including GAN-based methods [7, 8] and diffusion model

† Corresponding author.

based ones [9, 10, 11, 12, 13, 14, 15, 16]. MRAA [7] and TPS [8] are top-performing GAN-based methods, which utilize a flow predictor to estimate the optical flow of driving sequences, the outcome is used for warping the reference image, and then inpaint the occluded regions. However, GAN-based methods often fall short in training stability and generation quality.

Thanks to the exponential growth of diffusion models, many diffusion model based HIA methods have emerged. Prevalent network design often incorporate a denoising U-Net and a pose controller, typically a ControlNet or a lightweight ConvNet, to enable pose controllability. Additionally, an appearance net, either a frozen image encoder [17, 18], or a hybrid of the two is employed to inject features from the reference image into the U-Net. To illustrate, DisCo [9], MagicAnimate [10], MagicPose [12] and AnimateAnyone [11] are several distinct examples of the aforementioned network design. The aforementioned methods rely exclusively on pose maps to guide human animation, while Champ [13] utilizes the SMPL [19], a 3D parametric human model, to render depth maps, normal maps, and semantic maps as extra conditioning. Recently proposed MimicMotion [15] employs a confidence-aware pose guidance strategy to mitigate artifacts from pose inaccuracies, especially in regions with intricate motions. StableAnimator [16] adopts a similar network architecture to MimicMotion but introduces a distribution-aware ID Adapter and an HJB equation-based face optimization method to further enhance the face quality of the generated video frames. Despite producing visually plausible results, existing methods typically require large-scale training data and significant computational resources. In contrast, FreeAnimate achieves training-free human image animation by leveraging pre-trained models and preview frames.

B.2. Training Efficient Human Image Animation

Several zero-shot and one-shot HIA methods have recently emerged. PoseCrafter [20], a one-shot approach, uses reference image inversion and temporal attention for temporal consistency, and applies latent editing to mitigate facial and hand degradation. PoseAnimate [21] introduces a zero-shot framework that features a pose-aware control module for precise pose alignment, a dual-consistency attention module for maintaining identity and temporal consistency, and a mask-guided decoupling module for effective foreground-background separation. However, PoseAnimate [21] primarily showcases cartoon character animation and involves a complex network structure. PoseCrafter [20] shares a similar concept with our approach, but it relies on a heuristic search to find the preview frame in training set that best matches the inference pose. In contrast, FreeAnimate employs a more generalizable Preview Generation Strategy to generate preview frames, yielding superior results.

C. DIFFUSION MODELS FOR VIDEO GENERATION

Spurred by the introduction of DDPM [1] and DDIM [2], diffusion models have experienced exponential growth. The advent of Stable Diffusion [3], coupled with the development of techniques like ControlNet [6] and T2I-Adapter [22], has enabled unprecedented control over the generative process, fostering a wide range of creative applications. Based on the breakthroughs in image generation, the video generation paradigm has also witnessed great progress.

Make-A-Video [23] and Imagen [24] exploit pretrained text-to-image models and temporal attention or convolution layers to perform low-resolution video generation tasks. Coupled with super-resolution networks, Make-A-Video and Imagen can generate video clips of remarkable visual quality. However, these methods model video distribution in the image space. To reduce computational cost, ModelScopeT2V [25], Show-1 [26], Video-LDM [27], and AnimateDiff [28] operate in the latent space, using U-Net with trainable temporal blocks that apply attention along the temporal axis to ensure temporal consistency.

To further harness the capabilities of image diffusion models, various training-free techniques have been proposed, such as Text2Video-Zero [29], Pix2Video [30], ControlVideo [31], Freenoise [32] and FateZero [33]. Text2Video-Zero [29] performs self-attention between current frame and the first frame and encodes motion dynamics in the latent codes to keep temporal consistency. By applying DDIM inversion to obtain the initial noise, Pix2Video [30] refines the current latent code with guidance from previous and first frames. ControlVideo [31] is inflated from ControlNet by adding fully cross-frame interaction to ensure structure and content consistency. Through a noise rescheduling strategy, Freenoise [32] enables novel content generation while maintaining temporal coherence. FateZero [33] employs intermediate attention maps from DDIM inversion to preserve spatial and dynamic details and fuses these maps directly during editing. These methods provide significant insights into designing training-free frameworks in HIA tasks.

D. MORE DETAILS FOR PREVIEW GENERATION STRATEGY

The preview generation pipeline is illustrated in Figure 3 in the main paper. Specifically, we use MasaCtrl [34] with the T2I-Adapter [22] to generate the pose-conditioned image (a). MasaCtrl is a general image editing model capable of real image editing without fine-tuning. T2I-Adapter, similar to ControlNet, is a plug-and-play module that provides additional guidance to pre-trained text-to-image models without altering their original network topology or generation capabilities. The initial noise of MasaCtrl with T2I-Adapter is obtained from the reference image I_{ref} via DDIM inversion, while the T2I-Adapter's conditioning image comes from the pose sequence p_i. In the generated image (a), the identity's appearance closely resembles I_{ref}, but the background differs noticeably from the original. To address this, we use the Grounded-SAM model [35] to retain the foreground, producing image (b), referred to as the preview foreground. To obtain the clean background from I_{ref}, we first use the Grounded-SAM model to remove the foreground identity, resulting in a background image with a missing region (c). Next, we apply an image inpainting method, MAT [36], to fill the missing region, generating image (d), referred as the preview background. However, several unrealistic pixel areas appear in the previously missing region of the preview background. We attribute this to a domain gap between the MAT algorithm and our specific application scenario. However, the suboptimal regions in the preview background do not degrade our preview frames, as we perform a pixel-wise mix-up of the preview foreground and preview background. The foreground, placed on top, effectively masks the suboptimal areas in the background. The final generated preview frames, referred to as image (e), are thus produced.

MasaCtrl with T2I-Adapter. MasaCtrl [34] is a tuning-free image editing method that enables edited images to closely follow text prompts while preserving content details. T2I-Adapter [22] enhances text-to-image models by aligning outputs with target layouts from structure guidance, but it may alter the details of the original content. MasaCtrl with T2I-Adapter [34] combines these strengths by integrating T2I-Adapter's layout guidance with MasaCtrl's content fidelity, producing images that adhere to the target layout while

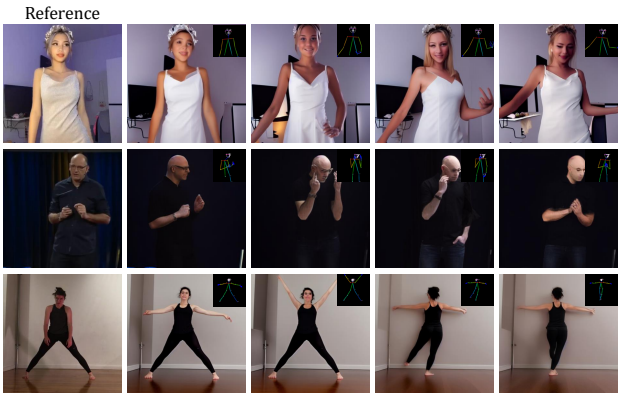

Fig. 1. Visualizations of the results produced by MasaCtrl with the pose-conditioned T2I-Adapter on TikTok (top row), TED-Talks (middle row) and EverybodyDanceNow (bottom row) datasets.

preserving original details for precise, fine-grained image editing. These features make MasaCtrl with T2I-Adapter well-suited for generating the preview foreground needed by FreeAnimate. Figure 1 shows the generation results of MasaCtrl with the pose-conditioned T2I-Adapter on several datasets.

MAT. The Mask-Aware Transformer (MAT) [36] algorithm tackles large-hole image inpainting by generating semantically accurate content for missing regions. It leverages transformers to capture long-range dependencies, alongside convolutional layers for local detail refinement, enabling high-quality, context-aware inpainting results. To generate the preview background, we use the MAT model trained on the Places365-Standard dataset [37] at a resolution of 512 × 512. Figure 2 shows preview background results across multiple datasets.

CFLD. We further explore Pose-Guided Person Image Synthesis for generating high-quality preview foregrounds, with CFLD [38] as a leading approach in this domain. CFLD, or Coarse-to-Fine Latent Diffusion, addresses Pose-Guided Person Image Synthesis by progressively refining image generation to capture pose and fine-grained appearance details, producing sound results in pose-driven synthesis. However, images generated by CFLD often showcase strong stylistic bias toward the training data, causing the synthesized identity's appearance to align with the Deepfashion dataset [39] more closely than the reference image. Figure 3 shows the preview foreground generated by CFLD.

E. EVALUATION METRICS AND BENCHMARK DETAILS

Evaluation Metrics. We compare the models on both image quality and video fidelity metrics. Image quality evaluation metrics include the Structural Similarity Index (SSIM) [40], Learned Perceptual Image Patch Similarity (LPIPS) [41], Peak Signal-to-Noise Ratio (PSNR) [42], Fréchet Inception Distance (FID) [43] and L1 error. Video fidelity is evaluated through the Fréchet Video Distance (FVD) [44].

More Details for Baselines. Most quantitative comparisons were based on statistical data from the original literature and FYPv2 [45], except StableAnimator, Champ and MimicMotion on TED-Talks dataset, which were obtained by re-running the official code. All visualization results of AnimateAnyone were obtained using the

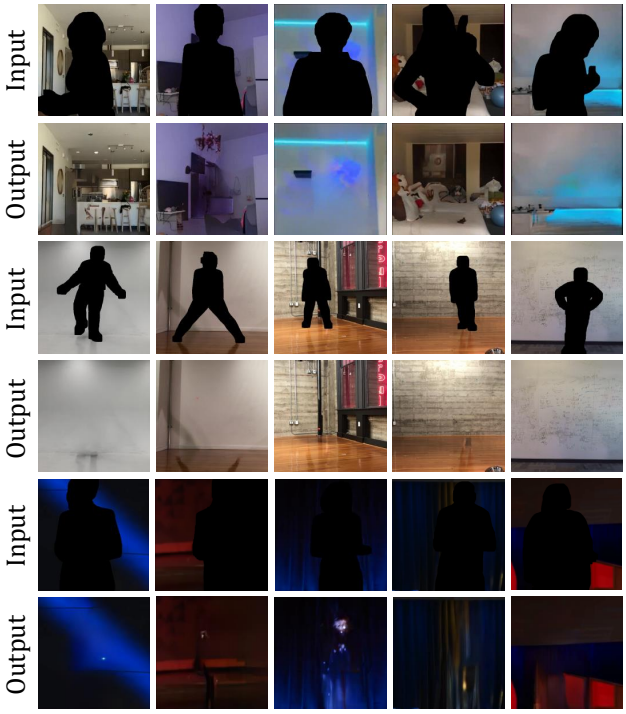

Fig. 2. Visualizations of the results produced by MAT on TikTok (Top two rows), EverybodyDanceNow (middle two rows) and TED-Talks (bottom two rows) datasets.

open-source code provided by Moore Threads [1] and the DensePose [46] conditioning maps required for MagicAnimate [10] are generated using the codebase provided by Flode-Labs [2].

More Details for Benchmark Datasets. TikTok dataset [47] comprises 350 short-duration (10∼15 seconds) dance videos, mostly of videos featuring a single individual. The majority of these videos focus on the human face and upper body. To ensure a fair comparison with SOTA methods, we employ the same test set comprising 8 TikTok-style videos from the web and 2 from TikTok dataset as DisCo [9]. EverybodyDanceNow dataset [48] consists of full-body videos of five subjects, including training set and validation set. The training set for each subject includes roughly 10k to 30k frames, while the validation set consists of approximately 1k to 5k frames. Using this dataset, we aim to evaluate the generalization of our method to full-body motions captured in unconstrained environments. TED-Talks [7] dataset comprises a large collection of video clips featuring primarily the upper body of speakers. The dataset is split into 1134 training video clips and 131 testing video clips, with each clip containing between 100 and 200 frames. Compared to the previous two datasets, TED-Talks features fewer dynamic pose variations and simpler background elements, making it a less challenging dataset.

More Details for Implementation. We use DWPose [49] to extract the pose sequence from the driving video, and adopt the pose alignment algorithm proposed in StableAnimator [16] to align the driving pose sequence with the reference pose. For each video, we sample the first frame of the video clip as the reference and all others as driv-

[1] *https://github.com/MooreThreads/Moore-AnimateAnyone*
[2] *https://github.com/Flode-Labs/vid2densepose*

Fig. 3. Visualizations of the results produced by CFLD model on TikTok dataset.

ing frames. Both reference and driving frames are center-cropped at the same position along the height dimension with the aspect ratio of 1 before resizing to 512 × 512. Due to memory and VRAM constraints, each animation sequence is generated with 8 frames at a resolution of 512 × 512 pixels. To produce longer videos, we followed the long-video animation strategy proposed in MagicAnimate [10]. For our experiments, we configured the DDIM sampler with the default settings of denoising steps T = 50 and a classifier-free guidance scale of 7.5. We apply Inversion-Boosted Attention in the interval $t \in [0.5 \times T, T]$ of the denoising step with total timestep $T = 50$, while the Reference-Anchored Self-Attention is applied across the entire denoising step. All experiments are conducted on an NVIDIA 3090 GPU, with an approximate time of 5 seconds required to complete 50 denoising steps for a single image. Furthermore, xFormers [50] are employed to optimize memory efficiency.

F. MORE VISUALIZATION RESULTS

More Qualitative Comparison. Additional qualitative comparisons of FreeAnimate with other methods are provided in Figure 4. As shown in Fig. 4, FreeAnimate demonstrates competitive, or even superior, video frame quality compared to other training-based methods.

More Video Results. Figure 5, Figure 6, and Figure 7 show additional video frames generated by FreeAnimate on the TikTok, TED-Talks, and EverybodyDanceNow datasets, respectively. As can be seen, we demonstrate consistent image animation results across different datasets.

Generalization Ability. To validate the generalization capability of our method, we present visual results on some animated images generated by diffusion models. As shown in Figure 8, FreeAnimate demonstrates strong generalization since it does not require training, thus avoiding the data bias introduced by the TikTok and TED-Talks datasets.

More Preview Frames. In Figures 9 and Figure 10, we present additional preview frames, which are already closely aligned with the target frames in terms of overall structure. These visual results demonstrate the effectiveness of the proposed Preview Generation Strategy.

G. LIMITATIONS AND FUTURE WORK

While FreeAnimate performs well in human pose alignment and background stability, it can struggle with generating fine details, particularly in complex facial expressions and hand movements within dynamic scenes. One possible solution is to leverage efficient training methods on a small-scale dataset, which could preserve the image diffusion model's ability to model complex backgrounds as well as hand and facial areas, while achieving high-fidelity image animation. Future work will focus on utilizing affordable training methods with small-scale datasets to achieve better generation details, especially for image animation in dynamic backgrounds.

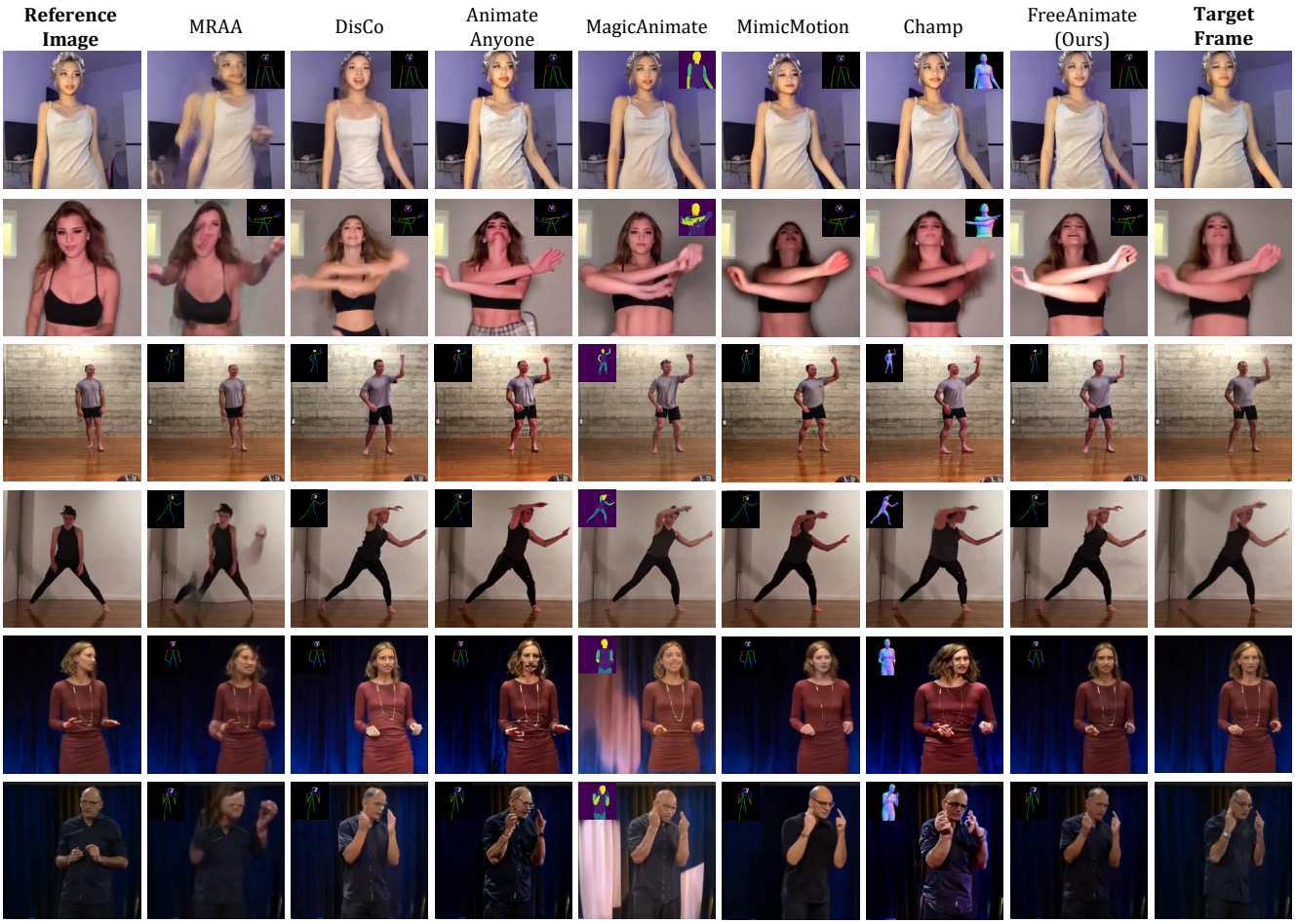

Fig. 4. Qualitative comparisons between FreeAnimate and baselines on TikTok (top two rows), EverybodyDanceNow (middle two rows) and TED-Talks (bottom two rows) datasets. The conditioning pose map is overlaid in the top corner of the generated frames.

Fig. 5. Visualization results of FreeAnimate on TikTok dataset.

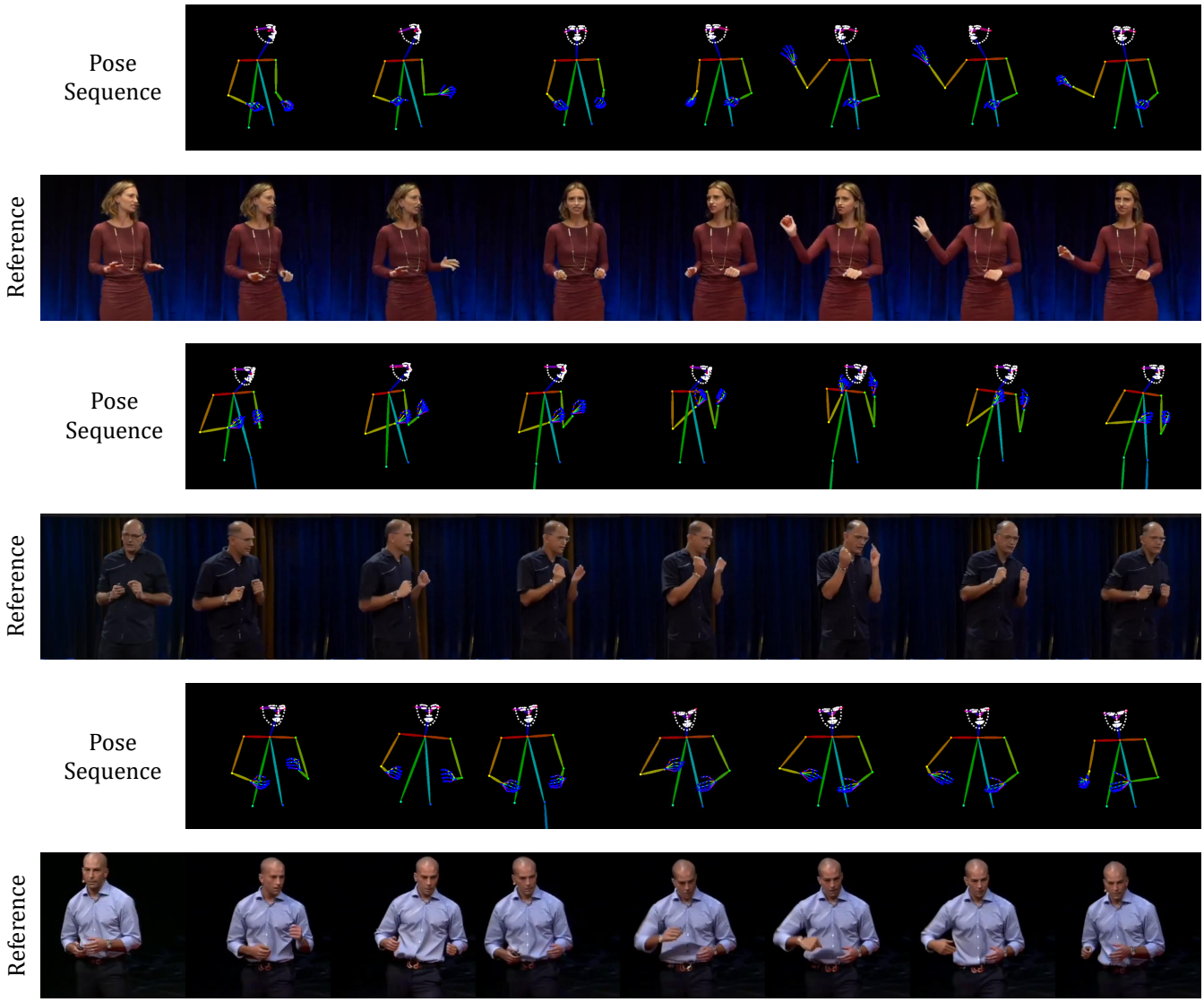

Fig. 6. Visualization results of FreeAnimate on TED-Talks dataset.

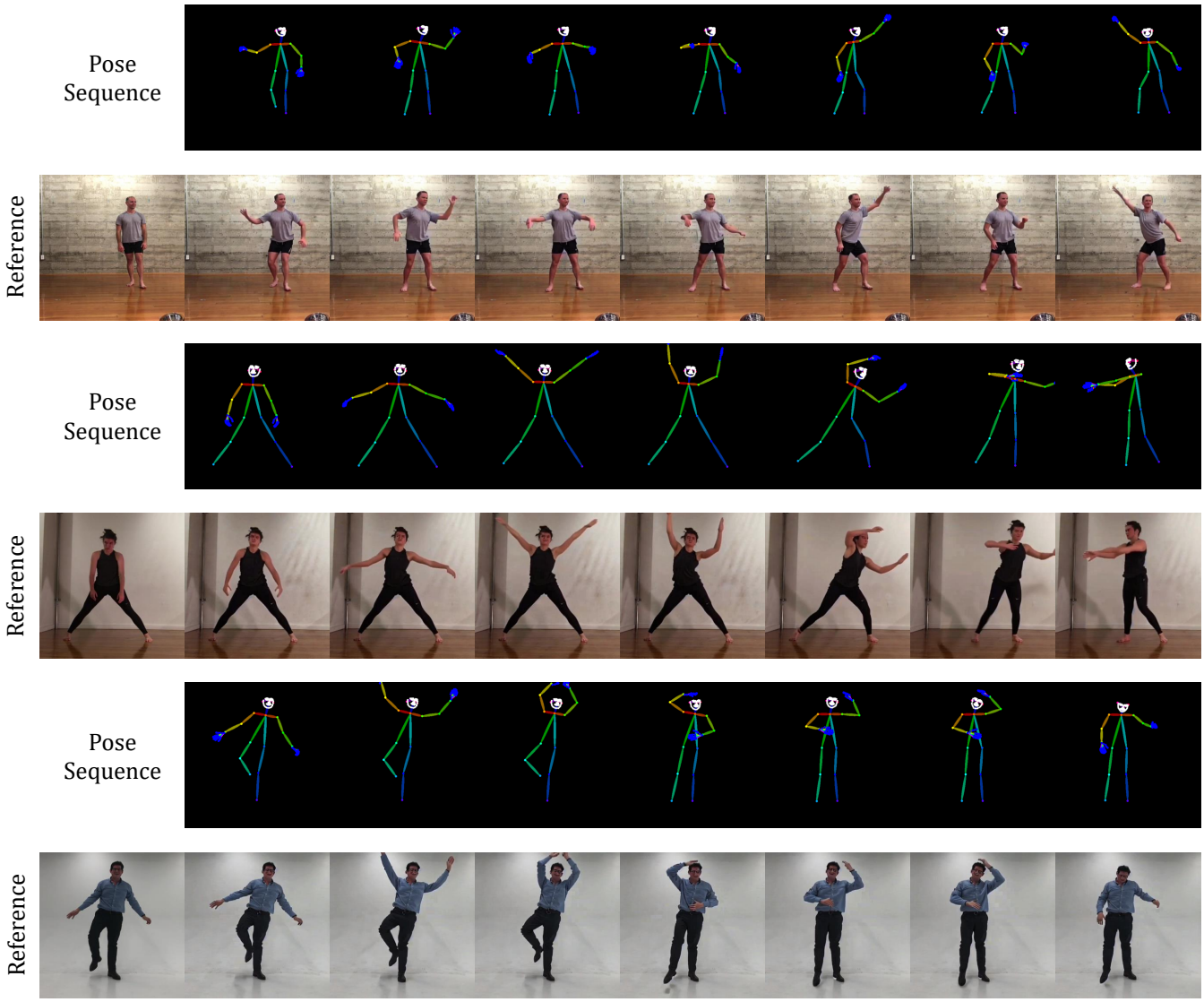

Fig. 7. Visualization results of FreeAnimate on EverybodyDanceNow dataset.

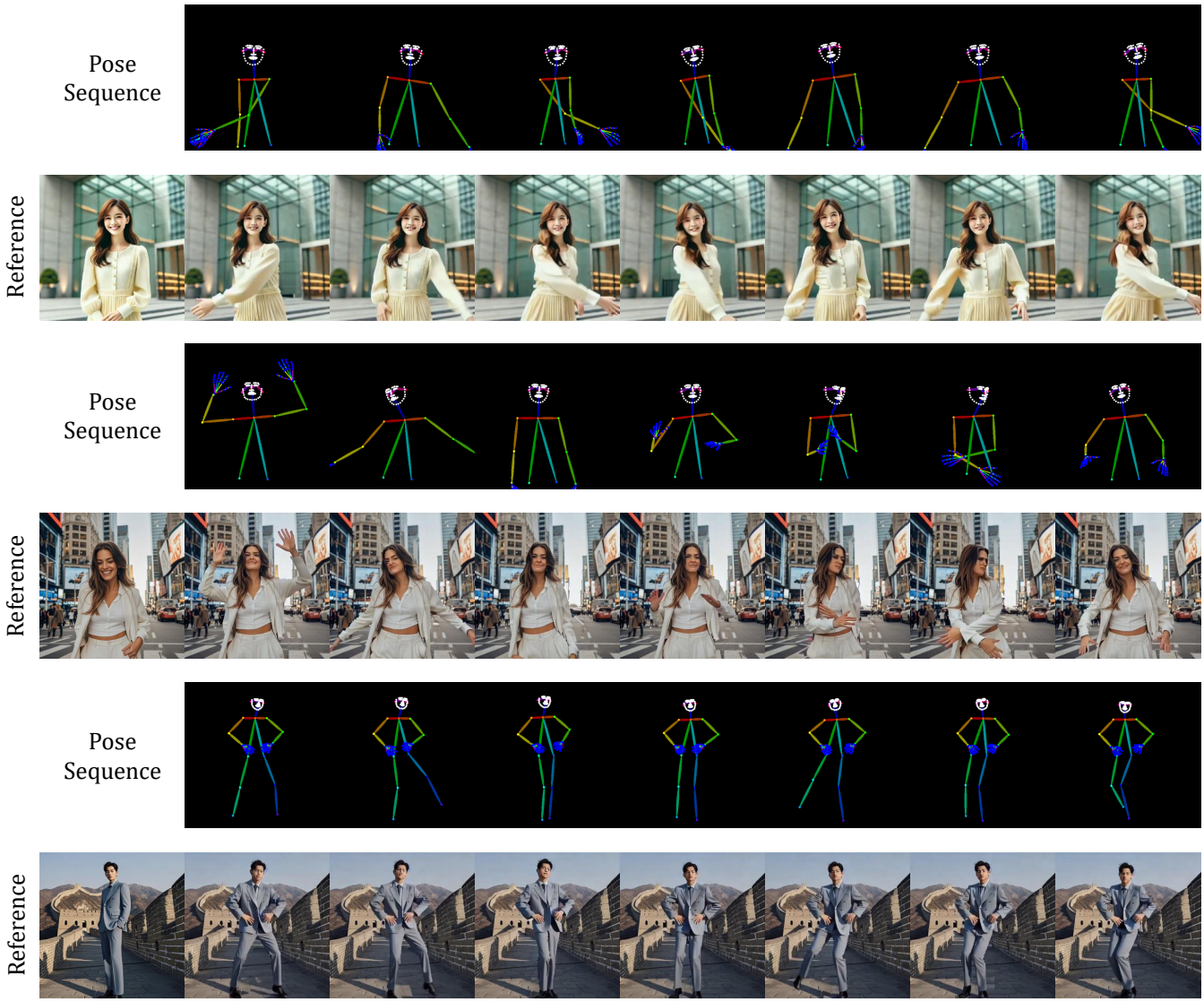

Fig. 8. Visualization results of FreeAnimate on animated images.

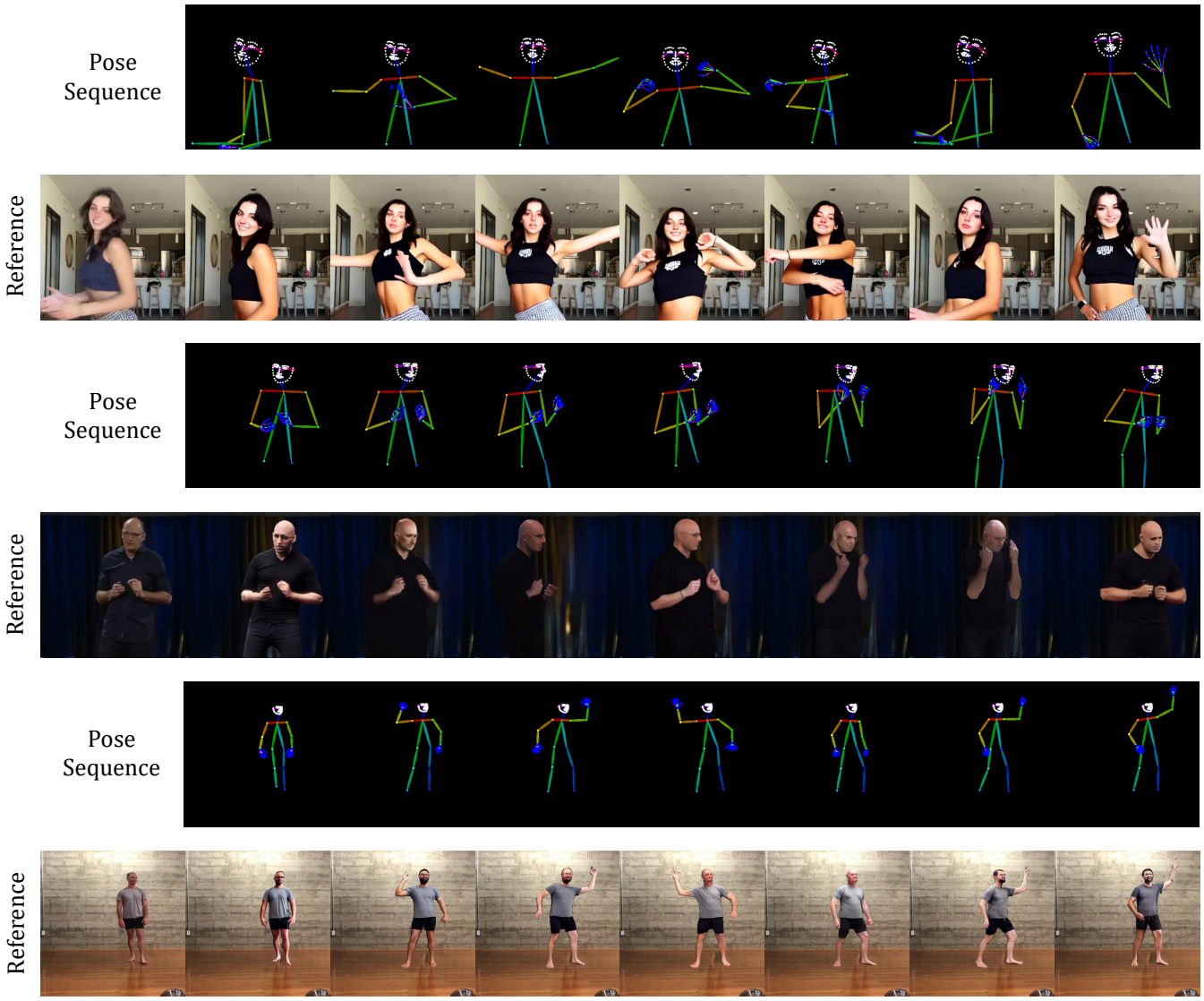

Fig. 9. Visualization results of preview frames(1/2).

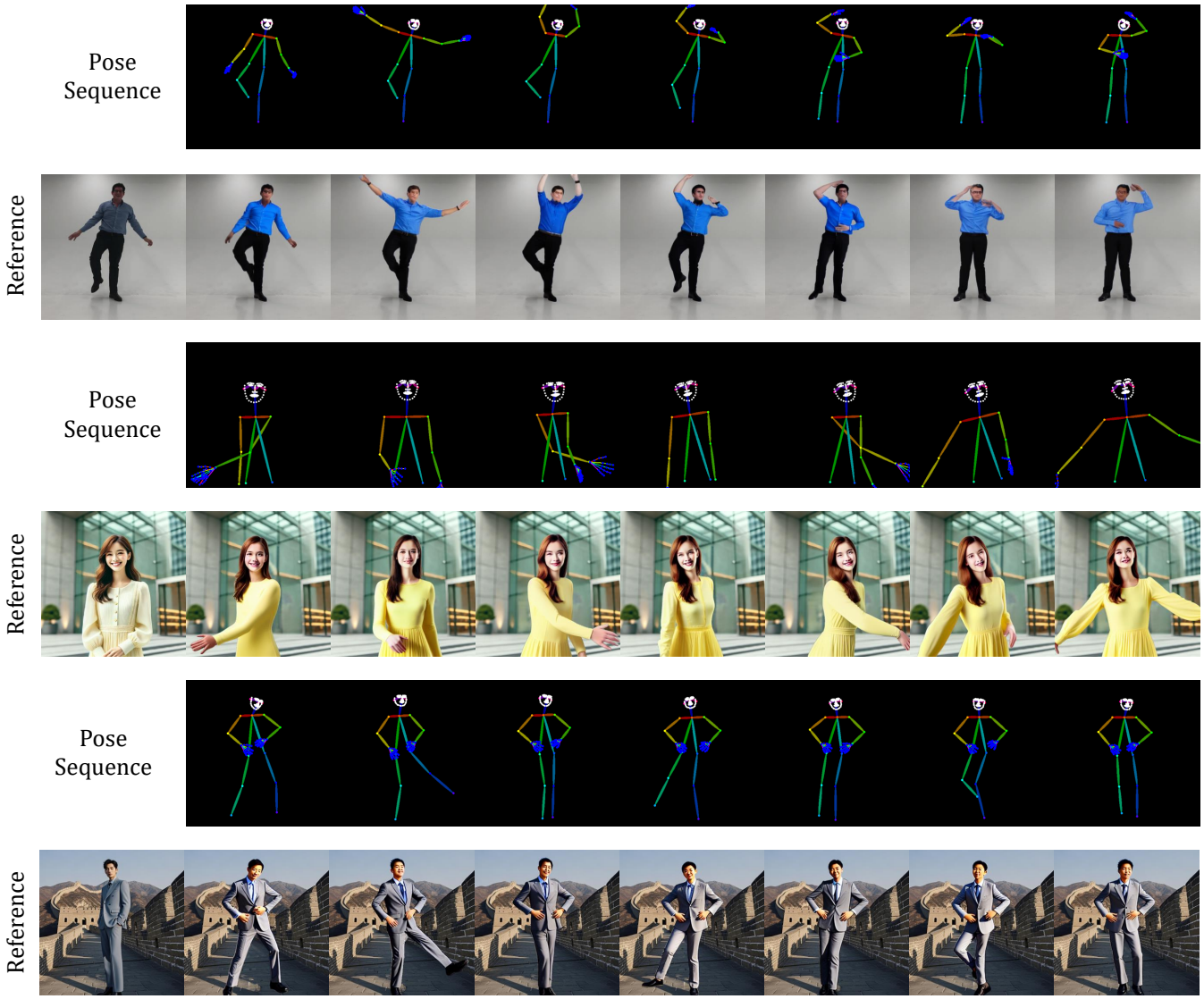

Fig. 10. Visualization results of preview frames(2/2).

H. REFERENCES

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.

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

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10684–10695.

[4] Aaron Van Den Oord, Oriol Vinyals, et al., "Neural discrete representation learning," *NeurIPS*, vol. 30, 2017.

[5] 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*. Springer, 2015, pp. 234–241.

[6] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023, pp. 3836–3847.

[7] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov, "Motion representations for articulated animation," in *CVPR*, 2021, pp. 13653–13662.

[8] Jian Zhao and Hui Zhang, "Thin-plate spline motion model for image animation," in *CVPR*, 2022, pp. 3657–3666.

[9] 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 *CVPR*, 2024, pp. 9326–9336.

[10] 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 *CVPR*, 2024, pp. 1481–1490.

[11] Li Hu, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," in *CVPR*, 2024, pp. 8153–8163.

[12] 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 *ICML*, 2023.

[13] 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.

[14] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman, "Dreampose: Fashion image-to-video synthesis via stable diffusion. in 2023 ieee," in *ICCV*, 2023, pp. 22623–22633.

[15] 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," *arXiv preprint arXiv:2406.19680*, 2024.

[16] 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.

[17] 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*. PMLR, 2021, pp. 8748–8763.

[18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al., "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[19] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. 2023.

[20] Yong Zhong, Min Zhao, Zebin You, Xiaofeng Yu, Changwang Zhang, and Chongxuan Li, "Posecrafter: One-shot personalized video synthesis following flexible pose control," in *ECCV*. Springer, 2025, pp. 243–260.

[21] Bingwen Zhu, Fanyi Wang, Tianyi Lu, Peng Liu, Jingwen Su, Jinxiu Liu, Yanhao Zhang, Zuxuan Wu, Guo-Jun Qi, and Yu-Gang Jiang, "Zero-shot high-fidelity and pose-controllable character animation," *arXiv preprint arXiv:2404.13680*, 2024.

[22] 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 *AAAI*, 2024, vol. 38, pp. 4296–4304.

[23] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al., "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

[24] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al., "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[25] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang, "Modelscope text-to-video technical report," *arXiv preprint arXiv:2308.06571*, 2023.

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

[27] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *CVPR*, 2023, pp. 22563–22575.

[28] 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," *ICLR*, 2024.

[29] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *ICCV*, 2023, pp. 15954–15964.

[30] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra, "Pix2video: Video editing using image diffusion," in *ICCV*, 2023, pp. 23206–23217.

[31] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian, "Controlvideo: Training-free controllable text-to-video generation," *arXiv preprint arXiv:2305.13077*, 2023.

[32] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu, "Freenoise: Tuning-free longer video diffusion via noise rescheduling," *arXiv preprint arXiv:2310.15169*, 2023.

[33] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen, "Fatezero: Fusing attentions for zero-shot text-based video editing," in *ICCV*, 2023, pp. 15932–15942.

[34] 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, pp. 22560–22570.

[35] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al., "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[36] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *CVPR*, 2022, pp. 10758–10768.

[37] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[38] Yanzuo Lu, Manlin Zhang, Andy J Ma, Xiaohua Xie, and Jianhuang Lai, "Coarse-to-fine latent diffusion for pose-guided person image synthesis," in *CVPR*, 2024, pp. 6420–6429.

[39] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *CVPR*, 2016, pp. 1096–1104.

[40] 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*, vol. 13, no. 4, pp. 600–612, 2004.

[41] 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, pp. 586–595.

[42] Alain Hore and Djemel Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369.

[43] 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," *NeurIPS*, vol. 30, 2017.

[44] 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 preprint arXiv:1812.01717*, 2018.

[45] Jingyun Xue, Hongfa Wang, Qi Tian, Yue Ma, Andong Wang, Zhiyuan Zhao, Shaobo Min, Wenzhe Zhao, Kaihao Zhang, Heung-Yeung Shum, Wei Liu, Mengyang Liu, and Wenhan Luo, "Follow-your-pose v2: Multiple-condition guided character image animation for stable pose control," 2024.

[46] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos, "Densepose: Dense human pose estimation in the wild," in *CVPR*, 2018, pp. 7297–7306.

[47] Yasamin Jafarian and Hyun Soo Park, "Learning high fidelity depths of dressed humans by watching social media dance videos," in *CVPR*, 2021, pp. 12753–12762.

[48] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros, "Everybody dance now," in *ICCV*, 2019, pp. 5933–5942.

[49] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li, "Effective whole-body pose estimation with two-stages distillation," in *ICCV*, 2023, pp. 4210–4220.

[50] Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov, "xformers: A modular and hackable transformer modelling library," `https://github.com/facebookresearch/xformers`, 2022.