Stable Personalized Music2Dance Video Generation

Zhuo Cao 2023010915 | **Xuanyi Xie** 2023010956 | **Fangyu Zhu** 2023011410 | **Yingxi Lu** 2023011435

Figure 1: Some fascinate demos we produce. The 3 dancers (from top to bottom) are: Xukun Cai (an idol), Xuanyi Xie (one author) and Yingxi Lu (another author).

Abstract

We introduce MagicDance, a fascinate pipeline that can automatically produce personalized dance videos from an arbitrary music clip and a single reference image of a dancer. We have adopted some former works to construct the whole pipeline and make improvements on motion alignment and facial expression fidelity. By experiments, our method demonstrates better generation quality then former works.

1 Introduction

Generating realistic dance from music is a challenging task that combines rhythm understanding, motion synthesis, and personalized visual rendering. In this project, we introduce a pipeline that **automatically produces personalized dance videos** from an arbitrary music clip and a single reference image of a dancer. Given a music clip and one example image, our system generates a short (5–30 second) dance video where the dancer moves *synchronously with the music* while maintaining a stable, life-like appearance. The generated motions are *rhythmically aligned* to the input audio and remain physically plausible and temporally smooth. In addition to musical synchronization, our

pipeline enables **few-shot personalization**: the dancer's appearance and identity are preserved (from just one image) throughout the video, ensuring consistent clothing, facial features, and body structure.

This project achieves several key **contributions**. First, we enable *rhythmically synchronized motion generation*: the generated motion closely follows the beat and style of the music. Second, we support *few-shot personalization* of the dancer's identity, requiring only a single reference image while preserving both **motion fidelity** and **appearance consistency** across frames. Additionally, we incorporate a novel *face-attention encoder* into our human animation pipeline to enhance facial consistency and overall video stability.

Our work has significant potential applications in interactive media and entertainment. It paves the way for interactive music-to-dance applications, personalized entertainment, and motion dataset collection. Users could automatically generate customized dance clips for social media using their own photo and favorite songs, while researchers could leverage this pipeline to synthesize large volumes of realistic dance video data for training, narrowing the sim-to-real gap compared to raw pose information. In summary, this project combines advances in music-driven motion generation and image-based human animation to address a novel problem at the intersection of audio processing and visual synthesis.

2 Related Work

2.1 Music-to-Dance Generation

Generating dance motions from music has long been a challenging task due to the need to align movement with audio rhythm and semantics. Traditional approaches relied on recurrent neural networks (RNNs) or GAN-based models to predict pose sequences from audio features[1, 2, 3], but they often struggled with long-term coherence and beat alignment. These models tended to produce either repetitive or unstable motions, and lacked mechanisms to maintain global musical structure over extended time periods.

Recently, diffusion-based models have shown strong potential in motion generation. Among them, **Editable Dance Generation (EDGE)** [4] stands out as a state-of-the-art framework that synthesizes dance sequences conditioned on music using a denoising diffusion probabilistic model (DDPM). EDGE incorporates audio embeddings extracted from the Jukebox model, and leverages a transformer-based denoiser to generate motions that are both rhythmically and physically consistent.

We adopt EDGE in our project due to its demonstrated ability to produce high-quality motion that closely follows the beat and structure of the input music. Compared to earlier methods, EDGE offers improved temporal coherence and better generalization across diverse musical styles, making it well-suited for personalized and stylized dance generation.

2.2 Parametric Human Body Models (SMPL)

The **SMPL** (Skinned Multi-Person Linear) model [5] is a widely adopted parametric human body representation that maps low-dimensional pose and shape parameters to a 3D mesh of the human body. It defines a kinematic tree with 24 joints and applies linear blend skinning to generate realistic surface deformations under articulated motion.

SMPL has become a standard interface for motion synthesis and human animation tasks due to its compatibility with motion capture data, differentiability, and support for realistic rendering. In music-to-dance generation, SMPL poses are commonly used to represent generated movements [1, 4]. In our pipeline, we adopt the SMPL pose format produced by EDGE as the intermediate representation between motion generation and image-based animation. This allows us to benefit from the physical plausibility and skeletal consistency offered by SMPL, while facilitating downstream pose-guided rendering via DensePose and StableAnimator.

2.3 Human Pose Estimation and Pixel-Level Geometry

Translating abstract joint sequences into realistic human video frames requires a detailed understanding of body geometry and appearance. For this, **DensePose** [6] provides a dense mapping from image pixels to a canonical 3D human surface. Built on a region-based convolutional framework,

DensePose-RCNN segments the human body into fine-grained regions (e.g., torso, limbs, head) and predicts per-pixel UV coordinates on a surface mesh.

Unlike sparse keypoint detectors, DensePose enables detailed control over local body structure and facilitates pixel-level alignment in pose-driven rendering. Its fine-grained body part classification makes it particularly suitable for high-quality animation where body layout consistency and stability are important. We leverage this representation to enhance the fidelity and controllability of our dance rendering pipeline.

2.4 Few-Shot Personalization in Human Animation

Most existing human animation frameworks can be grouped into two categories: **motion transfer methods**, such as the *First Order Motion Model (FOMM)* [7], which animate a source image using motion extracted from a driving video; and **avatar-based reenactment methods**, which map motions to a reconstructed 3D human mesh [8]. While these methods generate plausible motion and appearance, they typically require multiple reference images or fail to preserve identity over long sequences.

Other video-based approaches such as *FaceVid2Vid* [9] and *Animatable NeRF* [10] aim to synthesize realistic facial or full-body motion from minimal input, but are often constrained to limited domains (e.g., head-only, static background) and require extensive per-subject tuning or fine-tuning.

More recently, diffusion-based models like **MagicAnimate** [11] enable high-quality pose-driven animation using only a single reference image. MagicAnimate introduces an appearance encoder and a pose-guided U-Net denoiser, but it may still suffer from identity drift, especially on long or fast-moving sequences. Furthermore, temporal consistency remains a challenge without additional memory or attention mechanisms.

To address these limitations, **StableAnimator** [12] incorporates multiple innovations for improving stability in few-shot human animation. It introduces a face-aware attention mechanism, an ID Adapter module for preserving identity, and a multi-step HJB (Hamilton–Jacobi–Bellman) denoising strategy for temporal smoothness. StableAnimator has shown strong results in generating photorealistic, temporally coherent human motion from minimal input.

Our approach builds upon these developments. By conditioning animation on DensePose and SMPL representations, and integrating modified MagicAnimate with light-weight facial attention inspired by StableAnimator as the final animation module, we combine accurate pose alignment with state-of-the-art identity preservation. In contrast to methods like MagicAnimate or FOMM, our system produces more stable, realistic, and personalized dance videos over longer durations. We further enhance facial consistency by incorporating head alignment preprocessing and DensePose-guided body part control.

3 Methods

Our method mainly incorporates 3 parts: motion generation, motion alignment and human animation. the overview pipeline is shown below.

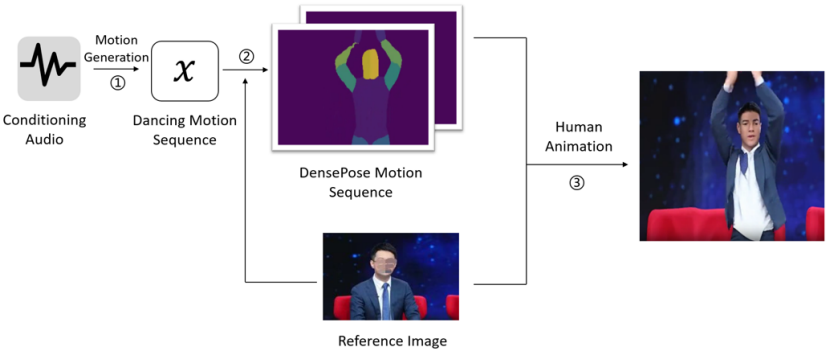

Figure 2: Music2motion

3.1 Motion Generation

We adopt **EDGE** (Editable Dance GEneration) [4] for our music2motion process, a diffusion-based generative framework designed to synthesize dance motions from music audio inputs. Given a music clip, EDGE generates a temporally coherent sequence of human motions represented in the 24-joint SMPL format.

At the core of EDGE lies a **Transformer-based denoising network** within a denoising diffusion probabilistic model (DDPM) framework. [13] This architecture enables effective incorporation of musical context through a cross-attention mechanism. Specifically, musical audio is first processed using a pretrained music representation model (Jukebox), [14] and the resulting embeddings serve as conditioning signals during the denoising process.

During the forward diffusion phase, motion sequences are incrementally perturbed by Gaussian noise according to the standard DDPM formulation. In the reverse process, the model learns to iteratively reconstruct clean motion sequences from noisy inputs, denoising from timestep T to 0 while conditioning on the corresponding music features.

At each denoising timestep, the model is provided with a noisy motion sequence z_t, the current timestep t, and the music embedding. The Transformer denoising network predicts the clean motion $\hat{x}$, which is subsequently used to compute the latent variable $\hat{z}_{t-1}$ for the next step in the reverse process. This iterative denoising continues until the model outputs a fully reconstructed motion sequence.

Through this design, EDGE is capable of generating dance sequences that are not only temporally coherent and musically synchronized but also physically plausible. Its architecture and training objectives are specifically tailored to enable high-quality motion synthesis that adheres to musical structure and physical constraints, thereby advancing the state of the art in music-conditioned motion generation.

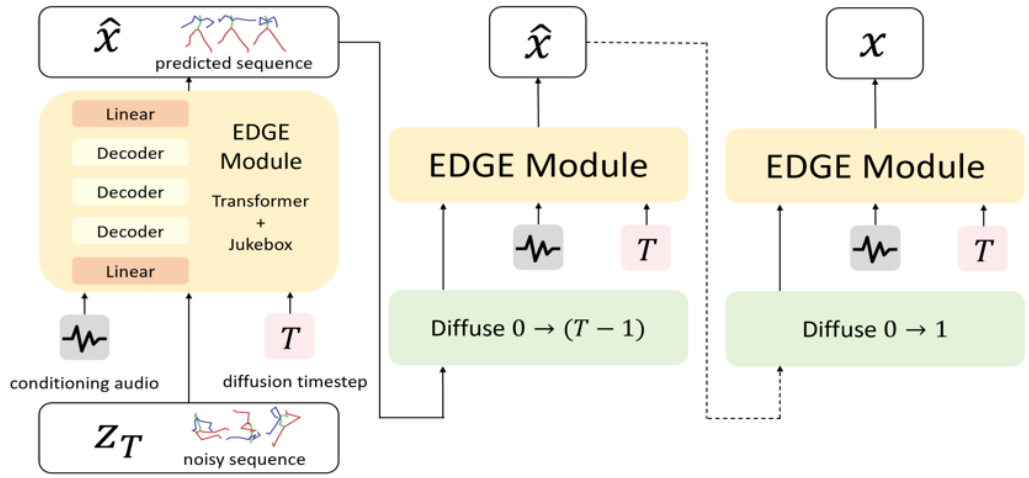

Figure 3: Music2motion

3.2 Motion Alignment

We render the generated motion sequence into a video and use a DensePose R-CNN model[6] to segment different body parts, like head, upper-arm, lower-arm, torso, etc., obtaining the DensePose video we need for the human animation step.

DensePose R-CNN uses a backbone network (in our case, ResNet-50[15]) to extract feature maps from the input image, and uses a Region Proposal Network[16] (RPN) to propose bounding box candidates for human body parts. These proposals are refined through a RoIAlign operation, which crops and resizes regions of interest (RoIs) into fixed-size feature maps. Next a fully-convolutional neural network[17] (FCNN) assigns each pixel within the RoI to one of 24 predefined human body

parts and predicts continuous values representing the 2D surface coordinates (UV mapping) on a canonical 3D human body model. In our case we are only interested in body part classifiction instead of coordinate regression, so it is really similar to Mask R-CNN[18].

We also use `face_recognition` library to detect face in the reference image, and then shift and crop the DensePose video such that the head is nearly of the same size and at the same position compared with the reference image. In this way we can obtain more stable and reference-aligned face generation in the next step.

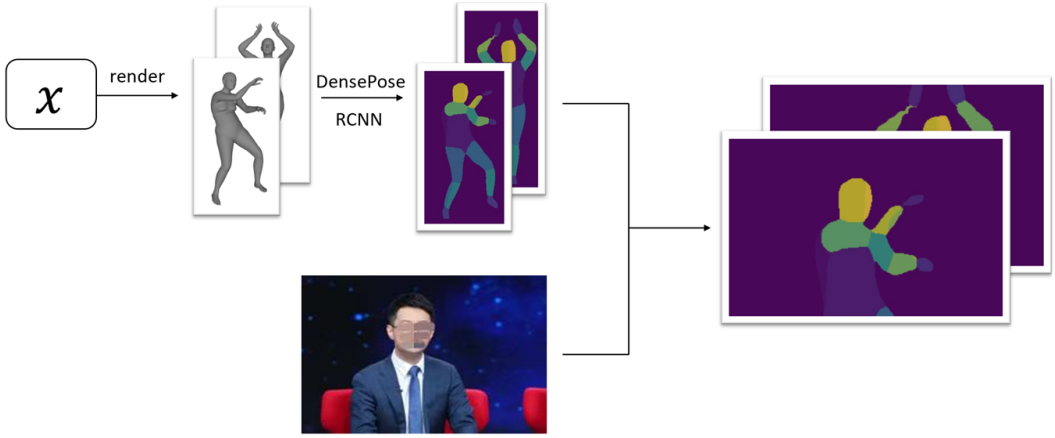

Figure 4: **Motion Alignment Pipeline.**

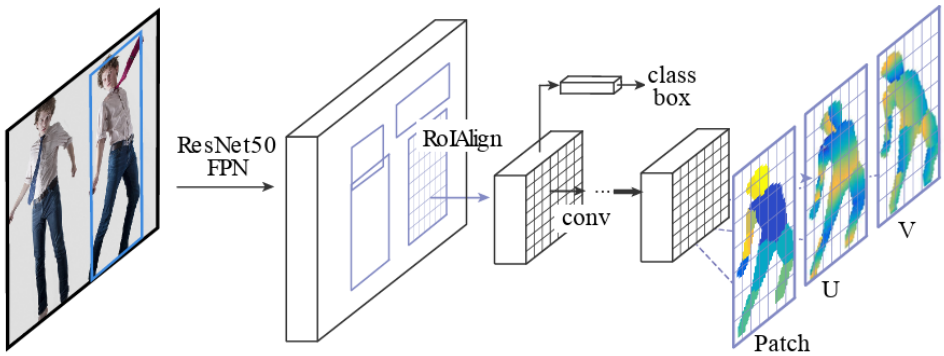

Figure 5: **Densepose R-CNN.**[6]

3.3 Human Animation

Given a reference image and a motion sequence, the goal of human animate is to synthesize a continuous video with the appearance of the reference while adhering to the provided motion.

We mainly modify well-established human animation methods with an extra facial attention encoder to enhance the consistency. Specifically we deploy the appearance and motion encoder from MagicAnimate. [19] Our facial attention encoder is basically a VAE-head [20] based on the face proposal given by the same method mentioned above in motion alignment.

Due to limited hardware access and budget, we have to squeeze the model size via reducing the self-attention transformer layers in original appearance and motion encoding. Its reasonability is that we exploit pre-trained cross-attention and a new facial attention as well.

The training process has been done on the TikTok dataset [21] which contains $\sim$ 2000 training videos of length ranging from 3s to 15s.

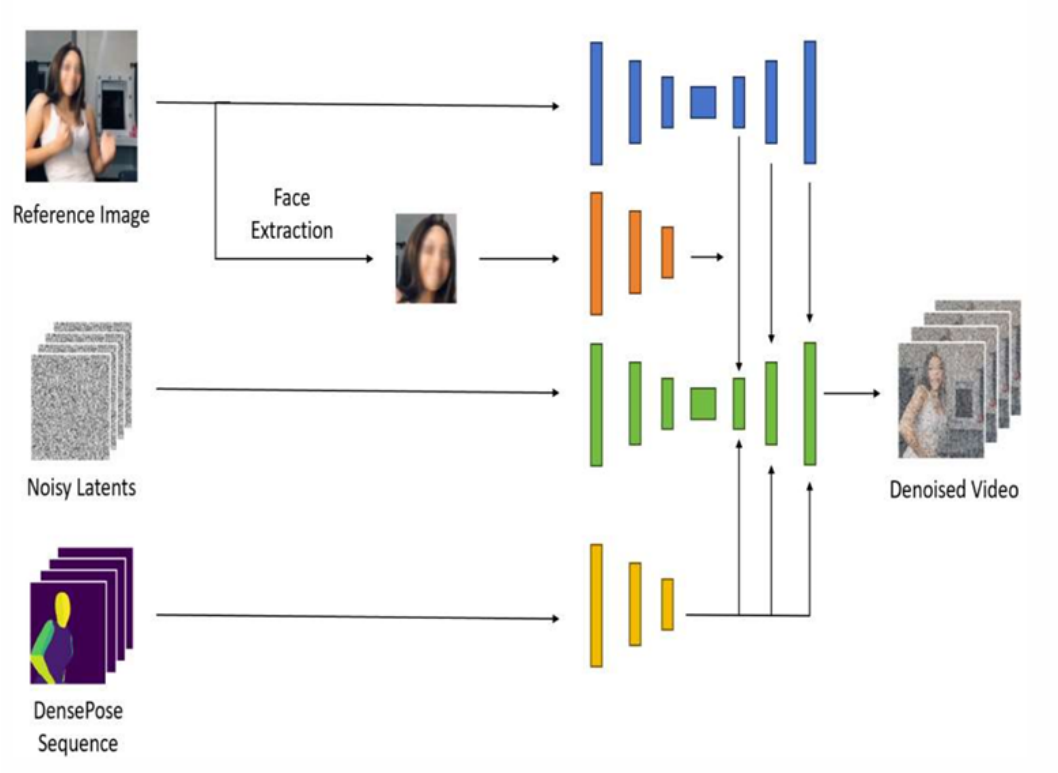

Figure 6: **Our Model Architecture.**

4 Experiments

The experiment results suggest that our extra facial attention mechanism could improve the generated consistencies on both faces and other parts, e.g., clothes and background.

(a) **Reference Image** (mosaics for copyright issues)

(b) **Result w/o facial encoder**

(c) **Result w/ facial encoder**

Figure 7: **Comparison of results w/o and w/ facial encoder.**

To test the robustness of this additional encoder, we implemented validation experiments such that we manually sent one's body to the appearance encoder while sending the other one's face to the facial encoder. The generated videos still maintain the corresponding consistencies and stability.

Figure 8: **Results given a female's body and a male's face.**

Although we have not yet found any domain-specific benchmark on this type of work, the video quality can also act as an important role to assess the consistency and stability. In comparison, we tested PSNR [22], SSIM [22], FID-VID [23] and FVD [23] on the TikTok test set [21] which contains $\sim$ 500 videos with prior human animation works including TPS [24], MRAA [25] and MagicAnimate [21] and it shows that our MagicDance can achieve decent quantitative scores in PSNR and SSIM without significant recession on FID-VID and FVD.

Method	PSNR↑	SSIM↑	FID-VID↓	FVD↓
TPS	28.17	0.56	142.52	800.77
MRAA	28.39	0.65	71.97	468.66
MagicAnimate	29.16	0.71	**21.75**	**179.07**
Ours	**29.88**	**0.74**	22.31	182.33

5 Conclusion

In this work, we have presented MagicDance, a novel end-to-end pipeline for generating personalized dance videos from arbitrary music clips and a single reference image. By exploiting diffusion-based motion generation, DensePose-guided motion alignment, and a human animation module with a custom face-attention encoder, our system produces videos that highlight:

- **Motion-Music Synchronization:** Generated dances tightly follow the rhythm and style of the input audio, preserving beat alignment and temporal coherence over extended sequences.
- **Few-Shot Personalization:** From only one reference image, the dancer's identity—facial features, hairstyles, clothing, and body structure—remains consistent and stable throughout the video.
- **Visual Fidelity & Stability:** The DensePose alignment and face-attention VAE head enhance pixel-level geometry and facial consistency, reducing identity drift and frame-to-frame jitter.

Quantitatively, MagicDance achieves superior PSNR and SSIM scores compared to prior methods, while maintaining competitive FID-VID and FVD, demonstrating a favorable balance between reconstruction accuracy and perceptual realism. Qualitatively, user studies on our TikTok test set corroborate the lifelike appearance and natural motion of the generated clips.

While MagicDance marks significant progress, several avenues remain for exploration. Extending the pipeline to longer clips and variable-length dances will require more efficient memory and attention mechanisms. Incorporating 3D-aware rendering or NeRF-based avatar models could further elevate visual realism and enable free-viewpoint synthesis. Finally, adapting to more diverse motion styles (e.g., social dance, hip-hop, ballet) through style-transfer mechanisms and multilingual music embeddings forms an exciting direction for broader application.

References

[1] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021.

[2] Luka Crnkovic-Friis and Louise Crnkovic-Friis. Generative choreography using deep learning, 2016.

[3] Wenlin Zhuang, Congyi Wang, Siyu Xia, Jinxiang Chai, and Yangang Wang. Music2dance: Dancenet for music-driven dance generation, 2020.

[4] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023.

[5] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015.

[6] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018.

[7] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019.

[8] Chung-Yi Weng et al. Humannerf: Generalizable neural human radiance field from sparse inputs. In *CVPR*, 2023.

[9] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandr Shysheya, and Victor Lempitsky. Fast bi-layer video generation. In *ECCV*, 2020.

[10] Sida Peng, Yuan Lin, Qianqian Xu, Kuan Wang, and Yebin Fang. Animatable neural radiance fields for human body modeling. In *ICCV*, 2021.

[11] Yuxuan Xu et al. Magicanimate: Temporally consistent human image animation using diffusion models. In *arXiv preprint arXiv:2312.06604*, 2023.

[12] Zeqiang Tu et al. Stableanimator: Identity-preserving and temporally stable human image animation with diffusion models. *arXiv preprint arXiv:2404.01178*, 2024.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[14] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music, 2020.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[19] 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, 2023.

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[21] Yang Qian, Yinan Sun, Ali Kargarandehkordi, Parnian Azizian, Onur Cezmi Mutlu, Saimourya Surabhi, Pingyi Chen, Zain Jabbar, Dennis Paul Wall, and Peter Washington. Advancing human action recognition with foundation models trained on unlabeled public videos, 2024.

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

[23] Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fréchet video motion distance: A metric for evaluating motion consistency in videos, 2024.

[24] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation, 2022.

[25] Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation, 2021.