

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

Supplementary Material

This supplementary document provides additional details and insights into our work. In Sec. 1, we present a comprehensive overview of our dataset, *TalkingPose*. Sec. 2 describes the modified TED-talk[#] dataset and further clarification regarding PSNR calculation. Sec. 3 provides further ablation studies and comparative analyses, Sec. 4 highlights the limitations of our model and finally, Sec. 5 showcases additional visualizations produced by our framework.

1. TalkingPose Dataset

Existing datasets [2, 18] largely fail to capture expressive upper body movements, particularly detailed facial expressions and hand gestures. The TED-Talks dataset [14], while the closest existing dataset relevant to our problem, is limited in size and lacks diversity in backgrounds, making it insufficient for studying a wide range of expressive human animations.

To address these limitations and advance research in this area, we developed the TalkingPose dataset. This dataset comprises approximately 18K video samples — 43× larger than TED-talk [14] (411 videos) and 51× larger than TikTok [9] (350 videos) — and features individuals delivering presentations with expressive hand gestures. These videos were sourced from various YouTube channels under the “CC BY-NC-ND 4.0” license. TalkingPose represents a diverse range of individuals across different ages, genders, and backgrounds, offering a comprehensive resource for analyzing expressive upper body movements (see Fig. 1).

Dataset Collection and Processing Methodology:

We present a detailed description of the dataset collection and processing pipeline used in our work. Our goal was to construct a large-scale, diverse dataset of human upper-body videos to support the training of robust generative models. The dataset was carefully curated to ensure a wide range of ages, genders, ethnicities, clothing styles, and environmental contexts, while maintaining strict identity separation between training and testing partitions.

Data Acquisition and Initial Curation:

We collected 21,000 raw videos from diverse YouTube sources, focusing on human presenters with visible upper-body regions. The videos were selected to include a wide variety of ages, genders, and ethnicities, as well as diverse clothing styles and background environments. This diversity is critical for training generative models to handle variations in appearance, pose, and context. Additionally, the dataset emphasizes dynamic upper-body

motions, such as expressive gestures and pose variations, which are essential for synthesizing coherent and realistic avatars.

Identity Verification and De-Duplication:

To ensure identity uniqueness and prevent leakage between training and testing sets, we implemented a multi-stage identity verification pipeline. First, facial regions were extracted from each video using RetinaFace (ResNet-50 backbone) [4] with five-point landmark detection. Frames were filtered to retain only front-facing poses using a nose-tip alignment heuristic. Next, face embeddings were generated using ArcFace (ResNet-100) [3], and spectral clustering was applied to identify and remove duplicate identities across videos. A cosine similarity threshold of 0.4 was used for clustering, resulting in 18K unique identities.

Upper-Body Processing and Quality Control:

The upper-body regions were extracted using YOLO V10 [16] for human detection. Each frame was processed to localize the upper-body region at a resolution of 512×512 pixels. Frames were discarded if the detected region after cropping is less than the desired resolution while considering the proper aspect ratio, ensuring consistent spatial proportions. The resulting dataset comprises 1250 hours of video at 20 FPS.

Dataset Partitioning and Statistics:

The curated videos were segmented into 50-frame clips for our task and the frames where the pose extraction failed were discarded. The final dataset consists of 500K clips for training (derived from 16,200 identities) and 47K clips for testing (derived from 1,800 held-out identities). This partitioning ensures that the training and testing sets are completely disjoint in terms of identities.

Demographic Analysis:

To validate the demographic diversity of our dataset, we employed the `buffalo_l` model pack from the Insightface repository for age and gender analysis, and the DeepFace repository for expression recognition, which integrates state-of-the-art models for facial attribute analysis. The results, illustrated in Fig. 2 and Fig. 3, confirm a balanced and varied representation of facial attributes. The expression analysis (excluding the “disgust” category) reveals that the dataset comprises six main classes: neutral (35.8%), sad (17.5%), happy (17.1%), angry (14.9%), fear (11.1%), and surprise (3.6%). The age distribution spans a broad range, with the majority of samples concentrated among young to middle-aged subjects. In particular, the 25–29, 30–34, and 35–39 age groups represent 22.8%,



Figure 1. **TalkingPose Dataset.** Example samples from the TalkingPose dataset, showcasing diversity in nationalities, age groups, and backgrounds.

Dataset	Real	Face	Hands	Full Body	#Identities	Resolution	Age Range	Diverse BG	Diverse App.
VoxCeleb2 [2]	✓	✓	✗	✗	6k	256×256	Wide	✓	✓
VFHQ [19]	✓	✓	✗	✗	-	512×512	Wide	✓	✓
TalkingHead-1KH [18]	✓	✓	✗	✗	-	512×512	Wide	✓	✓
AVspeech [5]	✓	✓	✗	✗	150k	Var.	Wide	✓	✓
HDTF [23]	✓	✓	✗	✗	362	512×512	-	✓	✓
TED-talk [14]	✓	✓	✓	✗	411	384×384	Adult	✗	✗
TalkShow [21]	✓	✓	✓	✗	4	1280×720	Adult	✗	✗
TikTok [9]	✓	✓	✓	partial	300	Var.	Young	✓	✓
TalkingPose (Ours)	✓	✓	✓	✗	18K	512×512	Wide	✓	✓

Table 1. **Comparison of face/human body datasets.** Columns indicate whether the dataset is Real (✓) or synthetic (✗), which body regions are included (partial indicates limited coverage), the approximate number of identities, total hours of video, typical resolution, gender balance, age range, and the diversity in background (BG) and appearance. Values are based on recent literature and can be adapted to each dataset’s specifics.

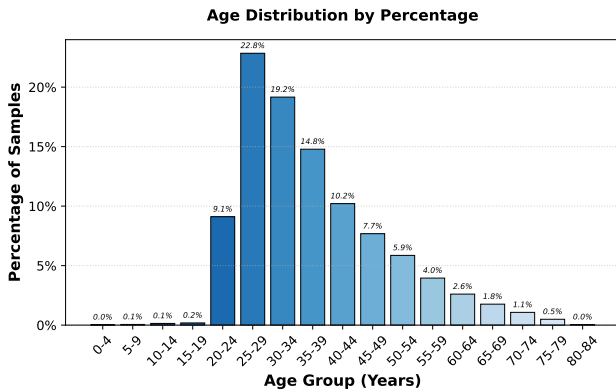


Figure 2. **Age distribution.**

19.2%, and 14.8% of the dataset, respectively, while both the very young (0–4) and older age groups (75–84) are less represented. Furthermore, the gender distribution is balanced, with males comprising 55.4% and females 44.6%

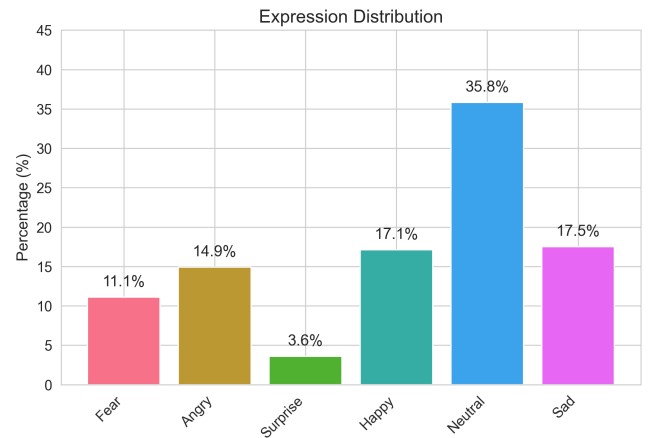


Figure 3. **Expression distribution.**

of the total.

This comprehensive demographic analysis demonstrates that the dataset adequately represents a wide array of

facial attributes, thereby ensuring its suitability for training inclusive and generalizable facial recognition models.

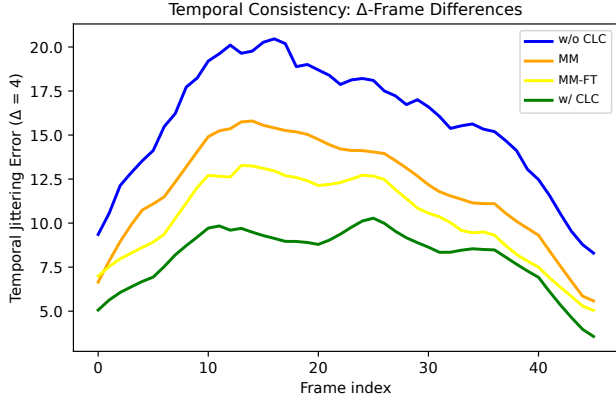


Figure 4. **Temporal Jittering Error (TJE).** "Base" refers to the model without a motion module, "MM" indicates the model with a pretrained motion module, "MM-FT" denotes the motion module fine-tuned, and "Ours" represents the Base model fitted with our CLC mechanism. Lower is better.

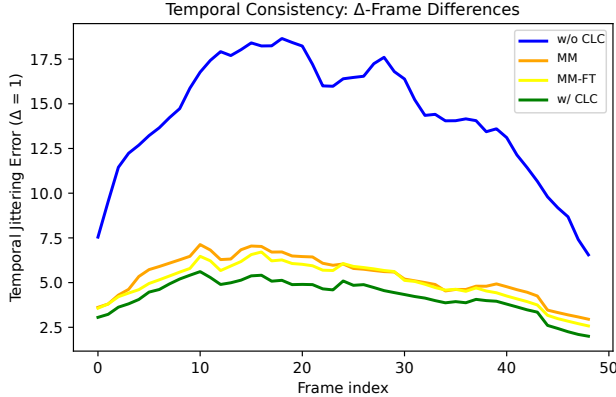


Figure 5. **Temporal Jittering Error (TJE).**

2. Experiments

TED-talk#. The TED-talk dataset was introduced in MRAA [14] (2021) using YouTube videos from the TED-Talk channel. When we attempted to download the full set of 411 videos, 62 videos were no longer downloadable. To address this, we replaced them with other videos from the same channel to maintain the total size of 411, resulting in a modified version we denote as TED-talk#. Importantly, all methods were evaluated on TED-talk# for fairness.

PSNR calculation. The discrepancy in reported PSNR values across prior works arises from differences in data type handling when computing Mean Squared Error (MSE)

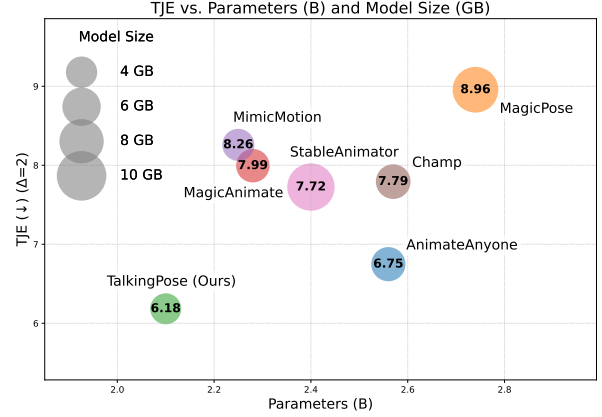


Figure 6. **TJE vs Model Complexity.** Temporal jittering error for $\Delta = 2$ compared across methods with respect to model parameters (B) and model size (GB).

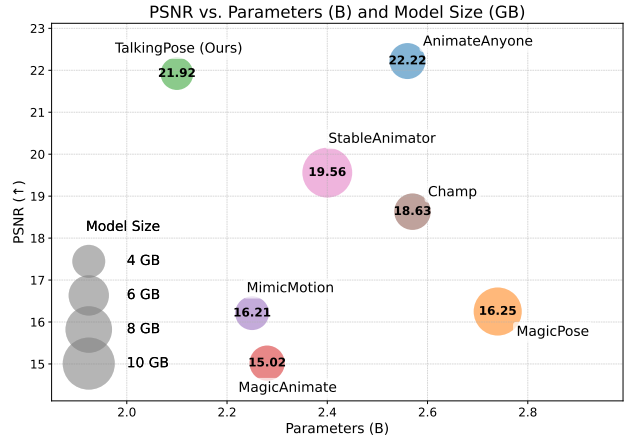


Figure 7. **PSNR vs Model Complexity.** PSNR compared across methods with respect to model parameters (B) and model size (GB).

for PSNR calculation. Specifically, using integer values can cause numerical overflow. This issue originates from the Disco [17] evaluation toolkit, which was widely used in earlier papers [1, 7, 24], and is also observable when comparing different versions of the Disco paper [17] (e.g., v1 on arXiv vs. v3). In the updated Disco toolkit, MSE is computed using floating point, and recent works (e.g., [15, 20, 22]) follow this update. StableAnimator [15] also reported both cases for comparison.

3. Additional Analysis

In addition to the image-based metrics (SSIM, PSNR, LPIPS) and temporal evaluations (FID-VID, FVD), we further measure performance using the Average Keypoint Distance (AKD) [6] computed with MediaPipe [11] and the face-based Cosine Similarity (CSIM) [8, 12], following [1]. During experimentation, we observed that AKD can

Method	TED-talk [#]				TalkingPose				Model	
	AKD (face)↓	AKD (Hands)↓	AKD (torso)↓	CSIM↑	AKD (face)↓	AKD (Hands)↓	AKD (torso)↓	CSIM↑	Param. (B)↓	Size (GB)↓
AnimateAnyone [7]	<u>0.45</u> (0.451)	1.45 (1.523)	3.38 (3.380)	<u>0.54</u>	0.43 (0.436)	1.43 (1.659)	<u>2.67</u> (2.670)	0.54	2.56	4.77
MagicPose [1]	0.60 (0.601)	2.14 (2.247)	4.48 (4.480)	0.51	0.55 (0.557)	1.98 (2.297)	3.49 (3.490)	0.49	2.74	8.53
MagicAnimate [20]	1.52 (1.523)	2.99 (3.140)	4.71 (4.710)	0.36	1.34 (1.358)	2.53 (2.935)	3.90 (3.900)	0.35	2.28	4.57
MimicMotion [22]	2.46 (2.465)	3.04 (3.193)	4.78 (4.780)	0.40	1.81 (1.835)	2.41 (2.795)	3.32 (3.320)	0.40	2.25	4.20
Champ [24]	3.89 (3.898)	5.74 (6.028)	8.62 (8.620)	0.27	3.80 (3.853)	5.69 (6.601)	8.81 (8.810)	0.29	2.57	4.91
StableAnimator [15]	2.36 (2.365)	2.93 (3.077)	4.15 (4.150)	0.78	1.77 (1.794)	2.38 (2.761)	2.85 (2.850)	0.75	2.40	9.16
TalkingPose (Ours)	0.43 (0.430)	<u>1.46</u> (1.533)	3.17 (3.170)	<u>0.54</u>	0.42 (0.425)	<u>1.51</u> (1.751)	2.25 (2.250)	<u>0.56</u>	2.10	3.93

Table 2. **Quantitative comparisons on TED-talk[#] and TalkingPose Dataset.** The reported AKD is shown alongside its adjusted value (in parentheses), which is obtained by dividing the original AKD by the fraction of frames where the corresponding landmark is detected. Bold values indicate the best performance and underlined values indicate the second-best performance.

Method	TikTok			
	AKD (face)↓	AKD (Hands)↓	AKD (torso)↓	CSIM↑
AnimateAnyone [7]	<u>0.63</u> (0.660)	2.91 (10.070)	4.60 (4.600)	0.45
MagicPose [1]	0.72 (0.754)	4.21 (14.568)	8.11 (8.109)	0.43
MagicAnimate [20]	2.10 (2.200)	5.34 (18.478)	6.34 (6.340)	0.42
MimicMotion [22]	3.25 (3.404)	5.36 (18.548)	6.74 (6.740)	0.26
Champ [24]	3.89 (4.075)	8.29 (28.687)	9.64 (9.640)	0.35
StableAnimator [15]	3.02 (3.163)	4.77 (16.506)	5.31 (5.310)	0.83*
TalkingPose (Ours)	0.61 (0.639)	<u>2.93</u> (10.139)	<u>4.78</u> (4.780)	<u>0.51</u>

Table 3. **Quantitative comparisons on TikTok dataset.** Values marked with * are reported directly from the original publications.

Method	$\Delta = 1$	$\Delta = 2$	$\Delta = 4$	$\Delta = 8$
AnimateAnyone w/o MM	5.264	5.882	6.792	<u>7.964</u>
AnimateAnyone	2.997	<u>4.382</u>	<u>6.263</u>	8.673
StableAnimator	<u>2.929</u>	4.457	6.440	8.865
Ours	2.336	3.366	4.677	6.197

Table 4. **Average Temporal Jittering Error (TJE)** at four frame-offsets Δ . Lower is better.

be inaccurate for frames with noticeable artifacts, because the landmark extractor often fails to detect landmarks in those frames. To address this, we first calculate AKD only on frames deemed valid across all methods (i.e., where faces, hands, or torsos are successfully detected). We then compute the average percentage of valid landmarks across all frames—so frames with artifacts that cause the extractor to fail for one method but not for others result in a lower detection fraction for that particular method. Finally, we divide the original AKD values by this detection fraction to achieve a fairer comparison. Tables 2 and 3 show that *TalkingPose* outperforms all baselines on most metrics, ranking second in AKD for hands and CSIM, indicating strong identity preservation. Moreover, our method is the most efficient: it adds no additional temporal layers and instead relies on a more effective CLC mechanism, resulting in fewer parameters and reduced computational overhead. Table 2 further quantifies this efficiency: because the model

does not include temporal parameters in its inference-time size, does not require a second training stage as in [7], and does not train on video stacks of frames like StableAnimator [15] or MimicMotion [22], its GPU memory requirement at inference is minimal—only one frame is held in memory at a time—whereas all other methods operate on stacks of frames.

To further analyze temporal consistency, we performed a temporal jittering error (TJE) evaluation across several Δ values (e.g., $\Delta=1$, Fig. 5; $\Delta=4$, Fig. 4). These plots show that our CLC mechanism consistently outperforms the motion module and, by a large margin, the base model, indicating strong temporal stability.

We additionally evaluated long videos (approximately 1k frames per sample) on a 50-video validation set from TED-talk[#]. We compared our method with AnimateAnyone [7] (a Stable Diffusion-based [13] method with a motion module) and StableAnimator [15] (based on Stable Video Diffusion [15]). Although these methods maintain reasonable temporal consistency over short durations, performance degrades on long sequences because concatenating short generated chunks introduces temporal drift, as reported in Tab. 4. Our method outperforms these approaches while also being significantly more efficient. Inspired by the representation style in [10], we plot the Temporal Jittering Error (TJE) against model parameters and size in Fig. 6. TJE measures the temporal consistency of generated videos with respect to the ground truth in same-identity animation on the TED-talk[#] dataset. In Fig. 7, we report the PSNR, which evaluates the photorealism of the generated frames. As shown, our proposed method consistently outperforms competing approaches in terms of TJE, and achieves the highest PSNR among all methods, performing on par with AnimateAnyone, the strongest baseline. Importantly, these improvements are obtained in a highly efficient manner: our CLC-based approach requires neither additional parameters to enforce temporal consistency nor training on temporal data, resulting in a significantly smaller and more parameter-efficient model.

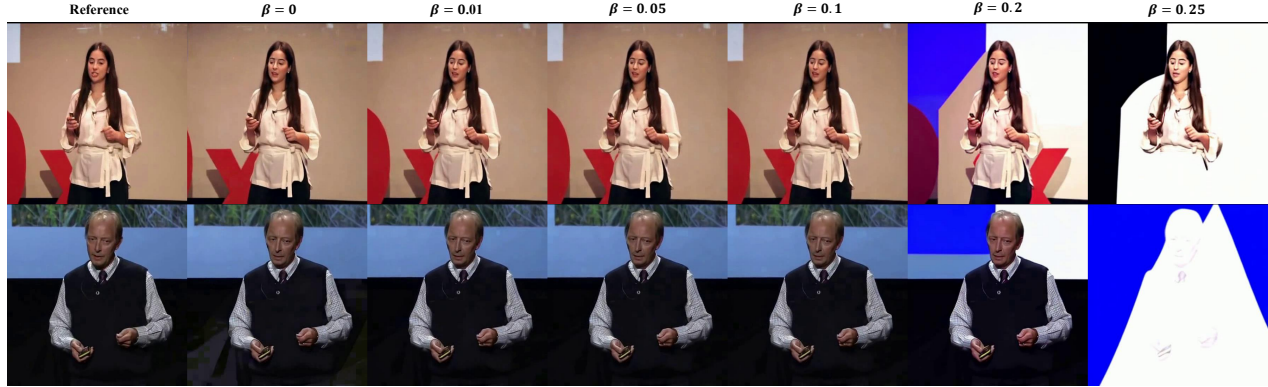


Figure 8. Ablation study on β values.

To further demonstrate the effectiveness of our proposed CLC, beyond the quantitative ablation on β values, we present two examples from the TalkingPose dataset in Fig. 8. When $\beta = 0$, the generated results exhibit high visual quality, but temporal consistency is not maintained. For instance, in the first row, the background is slightly misaligned compared to the ground truth. Such artifacts become more pronounced in videos, manifesting as jittering. Notably, even a small increase in β , which activates the feedback loop, can effectively suppress these artifacts. On the other hand, setting β too high (e.g., $\beta > 0.1$) introduces significant distortions into the output. Through our experiments, we found the optimal value to be $\beta = 0.05$.

4. Limitations

Our framework exhibits certain constraints when confronted with extreme poses, particularly those involving head or torso rotations beyond 90 degrees. In such scenarios, the accurate extraction of 3D poses becomes challenging (as illustrated in the top section of Fig. 9). Additionally, the system’s performance is compromised when the initial source image fails to adequately represent or partially occludes the hand. Consequently, in some instances, the character’s hand may appear slightly blurred, stemming from the diffusion backbone’s inability to faithfully generate the hand region. These limitations highlight areas for potential future improvements in our approach. Furthermore, the strong generative prior of stable diffusion can sometimes cause the framework to hallucinate certain aspects of the appearance in the generated motion, leading to inaccuracies.

5. Additional Results

We present additional results in Fig. 10 and Fig. 11 to further demonstrate the effectiveness of our framework. Additionally, Fig. 12 highlights our framework’s ability

to replicate the same pose across multiple identities with high precision, regardless of variations in age, ethnicity, or gender. We encourage readers to refer to the supplementary video for additional visualizations.

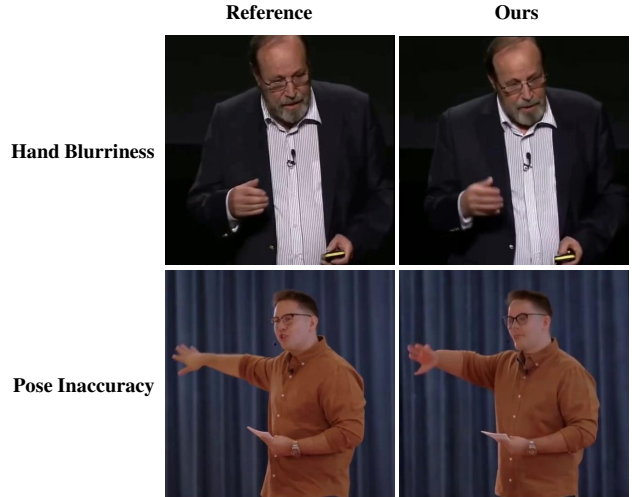


Figure 9. Limitations.

References

- [1] 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. 3, 4, 7, 8
- [2] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 1, 2
- [3] 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. 1

- [4] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 1
- [5] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37(4), 2018. 2
- [6] Michael Gashler. Waffles: A machine learning toolkit. *Journal of Machine Learning Research*, 12(69):2383–2387, 2011. 3
- [7] 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. 3, 4, 7, 8
- [8] 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. 3
- [9] 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. 1, 2
- [10] Daikun Liu, Lei Cheng, Teng Wang, and Changyin Sun. Edflow: Exploring temporally dense difference maps for event-based optical flow estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1984–1993, 2025. 4
- [11] 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. 3
- [12] 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. 3
- [13] 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. 4
- [14] 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. 1, 2, 3
- [15] 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. 3, 4, 7, 8
- [16] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024. 1
- [17] 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. 3
- [18] 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. 1, 2
- [19] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 2
- [20] 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. 3, 4, 7, 8
- [21] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–480, 2023. 2
- [22] Yuang Zhang, Jiayi 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. 3, 4, 7, 8
- [23] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 2
- [24] 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. 3, 4, 7, 8



Figure 10. **Qualitative Comparison.** Comparison of our framework with the state-of-the-art methods Champ [24], AnimateAnyone [7], MagicPose [1], MagicAnimate [20], MimicMotion [22] and StableAnimator [15] on the TalkingPose dataset (top six rows) and the TED-talk[#] dataset (bottom six rows).



Figure 11. **Qualitative Comparison.** Comparison of our framework with the state-of-the-art methods Champ [24], AnimateAnyone [7], MagicPose [1], MagicAnimate [20], MimicMotion [22] and StableAnimator [15] on the TikTok dataset.

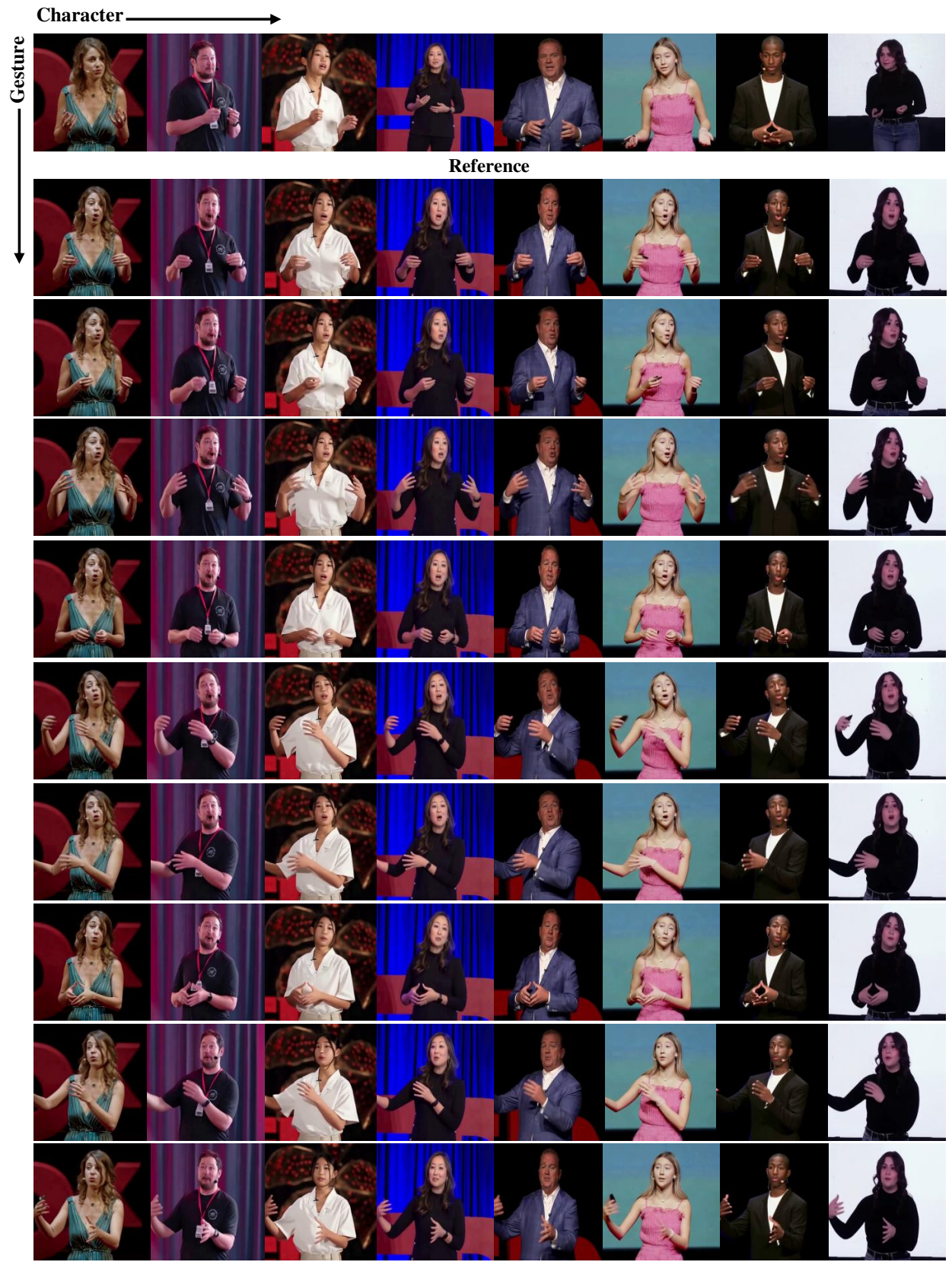


Figure 12. **Cross-identity Animation.** The rows represent the *same pose* with different identities while the columns represent the different poses within the *same identity*.