Supplementary Material for Submission 3319

DreamAnimate: Temporal Consistency and Detail Preservation for Character Animation

This supplementary material is organized as follows:

- Section A includes evaluation metrics and other details.
- Section B presents more qualitative and quantitative results to supplement the main paper.
- Section C presents an analysis of the user study.

A. Experimental Evaluation Metric

1) $\mathcal{L}_1$: The mean absolute pixel value difference between the reconstructed and original frames.

2) Average keypoint distance (AKD): The average distance between corresponding landmarks. A human body pose estimator [1] predicts keypoints for the reconstructed and original frames, and the average distance between corresponding keypoints is computed.

3) Missing keypoint rate (MKR): The proportion of keypoints that are missing in the reconstructed frames but present in the original frames.

4) Average Euclidean distance (AED): A metric to evaluate identity preservation in reconstructed videos. Identity embeddings are extracted from reconstructed and original frames using public identification networks for bodies [2] (applicable to TaiChiHD and TED-talks datasets) and faces [3]. The average Euclidean distance between corresponding embeddings is then calculated.

B. More experimental comparison results

We compare DreamAnimate with three baseline methods: First Order Motion Model (FOMM) [4], Motion Representations for Articulated Animation (MRAA) [5], and Thin-Plate Spline Motion Model (TPSMM) [6]. All models were retrained under identical conditions for fair comparison, and evaluations were conducted for both same-identity and cross-identity scenarios.

Same-identity video reconstruction. Figure 1 illustrates challenging cases on the TaiChiHD and TED-talks datasets, where large motion amplitudes lead to detail errors in synthesized videos. Compared to baseline methods, DreamAnimate successfully resolves artifacts in hand, arm, and head regions. Table I highlights that our method can superior reconstruction quality and temporal continuity for same-identity videos.

Same-identity animation quality. Fig. 5 compares same-identity animation results of DreamAnimate with FOMM, MRAA, and TPSMM on TED-talks and TaiChiHD datasets. DreamAnimate demonstrates superior motion transfer for full-body animations and preserves video details, including hands and faces. Baseline methods often suffer from ghosting artifacts, whereas our method ensures higher generation quality. Fig. 2 further showcases that our method has significant improvements in limb quality compared to TPSMM.

Fig. 1. Some bad cases of FOMM, MRAA, and TPSMM methods. The first and second rows show the results of the TED-talks dataset, while the third and fourth rows display the results of the TaiChiHD dataset. By observing, we can see that the problems of these methods are in the hand, arm, and head areas.

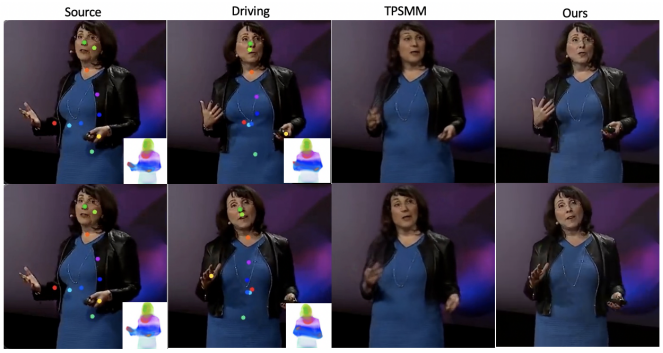

Fig. 2. Comparison between TPSMM and our method for synthesizing same-identity animations. The first and second columns display the detected keypoints and area heatmaps (inset) in the source image and driving video, the third column shows the results synthesized by TPSMM method, and the column on the right shows our results.

Cross-identity video reconstruction. Cross-identity animation performance depends on the initial postures of the source image and driving video. Table I demonstrates DreamAnimate's effectiveness in generating body parts at desirable positions, outperforming baseline methods. FOMM struggles with keypoint accuracy due to its unsupervised learning approach, MRAA fails to preserve body shape, and TPSMM often compromises identity. DreamAnimate excels in synthesizing

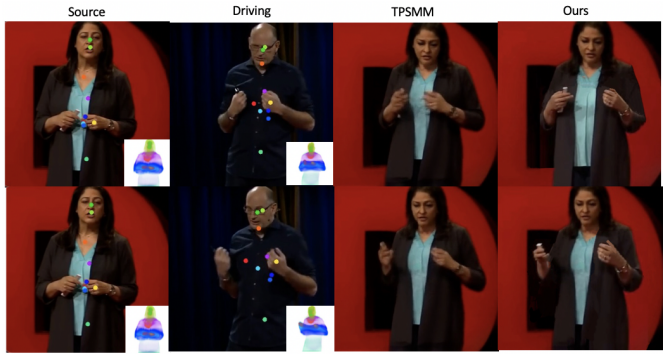

Fig. 3. The comparison between our method and the TPSMM method for cross-identity animation reconstruction. The first and second columns show the detected keypoints and area heatmaps (inset), the third column shows the results of the TPSMM method, and the fourth column shows our results. Note that the source image we selected has an object in hand, which increases the difficulty of video reconstruction.

Fig. 4. Some bad cases of baseline methods in cross-identity video reconstruction task. We evaluated these methods on TED-talks and TaiChiHD datasets, and the primary problems with the synthesized videos are blurred fingers, unnatural facial expressions, and inaccurate motion transfer (especially on the TaiChiHD dataset, as shown in the third and fourth lines).

realistic videos, especially in hand and finger details. Figure 4 highlights improvements over baseline methods in challenging cross-identity scenarios.

Cross-identity animation quality. Figure 7 compares cross-identity animation results, showing DreamAnimate's ability to maintain hand and face details while ensuring consistency with driving videos. Baseline methods suffer from finger loss, motion artifacts, and poor movement transfer. Figure 3 shows DreamAnimate's superior performance when the source image includes objects like pens, emphasizing significant improvements in finger quality and overall animation fidelity.

C. User Study

To comprehensively evaluate animation quality, we conducted a user study with 100 participants who assessed 100

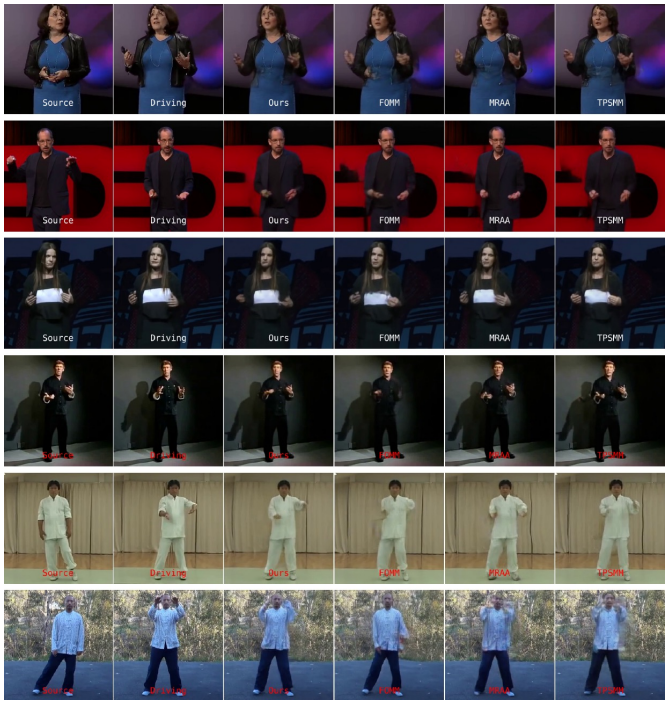

Fig. 5. Same-identity reconstruction qualitative results. From left column to right: source, driving, results of our method, FOMM, MRAA, and TPSMM.

TABLE I
CROSS-IDENTITY MOTION TRANSFER QUANTITATIVE RESULTS.

Method	TED-talks	
	Average keypoint distance	Missing keypoint rate
MRAA(100 epochs)	4.55683	0.00647
TPSMM(100 epochs)	4.06219	0.00465
Ours w/o. finetuning (10 epochs)	4.89683	0.01075
Ours w. finetuning (85 epochs)	4.09754	0.00435
Ours w. finetuning (100 epochs)	**8.84720**	**0.00414**

TABLE II
SAME-IDENTITY MOTION TRANSFER QUANTITATIVE RESULTS.

Method	TED-talks			
	$\mathcal{L}_1$	AKD	MKR	AED
MRAA	0.02583	3.24598	0.00536	0.10767
TPSMM	0.02547	2.73063	0.00609	0.12049
Ours w/o. bg	0.03027	3.12230	0.00552	0.13023
Ours	**0.02284**	**2.00547**	**0.00341**	**0.10604**

randomly selected pairs of results. Participants compared the authenticity and temporal coherence of animations generated by DreamAnimate and baseline methods. As shown in Table III, DreamAnimate received significantly higher positive evaluations. Its superior performance stems from effectively handling complex actions and scenarios, such as characters holding objects, while preserving realistic details and temporal consistency.

REFERENCES

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

Fig. 6. Compare w-o facial keypoints 79 vs 149 epoch

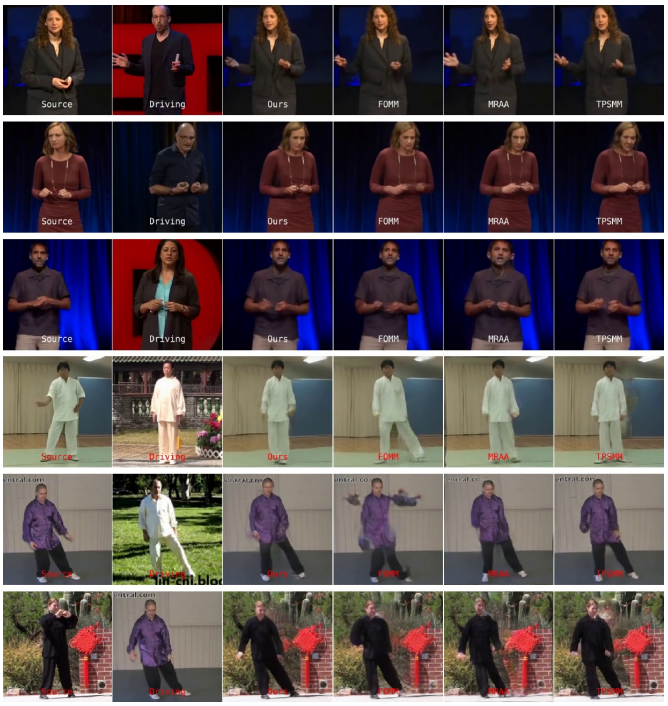

Fig. 7. Cross-identity motion transfer qualitative results. From left column to right: source, driving, results of our method, FOMM, MRAA, and TPSMM.

TABLE III
USER STUDY ON CHARACTER ANIMATION, NUMBERS RESPECT THE PROPORTION (%) OF USERS THAT PREFER OUR METHOD OVER THE UNIANIMATE METHOD.

Dataset	Ours vs TPSMM [6]		Ours vs MagicAnimate [7]		Ours vs Unianimate [8]	
	Authenticity	Continuity	Authenticity	Continuity	Authenticity	Continuity
TaiChiHD [4]	99%	90%	82%	85%	75%	84%
TED-talks [5]	97%	87%	85%	94%	74%	83%
TikTok [9]	96%	98%	86%	85%	78%	76%

[2] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[3] Olivia Wiles, A Koepke, and Andrew Zisserman, "X2face: A network for controlling face generation using images, audio, and pose codes," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 670–686.

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

[5] 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*, 2021, pp. 13653–13662.

[6] Jian Zhao and Hui Zhang, "Thin-plate spline motion model for image animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3657–3666.

[7] 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*, 2024, pp. 1481–1490.

[8] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang, "Unianimate: Taming unified video diffusion models for consistent human image animation," *arXiv preprint arXiv:2406.01188*, 2024.

[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*, 2021, pp. 12753–12762.