TokenMotion: Decoupled Motion Control via Token Disentanglement for Human-centric Video Generation

In this supplementary material, we provide additional details and extensive video results for the paper, which is organized follows:

1. Details of datasets (Sec.1).
2. Additional results for joint camera and human motion control (Sec. 2).
3. Additional results for human motion control (Sec. 4).
4. Additional results for camera control (Sec. 3).
5. Discussions of limitations and future works (Sec. 5).

1. Details of Datasets

For training, we use both HumanVid [6] and RealEstate10K [14], which satisfies the need of both scenarios of rich human motion and rich camera motion. This dataset combination aligns with approaches adopted in related literature [7, 12]. We first aligned the two datasets' control signals to ensure their compatibility: For human motion, we extracted different motion cues as detailed in Sec.4 in the main paper. For camera motion, we normalized and transformed camera parameters from relative to world coordinates in HumanVid samples to match RealEstate10K's format. Then, we further augmented the datasets by applying a central crop with a 2:3 height-to-width ratio to all video frames, followed by resizing to 256 $\times$ 384 pixels. To further validate the performance of proposed model on unseen scenarios, we collected 12 images from Grammy GlamBot Shots [9] collections on Youtube as shown in Fig. 6.

For evaluation, we filtered the test data to ensure a rich diversity of actions, enhancing the comprehensiveness and effectiveness of our assessment. To verify this diversity, we performed action classification using InternVL-8B [2], demonstrating that our test set encompasses approximately 110 action types across 500 samples, with walking (74), dancing (53), and waving hands (32) being the most frequent.

2. Additional Results: Joint Control

We provide additional qualitative comparisons of jointly controlling camera and human motion for T2V generation in Fig. 1, Fig. 2 and Fig. 3, and for I2V generation in Fig. 5. For T2V generation, we compare our approach with Direct-A-Video [12], MotionBooth [8] and MotionCtrl [7]. For I2V generation, we compare our approach with ImageConductor [4].

3. Additional Comparisons: Camera Control

Quantitative Results The widely used COLMAP metrics are known to introduce randomness to its pipeline, and sensitive to inconsistent features that easily leads to reconstruction failure [10, 13]. Following CamCo [10], we consider the number of 2D points available in the reconstructed 3D point clouds as a reflection of the 3D-reconstruction consistency of the video samples, and additionally compute this reconstruction error rate (Recon-err) to complement results in the main paper in Tab. 1.

Compared to baselines, our approach achieves the least reconstruction error across camera-only control, and joint-control in both T2V and I2V generation. This demonstrates that our approach outperforms other baselines in the overall camera-control following, despite minor compromises at select key points.

We note that CamCo[10] and VD3D [1] are excluded for comparison, as their codes and model weights are not available. While DimensionX [5] is also built upon CogVideoX and performs camera control, it only offers weights for two types of orbital camera controls by the submission deadline, which prevents comprehensive comparison on RealEstate10K test set, and thus is not included for quantitative comparison as well. Please refer the videos in the attachments for DimensionX results.

Additional Qualitative Results We carefully selected different scenarios of camera control to demonstrate the generation capabilities of TokenMotion. Fig. 6 and Fig. 7 shows the results on our collected Grammy dataset. And Fig. 8 and Fig.9 shows the camera control results on RealEstate10k[14] dataset.

4. Additional Results: Human-motion Control

We further conduct comparisons with concurrent human image animation works, namely MagicAnimate [11] and AnimateAnyone [3] to show our model's capacities of incorporating human-motion control under the static camera

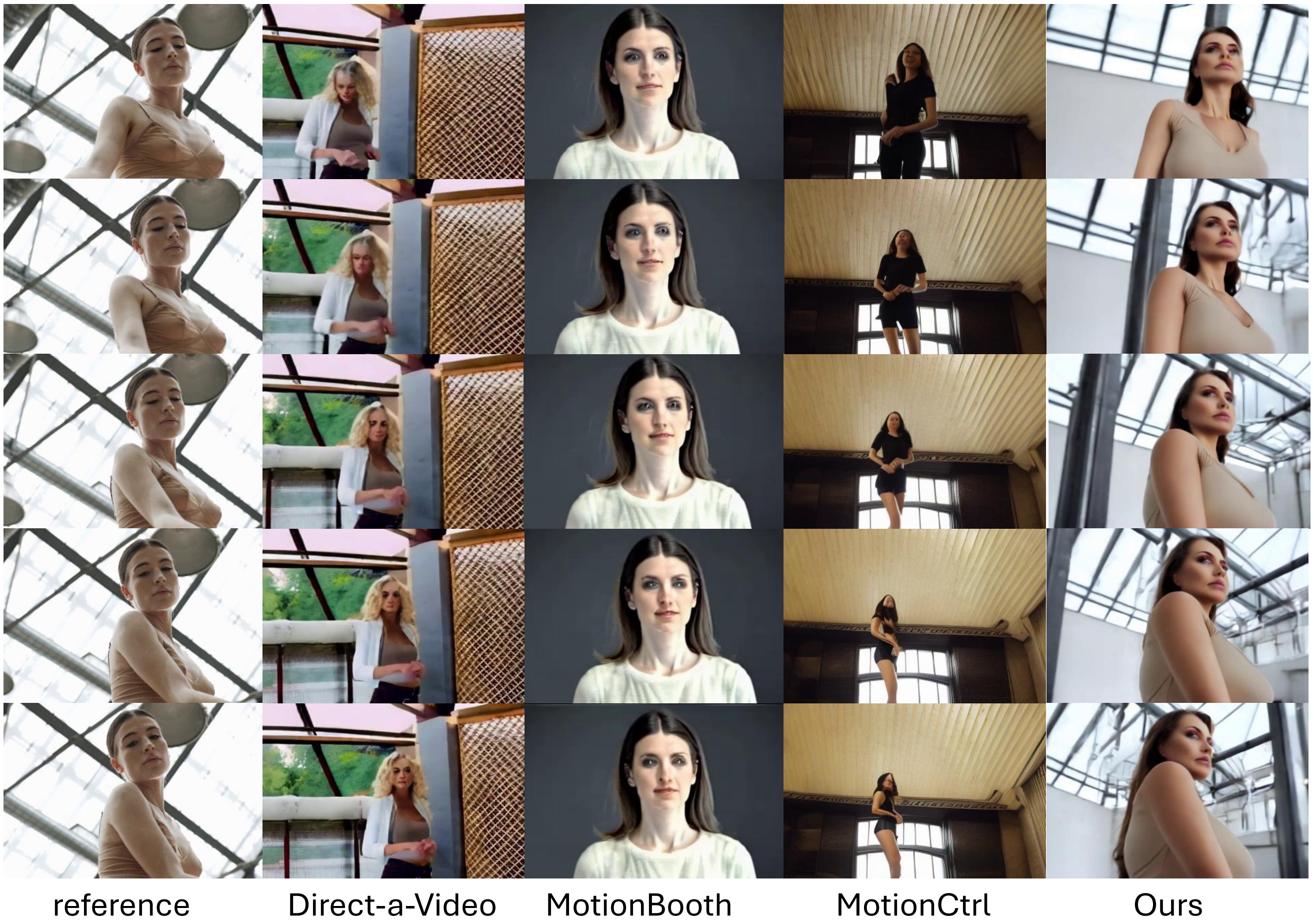


Figure 1. Qualitative comparison results between our approach and Direct-A-Video, MotionCtrl and MotionBooth for joint-control in T2V generation.

Methods	Recon-error	Kpts-error	Rot-error	Trans-Error
CameraCtrl	0.6	5.61	**0.27**	6.48
MotionCtrl-cam	0.47	**4.46**	0.33	5.83
Ours-Cam	**0.41**	**4.36**	0.28	**5.39**
MotionCtrl	0.81	6.72	0.47	8.66
Direct-A-Video	0.71	**3.85**	**0.24**	**5.06**
MotionBooth	0.46	5.24	0.36	6.42
Ours(TokenMotion-T)	**0.39**	4.43	0.27	5.34
ImageConductor	0.78	4.32	0.44	5.44
Ours(TokenMotion-I)	**0.51**	**3.77**	**0.22**	**2.98**

Table 1. Additional camera control results for camera-only control on T2V generation, and joint camera and human motion control on both T2V and I2V generation.

settings with results shown in Tab. 2.

Results show that our model outperforms AnimateAnyone for all four metrics and yield strong results on par with the state-of-the-art MagicAnimate. This indicates that our approach achieves both strong performance in dynamic camera setings and static camera settings.

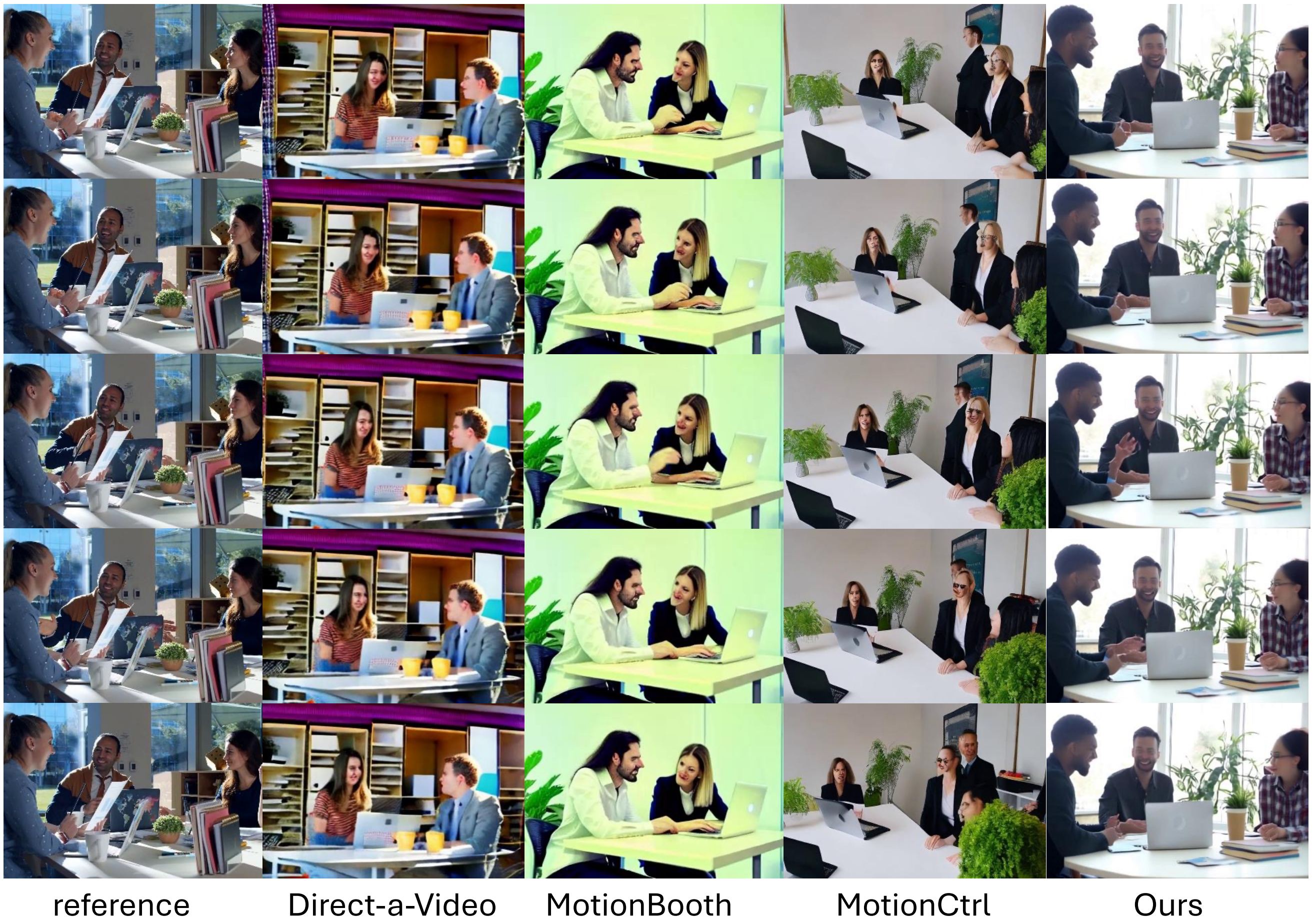


Figure 2. Qualitative comparison results between our approach and Direct-A-Video, MotionCtrl and MotionBooth for joint-control in T2V generation.

	FVD	FID	Pose-Err	DetErr
MagicAnimate	**126.87**	70.93	**11.92**	**0.42**
AnimateAnyone	257.46	88.37	39.09	1.13
Ours	143.78	**69.06**	26.27	0.43

Table 2. Additional human-motion control comparisons with human image animation works under static-camera settings on I2V generation.

5. Limitations and Future Works

While TokenMotion can precisely generate videos that follow the joint control signals, the visual quality is limited by the base model. We identify two representative limitations in Figure 12. First, the model shows difficulties in modeling finger movements. As shown in Figure 12a, while our model precisely generates the trajectory of the moving finger, the static figures lacking explicit motion cues seem unnatural in their positioning. Second, the model struggles with facial detail preservation, as demonstrated in Figure 12b, where facial features appear slightly blurred and exhibit geometric distortions. Future work could explore adapting TokenMotion's joint-control strategy to larger-scale backbones.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. VD3D: Taming Large Video Diffusion Transformers for 3D Camera Control. *arXiv preprint arXiv.2407.12781*, 2024. 1

[2] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 1

[3] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yao-

Figure 3. Qualitative comparison results between our approach and Direct-A-Video, MotionCtrl and MotionBooth for joint-control in T2V generation.

Figure 4. Qualitative comparison results between our approach and ImageConductor in I2V generation.

hui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024. 1

[4] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang,

Figure 5. Qualitative comparison results between our approach and ImageConductor in I2V generation.

Ziyang Yuan, Liangbin Xie, Yuexian Zou, and Ying Shan. Image Conductor: Precision Control for Interactive Video Synthesis. *arXiv preprint arXiv:2406.15339*, 2024. 1

[5] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024. 1

[6] Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, and Dahua Lin. HumanVid: Demystifying Training Data for Camera-controllable Human Image Animation. *arXiv preprint arXiv:2407.17438*, 2024. 1, 10, 11

[7] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers, SIGGRAPH 2024, Denver, CO, USA, 27 July 2024- 1 August 2024*, page 114, 2024. 1

[8] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *arXiv preprint arXiv.2406.17758*, 2024. 1

[9] www.youtube.com/@GlambotOfficial. Glambot official. 1

[10] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. CamCo: Camera-Controllable 3D-Consistent Image-to-Video Generation. *arXiv preprint arXiv.2406.02509*, 2024. 1

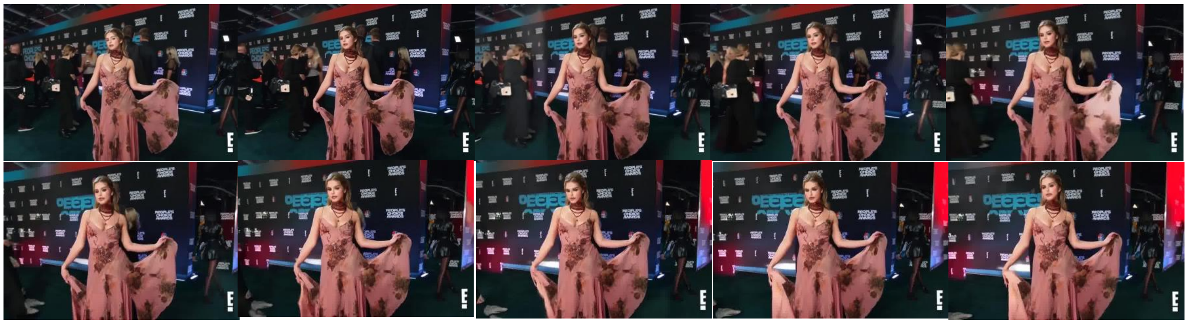
[11] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pages 1481–1490, 2024. 1

[12] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers, SIGGRAPH 2024, Denver, CO, USA, 27 July 2024- 1 August 2024*, page 113, 2024. 1

[13] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 1

[14] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4): 65, 2018. 1, 8, 9

Figure 6. Visualization of different control type on Grammy dataset. The camera motion trajectories are simplified as directions(Tilt Up, Tilt Down, Orbit Right, Orbit Left, Zoom in)



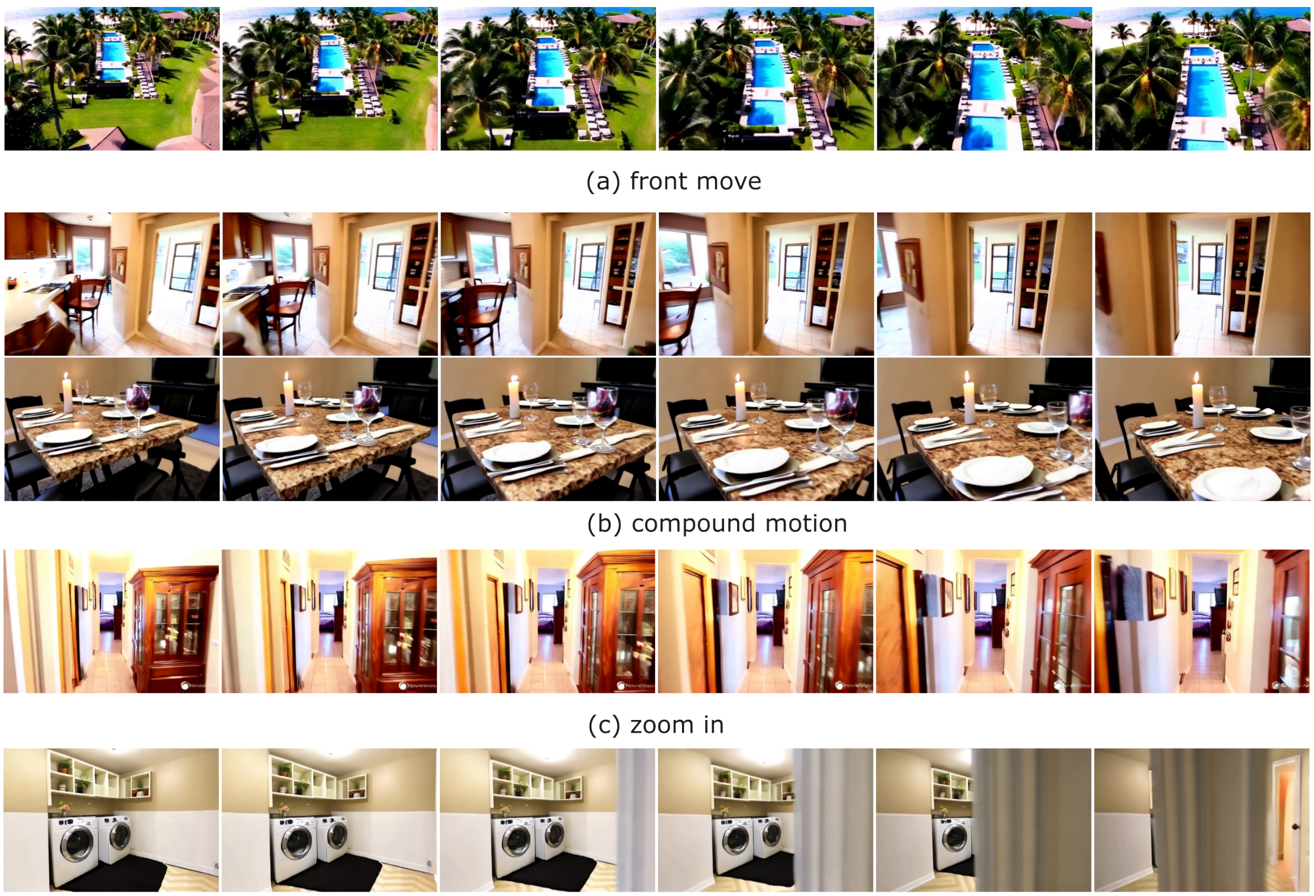
The video depicts a woman standing on a red carpet at an event. She is wearing a long, elegant pink dress with a plunging neckline and floral patterns. The dress has a fitted bodice and a flowing skirt that extends to the floor. She has her hair styled in loose waves and is holding her dress with one hand, revealing the floral design on the fabric.

The video depicts a scene from a fashion event, likely a red carpet or a similar high-profile occasion. A woman is walking down the runway, dressed in a stunning, floor-length gown. The dress is golden and has a textured, possibly floral or paisley pattern. The gown has a fitted bodice and a flowing skirt that flares out slightly at the bottom, creating a dramatic silhouette.

The video is a still from a television show or movie, featuring a person standing on a red carpet. The person is a woman with short, white hair, dressed in a black dress with a sheer overlay. She is smiling and leaning on a railing with her left arm. The background shows a large, white tent with a green hedge in the foreground, and there are people standing and watching the scene. The text "JAMIE LEE CURTIS" is displayed prominently in the bottom left corner of the video.

The video appears to be a still from a television show or movie, featuring a man standing on a balcony. He is dressed in a black leather jacket and has a confident posture. The setting seems to be an outdoor event, possibly a fashion show or a similar high-profile gathering, as indicated by the presence of a large, modern structure in the background with a metal framework and lighting equipment.

Figure 7. Additional visualization of camera control on more Grammy dataset



(a) front move

(b) compound motion

(c) zoom in

(d) zoom out

Figure 8. Visualization of camera control on RealEstate10K[14] dataset

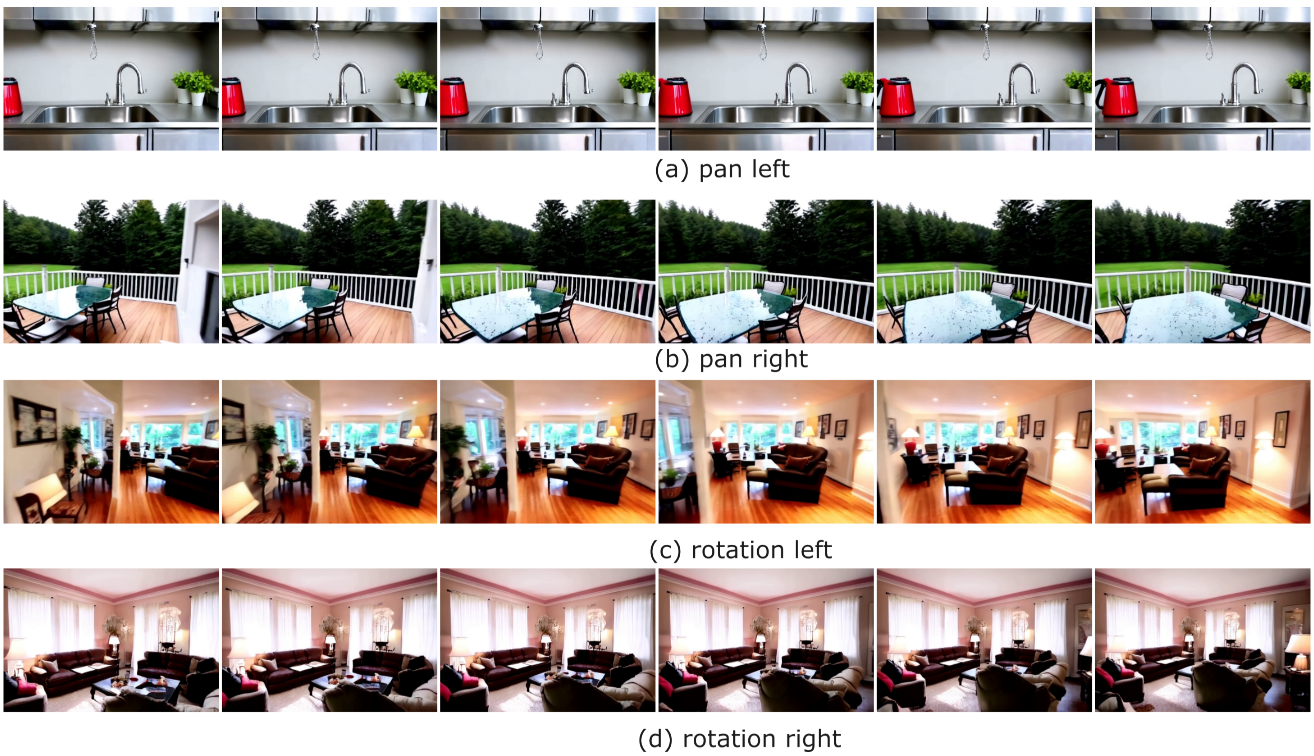


Figure 9. Visualization of camera control on RealEstate10K[14] dataset

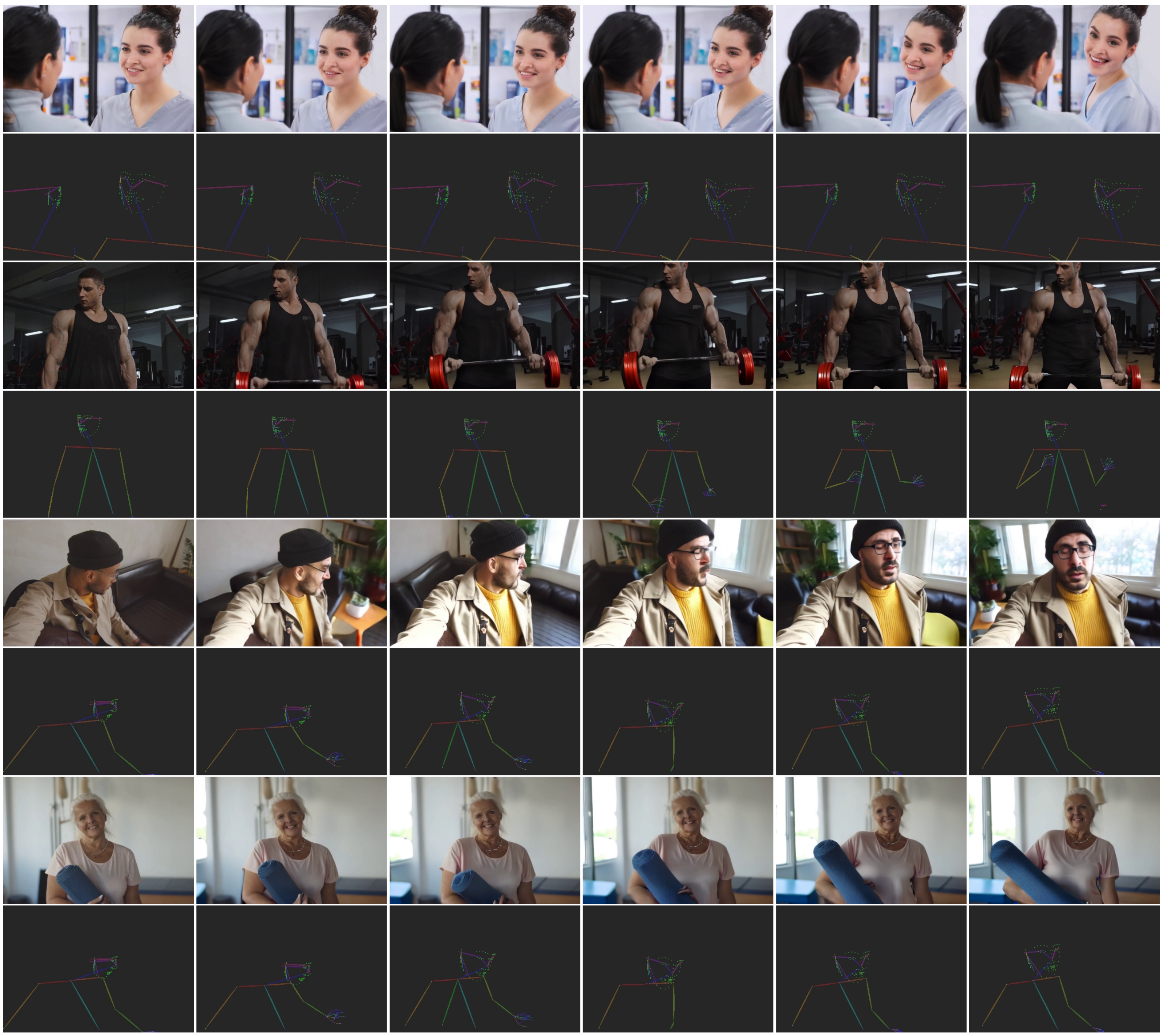


Figure 10. Visualization of camera control on HumanVid[6] datasets

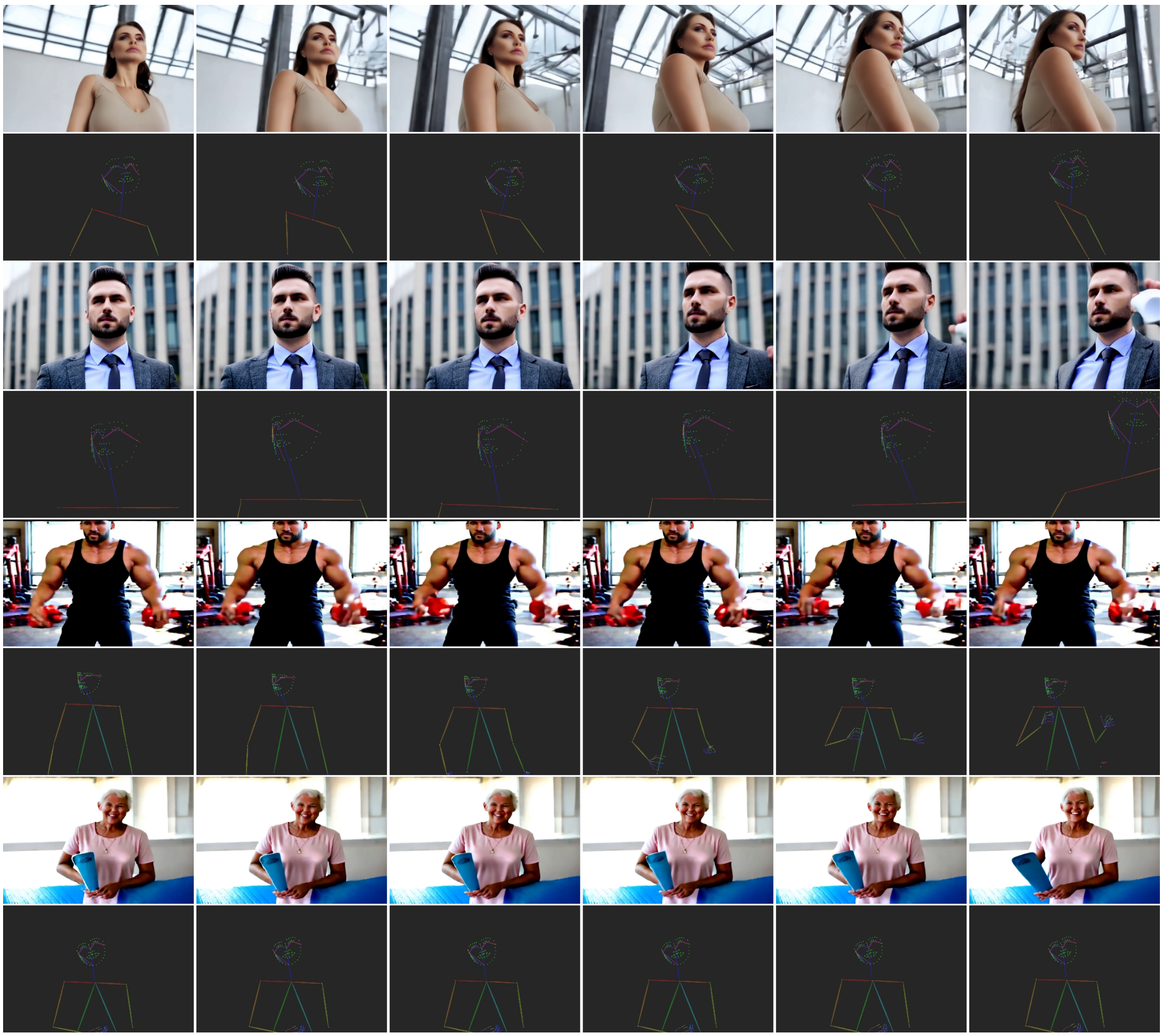
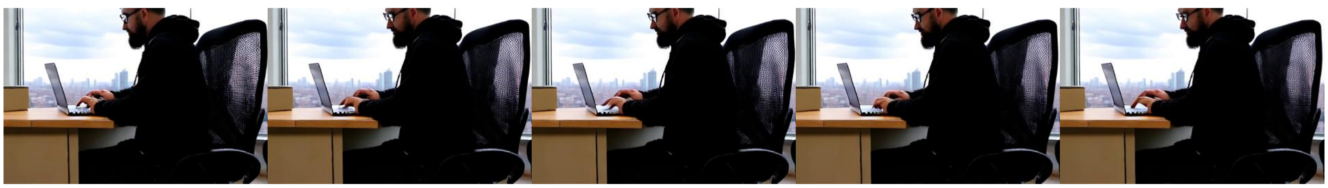


Figure 11. Visualization of camera control on HumanVid[6] datasets



(a) Limitation in finger movements, where some fingers are flickering.

(b) Limitation in facial details, where the man's facial features are vague.

Figure 12. Limitations of TokenMotion