Instructional Video Generation
—
Supplementary Materials

Yayuan Li
yayuanli@umich.edu

Zhi Cao
zhicao@umich.edu

Jason J. Corso
University of Michigan
jjcorso@umich.edu

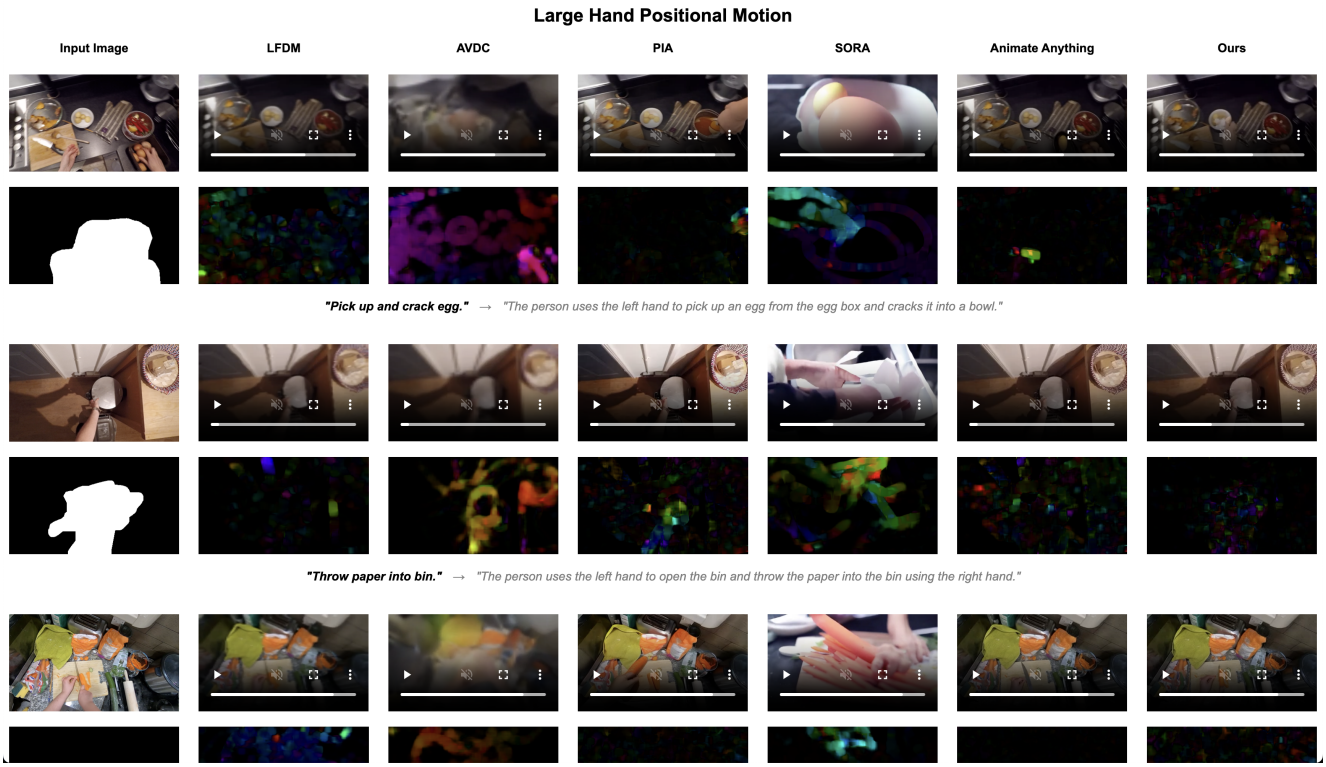

Figure 1. An overview of the HTML page that shows more qualitative results in videos. Please view the full page in attachment baseline_comparison/index1.html

1. More Qualitative Results Shown in Videos

Most importantly, we invite readers to view more qualitative results in videos in the off-line HTML page attached baseline_comparison/index1.html. An illustration of the page is shown in Fig. 1 for readers' reference.

By comparing with all baselines, we emphasize three key capabilities of our method that are accountable for the two key designs. Specifically, we observe good performance on generating instructional videos for actions with (i) large hand positional motion (without hallucination in the background); (ii) object state changing; (iii) subtle dexterous fingertip motion. Our proposed learnable automatic Region of Motion mask generation helps the model to focus on the accurate region for the action ("i" and "ii"), avoiding distraction from the cluttered background commonly seen in instructional videos. Our novel hand structure loss helps the model handle the actions with complex but essential fingertip motion ("iii"), which is rarely seen in previous non-instructional video generation benchmarks.

Dataset	Method	HS-Err. ↓	GT-Frame		GT-Video		Consistency	Semantic	
			FID ↓	CLIP ↑	FVD ↓	EgoVLP ↑	CLIP ↑	CLIP ↑	BLIP_large ↑
EpicKitchens	Open Sora [3]	0.01968	135.34	93.823	124.52	0.187	0.9573	24.46	0.186
	Ours	**0.01512**	**5.27**	**95.904**	**101.89**	**0.377**	**0.9896**	**31.14**	**0.298**
Ego4D	Open Sora [3]	0.02142	141.90	87.160	117.87	0.252	0.9753	24.12	0.172
	Ours	**0.01939**	**21.51**	**96.506**	**103.15**	**0.664**	**0.9873**	**28.63**	**0.263**

Table 1. Comparison our method with a State-of-the-Art Text-to-Video generation method — Open Sora [3]. We show quantitative results on EpicKitchens [1] and Ego4D [2]. Open Sora, representing Text-to-Video methods, is not appropriate for the problem of Instructional Video Generation (IVG) since it fails to generate action demonstration in the specific visual environment. The problem setting of Text-Image-to-Video (TI2V) generation fits more for IVG. Full comparison between our method and SOTA TI2V methods are presented in the main paper.

References

[1] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 2

[2] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[3] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 2