Dexterous World Models

Supplementary Material

A. Implementation Details.

We adopt CogVideoX-Fun-V1.5-5B-InP [1] as our base video diffusion model, generating output simulations at a resolution of 720 × 480 with 49 frames. For model fine-tuning, we apply LoRA [3] to the diffusion transformer with a rank of 64 and $\alpha = 64$. During training, only the added LoRA layers and the image projection layer are optimized, while all other parameters remain frozen. We train the model using the AdamW [4] optimizer with a learning rate of 1×10^{-4} and an effective batch size of 56. Training approximately takes 10 days on 4 NVIDIA A100 GPUs.

B. Additional Qualitative Results

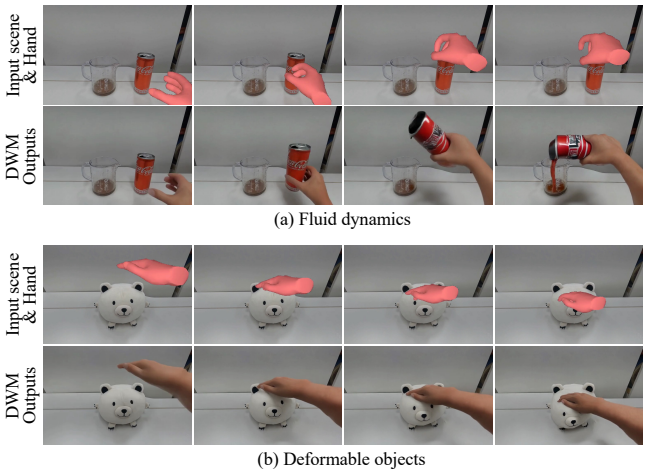
We additionally showcase DWM's ability to capture complex real-world dynamics, including fluid motion and deformable object interactions in Fig. 1. We further present qualitative results in Fig. 6–Fig. 10, along with their corresponding input static scenes and hand actions. As shown in Figs. 6 and 7, DWM demonstrates strong generalization capability, producing realistic and physically consistent interactions in unseen real-world environments with dynamic viewpoint changes, despite the absence of such data during training. Fig. 8 presents additional results on the synthetic dataset under a static camera, while Figs. 9 and 10 show additional results on the real-world dataset under a static camera.

C. Real-World Dataset with Dynamic View

To evaluate DWM under realistic embodied-view conditions, we collect a real-world dataset with dynamic egocentric camera motion. Since existing egocentric datasets do not provide paired static scene videos for dynamic views, we follow the capture protocol introduced in the main paper, using Aria to record accurate camera trajectories and reconstruct a static 3D Gaussian scene from pre-action frames. Rendering this reconstruction along the interaction trajectory produces the paired static–dynamic video pairs required by our task. Our current dataset contains 48 paired samples covering diverse interactions, including pick-and-place, articulated object manipulation, and counterfactual dynamics. Examples are shown in Figs. 4 and 5.

D. Ablation Study on Base Models

As discussed in Sec. 3.2, we adopt the inpainting variant of CogVideo-X [1] as our base model. This choice is motivated by the observation that a pretrained inpainting video diffusion model with a full mask $m = 1$ (all pixels known) behaves as a near-identity operator while retaining a rich generative prior. To validate this hypothesis, we compare our model initialized from the inpainting model with the same architecture initialized from the Image-to-Video (I2V) model [5]. Table 1 reports the DreamSim [2] metric evaluated every 1000 training steps up to 4000 iterations for the two models. The results show that the inpainting-based initialization achieves consistently better DreamSim scores across training iterations. Figure 2 further presents qualitative comparisons. While the I2V-initialized model struggles to properly capture action-conditioned dynamics, our inpainting-initialized model more effectively learns the motion patterns induced by manipulation. These observations provide further empirical support for our hypothesis that full-mask inpainting priors offer a more suitable initialization for residual dynamics learning than I2V-based initialization.



(a) Fluid dynamics

(b) Deformable objects

Figure 1. **Qualitative results for complex real-world dynamics.**

Initialize	DreamSim↓ @ Iterations				
	0	1000	2000	3000	4000
From I2V	0.337	0.240	0.221	0.150	0.103
Ours	**0.110**	**0.166**	**0.098**	**0.103**	**0.088**

Table 1. **Ablation study on base model initialization.** DreamSim comparison between I2V-based and inpainting-based initialization across training iterations.

E. Effect of Text Prompts

Since our model is initialized from a pretrained text-guided video inpainting diffusion model [1], it takes a text prompt as an additional conditioning signal. In Fig. 3, we compare results with and without text prompts on real-world interaction

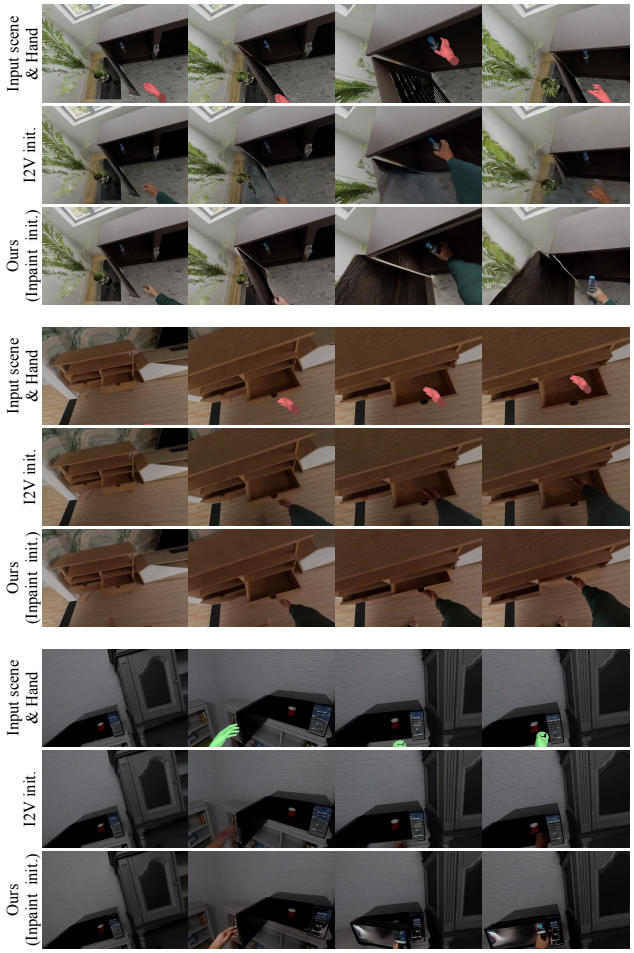


Figure 2. **Qualitative comparison for the base model ablations.** Our inpainting-based initialization consistently achieves better DreamSim scores compared to I2V initialization.

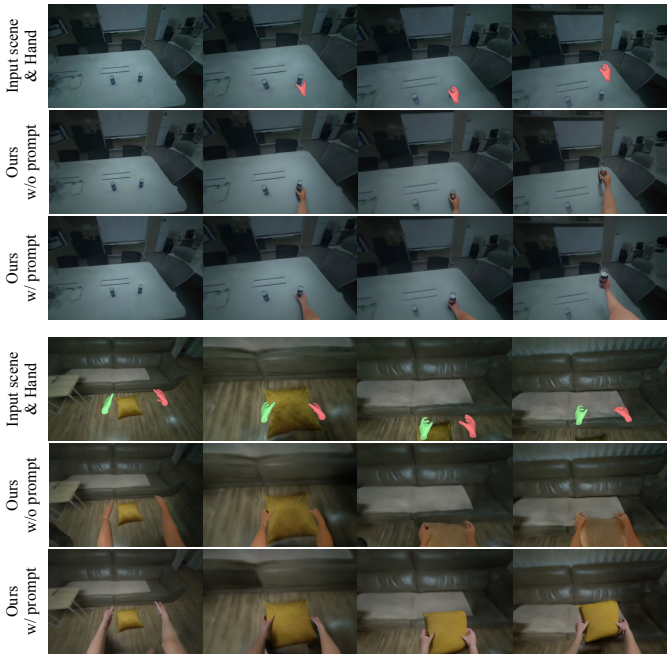


Figure 3. **Qualitative comparison of DWM outputs for real-world sequences with prompt and without prompt.**

sequences. Even without textual input, our model generates object motions induced by the given hand manipulation. However, the absence of text prompts leads to weaker object consistency and degraded motion accuracy. Incorporating text prompts helps stabilize object identity and improves the fidelity of manipulation outcomes, suggesting that textual priors complement our action-driven conditioning by providing high-level semantic guidance.

F. Limitations

While our model opens a new direction for simulating embodied dexterous actions in 3D scenes, several limitations remain. As discussed in Sec. E, our framework still relies on text prompts to achieve the best visual quality, since current video diffusion models heavily depend on textual guidance for semantic grounding. Distilling such semantic priors into purely action-based control without text supervision remains an important future research direction. In addition, the model struggles with non-rigid or highly deformable objects, occasionally failing to maintain object rigidity and visual consistency during complex manipulations. This limitation likely stems from the limited diversity of deformable-object interactions in our training data and could be alleviated with richer and more varied manipulation datasets. Moreover, our current pipeline for constructing the real-world dataset with dynamic views lacks scalability. While it is sufficient for use as an evaluation set, scaling it to a size suitable for training remains challenging. Data augmentation may help increase variability, but a truly scalable solution would require automating both acquisition and reconstruction to reduce dependence on manual scene exploration and preprocessing. Finally, our model does not explicitly reason about 3D structure, depth, or physical contact. This restricts its ability to enforce strict physical constraints, which is particularly important for downstream applications such as robotic policy learning. Future work includes incorporating depth-aware priors, modeling contact dynamics more explicitly, and using DWM as a differentiable simulator for action optimization and policy learning.

References

[1] aigc apps, 2024. https://https://github.com/aigc-apps/VideoX-Fun. 1

[2] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *NeurIPS*, 2023. 1

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Proc. ICLR*, 2022. 1

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *Proc. ICLR*, 2019. 1

[5] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *Proc. ICLR*, 2025. 1

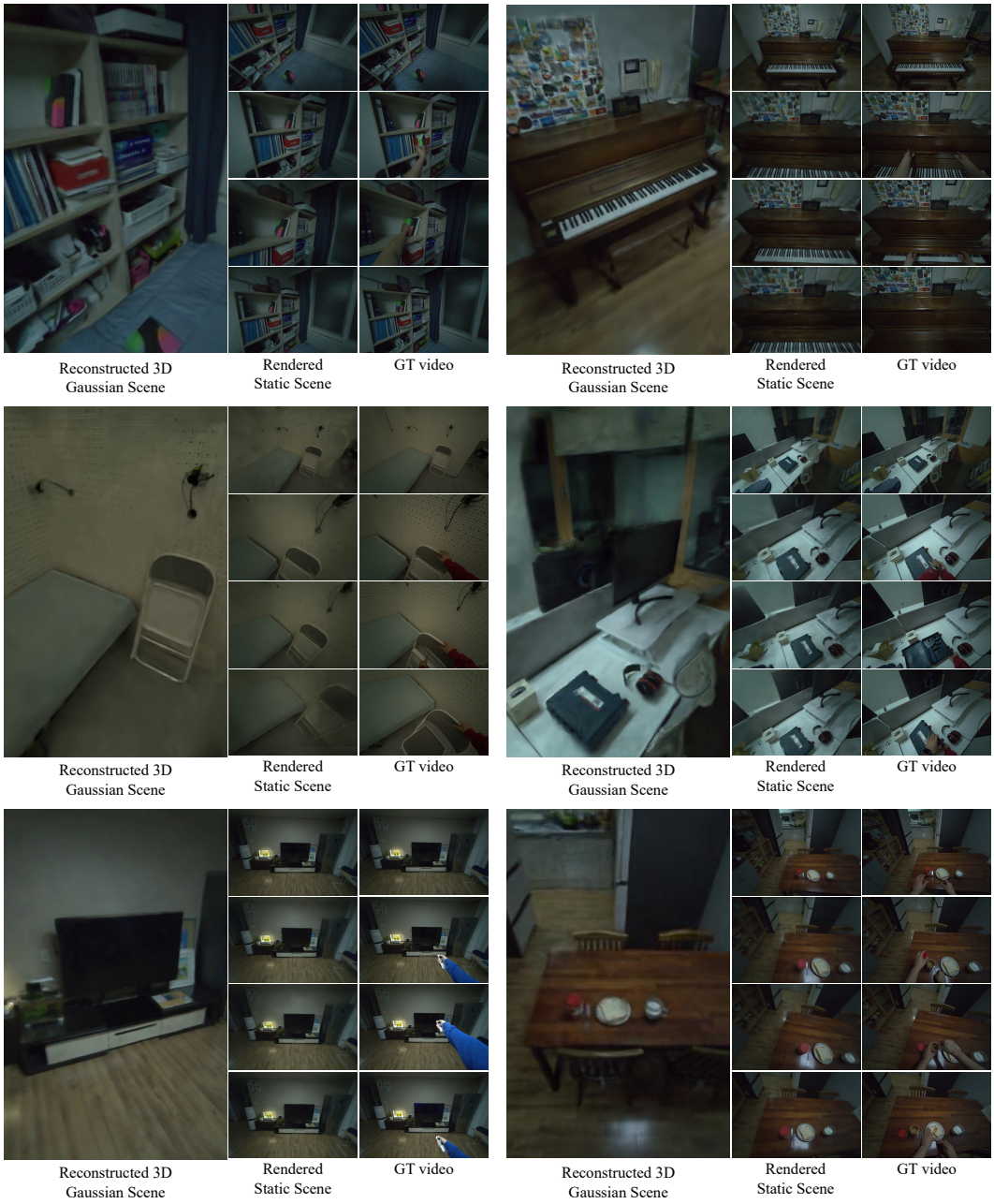


Figure 4. **Examples of paired data of our real-world dataset with dynamic view.** We collect paired sets of static scene video and GT interaction video with our custom data capture protocol using Aria Glasses.

Figure 5. **Examples of paired data of our real-world dataset with dynamic view.** We collect paired sets of static scene video and GT interaction video with our custom data capture protocol using Aria Glasses.

Figure 6. **Additional qualitative results on real-world scene with dynamic view.** We show the input static scene video with hand action in the first row of each pair. The resulting egocentric simulation generated via DWM is demonstrated in the second row.

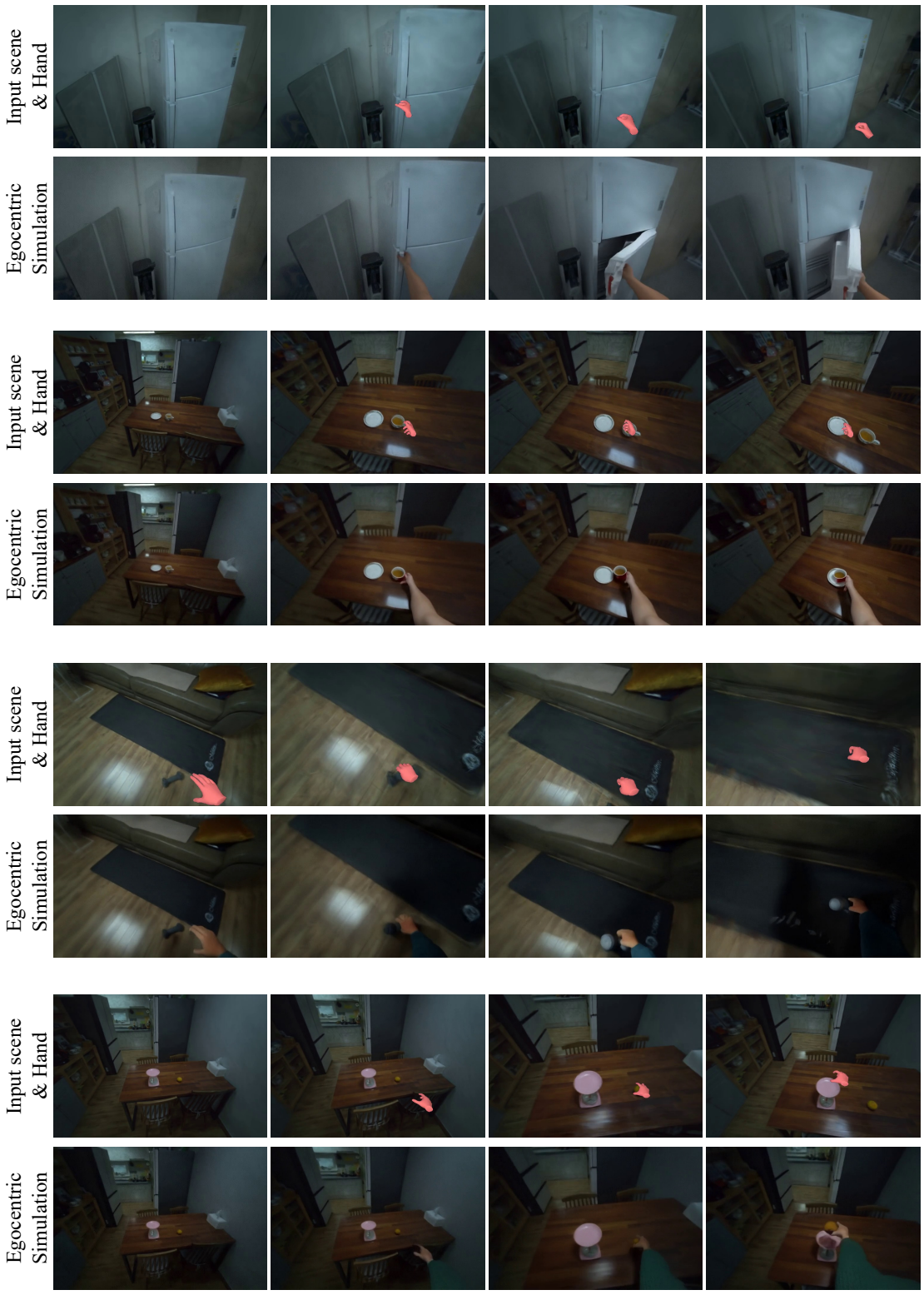


Figure 7. **Additional qualitative results on real-world scene with dynamic view.** We show the input static scene video with hand action in the first row of each pair. The resulting egocentric simulation generated via DWM is demonstrated in the second row.

Figure 8. **Additional qualitative results on synthetic scene with dynamic view.** We show the input static scene video with hand action in the first row of each pair. The resulting egocentric simulation generated via DWM is demonstrated in the second row.

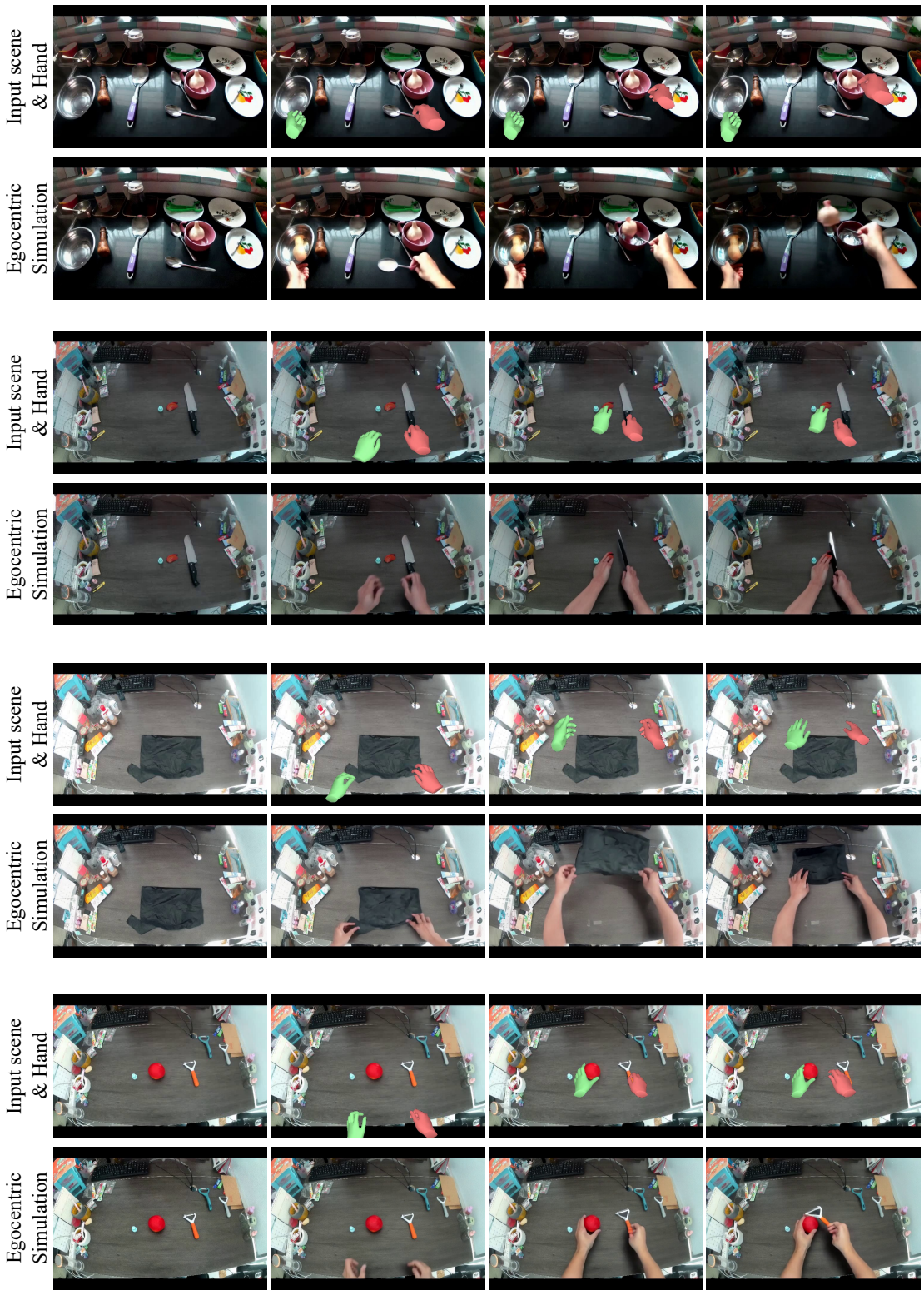


Figure 9. **Additional qualitative results on real-world scene with static view.** We show the input static scene video with hand action in the first row of each pair. The resulting egocentric simulation generated via DWM is demonstrated in the second row.

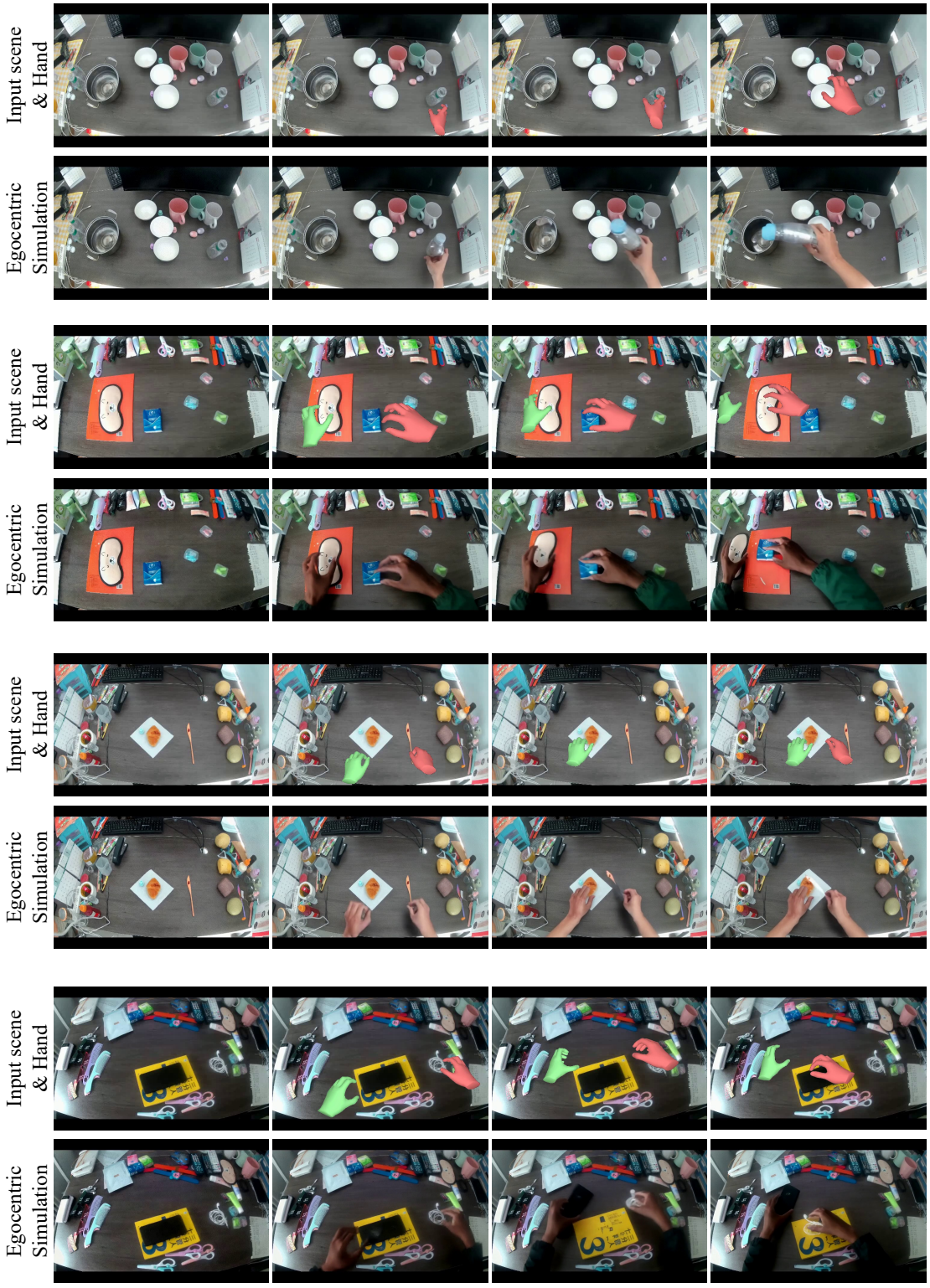


Figure 10. **Additional qualitative results on real-world scene with static view.** We show the input static scene video with hand action in the first row of each pair. The resulting egocentric simulation generated via DWM is demonstrated in the second row.