VideoWeaver: Multimodal Multi-View Video-to-Video Transfer for Embodied Agents

Supplementary Material

This supplementary is structured in the following way. In Sec. 6, we show visualizations of pointclouds from Pi3. Finally, in Sec. 9, we show a sample of generated videos from our model. For a more extensive visualizations, please refer to the video supplementary.

6. 4D Pointcloud from Pi3

We visualize several estimated point clouds obtained from multi-view videos in Figure 6. Leveraging its permutation-invariant design, Pi3 can reconstruct the underlying scene in 4D. To do so, we sample the same number of frames from each viewpoint and provide all frames to Pi3 simultaneously, without requiring any specific ordering.

7. Generalization to Unseen Datasets

We show out-of-distribution results on a single-view simulator dataset and two multi-view real datasets (Berkeley-Autolab and Robomind) in Figure 7. Berkeley-Autolab has only two views. The three datasets present different tasks, embodiments (robot types), camera poses and texture (synthetic data included) unseen during training. VideoWeaver exhibits strong generalization ability across these scenarios. However, failure cases still happen and include mostly cluttered scenes or small objects in normal environments.

8. Text Prompts

We re-caption all videos using Qwen-2.5-VL [73]. For multi-view videos, we select the viewpoint that provides the clearest and most comprehensive visibility of the environment. The same caption strategy is applied to all videos: we first caption the action performed by the robot "[Action] A white and black robotic arm is holding a small container and pouring liquid into a bright red bowl on the table. [Description] The robotic arm has a white exterior with black joints and is positioned over a dark brown table. The bowl is bright red and sits near the center of the table. Nearby, there are small objects, including a light-colored cylindrical item and a thin tool. To the left, there is a light wooden box next to a black cabinet or machine. In the background, there are desks, chairs, and various equipment in black, gray, and metallic colors, along with visible cables and lab tools." For style optimization, we keep the same caption structure and only change how the objects look like. Since we only train on real data, artistic styles are not supported. When the caption is too long, or the scene is too cluttered, the prompt following is weakened.

9. Additional Qualitative Results

We show more qualitative results of the multi-view model with and without the 4D pointcloud injection from Pi3 in Figure 8. Moreover, in Figure 9, we show the ability of our model to generate diverse styles of the same multi-view video only by varying the text prompt (depth, sketch and pointcloud remain the same). VideoWeaver can perform localized multi-view edits in challenging viewpoints. More extensive viusalizationsare shown in the video supplementary.

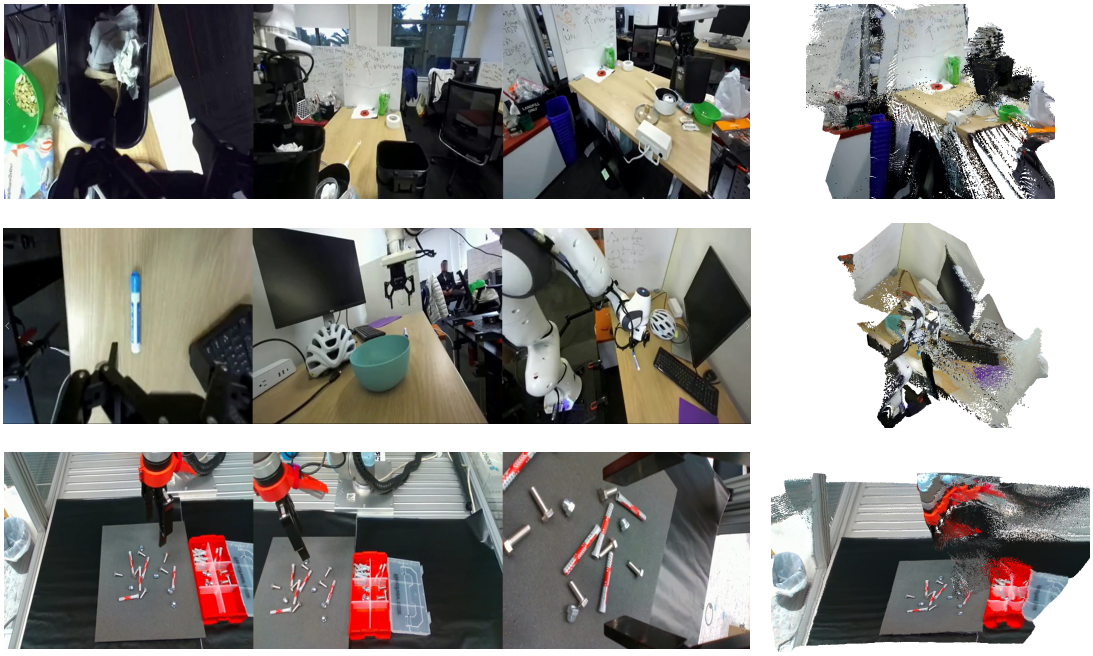

Figure 6. We show several visualizations of the estimated 4D pointcloud from Pi3 and some of their corresponding RGB frames.

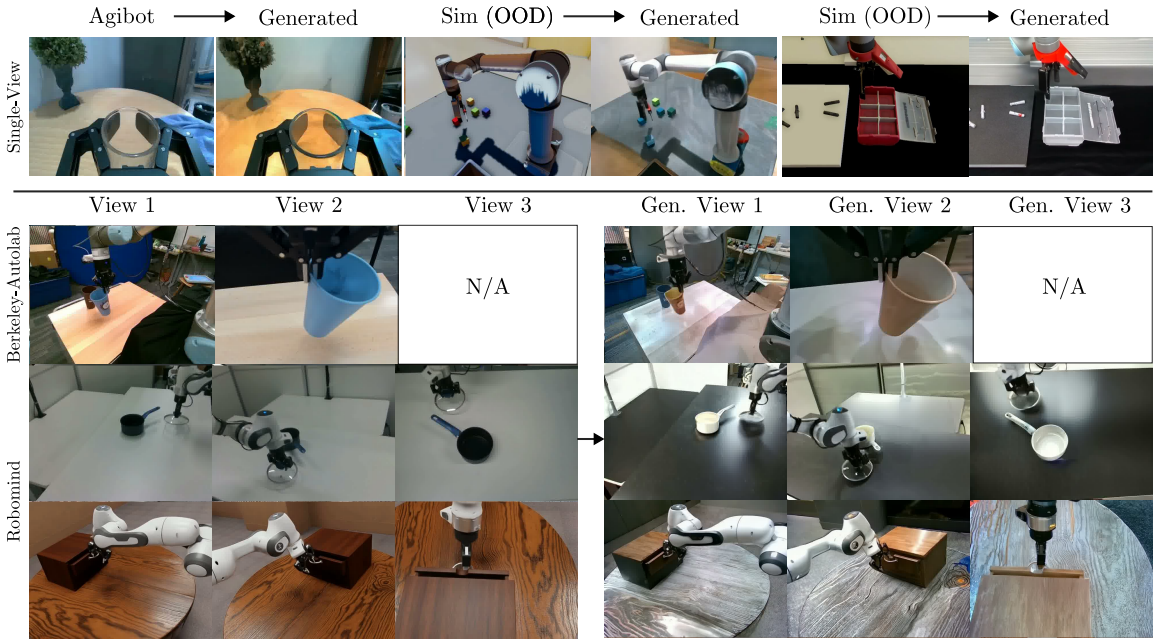

Figure 7. Inference on novel datasets, Berkeley, Robomind and an internal simulator data with reflective and transparent surfaces.

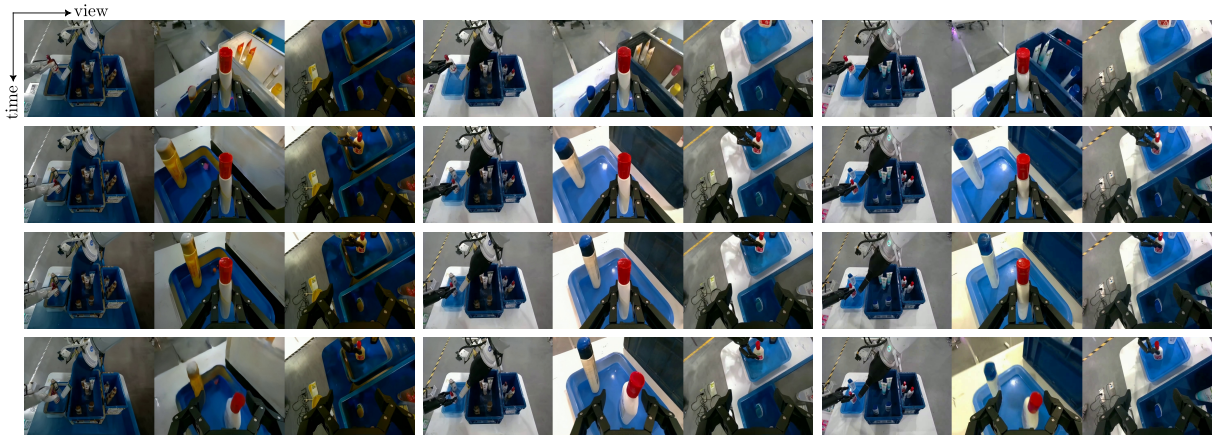

Figure 8. Additional qualitative comparisons between our full multi-view model and its ablated counterparts.

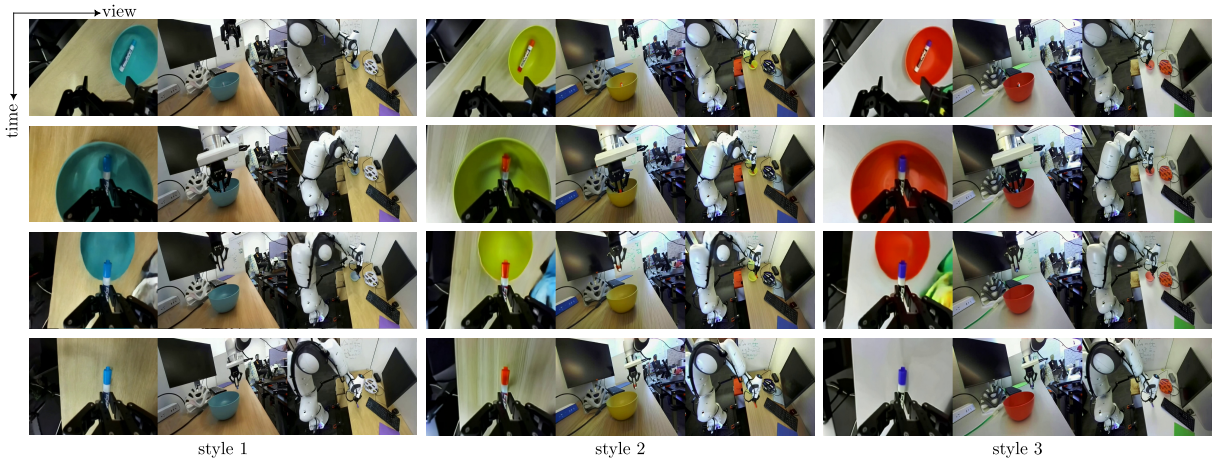

Figure 9. Additional qualitative results generated from our multi-view model by varying the prompt.