MonoMobility: Zero-Shot 3D Mobility Analysis from Monocular Videos

Supplementary Material

Summary

The supplementary material consists of the following four parts: (1) A detailed description of our method; see Section 1. (2) Limitation analysis and failure case; see Section 2. (3) A detailed introduction to the dataset, including both the synthetically simulated scenarios and the real-world captured scenarios; see Section 3. (4) Additional results, including quantitative and qualitative comparison experimental results with state-of-the-arts methods [3, 5, 6], as well as more qualitative experimental results of our method; see Section 4.

1. Details of Method

Correspondences between N point clouds are required both for point cloud registration in Motion Attributes Initialization and for motion loss computation in Optimization. Therefore, we provide a more detailed description of how to obtain correspondences:

(1) Optical flow estimation to obtain pixel-level correspondences between frames; (2) All estimated flows are then propagated to establish consistent mappings between each of the N frames and the reference (first) frame; (3) Depth values from corresponding pixels (using per-frame depth maps) are converted to discrete point clouds, yielding N mutually corresponding point clouds for motion estimation. This simple and efficient strategy ensures consistent inter-frame correspondences for motion estimation.

2. Limitation analysis & failure case

As noted in conclusion, our method relies on depth and optical flow estimation. While these components are typically robust, they may fail in extreme cases, e.g., textureless regions and highly dynamic scenes cause unstable optical flow estimation (Fig. 1), leading to suboptimal parsing.

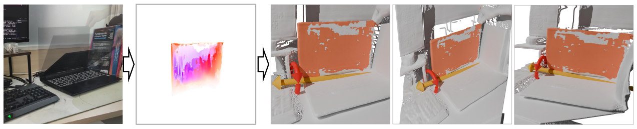

Figure 1. Suboptimal parsing result in extreme case.

3. Details of Dataset

Our goal is to analyze motion parts and their motion attributes from monocular videos. For effective evaluation of our algorithm, we have constructed a Motion Parsing Dataset that primarily comprises virtual simulation and real-world scenarios. It includes various common articulated object categories such as drawers, wardrobes, laptops, staplers, and liftchairs, which cover three main types of articulated motion: translation, rotation, and rotation+translation. Some scenes contain multiple motion parts with different motion types, aimed at verifying the effectiveness of relevant algorithms in solving complex tasks. The statistical details of the dataset are presented in Tab.1.

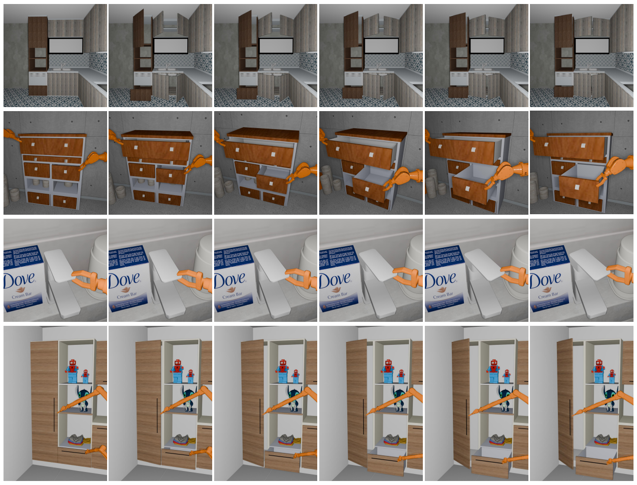

Figure 2. Virtual simulation scenarios. Each row represents a sequence of video frames captured in a virtual scene.

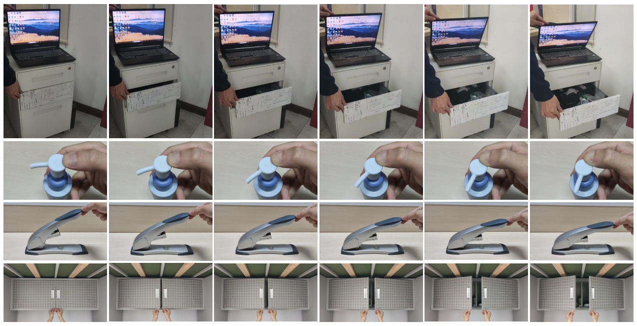

Figure 3. Real-world scenarios. Each row represents a sequence of video frames captured in a real-world scene.

3.1. Virtual Simulation Scenarios

For virtual simulation scenarios, we first collect 3D models from 3D Warehouse [2] and annotate the motion parts and motion attributes using Blender [1]. Subsequently, robots are introduced into the scenes to simulate the interaction operations of articulated object. Finally, the constructed

Data Type	Number of Scenes	Number of Motion Parts	Distribution of Motion Types	Object Categories
Virtual	15	18	10×rotation	Fridge(3), Door(1), Cupboard(4), Faucet(1), Laptop(1)
			7×translation	Drawer(6), Flatdoor(1)
			1×rotation-translation	Liftchair(1)
Real-world	11	13	8×rotation	Cupboard(3), Laptop(3), Wrench(1), Stapler(1)
			4×translation	Drawer(4)
			1×rotation-translation	Pumpbottle(1)

Table 1. The statistical of the dataset. ’10×rotation’ indicates that the rotation motion type contains 10 motion parts, ’Fridge(3)’ indicates that the articulated object category ‘’Fridge’ contains 3 motion parts, and similar symbols apply accordingly.

		rotation					translation		rotation-translation	
Metrics	Methods	Fridge*3	Door	Cupboard*4	Faucet	Laptop	Drawer*6	Flatdoor	Liftchair	Mean
AE(°)↓	Shape of Motion	2.995	**1.089**	3.841	8.785	9.001	12.525	**0.245**	2.363	6.721
	Ours	**1.077**	1.129	**1.341**	**0.298**	**1.681**	**1.183**	0.280	**1.506**	**1.262**
PE(cm)↓	Shape of Motion	8.481	**7.866**	15.043	1.752	25.665	-	-	9.863	11.887
	Ours	**4.938**	11.225	**5.547**	**0.057**	**2.889**	-	-	**2.093**	**4.843**

Table 2. Comparison with Shape of Motion. Our method outperforms Shape of Motion in the motion axis prediction.

dynamic scenes are rendered using Blender [1], capturing motion videos sequences. To quantitatively analyze the results of the motion parts and motion attributes parsing, we also captured 3D point clouds with annotations of motion parts and motion attributes. A total of 15 virtual motion simulation scenarios were built, some examples are shown in Fig.2.

3.2. Real-world scenarios

For real-world scenarios, we use the camera directly capture motion videos of articulated objects. The intrinsic parameters of camera are calibrated using COLMAP [4]. Unlike virtual simulation scenarios, we cannot obtain geometric and annotation information for real-world scenarios. Therefore, we only perform qualitative analysis on the real-world data. A total of 11 real-world scenarios were constructed, some examples are shown in Fig.3.

4. Additional Experimental Results

To further substantiate the superiority and effectiveness of our method, we also conducted a quantitative comparison with Shape of Motion [6], a dynamic scene reconstruction method, as shown in Tab.2. Moreover, we conducted qualitative comparisons in motion parameter prediction with the state-of-the-arts algorithms (PARIS-scene, PARIS-obj [3], DGMarbles* [5], Shape of Motion [6] and ours(w/o optim)) on additional data, as shown in Fig. 5 and Fig. 4; as well as more qualitative analysis results of motion parts and motion attributes of our method from more video sequences, detailed in Fig. 6–19, where each figure includes the analysis results of two scenes. For each scene, the first row represents the input (video sequence) to the algorithm, while the second to fourth rows show the continuous motion visualization of motion parts based on motion attributes from different viewpoints.

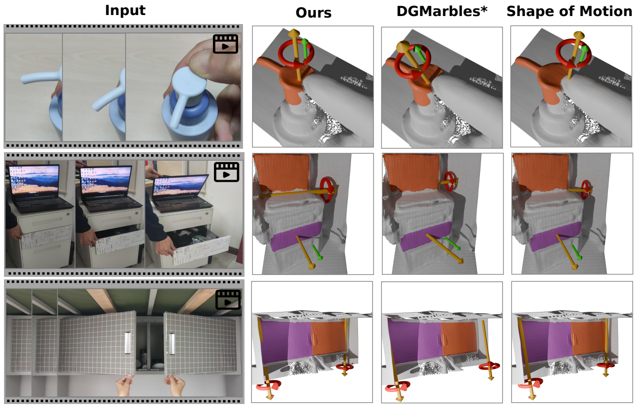

Figure 4. Comparison in real-world scenarios. All the motion parts segmentation results are from ours, with a focus on evaluating the motion axis predictions. The comparison clearly shows that our method outperforms other methods.

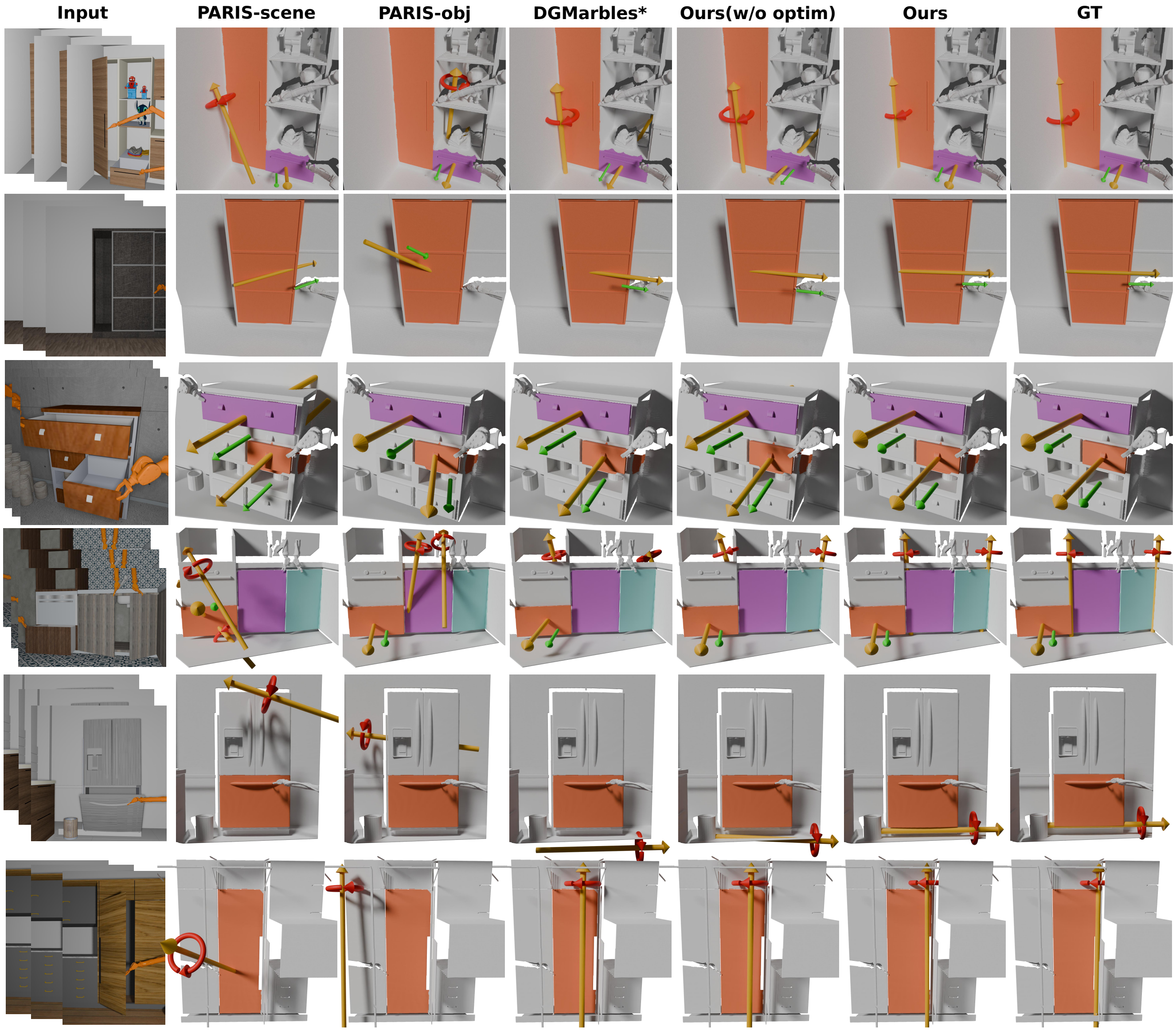

Figure 5. More qualitative comparison with state-of-the-arts methods (PARIS-scene, PARIS-obj, DGMarbles*) and Ours(w/o optim).

Real-world Data Results

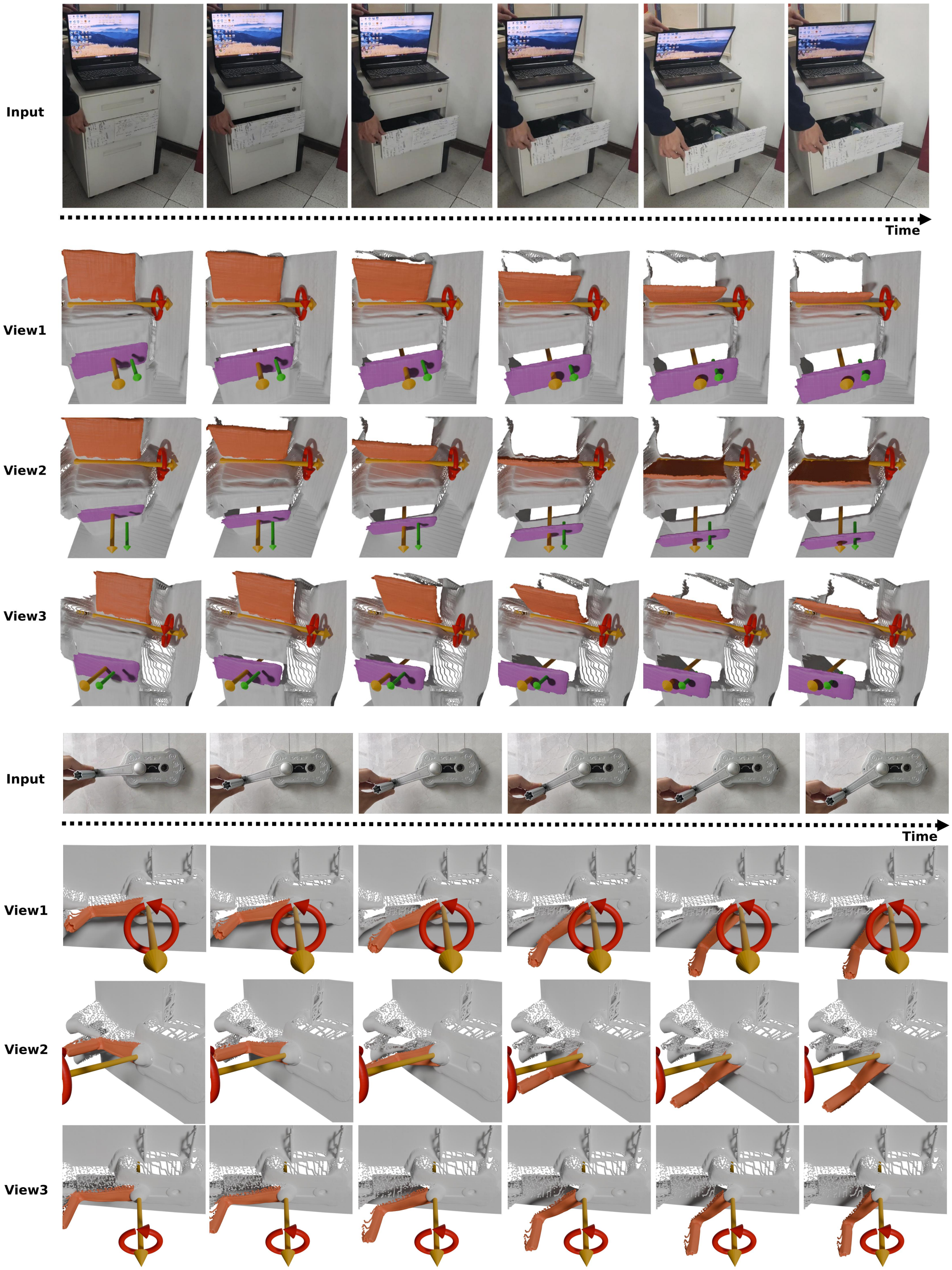

Figure 6. Analysis results of motion parts and their motion attributes for real-world scenes 1-2.

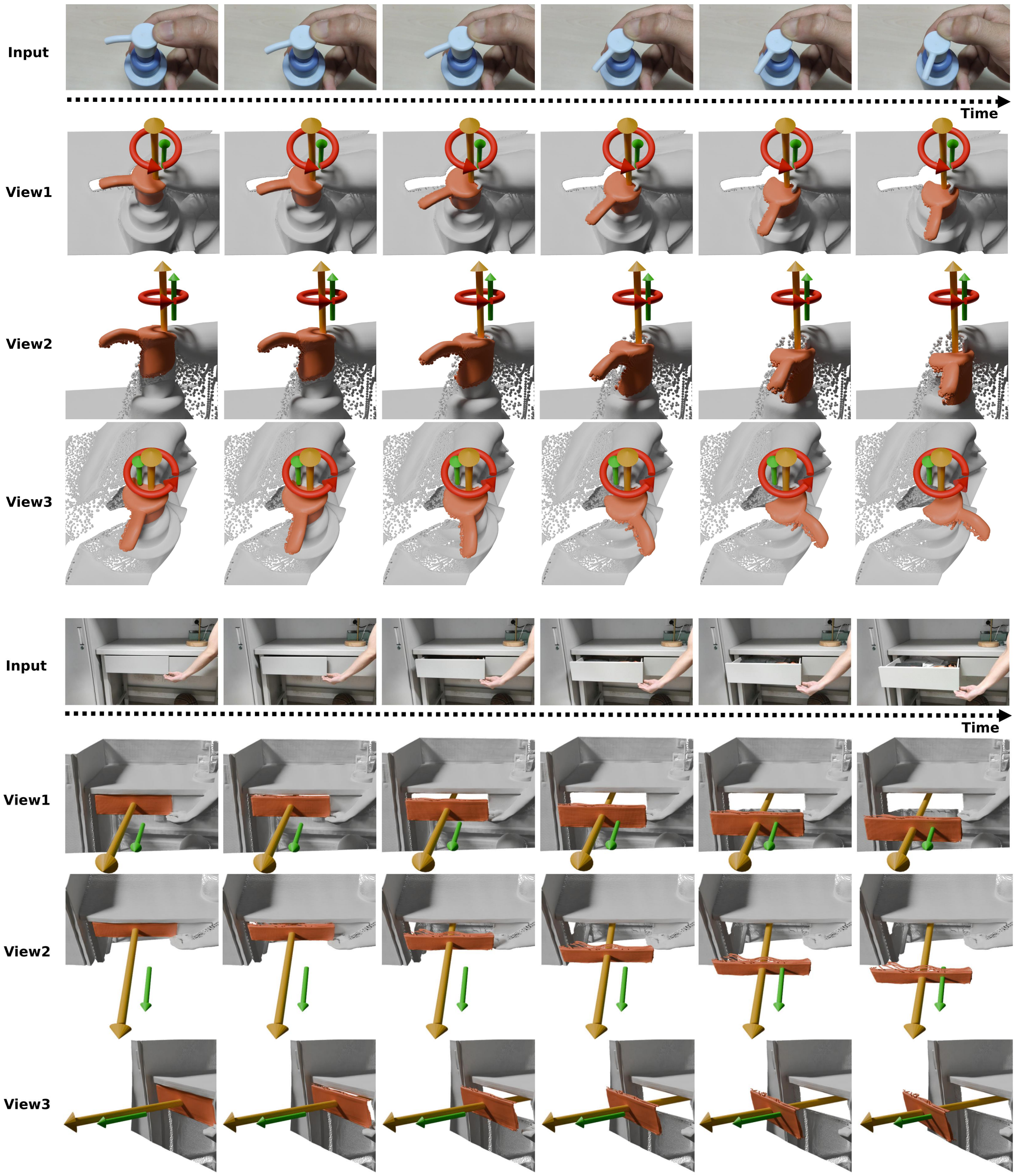

Figure 7. Analysis results of motion parts and their motion attributes for real-world scenes 3-4.

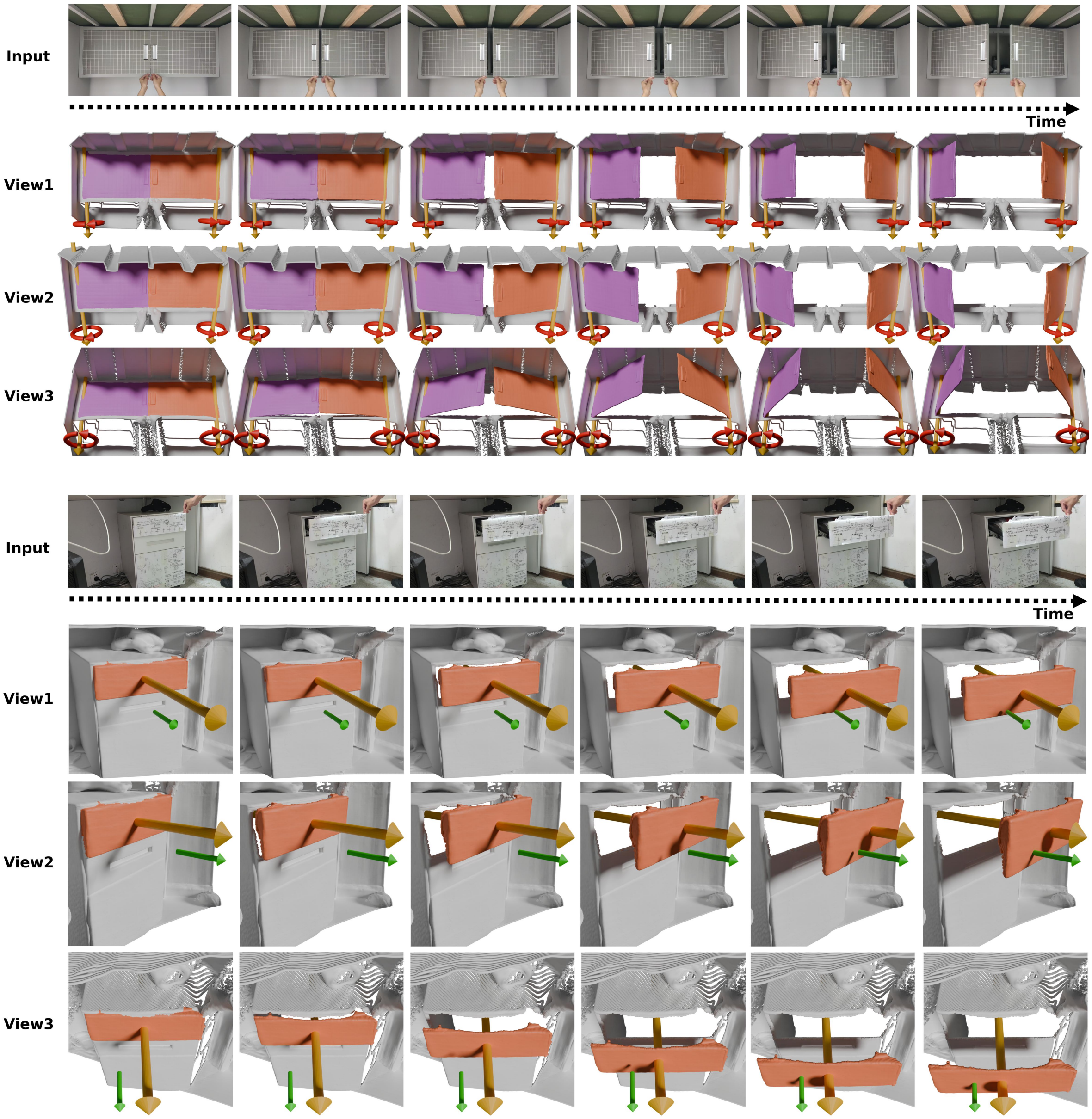

Figure 8. Analysis results of motion parts and their motion attributes for real-world scenes 5-6.

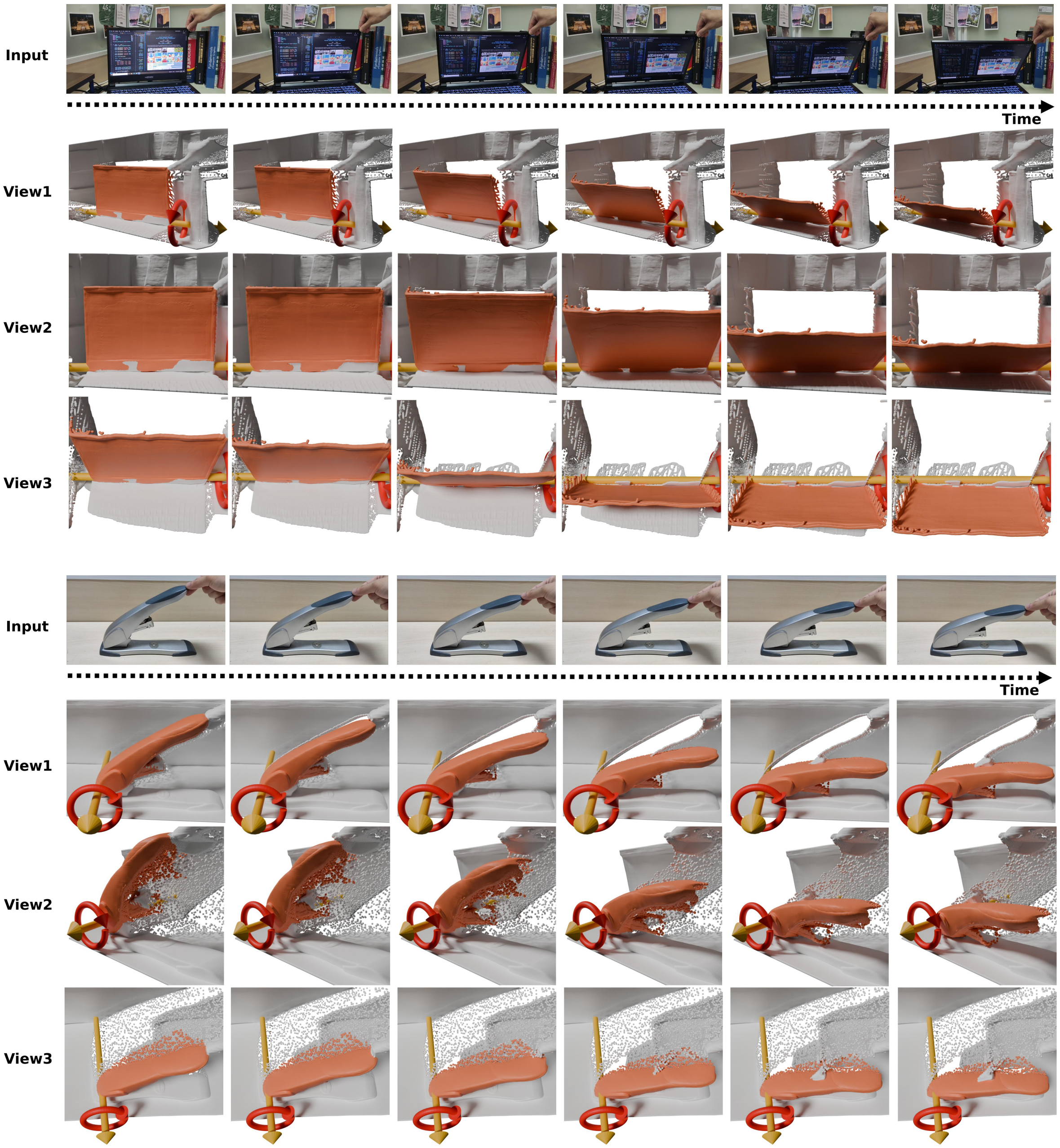

Figure 9. Analysis results of motion parts and their motion attributes for real-world scenes 7-8.

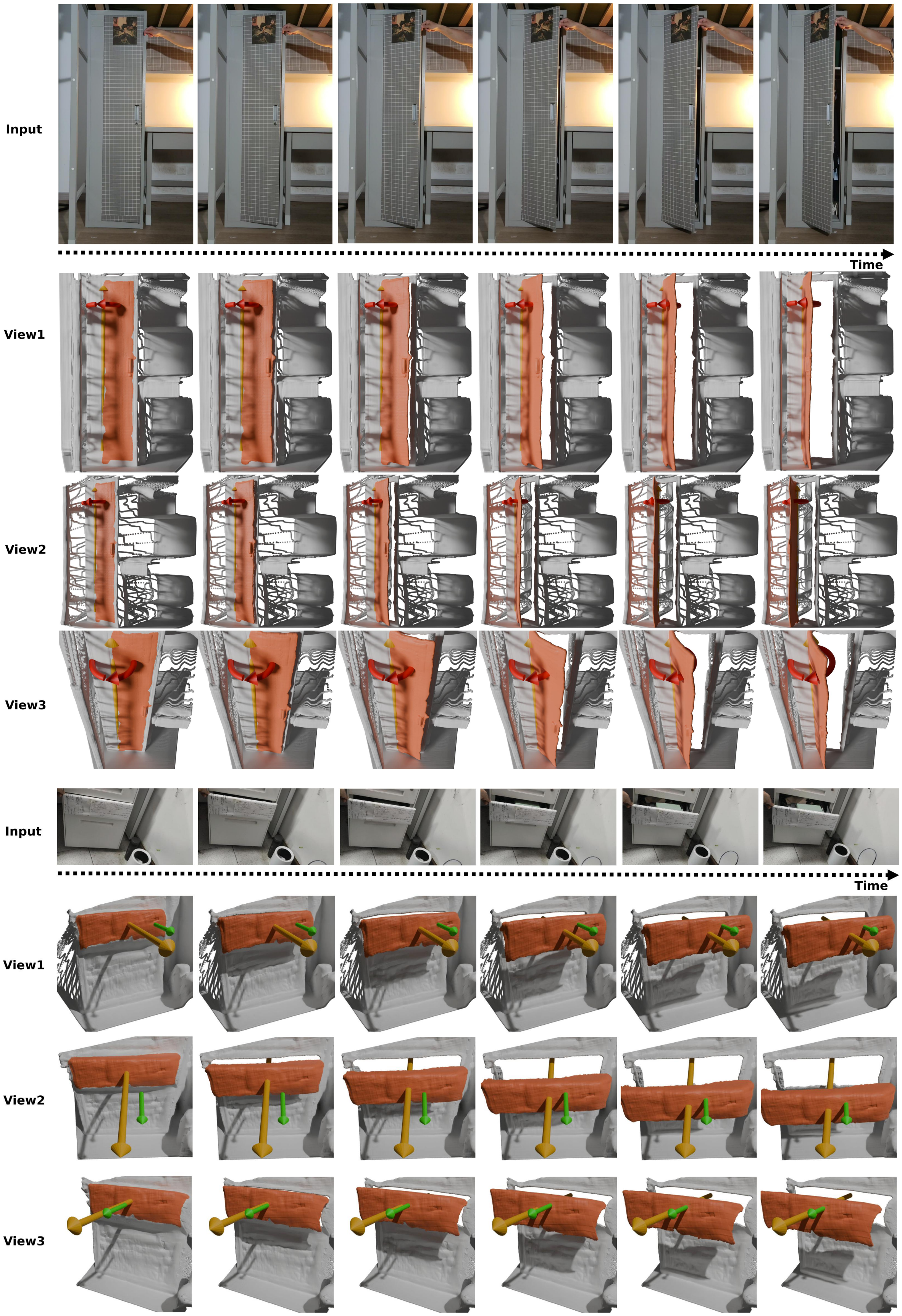

Figure 10. Analysis results of motion parts and their motion attributes for real-world scenes 9-10.

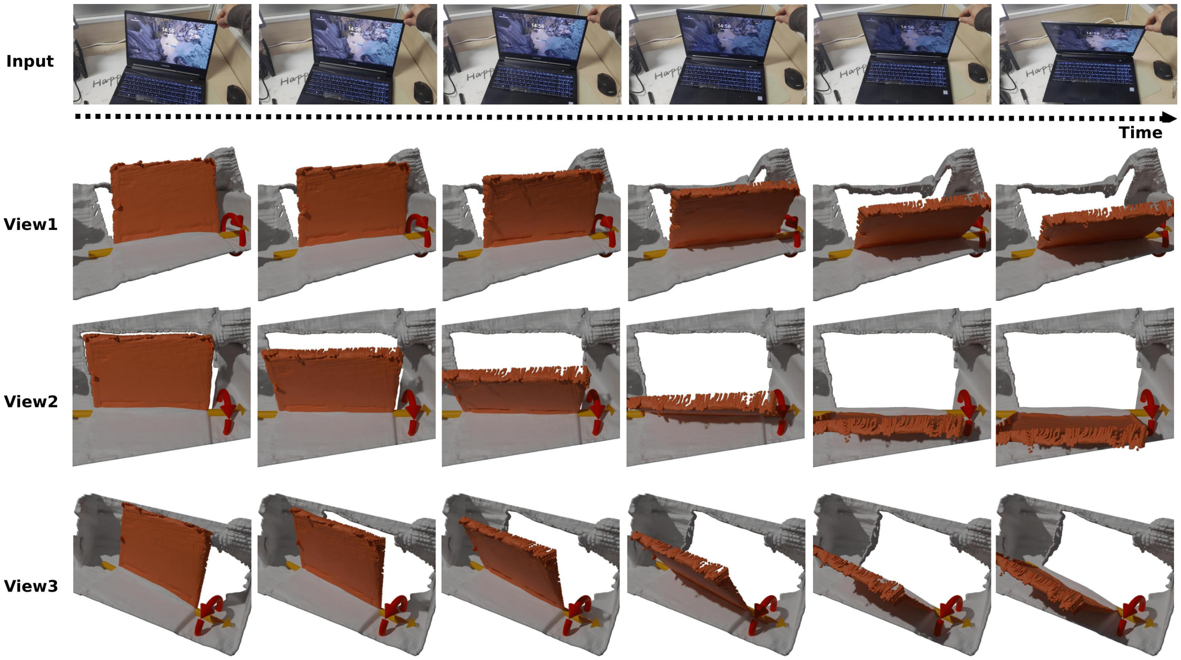

Figure 11. Analysis results of motion parts and their motion attributes for real-world scene 11.

Virtual Data Results

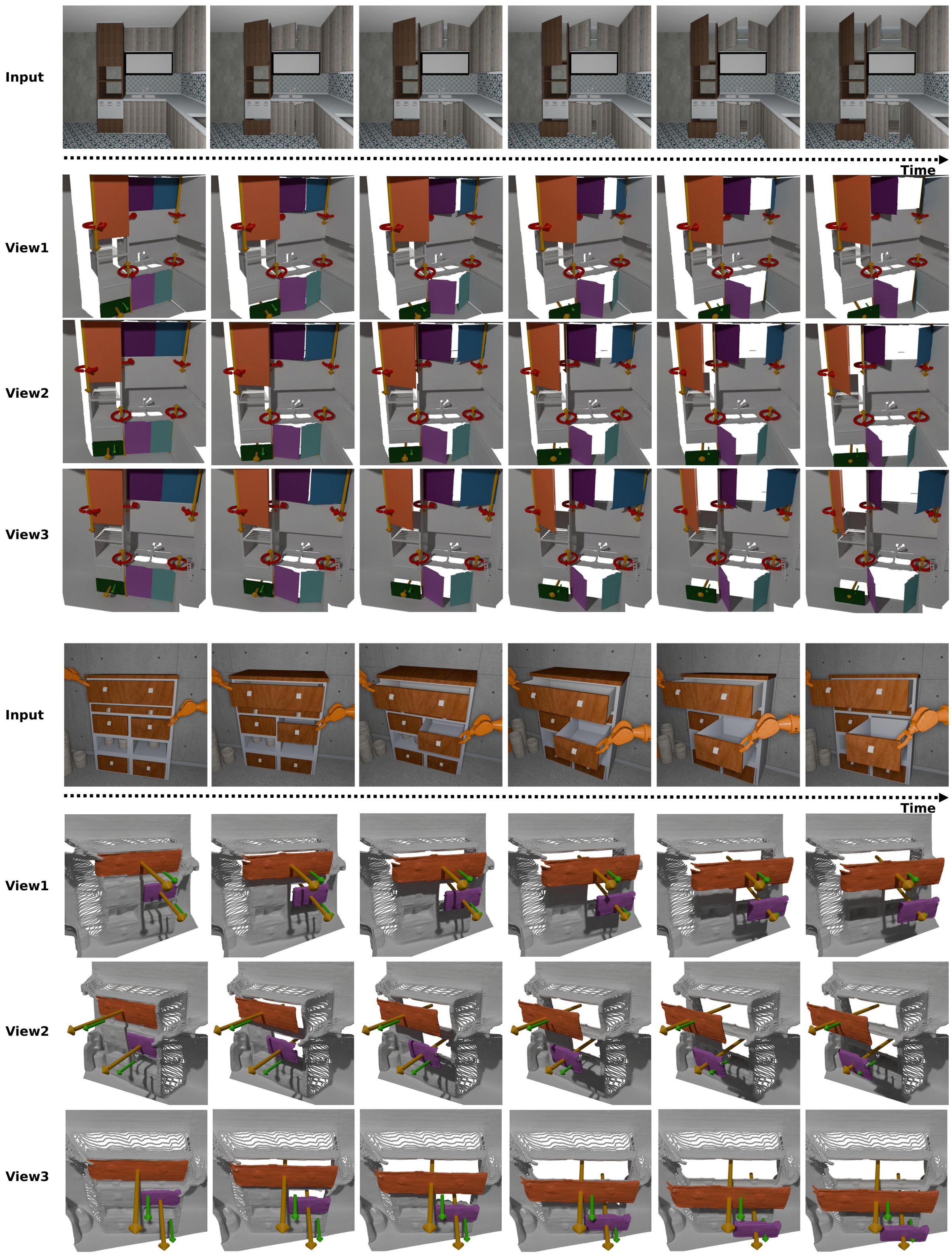

Figure 12. Analysis results of motion parts and their motion attributes for virtual simulation scenes 1-2.

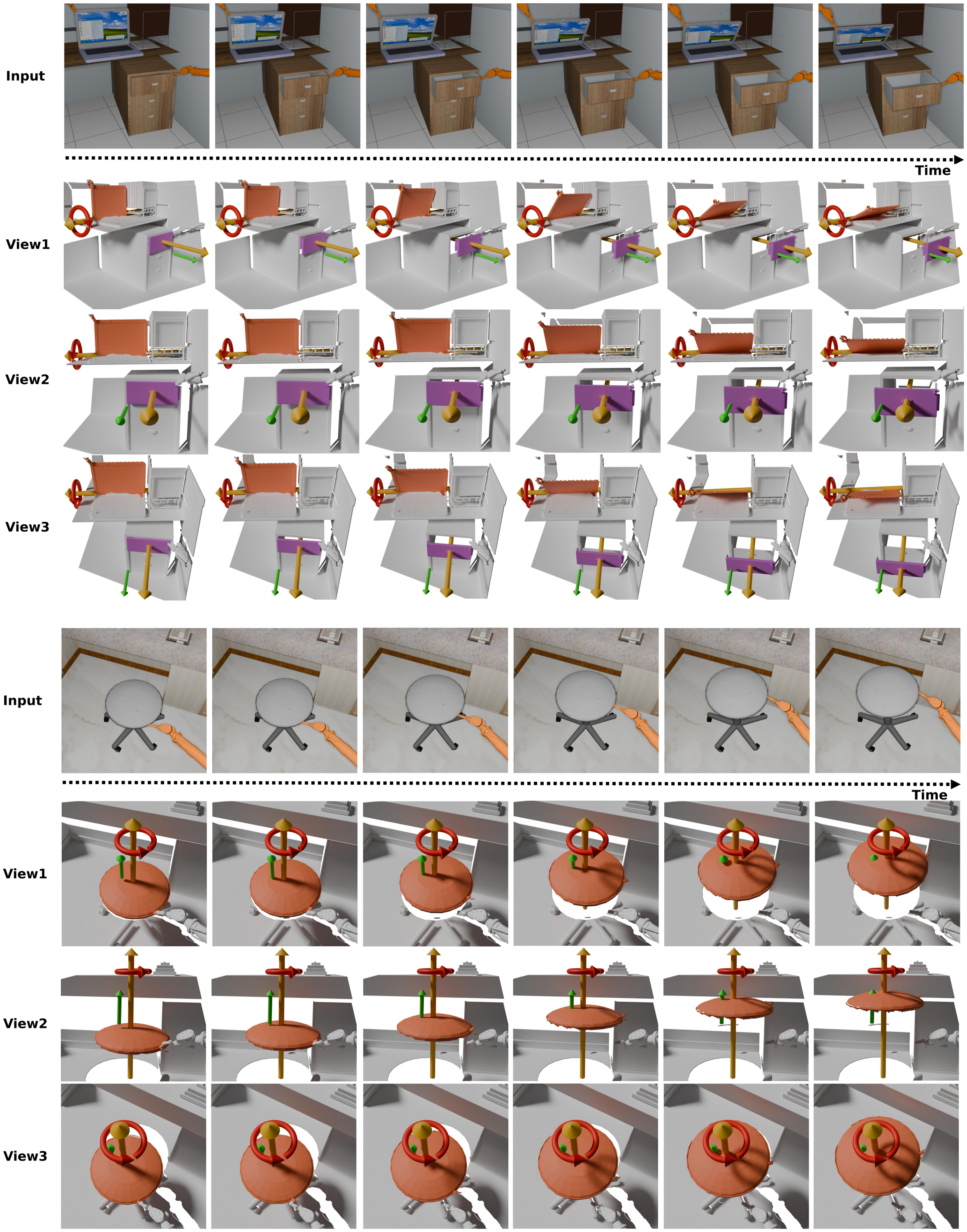

Figure 13. Analysis results of motion parts and their motion attributes for virtual simulation scenes 3-4.

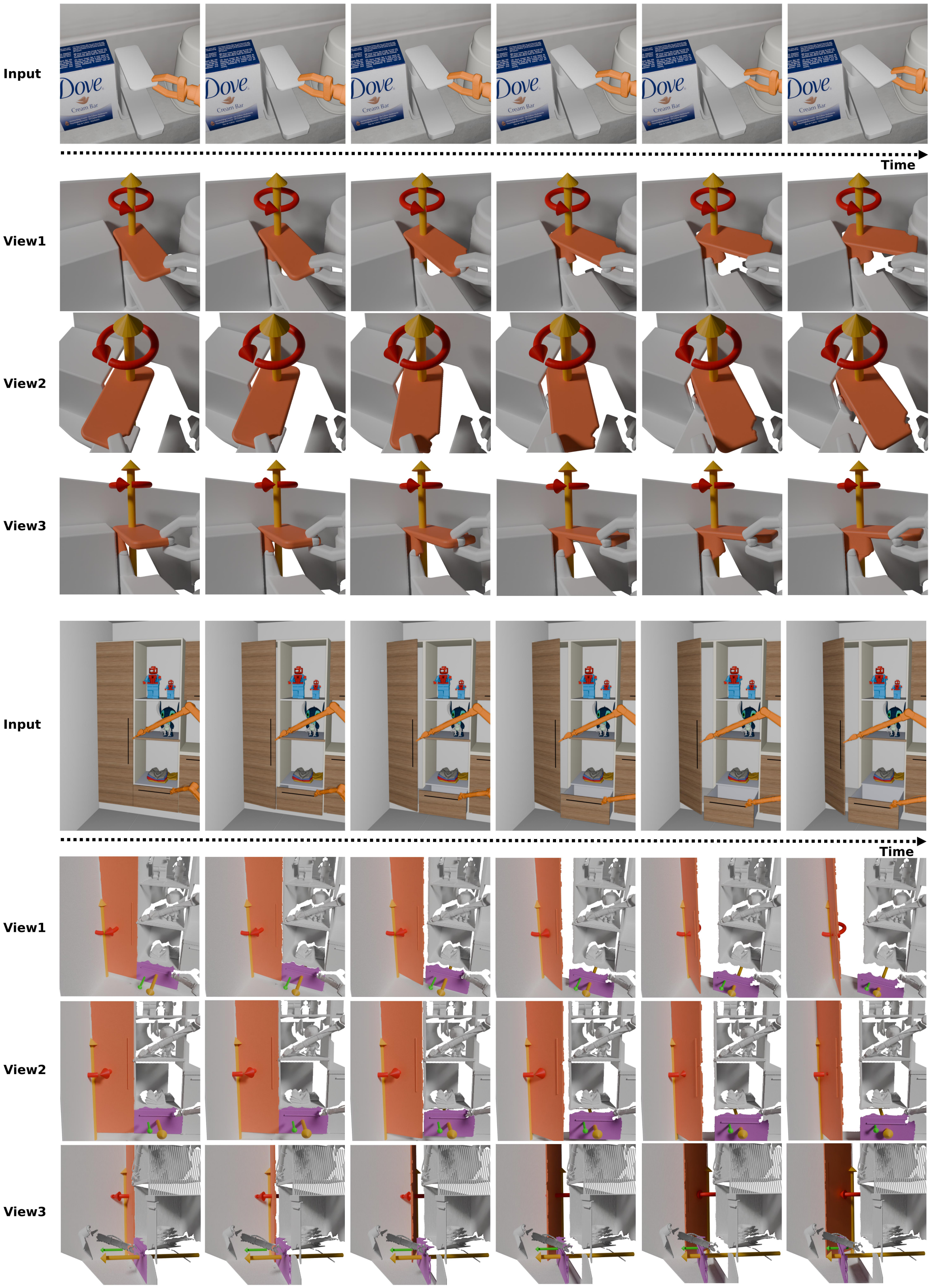

Figure 14. Analysis results of motion parts and their motion attributes for virtual simulation scenes 5-6.

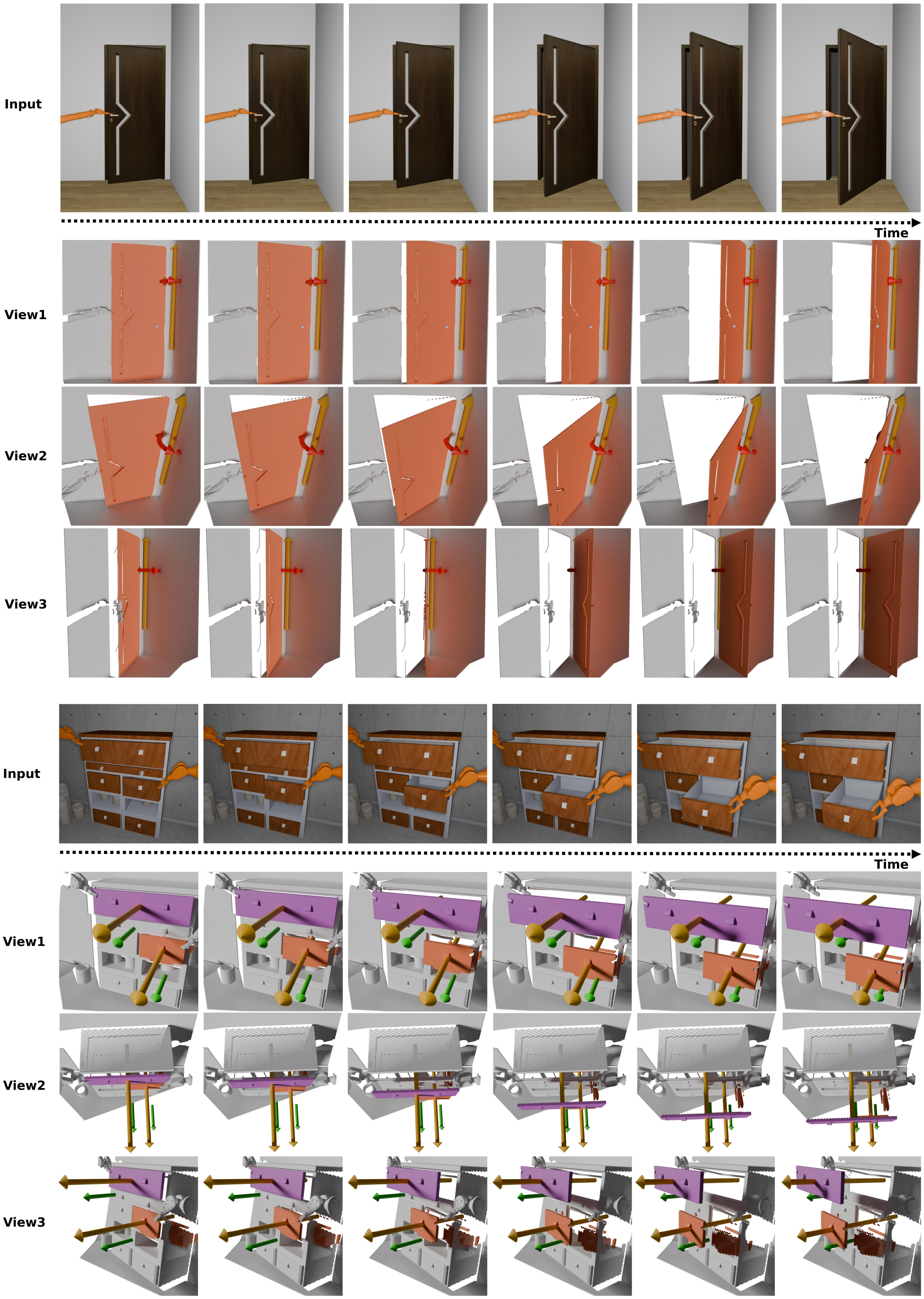

Figure 15. Analysis results of motion parts and their motion attributes for virtual simulation scenes 7-8.

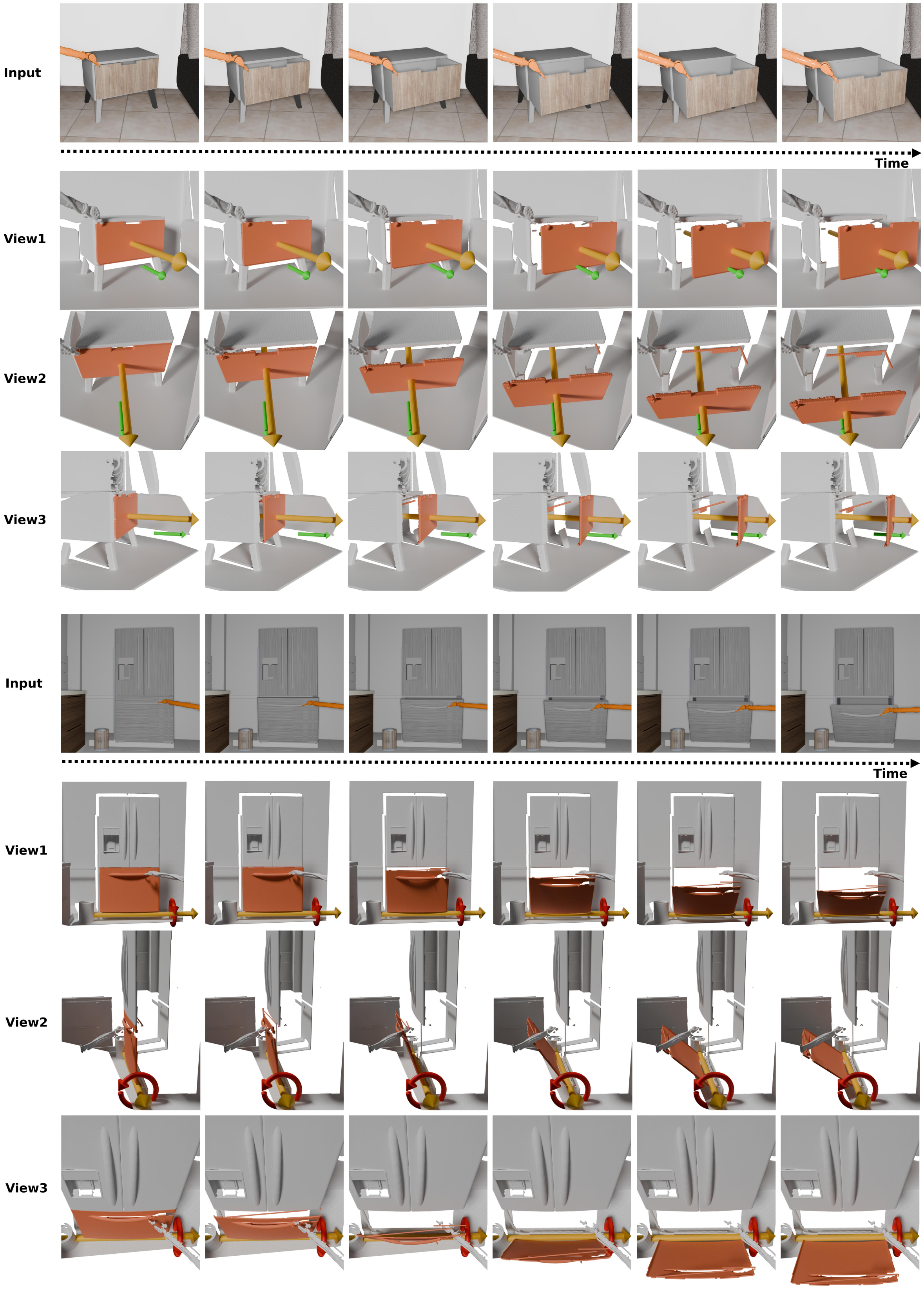

Figure 16. Analysis results of motion parts and their motion attributes for virtual simulation scenes 9-10.

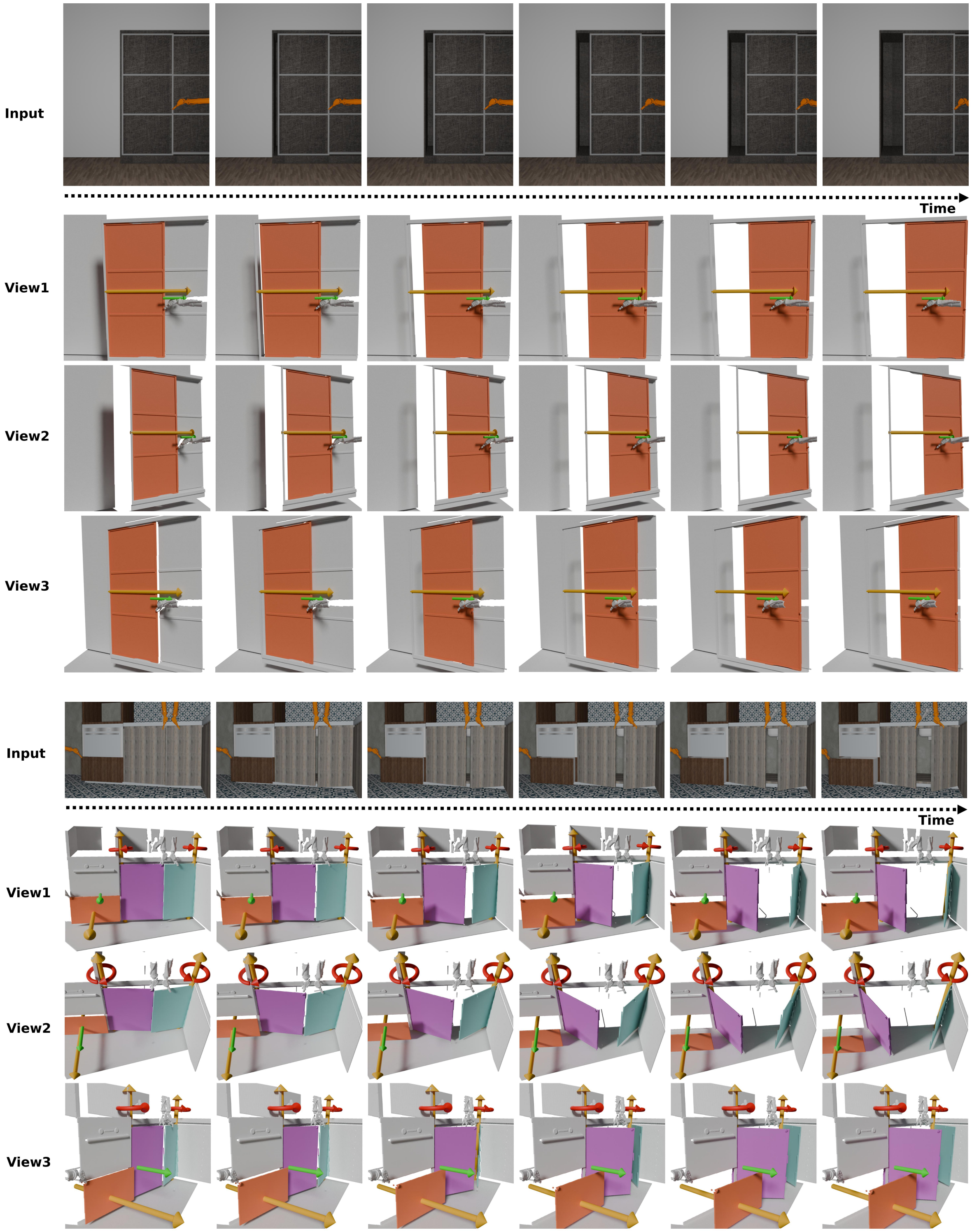

Figure 17. Analysis results of motion parts and their motion attributes for virtual simulation scenes 11-12.

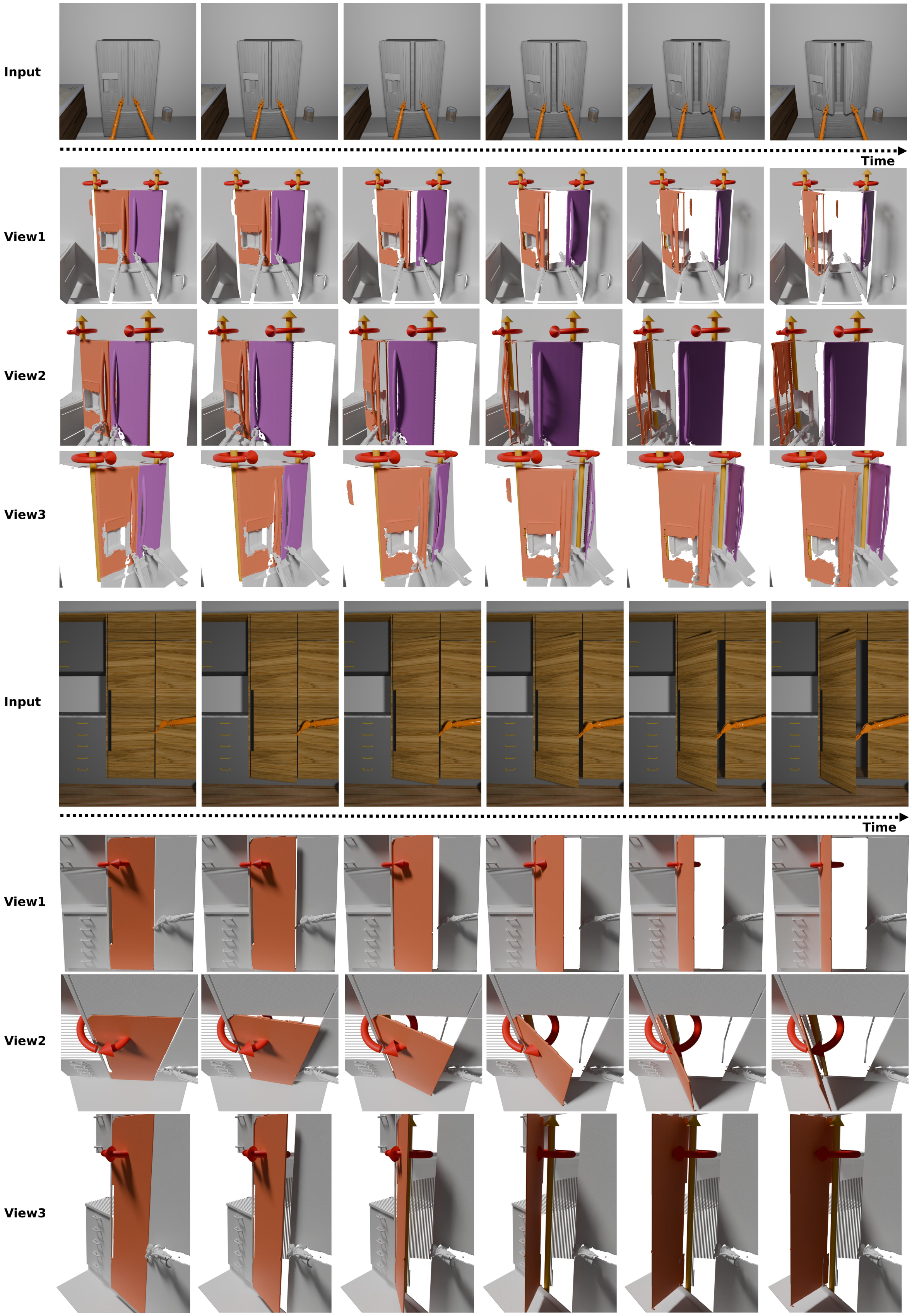

Figure 18. Analysis results of motion parts and their motion attributes for virtual simulation scenes 13-14.

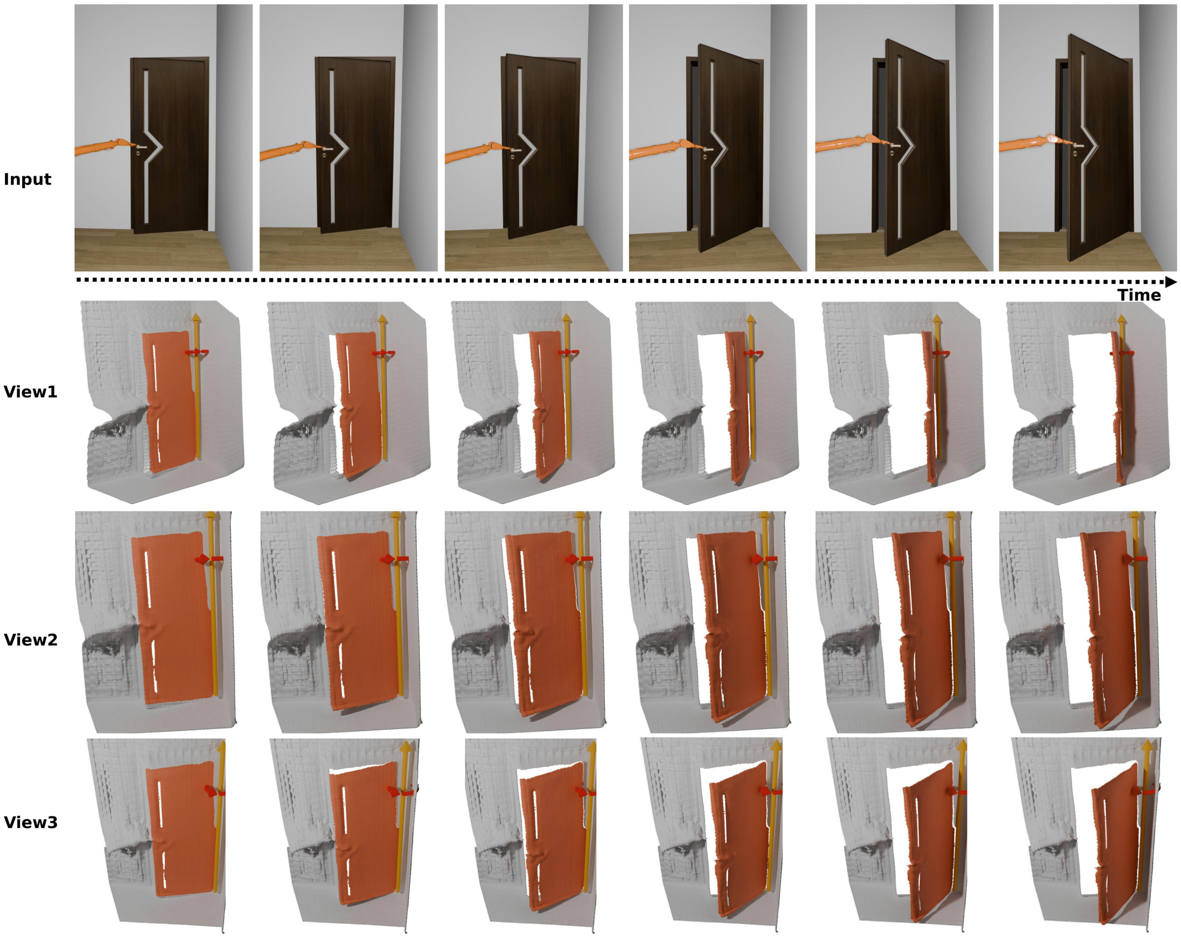

Figure 19. Analysis results of motion parts and their motion attributes for virtual simulation scenes 15.

References

[1] Blender Foundation. Blender: The free and open source 3d creation suite, 2024. 1, 2

[2] Trimble Inc. 3d warehouse: The largest library of free 3d models, 2024. 1

[3] Jiayi Liu, Ali Mahdavi-Amiri, and Manolis Savva. Paris: Part-level reconstruction and motion analysis for articulated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 352–363, 2023. 1, 2

[4] Johannes L. Schoenberger and Jan-Michael Frahm. Colmap: A general-purpose structure-from-motion and multi-view stereo pipeline, 2024. 2

[5] Colton Stearns, Adam Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. *arXiv preprint arXiv:2406.18717*, 2024. 1, 2

[6] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024. 1, 2