PartRM: Modeling Part-Level Dynamics with Large Cross-State Reconstruction Model

1. Additional Details

1.1. Details of PartDrag-4D Dataset

Category Details: We extract eight object categories from PartNet-Mobility [8]: dishwasher, laptop, microwave, oven, refrigerator, storage-furniture, washing-machine, and trashcan. Notably, the trashcan category is excluded from the training phase, serving as an unseen category. The extraction of specific categories is guided by the PartNet-Mobility dataset's detailed part-level annotations.

Rendering Details: We utilize Blender 3.5.0 for rendering multi-view images. The camera is fixed at a distance of 2.4 meters and a height of 1.5 meters, with 12 views rendered uniformly across the scene. The rendered images are produced at a resolution of 512 × 512 resolutions and stored in RGBA format. Both the images and their corresponding camera parameters are saved for subsequent processing.

1.2. Details of Training and Evaluation

Training Details: Our model was trained on a setup consisting of 4 × NVIDIA A800 GPUs, with a batch size of 4×8 and a learning rate of 5×10^{-4}. The training process included 100,000 iterations for the motion-learning phase and 50,000 iterations for the appearance-learning phase, utilizing the AdamW optimizer.

Evaluation Details: For evaluation, we randomly sample an initial state (e.g. a partially opened drawer) and apply drag-based deformation to transition to other states. This enables us to assess whether our model, as well as the baselines, can accurately perform the deformations.

1.3. Details of Applications in Manipulation Task

Details of Mesh and Axis Extraction: To facilitate robotic manipulation in Isaac Gym [6], the object's mesh is first derived by extracting it from generated Gaussian Splattings [1] as detailed in LGM [7]. Leveraging the meshes of these two states, the moving part and its axis can be extracted following [9] by uniformly sampling 10,000 points from the meshes as input.

2. Additional Experiment Results

2.1. Ablation Study on Drag Propagation Method

The drag propagation module is designed to enable PartRM to better understand the regions of moving parts. PartRM learns the translation and rotation of these parts utilizing a large amount of data, which can be considered as a data-driven approach. To inspect whether the rule-based propagation module can have a better performance, we have developed a pipeline that incorporates a drag classification model which takes in the input images and drags (Fig 1) to categorize the type of drag (i.e., translation or rotation), coupled with a rule-based method for drag propagation. Given input drags parameterized by its start and end points projection, i.e., $a_t = (a_{t,src}(x,y), a_{t,dst}(x,y))$. We denote $\Delta a_t = a_{t,dst} - a_{t,src}$ and the i-th propagated drag as $a_{t,i}$. For translation, our propagated rules can be formulated as:

$$a_{t,i} = (a_{t,i,src}, a_{t,i,src} + \Delta a_t) \tag{1}$$

where $a_{t,i,src}$ is the i-th point sampled from segmentation mask generated by Segment Anything [2]. For rotation, we define the propagated rules:

$$a_{t,i} = (a_{t,i,\text{src}}, \\ a_{t,i,\text{src}} + \Delta a_t \left(1 - \frac{\Delta a_t * (a_{t,i,\text{src}} - a_{t,\text{src}})}{\max_{a_{t,j,\text{src}}} \Delta a_t * (a_{t,j,\text{src}} - a_{t,\text{src}})}\right)) \tag{2}$$

where $a_{t,j,\text{src}}$ is the point sampled from the part segmentation mask, and $*$ represents the inner product of two vectors.

From our experiment, the drag classification model achieves **69.1%** accuracy on the PartDrag-4D test set. We also conduct an ablation study on the propagation methods, as shown in Table 1. The results demonstrate that PartRM outperforms this new method by effectively capturing both translational and rotational deformations through the synergistic learning of geometry and drag deformation.

2.2. Ablation Study on NVS Method

To assess the impact of the novel view synthesis (NVS) methods, we utilize an alternative NVS technique [3], while maintaining all other experimental conditions constant. As

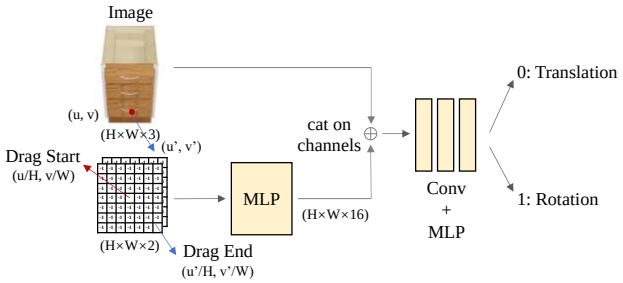

Figure 1. Drag classification network structure.

Prop. Method	PSNR (↑)	SSIM (↑)	LPIPS (↓)
Rule-Based	27.04	0.9368	0.0402
Data-Driven (Ours)	**28.15**	**0.9531**	**0.0356**

Table 1. Ablation on propagation methods.

NVS Method	PSNR (↑)	SSIM (↑)	LPIPS (↓)
EscherNet	**28.22**	**0.9574**	**0.0352**
Zero123++ (Ours)	28.15	0.9531	0.0356

Table 2. Ablation on NVS methods.

Embedding Method	PSNR (↑)	SSIM (↑)	LPIPS (↓)
DragAPart	27.51	0.9507	0.0403
Puppet-Master	27.74	**0.9545**	0.0411
PartRM (Ours)	**28.15**	0.9531	**0.0356**

Table 3. Ablation on Drag Embedding methods.

demonstrated in Table 2, the application of this more advanced NVS method results in a modest improvement in performance.

2.3. Ablation Study on Drag Embedding Method

To assess the impact of various drag embedding methods, we conducted an ablation study focusing on the drag embedding techniques and injection approaches described in DragAPart [5] and Puppet-Master [4]. All other experimental conditions were kept consistent with those in PartRM. Notably, neither DragAPart nor PartRM utilizes Fourier Embedding for encoding input drags. In DragAPart, the encoded drags are directly concatenated along the channels of the UNet features, whereas Puppet-Master divides the encoded drags into two blocks, treating them as the scale and shift parameters to apply on the UNet feature. As demonstrated in Table 3, PartRM outperforms other methods, with the incorporation of Fourier embeddings further enhancing its performance.

2.4. Qualitative Study on the Ambiguity of Drags

To inspect whether PartRM can handle the drags ambiguity well, we conduct a qualitative study as shown in Figure 2. PartRM can handle minor perturbations ((a)) due to noise introduced during training. When drags significantly differ from the training data, the model fails to produce the

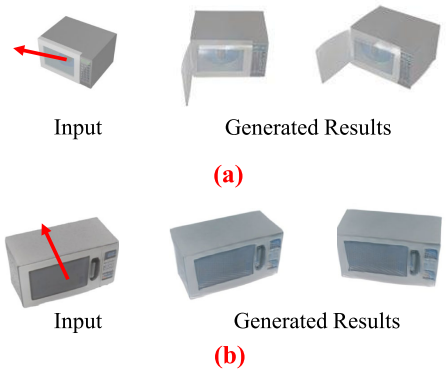

Figure 2. Qualitative Study on the Ambiguity of Drags.

expected outcome (b), where the drags attempt to push the entire microwave inside (the microwave's fully closed door cannot be articulated). This is because we don't model the motion of the whole object (only part motion) and there is lack of related training data.

3. Additional Visualization Results

We provide more qualitative results. Please refer to Figure 3, Figure 4 and Figure 5 for details.

References

[1] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 1

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1

[3] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9503–9513, 2024. 1

[4] Ruining Li, Chuanxia Zheng, Christian Rupprecht, and Andrea Vedaldi. Puppet-master: Scaling interactive video generation as a motion prior for part-level dynamics. *arXiv preprint arXiv:2408.04631*, 2024. 2

[5] Ruining Li, Chuanxia Zheng, Christian Rupprecht, and Andrea Vedaldi. Dragapart: Learning a part-level motion prior for articulated objects. In *European Conference on Computer Vision*, pages 165–183. Springer, 2025. 2

[6] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. 1

[7] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian

model for high-resolution 3d content creation. In *ECCV*, pages 1–18. Springer, 2025. 1

[8] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *CVPR*, pages 11097–11107, 2020. 1

[9] Chengliang Zhong, Yuhang Zheng, Yupeng Zheng, Hao Zhao, Li Yi, Xiaodong Mu, Ling Wang, Pengfei Li, Guyue Zhou, Chao Yang, et al. 3d implicit transporter for temporally consistent keypoint discovery. In *ICCV*, pages 3869–3880, 2023. 1

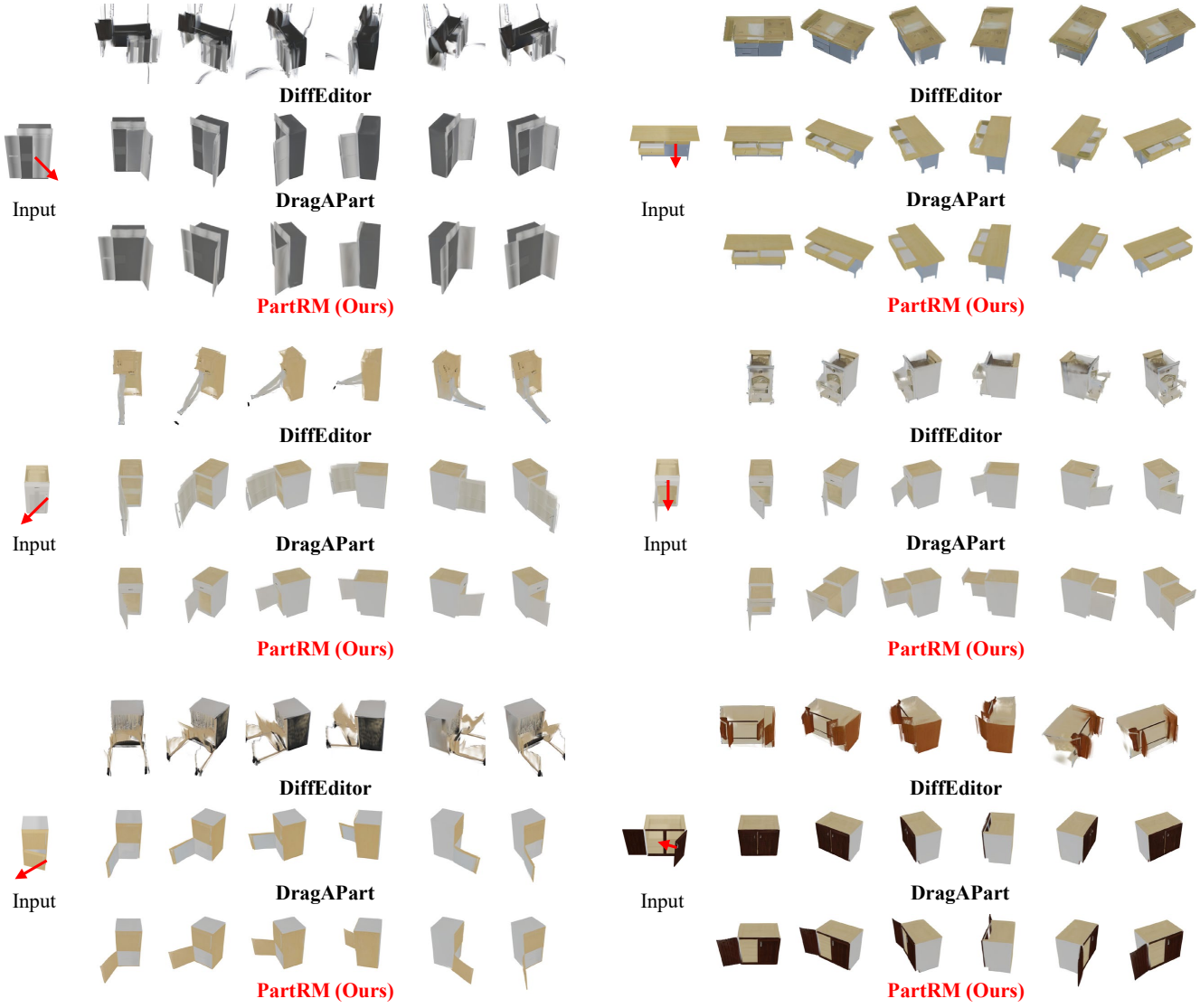

Figure 3. Qualitative comparisons between PartRM and baselines on PartDrag-4D dataset.

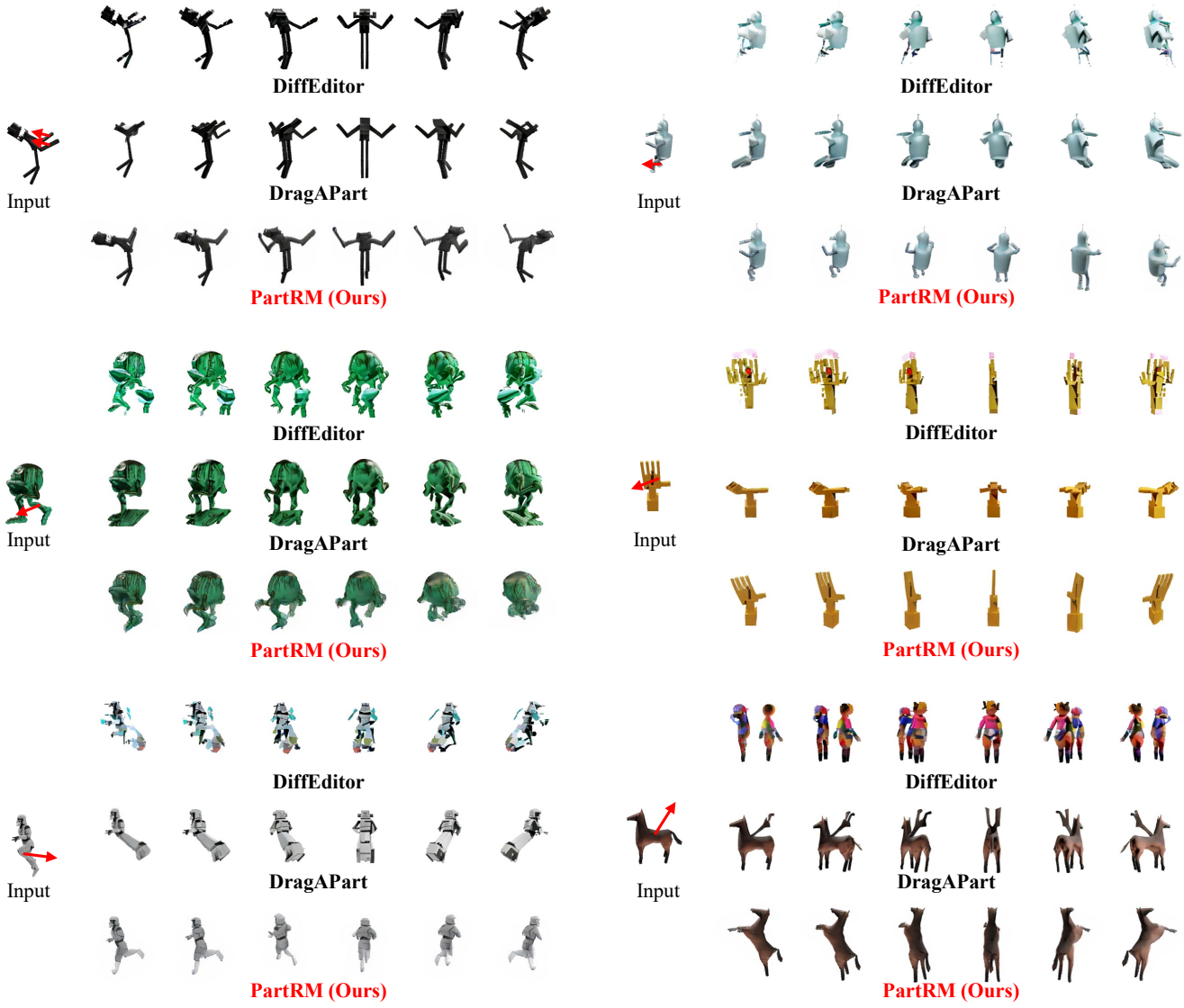

Figure 4. Qualitative comparisons between PartRM and baselines on Objaverse-Animation-HQ dataset.

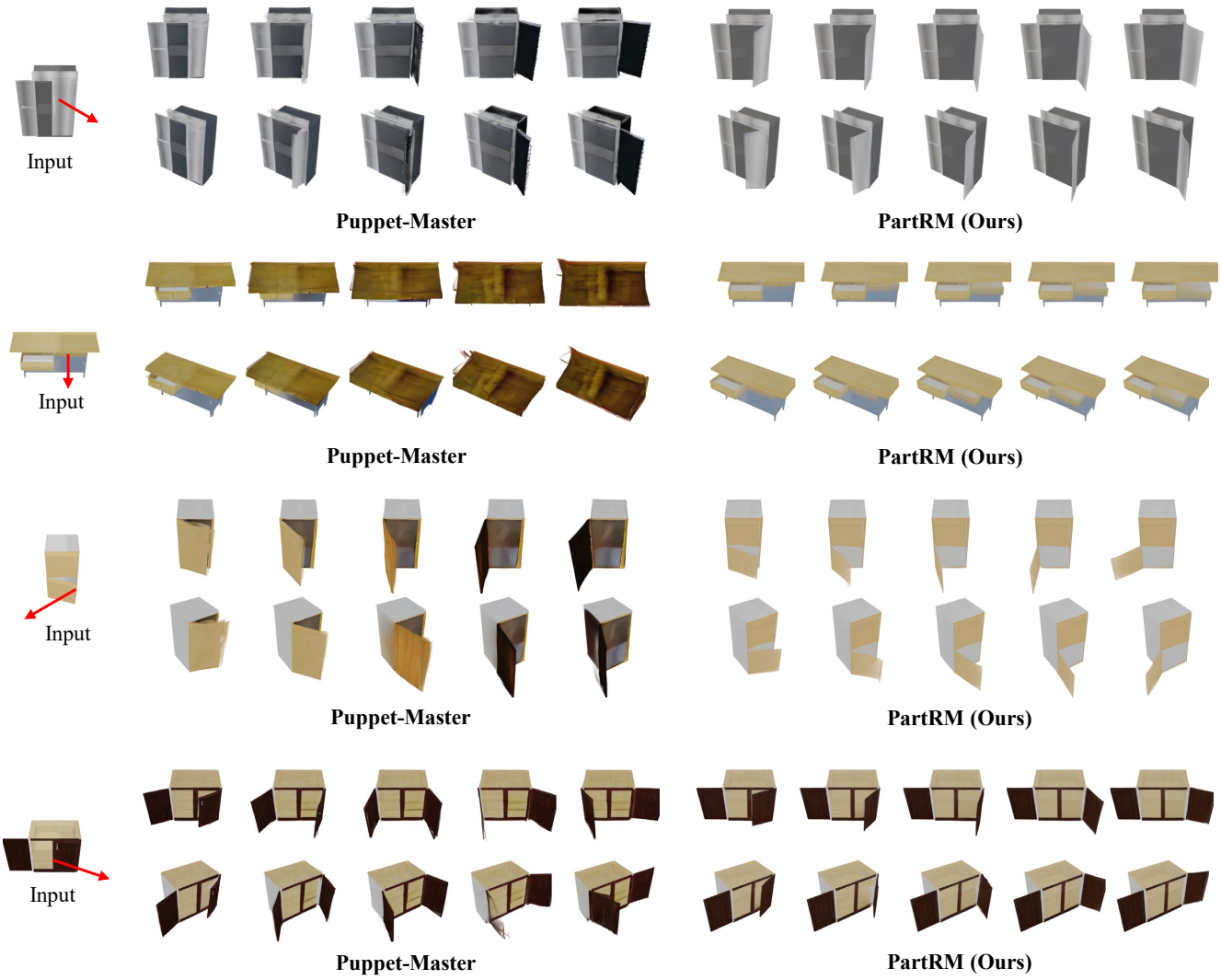

Figure 5. Qualitative comparisons between PartRM and Puppet-Master on PartDrag-4D dataset.