REArtGS++: Generalizable Articulation Reconstruction with Temporal Geometry Constraint via Planar Gaussian Splatting –Supplementary Materials–

Di Wu[1,2,4*], Liu Liu [3,5,6†], Anran Huang[3], Yuyan Liu[3], Qiaojun Yu[7], Shaofan Liu[3], Liangtu Song[1] , Cewu Lu[7]
1 *Hefei Institutes of Physical Science, Chinese Academy of Sciences*
2 *University of Science and Technology of China*
3 *Hefei University of Technology* 4 *RoboticsX, Tencent*
5 *Jianghuai Advance Technology Center*
6 *Anhui Provincial Key Laboratory of Humanoid Robots*
7 *Shanghai Jiao Tong University*
Email: wdcs@mail.ustc.edu.cn, liuliu@hfut.edu.cn

1. Overview

We provide the following contents in the appendix:

1. A detailed illustration of part segmentation for Gaussians in Sec. 2.1
2. The proof for entangled representation of dual quaternion in Sec. 2.2.
3. Articulated modeling with multi-state observation in Sec. 2.3.
4. Detailed implementation in Sec. 3.1.
5. Metrics calculation in Sec. 3.2.
6. Additional qualitative results in Sec. 3.3.
7. Additional ablation studies for the number of input views and states, the motion distance between two input states and canonical state in Sec. 3.4.
8. Video and codes for reproducibility in Sec. 4.

2. Methodology, Extended

2.1. Part-Aware Planar Gaussian Representation, Extended

As described in the main draft, we calculate part segmentation mask $\mathbf{M}_i$ of $\mathcal{G}_i$ through the Mahalanobis distances γ and a residual term W_i^{Δ}. Formally, the calculation can be represented as following:

$$\mathbf{M}_i = \Phi\left(\frac{-\gamma_i + W_i^{\Delta}}{\eta}\right) \tag{1}$$

where Φ is GumbelSoftmax operator and η is a learnable temperature factor. The residual term W^{Δ} is learned by a two-layer multi-layer perceptron (MLP) f. Concretely, the residual term of $\mathcal{G}_i$ can be formulated as:

$$W_i^{\Delta} = f(\mu, H(\mu), \mathbf{L}_i) \tag{2}$$

where H is a learnable hash grid with 6 levels to encode Gaussian centers, μ is the Gaussian center and $\mathbf{L}$ is the relative distances from Gaussian center to each part's center, which is defined in the main draft.

In this way, the part-aware planar Gaussian can be represented as: $\mathcal{G}_i : (\mu_i, \mathbf{q}_{\mathcal{G}_i}, \mathbf{s}_i | \mathcal{L}_{\text{scale}}, \sigma_i, \mathbf{c}_i, \mathbf{M}_i)$, where μ_i and $\mathbf{q}_{\mathcal{G}_i}$ denote the center position and quaternion of $\mathcal{G}_i$ respectively.

2.2. Entangled Representation of Dual Quaternion

In the main draft, we discuss the limitation of dual quaternions to represent screw motion, using in ArtGS [2]. We provide a proof below.

Proof. Let a screw motion be defined by rotation axis $\mathbf{a}$ (unit vector), rotation angle θ, translation distance d along the axis, and pivot point $\mathbf{o}$. The corresponding unit dual quaternion is:

$$\mathbf{dq} = \mathbf{q}_r + \varepsilon \mathbf{q}_d$$

where $\mathbf{q}_r = \cos(\theta/2) + \mathbf{a}\sin(\theta/2)$ and

$$\mathbf{q}_d = -\frac{d}{2}\sin\left(\frac{\theta}{2}\right) + \left[\frac{d}{2}\cos\left(\frac{\theta}{2}\right)\mathbf{a} + \sin\left(\frac{\theta}{2}\right)(\mathbf{o}\times\mathbf{a})\right].$$

*Work done as an intern at Tencent RoboticsX
†Corresponding author

Decomposing the dual part $\mathbf{q}_d = w_d + \mathbf{v}_d$:

$$w_d = -\frac{d}{2}\sin\left(\frac{\theta}{2}\right)$$
$$\mathbf{v}_d = \frac{d}{2}\cos\left(\frac{\theta}{2}\right)\mathbf{a} + \sin\left(\frac{\theta}{2}\right)(\mathbf{o}\times\mathbf{a})$$

The vector part $\mathbf{v}_d$ contains two entangled components: one proportional to $\mathbf{a}$ (translation-dependent) and one perpendicular to $\mathbf{a}$ (pivot-dependent). This creates an underdetermined system: for fixed $\mathbf{v}_d$, the pivot $\mathbf{o}$ can be arbitrarily shifted along $\mathbf{a}$ while compensating with translation d, producing identical dual quaternions. □

2.3. Learning With Multi-State Observations

The proposed approach can be readily extended to handle multi-state observation inputs. Specifically, $\mathcal{L}_{\text{vote}}$ and $\mathcal{L}_{\text{center}}$ are calculated without time. $\mathcal{L}_{\text{render}}$ can be represented as following:

$$\mathcal{L}_{\text{render}} = (1-\lambda_{\text{D-SSIM}})\mathcal{L}_1 + \lambda_{\text{D-SSIM}}\mathcal{L}_{\text{D-SSIM}} \tag{3}$$

where $\mathcal{L}_1$ and $\mathcal{L}_{\text{D-SSIM}}$ use input images $\mathbf{I}_t$ and rendered images $\bar{\mathbf{I}}(\omega,t)$. $t\in[0,1]$. The geometric loss is represented as:

$$\begin{aligned}\mathcal{L}_{geo} = (1-\nabla\mathbf{I}(t_0))\left(\left\|\bar{\mathbf{N}}(\omega,t_0)-\mathbf{N}(\omega,t_0)\right\|\right.\\ \left.+\left\|\nabla\bar{\mathbf{N}}(\omega,t_0)-\nabla\mathbf{N}(\omega,t_0)\right\|\right)\end{aligned} \tag{4}$$

Note that t_0 can be extended to any time in the interval $[0,1]$, excluding the canonical state. However, we observe that only using $(1-\nabla\mathbf{I}(t_0))\left\|\bar{\mathbf{N}}(\omega,t_0)-\mathbf{N}(\omega,t_0)\right\|$ in Eq. 4 at the canonical state also leads to an enhancement in performance, shown in the quantitative results of Table. 2. This is primarily because the rendering of the canonical state does not incorporate the learned motion, which performs a more fundamental optimization of Gaussian primitives.

In summary, Given any multi-state RGB images, our method only needs to normalize all states to $t\in[0,1]$, and then these inputs can be used for our pipeline. We also provide an ablation study of multi-state inputs in Sec. 3.4.

3. Experiment, Extended

3.1. Implementation

All experiments are conducted on a single RTX 4090 GPU. For initialization, we train two sets of Gaussians for 10,000 iterations respectively. Note that this process only takes few minutes. For optimization, we train Gaussian primitives for 30,000 iterations. We set $\lambda_{\text{render}}=1.0$, $\lambda_{\text{scale}}=1.0\times10^2$, $\lambda_{\text{center}}=0.1$, $\lambda_{\text{geo}}=1.5\times10^{-2}$, $\lambda_{\text{vote}}=1.0\times10^{-3}$. $\mathcal{L}_{\text{center}}$ is learned from 3,000 iterations. $\mathcal{L}_{\text{geo}}$ is learned from 7,000 iterations and $\mathcal{L}_{\text{vote}}$ is learned from 15,000 iterations. For initialization, we set the quarternion as $\bar{\mathbf{q}}+\varepsilon_1$, translation as ε_2 for each joint, where $\bar{\mathbf{q}}$ is a unit quarternion and ε is a number randomly sampled from $[1.0\times10^{-6}, 1.0\times10^{-5}]$. We perform pruning and densification for Gaussians from iteration 600 to 15,000. All methods are trained without depth supervisions. ArtGS is trained for 30,000 iterations, which is same as ours. REArtGS [3] is trained for 30,000 iterations in both the first and second stages.

3.2. Metrics

We use the same evaluation protocols as ArtGS. For mesh reconstruction, we generate surface meshes at both state and end states from canonical state, and compute the bi-directional Chamfer Distance (CD) between reconstructed mesh and ground truth mesh with 10,000 sampled points of each mesh. Specifically, CD is defined as the mean of the shortest distance between two point clouds, formulated as:

$$\begin{aligned}\text{CD} = \frac{1}{2}\Big(&\frac{1}{N_1}\sum_{\mathbf{x}\in N_1}\min_{\mathbf{y}\in N_2}\|\mathbf{x}-\mathbf{y}\|_2^2\\ &+\frac{1}{N_2}\sum_{\mathbf{y}\in N_2}\min_{\mathbf{x}\in N_1}\|\mathbf{y}-\mathbf{x}\|_2^2\Big)\end{aligned} \tag{5}$$

where N_1, N_2 denote the point number of sampled output and sampled ground truth respectively. Note that the CD results are multiplied by 1,000. We employ CD-w, CD-s, and CD-m for whole surfaces, static parts and dynamic parts respectively.

For joint parameter estimation, we evaluate the predicted joints using angular error (Axis Ang.) and pivot distance (Axis Pos., except for translation joints) between the predicted and ground-truth joint axes. Formally, the Axis Ang. $\mathbf{a}_{\text{err}}$ can be calculated as following:

$$\mathbf{a}_{\text{err}} = \min\left(\frac{\mathbf{a}_{\text{pred}}\cdot\mathbf{a}_{\text{gt}}}{|\mathbf{a}_{\text{pred}}|\,|\mathbf{a}_{\text{gt}}|}, 180^\circ - \frac{\mathbf{a}_{\text{pred}}\cdot\mathbf{a}_{\text{gt}}}{|\mathbf{a}_{\text{pred}}|\,|\mathbf{a}_{\text{gt}}|}\right) \tag{6}$$

and the Axis Pos. $\mathbf{o}_{\text{err}}$ can be represented by:

$$\mathbf{o}_{\text{err}} = \left|\frac{(\mathbf{a}_{\text{pred}}\times\mathbf{a}_{\text{gt}})\cdot(\mathbf{o}_{\text{pred}}-\mathbf{o}_{\text{gt}})}{\|\mathbf{a}_{\text{pred}}\times\mathbf{a}_{\text{gt}}\|}\right| \tag{7}$$

Note that $\mathbf{o}_{\text{err}}$ are multiplied by 1,000 for better comparison (i.e., measured by mm). We also report part motion error, which quantifies the geodesic rotation distance error θ_{err} of revolute joint (in degrees) and the Euclidean translation distance error $\mathbf{t}_{\text{err}}$ of translation joints (in meters). θ_{err} can be expressed as:

$$\theta_{\text{err}} = |\theta_{\text{pred}} - \theta_{\text{gt}}| \tag{8}$$

and $\mathbf{t}_{\text{err}}$ is formulated as following:

$$\mathbf{t}_{\text{err}} = \|\mathbf{t}_{\text{pred}}\| - d_{\text{gt}} \tag{9}$$

where d_{gt} is the ground truth prismatic distance.

Metrics are reported as mean results over 10 trials at the joint state with higher visibility, following ArtGS.

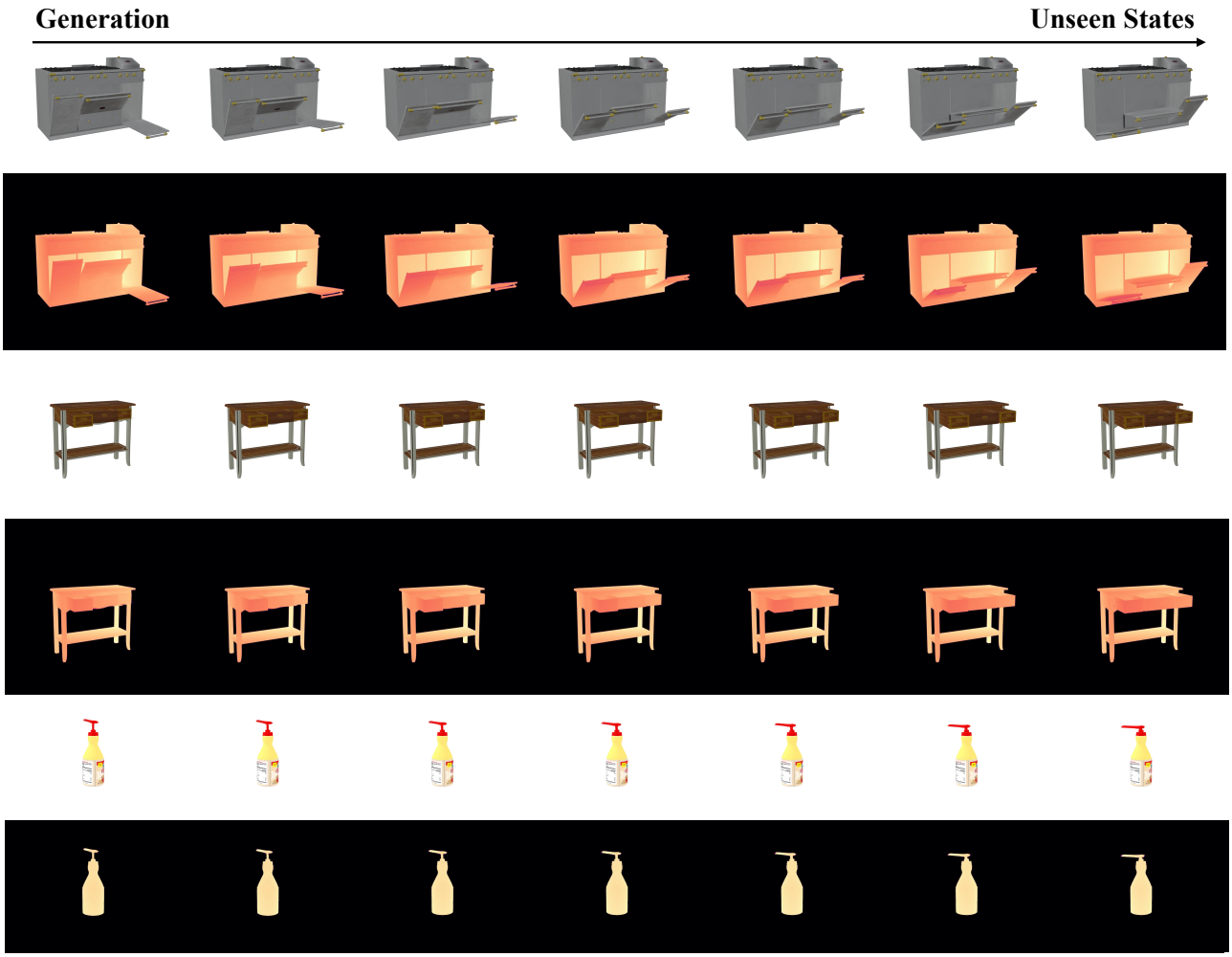

Figure 1. More qualitative results of generation at unseen states. We show both rendering and depth for better visualization.

3.3. Additional Qualitative Results

We provide additional qualitative results in Fig. 1, 2, 3 for generation at unseen states, more comparison of ArtGS-Multi and PARIS datasets respectively. It can be observed that our method achieves high-fidelity part-level surface reconstruction and accurate joint estimation. Please refer to the video of Sec. 4 for more qualitative results.

Table 1. **Ablation for the number of input views**. We report ArtGS-Multi dataset and 2 additional screw-joint objects. Axis Pos results are multiplied by 1,000.

Input views	Axis Ang	Axis Pos	Part Motion	CD-s	CD-m	CD-w
8	41.63	185.30	35.74	48.23	140.65	30.41
16	5.88	26.53	7.15	7.71	10.45	7.24
32	2.76	12.82	4.68	4.15	3.87	4.52
60-100	**0.41**	**1.18**	**0.22**	**1.41**	**2.38**	**2.13**

3.4. Ablation Experiments, Extended

Ablation for the number of input views. To verify our robustness to the number of input views, we conduct ablation studies of input views for each state on ArtGS-Multi dataset [2] and 2 additional screw-joint objects. The experimental results in Table. 1 prove that even if the number of input views is reduced to 16, our method still maintains competitive results for part-level surface reconstruction and joint parameter estimation.

Ablation for the number of input states. To evaluate our performance using multi-state inputs, we conduct ablation experiments on the ArtGS-multi dataset. We provide a total of 200 images for each object and evenly distribute them according to the number of input states. For instance, in the 4-input state setting, each state includes 40 images. All images are rendered from randomly sampled viewpoints on the upper hemisphere of the object. In the case of 3 in-

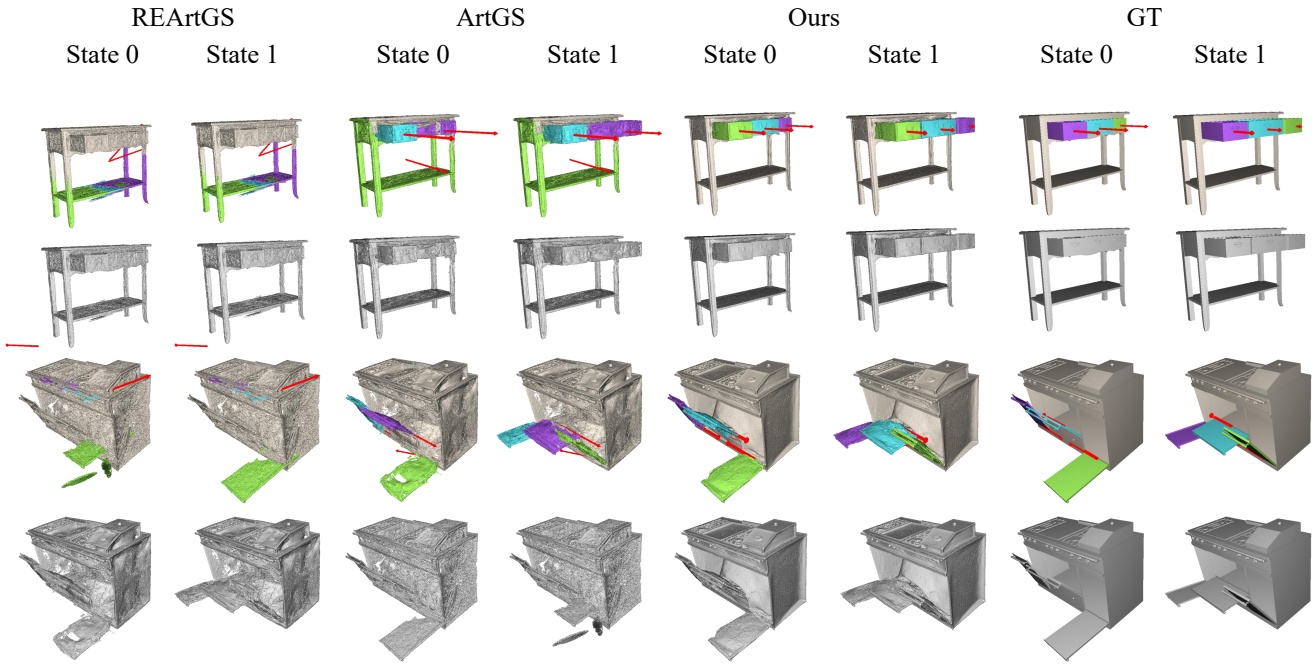

Figure 2. More qualitative results of dynamic surface reconstruction at start state and end state on ArtGS-Multi dataset. We show both articulated mode and surface meshes for better comparison. The red arrows represent joints.

Table 2. **Ablation for the number of input states**. We report the average results on ArtGS-Multi dataset. Axis Pos results are multiplied by 1,000.

Input States	Axis Ang	Axis Pos	Part Motion	CD-s	CD-m	CD-w
5	**0.23**	**0.86**	**0.10**	1.82	2.75	2.42
4	0.31	0.97	0.17	1.35	1.43	1.50
3	0.34	0.95	0.20	**1.18**	**1.64**	**1.32**
2	0.41	1.18	0.22	1.41	2.38	2.13

put states, we use states $t = 0, 0.5, 1$ with 67, 66, 67 images respectively. For 4 and 5 input states, we randomly sample additional states excluding $t = 0, 1$, and report the average results over 3 trials.

The experimental results in Table. 2 show that when using 3 input states, our method demonstrates performance improvements in both part-level surface reconstruction and joint parameter estimation. When the number of input states exceeds 3, although joint parameter estimation continues to improve, the quality of surface reconstruction slightly decreases, mainly due to the reduced image supervisions for each state, which leads to adverse effects on geometric learning in dynamic rendering. Moreover, our method achieves high-quality part-level surface reconstruction and joint parameter estimation with only 2 input states, which is sufficient to meet the requirements of downstream tasks.

Ablation of motion distance between two input states. To verify the robustness of our method for the movement distance between two input states, we conduct ablation experiments on *Storage 47648* (from ArtGS-Multi dataset) and *Fridge 10905* (from PARIS dataset [1]). We adjust the motion distance of each part of each object to 50%, 25% and 10% of original motion distance respectively, and render the same number of images as the input for end state. The experimental results in Table. 3 prove that on multi-part objects, our method still achieves high-quality surface reconstruction and accurate joint parameter estimation results when the motion distance is reduced to 25% of original data. For the 2-part object, our method maintains accurate articulated modeling results even when the motion distance is reduced to 15%, demonstrating our robustness for motion distance between two states.

Table 3. **Ablation of the motion distance between two states**. Axis Pos results multiplied by 1,000. 100% means using the same motion distance as original data.

Metrics	Storage 47468 (7 parts)				Fridge 10905 (2 parts)			
	100%	50%	25%	15%	100%	50%	25%	15%
Axis Ang	**2.58**	3.74	5.81	35.03	**0.01**	0.01	0.02	0.04
Axis Pos	**0.06**	0.73	3.45	116.72	**0.34**	0.52	1.13	3.04
Part Motion	**0.05**	0.08	0.08	17.63	**<0.01**	**<0.01**	**<0.01**	0.01
CD-s	**0.61**	0.81	0.85	4.31	0.97	**0.94**	1.05	0.97
CD-m	**3.19**	3.82	5.25	76.70	**1.25**	1.31	1.78	3.52
CD-w	**2.13**	2.17	2.41	6.57	1.20	**1.14**	1.26	1.83

Ablation of canonical state. We conduct ablation experiments for the selection of the canonical state on the

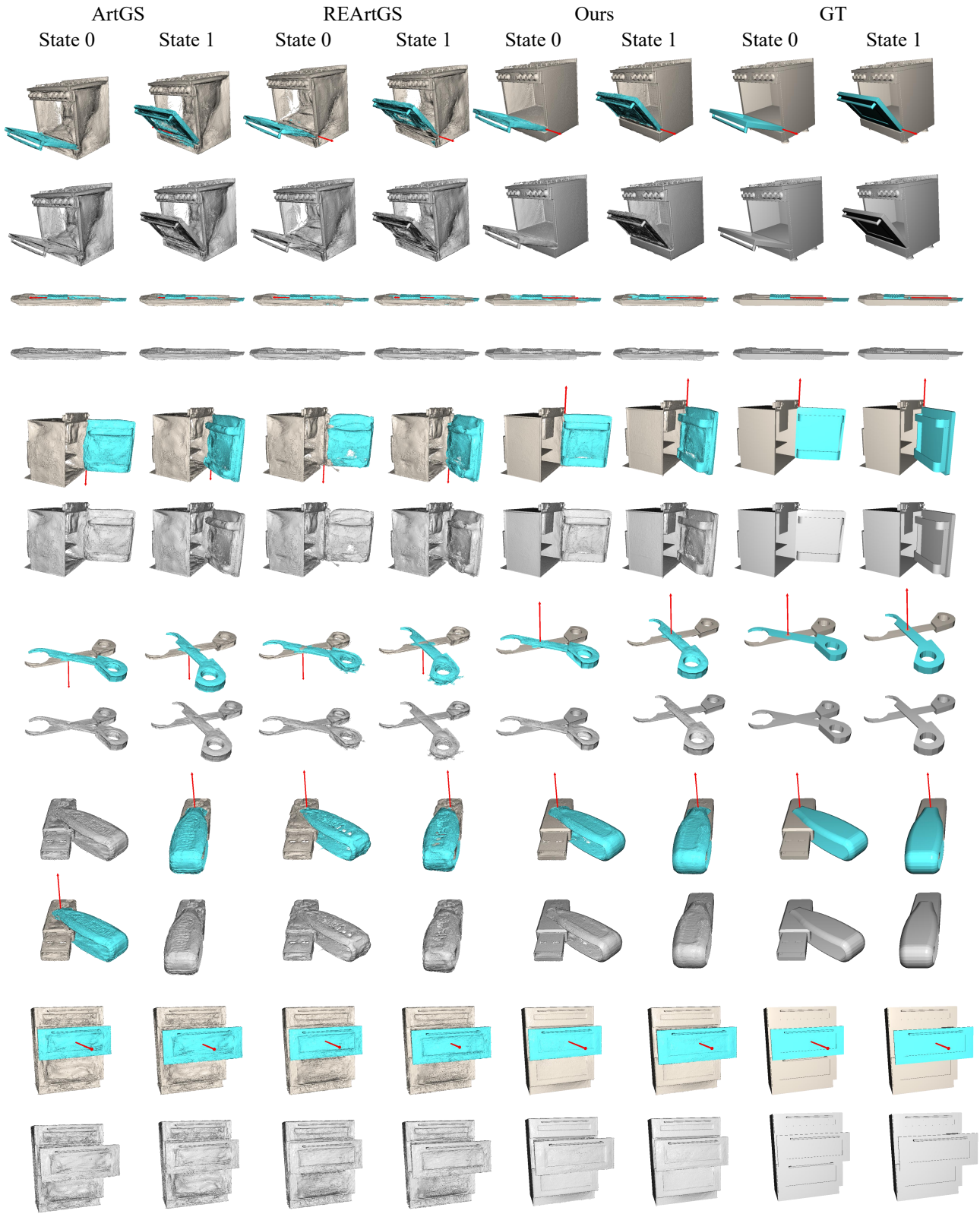

Figure 3. More qualitative results of dynamic surface reconstruction at start state and end state on PARIS dataset. We show both articulated modeling and surface meshes for better comparison. The red arrows represent joints.

Table 4. **Ablation for the number of input views**. We report ArtGS-Multi dataset and 2 additional screw-joint objects. Axis Pos results are multiplied by 1,000.

Canonical State	Axis Ang	Axis Pos	Part Motion	CD-s	CD-m	CD-w
$t^* = 0$	0.80	7.18	5.31	3.17	6.03	4.27
$t^* = 1$	1.54	13.75	10.48	3.25	9.84	6.54
$t^* = 0.5$	**0.41**	**1.18**	**0.22**	**1.41**	**2.38**	**2.13**

ArtGS-Multi dataset with 2 additional screw-joint objects. The quantitative results are presented in Table. 4. The experimental results show that when the canonical state is selected as 0.5, our method achieves the best performance. This is mainly attributed to the fact that the selection of the intermediate state helps to better optimize the entire motion process, especially in cases where the motion amplitude is large.

4. Reproducibility and Video

For the code and video, please refer to https://sites.google.com/view/reartgs2/home.

References

[1] 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. 4

[2] Yu Liu, Baoxiong Jia, Ruijie Lu, Junfeng Ni, Song-Chun Zhu, and Siyuan Huang. Building interactable replicas of complex articulated objects via gaussian splatting. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 3

[3] Di Wu, Liu Liu, Zhou Linli, Anran Huang, Liangtu Song, Qiaojun Yu, Qi Wu, and Cewu Lu. Reartgs: Reconstructing and generating articulated objects via 3d gaussian splatting with geometric and motion constraints. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2