IAAO: Interactive Affordance Learning for Articulated Objects in 3D Environments

Supplementary Material

6. Appendix

6.1. 3D-2D Correspondence Matching

We start with computing the feature similarity matrix α_{ip} between each pixel in part mask o and the sampled Gaussian point p^t. We then normalize similarity matrix α_{ip} using a softmax across the entire mask to obtain the weight β_{ip}. Finally, we identify the 2D point $s^{t'}_{p\to o}(I^{t'}_n)$ corresponding to the 3D point using a weighted sum. The computation steps are as follows:

1. Compute the feature distance:

$$\alpha_{ip} = \|F^{t'}_o(I^{t'}_n)[u_i] - F^t_{3D,o}(p)\|_2,$$

between the i-th pixel u_i of $I^{t'}_n$ and sampled Gaussian point $g^t_p(o)$.

2. Normalize α_{ip} using a softmax across the entire image to obtain the weight:

$$\beta_{ip} = \frac{\exp(-s\alpha_{ip})}{\sum_{i=1}^{|M^{t'}_o(n)|}\exp(-s\alpha_{ip})}.$$

3. Identify the 2D point:

$$s^{t'}_{p\to o}(I^{t'}_n) = \sum_{i=1}^{|M^{t'}_o(n)|}\beta_{ip}u_i,$$

corresponding to the 3D point using a weighted sum. $F^{t'}_o(I^{t'}_n)$ represents the DINOv2 features extracted from $I^{t'}_n$, and s is a hyperparameter that adjusts the smoothness of the heatmap β_{ij}.

6.2. Scene State Fusion

After obtaining the transformation $\xi^t = (s^t, R^t, T^t)$ (or its inverse function $\xi^{t'} = (s^{t'}, R^{t'}, T^{t'})$) for each part in scene state t, the next step is merging the two Gaussian Splatting (GS) models in the two states. We adopt the Gaussian splatting fusion and filtering strategy from [1]. To transform the Gaussians from the coordinate system of $G^{t'} = \{g^{t'}_p\}^{P^{t'}}_{p=1}$ to $G^t = \{g^t_p\}^{P^t}_{p=1}$, the position of each 3D Gaussian $g^{t'}_p$ is transformed as follows:

$$(x^{t'\to t}_p, y^{t'\to t}_p, z^{t'\to t}_p)^\top = s^{t'}R^{t'}(x^{t'}_p, y^{t'}_p, z^{t'}_p)^\top + T^{t'}.$$

The opacity remains unchanged during this transformation, i.e., $\alpha^{t'\to t}_p = \alpha^{t'}_p$. The rotation matrix $R^{t'\to t}_p \in \mathbb{R}^{3\times3}$ and scale $S^{t'\to t}_p \in \mathbb{R}^3$ are computed as:

$$R^{t'\to t}_p = R^{t'}R^{t'}_p, \quad S^{t'\to t}_p = s^{t'}S^{t'}_p.$$

Spherical harmonics (SH) coefficients undergo a linear transformation based on their rotation, which can be handled independently for each order. To this end, for any i-th order of SH coefficients, the following steps are performed:

1. Choose $2i+1$ unit vectors $u_0, \ldots, u_{2i+1}$ and compute their corresponding SH coefficients as $Q = (\mathrm{SH}(u_0), \ldots, \mathrm{SH}(u_{2i+1}))$.
2. Apply the transformation $\xi^{t'} = (s^{t'}, R^{t'}, T^{t'})$ to the vectors $u_0, \ldots, u_{2i+1}$ to obtain transformed vectors $\hat{u}_0, \ldots, \hat{u}_{2i+1}$.
3. Compute the transformation matrix for SH coefficients as:

$$(\mathrm{SH}(\hat{u}_0), \ldots, \mathrm{SH}(\hat{u}_{2i+1}))Q^{-1}.$$

Finally, the 3D Gaussians in $G^t = \{g^t_p\}^{P^t}_{p=1}$ closer to the center of scene t are merged with those in $G^{t'} = \{g^{t'}_p\}^{P^{t'}}_{p=1}$ near the center of scene t', producing $G^{t+t'}$.

6.3. More Visualization

Fig. 7 and Fig. 8 show the qualitative results on the "PARIS Two-Part Object Dataset". Column 4 shows the results of our IAAO and Column 3 shows the comparison with DigitalTwinArt. The input states at t and t' are shown in Columns 1 and 2. The part segments are shown in different colors and the red arrows indicate the joint prediction. Rotation is around the arrow and translation is along the arrow. Generally, as compared with DigitalTwinArt on both datasets, we can see from all figures that our IAAO produces better shape reconstruction with clearer details, more precise part segmentation with lesser erroneous labels, and more accurate joint predictions with correct motion directions. Fig. 9, 10 and 11 show snapshots of the generated motions on several examples at object and scene levels in a simulator. It shows IAAO can produce smooth interpolations of articulated objects with good part geometry modeling. Fig 12 illustrates the qualitative results on a sample scene from the Indoor Scene OmniSim dataset. Unlike existing baselines, which lack semantic meaning in their reconstructed neural fields, our model demonstrates the ability to perform object-level and fine-grained part localization based on prompts within complex indoor environments.

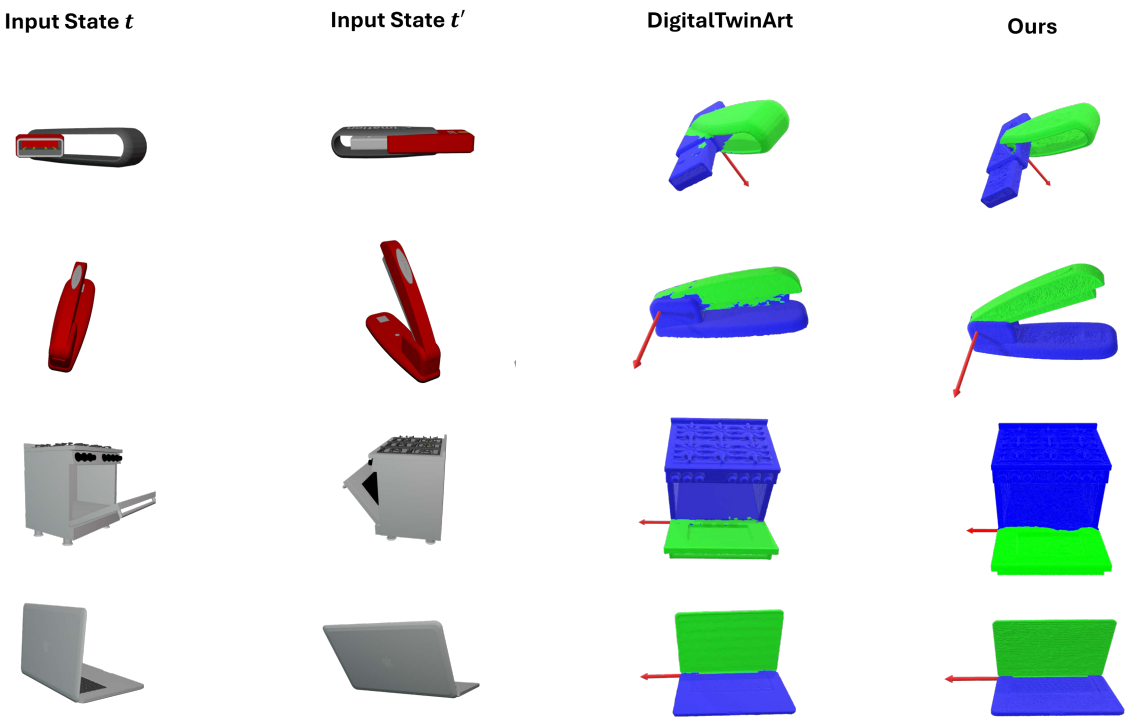

Figure 7. Qualitative analysis of shape reconstruction, part segmentation, and joint prediction results on the PARIS dataset. Our method shows better shape reconstructions and precise part segmentations on the thumbdrive and stapler shown in row 1 and 2.

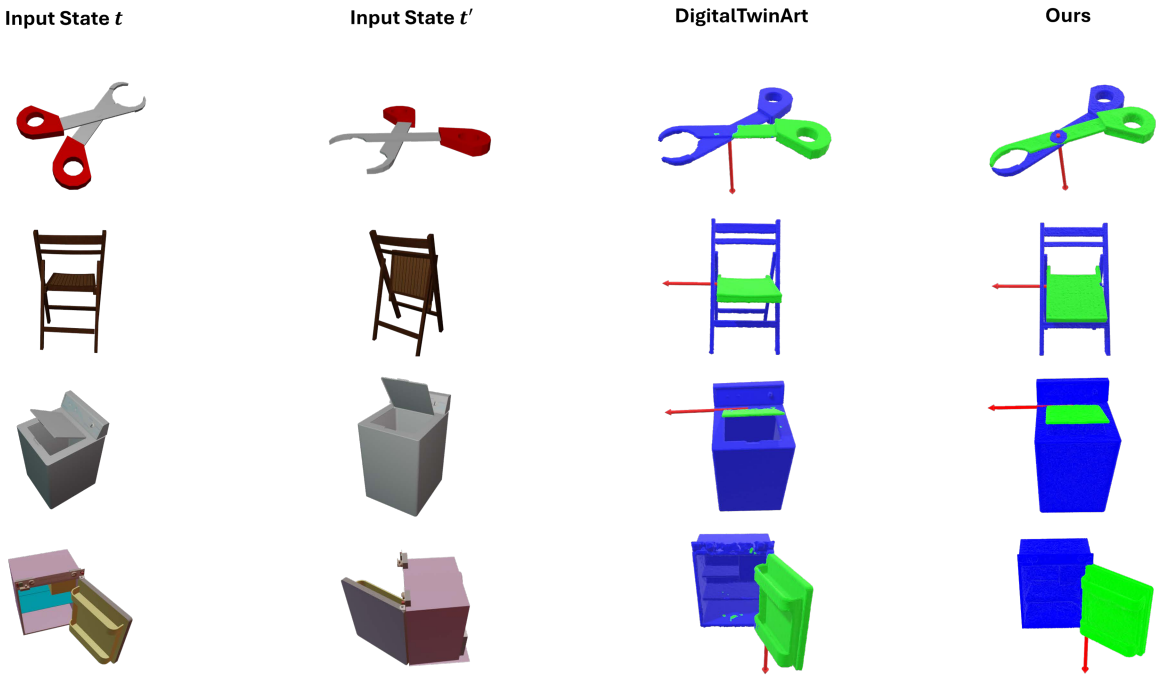

Figure 8. Qualitative analysis of shape reconstruction, part segmentation, and joint prediction results on the PARIS dataset. We can see that our method produces a better part segmentation result compared to DigitalTwinArt on the scissors shown in row 1.

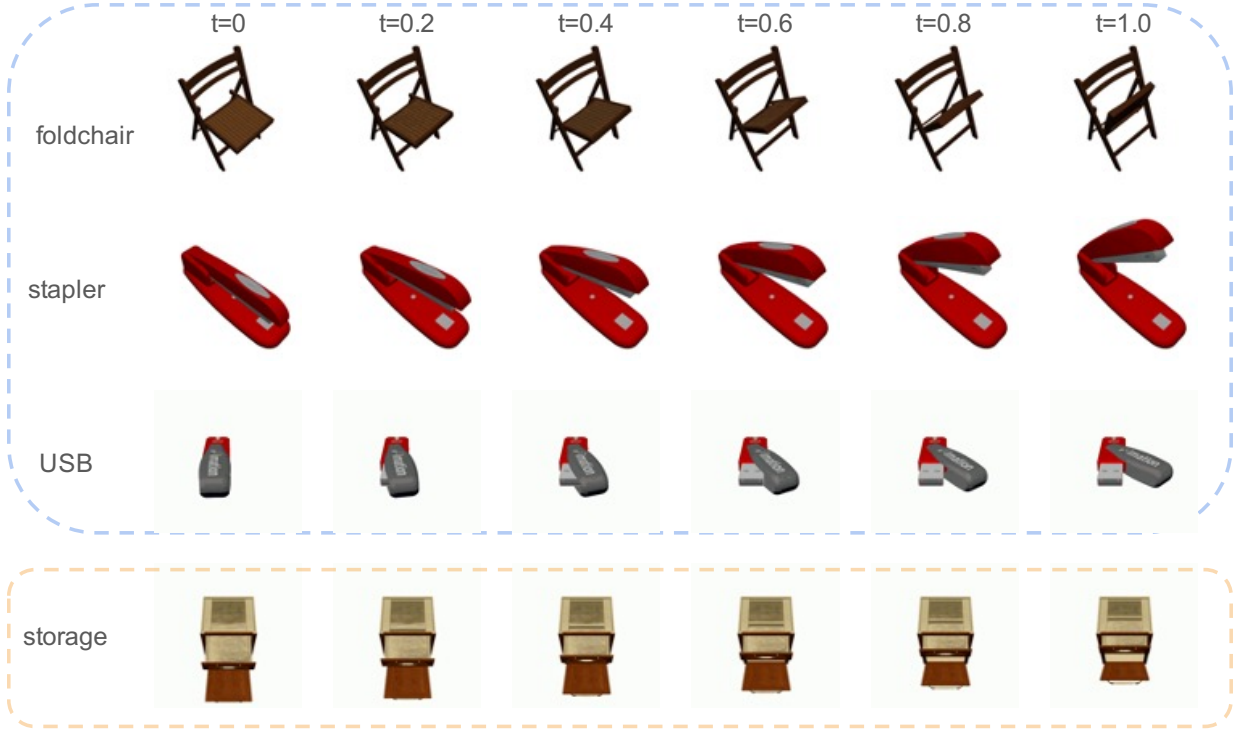

Figure 9. Qualitative analysis of scene interpolation on PARIS and PartNet-Mobility datasets.

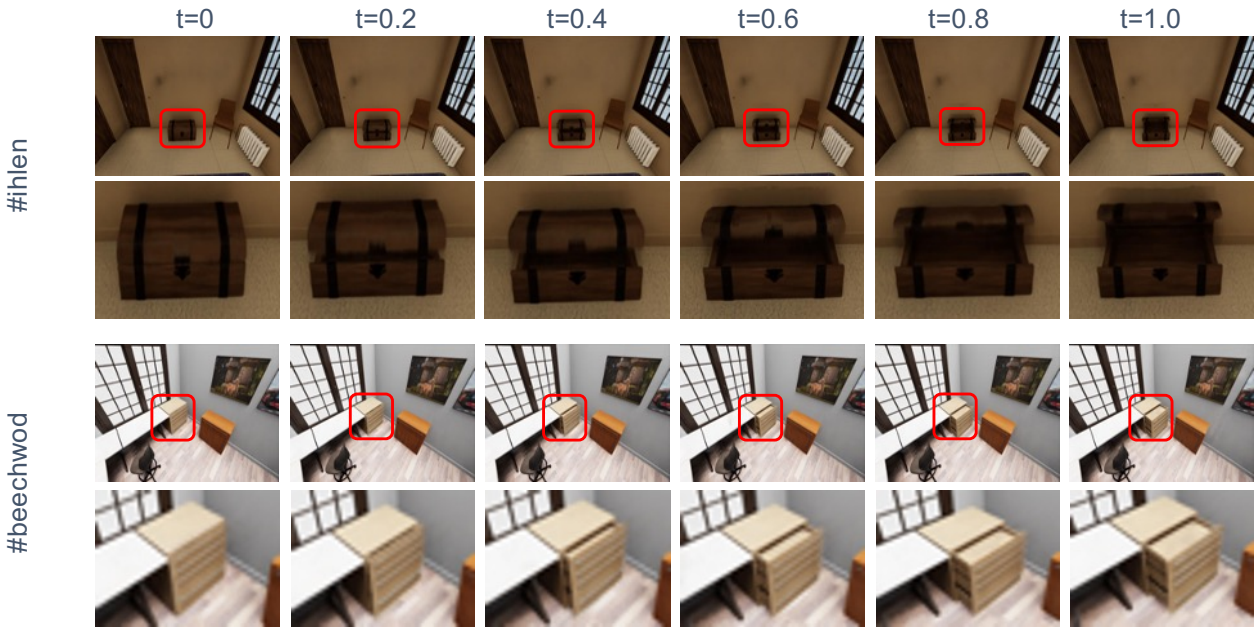

Figure 10. Motion snapshots on *#ihlen* and *#beechwod* from Indoor scene OmniSim dataset.

Figure 11. Motion snapshots on *#merom* and *#wainscott* from Indoor scene OmniSim dataset.

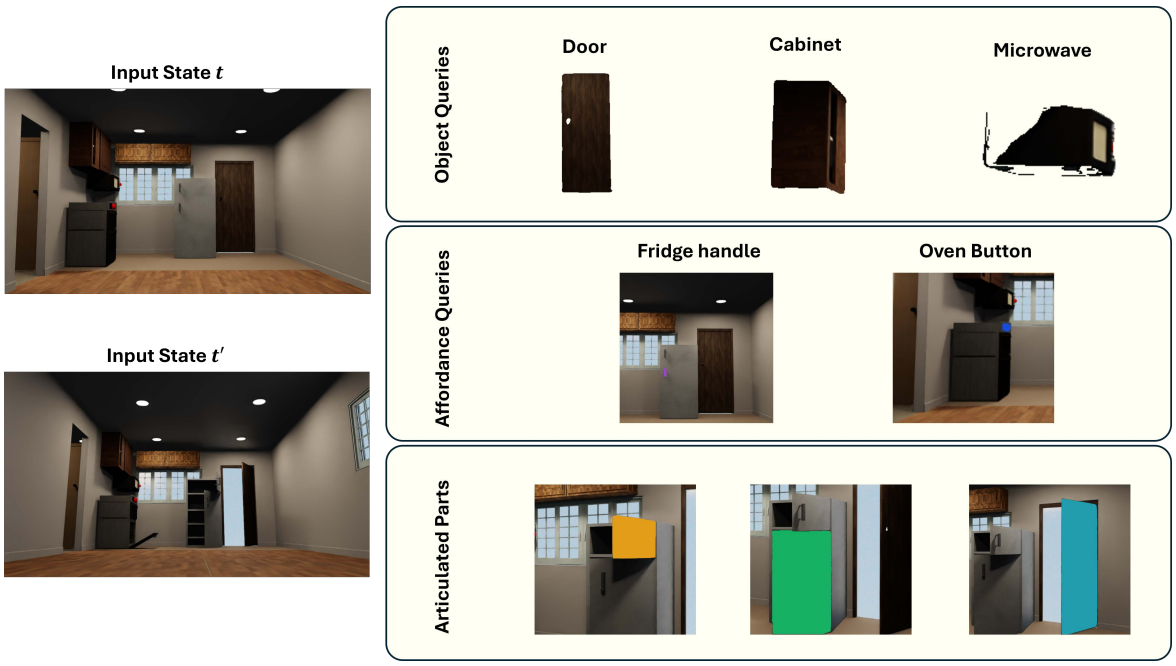

Figure 12. Qualitative analysis of object and affordance retrieval on one example scene from the Indoor scene OmniSim dataset.