

Proactive Scene Decomposition and Reconstruction

Supplementary Material

7. Supplementary Results

7.1. Interacted Object Trajectory Visualization

To visually demonstrate the accuracy of our method in object pose tracking, we visualize the trajectory of the interacted object. As shown in Fig. 6, the red dots in the second row of images represent the position of the object at each time step. From this visualization, it is evident that our method accurately estimates the object motion, which also indicates the reliability of our decomposition and scene reconstruction.

7.2. Progressive Decomposition and Online Reconstruction

Our method progressively achieves scene decomposition and reconstruction through proactive user interaction. Here, we provide Fig. 7 as an extra qualitative demonstration, which shows that our resulting map achieves accurate decomposition and high-quality rendering.

7.3. Run-time

We evaluate the average run-time of our system on the four sequences from HOI4D dataset, with the detailed timing of each module presented in Tab. 4. Our approach achieves a good balance between system performance and run-time, ensuring accurate decomposition and reconstruction while demonstrating significant efficiency advantages over offline methods.

| | Object Segmentation | Camera Tracking | Object Tracking | Joint Optimization |
|-----------|------------------------|--------------------|--------------------|-----------------------|
| Time (ms) | 429 | 371 | 344 | 493 |

Table 4. Per-frame time consumption of modules in our system.

7.4. Object Pose Evaluation

We also evaluate 6-DoF object pose tracking results on the HOI4D dataset (Tab. 5). For evaluation, we use the Area Under the Curve (AUC) of the ADD and ADDS metrics, which are defined as follows:

$$\text{ADD} = \frac{1}{m} \sum_{v \in \mathcal{O}} \|(Rv + T) - (R^*v + T^*)\|$$

$$\text{ADDS} = \frac{1}{m} \sum_{v_1 \in \mathcal{O}} \min_{v_2 \in \mathcal{O}} \|(Rv_1 + T) - (R^*v_2 + T^*)\|$$

where \mathcal{O} is the set of object vertices, R, T and R^*, T^* are the predicted and ground truth rotation and translation.

| Metric | ADD-S | ADD |
|----------|-------|-------|
| Mean (%) | 86.44 | 80.35 |

Table 5. The ADD and ADD-S metrics are reported as AUC percentages (0 to 0.1 m) on the tested sequences in HOI4D dataset.

7.5. Ablation Study

In Sec. 4.3, we discuss three strategies for mask refinement: A) flexible memory bank, B) absence check, and C) inter-frame inconsistency check. Here, we illustrate their effectiveness through ablation studies in Tab. 6. We also include an ablation study on hyperparameters. The results in Tab. 7 and Tab. 8 demonstrate the appropriateness of the hyperparameter values we chose for mask refinement and dynamic object identification.

| | w/o A | w/o B | w/o C | Ours |
|---------------------|-------|-------|-------|-------|
| Refined mask (mIoU) | 0.904 | 0.885 | 0.723 | 0.917 |

Table 6. The impact of different strategies on mask refinement.

| t_{m_1, m_2} | t_{m_3} | 0.6 | 0.7 | 0.8 | t_d | t_p | 0.4 | 0.5 | 0.6 |
|----------------|-----------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.8 | 0.861 | 0.912 | 0.912 | 0.912 | 0.2 | 0.836 | 0.895 | 0.908 | 0.908 |
| 0.9 | 0.898 | 0.917 | 0.915 | 0.915 | 0.3 | 0.902 | 0.917 | 0.883 | 0.883 |
| 0.95 | 0.903 | 0.917 | 0.910 | 0.910 | 0.4 | 0.917 | 0.870 | 0.745 | 0.745 |

Table 7. Ablation on hyperparameters $t_{m_1}, t_{m_2}, t_{m_3}$.

Table 8. Ablation on hyperparameters t_d, t_p .

8. Decomposition and Reconstruction of Articulated Objects

We assume that object motion follows a 6-DoF rigid body motion, allowing our method to model articulated objects as well. Specifically, after completing the decomposition and reconstruction using an input sequence, we examine each object individually. As shown in Fig. 8, if the object’s movement is restricted to translation or rotation only, which corresponds to a prismatic joint or a revolute joint respectively, we classify it as an articulated object and proceed to calculate its kinematics. The detailed evaluation process is as follows:

Pure Translation (prismatic joint) occurs when the rigid body moves along a single direction without rotation. To

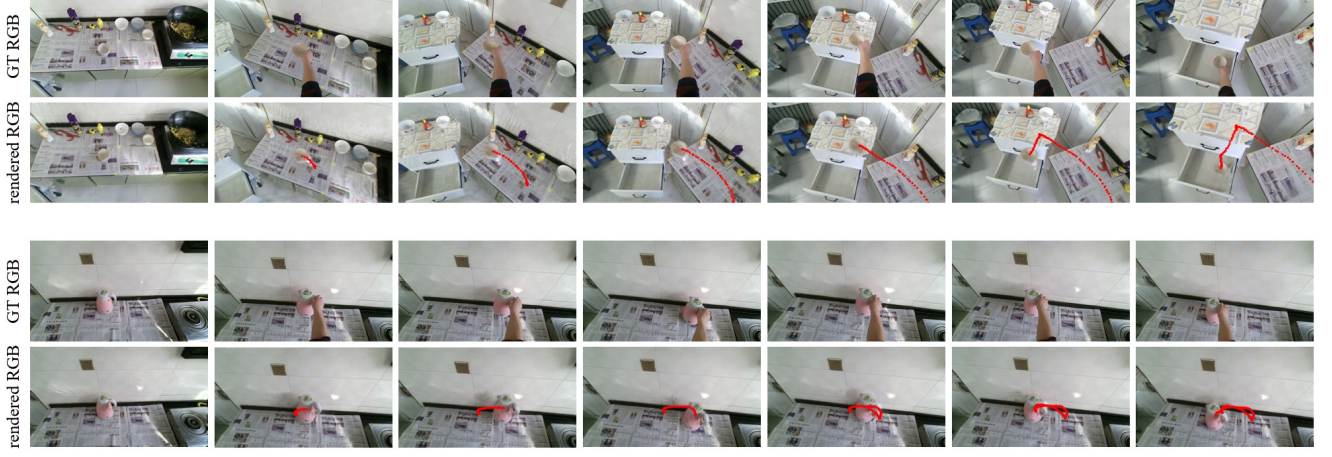


Figure 6. Visualization of the interacted object trajectory in HOI4D dataset.

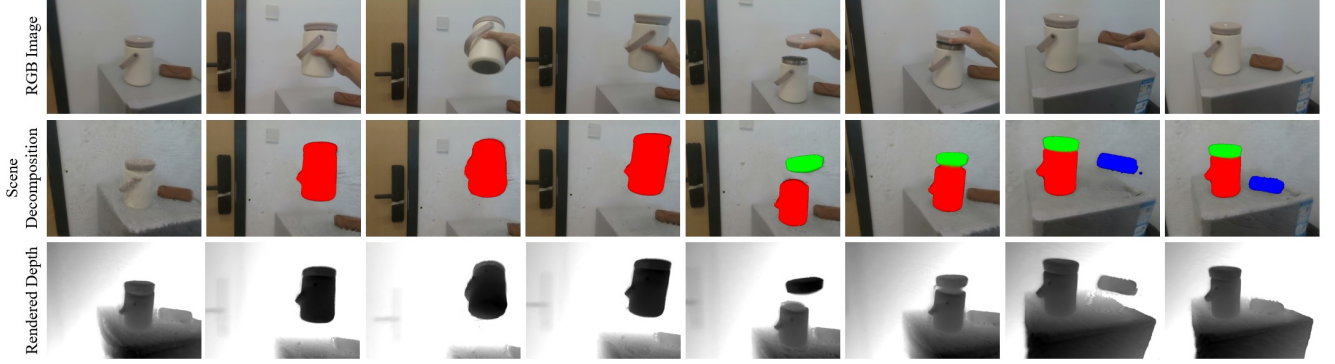


Figure 7. Qualitative results of progressive scene decomposition and reconstruction via proactive interaction.



Figure 8. Visualization of the revolute axis and prismatic joint of articulated objects.

identify such motion, two key properties are evaluated: (1) the rotation angles derived from the quaternions should be negligible, and (2) the translation vectors should exhibit approximate collinearity.

The quaternion $\mathbf{q}_t = [w_t, x_t, y_t, z_t]$ represents the rotation at time t . The corresponding rotation angle θ_t is com-

puted as:

$$\theta_t = 2 \arccos(w_t),$$

where w_t is the scalar component of the quaternion. For pure translation, θ_t should be close to zero within a predefined threshold ϵ_{rot} :

$$|\theta_t| < \epsilon_{\text{rot}}, \quad \forall t \in \{1, \dots, N\}.$$

Additionally, the translation vectors \mathbf{T} should align approximately along a single direction. This property can be evaluated by first centralizing the translations to remove the mean displacement:

$$\mathbf{T}_{\text{centered}} = \mathbf{T} - \bar{\mathbf{T}}, \quad \bar{\mathbf{T}} = \frac{1}{N} \sum_{t=1}^N \mathbf{T}_t.$$

The covariance matrix of the centralized translations is then computed as:

$$\mathbf{C} = \frac{1}{N} \mathbf{T}_{\text{centered}}^\top \mathbf{T}_{\text{centered}}.$$

The eigenvalues of \mathbf{C} describe the variance of the translations along the principal axes. For approximately collinear

translations, the largest eigenvalue λ_1 should dominate, while the remaining eigenvalues λ_2, λ_3 should remain small relative to λ_1 . Specifically, we define the collinearity condition as:

$$\frac{\lambda_2 + \lambda_3}{\lambda_1} < \epsilon_{\text{eig}}.$$

Thus, a motion is classified as pure translation if $|\theta_t| < \epsilon_{\text{rot}}$ for all t and the eigenvalue ratio satisfies $(\lambda_2 + \lambda_3)/\lambda_1 < \epsilon_{\text{eig}}$.

Pure Rotation Around a Fixed Axis (revolute joint) occurs when the rigid body moves about an axis without any significant translational displacement along the axis. To identify such motion, two conditions are evaluated: (1) the consistency of the rotation axis across all time steps, and (2) the translational displacement parallel to the axis must be negligible.

The rotation axis at time t can be extracted from the quaternion $\mathbf{q}_t = [w_t, x_t, y_t, z_t]$ as:

$$\mathbf{v}_t = \frac{(x_t, y_t, z_t)}{\sqrt{x_t^2 + y_t^2 + z_t^2}},$$

where x_t, y_t, z_t are the vector components of \mathbf{q}_t . The mean rotation axis is computed and normalized as:

$$\mathbf{v} = \frac{\sum_{t=1}^N \mathbf{v}_t}{\left\| \sum_{t=1}^N \mathbf{v}_t \right\|}.$$

The consistency of \mathbf{v}_t is evaluated using the angular deviation:

$$\Delta\theta_t = \arccos(\mathbf{v}_t \cdot \mathbf{v}),$$

In addition to the axis consistency, the translational displacement parallel to the axis is evaluated. For each time step t , the translation vector \mathbf{T}_t is decomposed as:

$$\mathbf{T}_t^{\parallel} = (\mathbf{T}_t \cdot \mathbf{v})\mathbf{v}, \quad \mathbf{T}_t^{\perp} = \mathbf{T}_t - \mathbf{T}_t^{\parallel}.$$

Here, \mathbf{T}_t^{\parallel} represents the displacement along the axis. The motion is classified as pure rotation around a fixed axis if and only if both conditions are satisfied:

$$\max_t |\Delta\theta_t| < \epsilon_{\text{axis}} \quad \text{and} \quad \max_t \|\mathbf{T}_t^{\parallel}\| < \epsilon_{\text{trans}}.$$

9. Potential Downstream Applications

Our method eliminates the ambiguity of segmentation granularity in a proactive manner. It incorporates priors from interactions and introduces structured representations through object decomposition. Additionally, it achieves photorealistic rendering results via Gaussian splatting. Consequently, compared to other approaches, our method delivers a structured scene map with accurate decomposition and high-fidelity rendering. As a result, the generated map can be effectively applied to zero-shot mobile manipulation, robot learning, photorealistic simulation, and immersive VR and AR applications.

10. Limitations and Future Work

Our method has some limitations. Currently, object tracking in our approach is restricted to rigid body motion, which limits its ability to model more flexible dynamics. Extending the method to handle non-rigid objects, such as deformable surfaces, would improve its applicability to more complex real-world interactions.

Additionally, while the core insight—resolving decomposition ambiguity through proactive interaction—is generalizable, the reliance on hand guidance makes it more suited for egocentric views. Future work could explore alternative guidance mechanisms to expand its use beyond egocentric settings and adapt to a broader range of applications.