# Articulated Object Estimation in the Wild: Supplementary Material

In this supplementary material, we provide additional insights on the evaluation protocol and utilized metrics in Sec. S.1.1, shed light on the necessity of ground truth camera poses to make ArtiPoint work in real-world deployment in Sec. S.1.2, evaluate to what degree ground truth interaction segments benefit the prediction performance in Sec. S.1.3 compared to predicted ones, and provide additional ablations regarding the hand detection as well as the point tracking in Sec. S.1.4. Furthermore, we provide qualitative insights in Sec. S.1.6 and in-depth explanations regarding the introduced Arti4D dataset in Sec. S.2.

## S.1 Experimental Evaluation

In this section, we present additional details on our experiments, the metrics we employ, and the additional ablation study.

### S.1.1 Evaluation Protocol & Metrics

In Sec. 5, we report quantitative results of our approach and the baselines. In the following, we detail the prediction to ground truth association procedure as well as the definition of the metrics that we employ to quantify performance.

In order to account for the fact that multiple interactions with the same object instance are likely throughout a single sequence, we match each obtained axis prediction against the corresponding ground truth axes based on the underlying interaction windows. This entails computing the 1-D intersection-over-union (IoU) between all predicted interaction segments and all ground truth interaction segments, as labeled as part of Arti4D. We consider a match whenever an $IoU > 0.5$ is exceeded.

As stated in the main manuscript, our metrics consist of positional and angular error. We compute the positional error $d_i$ for the placement of an articulation's rotation axis $\hat{\mathbf{a}}$ and the ground truth axis $\mathbf{a}_{gt}$ using the help of supporting points $\hat{\mathbf{p}}, \mathbf{p}_{gt}$ as

$$d_i = \begin{cases} \frac{(\hat{\mathbf{p}}_i - \mathbf{p}_{gt})^\top (\hat{\mathbf{a}}_i \times \mathbf{a}_{gt})}{\|\hat{\mathbf{a}}_i \times \mathbf{a}_{gt}\|} & \text{if } \|\hat{\mathbf{a}}_i \times \mathbf{a}_{gt}\| > \epsilon \\ \|(\hat{\mathbf{p}}_i - \mathbf{p}_{gt}) \times \mathbf{a}_{gt}\| & \text{else} \end{cases}, \tag{1}$$

where the first case covers the case in which the axes are not parallel with $\epsilon = 10^{-4}$. If a model does not provide a point on the axis directly, but a twist, we compute $\hat{\mathbf{p}}_i = \frac{\omega_i \times v_i}{\|\omega_i\|^2}$. The angular error of a prediction, we compute simply by using the normalized dot-product of the axes

$$\Theta_{err,i} = \cos^{-1}\left(\frac{\mathbf{a}_{gt}^\top \hat{\mathbf{a}}_i}{\|\mathbf{a}_{gt}\| \, \|\hat{\mathbf{a}}_i\|}\right). \tag{2}$$

We report only angular errors for prismatic joints, as the location of the axis does not have any effect on the motion of the parts of the articulated object.

### S.1.2 Performance Under Estimated Camera Poses

The experimental results reported in the main manuscript rely on ground truth camera odometry from the Arti4D dataset. However, in order to account for direct real-world deployment of our method, we have conducted additional experiments that do not require access to those. To do so, we have evaluated a number of prominent RGB-D SLAM approaches that produce metric map estimates: ORB-SLAM3 [56], Open-VINS [57], MAST3R-SLAM [58], and DROID-SLAM [59]. While all deep learning-based methods provide globally consistent maps, all traditional approaches fail due to

loss of camera tracking. Given their non-static character, the interaction segments pose the greatest challenge. As our method requires camera odometry instead of only keyframe-level pose estimates, we employ DROID-SLAM over MAST3R-SLAM.

First, we note that the maps produced by DROID-SLAM are registered against the ground truth point cloud using KISS-Matcher [60] in order to provide grounds for evaluation, which, in turn, may induce translational and rotational errors. When utilizing DROID-SLAM camera poses, we achieve reasonable results with only slightly increased angular and translational errors for both prismatic and revolute joints. Nonetheless, we observe a slight increase in prismatic type prediction accuracy when employing DROID-SLAM poses.

Table S.1: Comparison of estimated object axis under estimated and ground truth camera poses

| Method | Prismatic joints | | Revolute joints | | Type accuracy [%] | |
|---|---|---|---|---|---|---|
| | $\theta_{err}$[deg] | $d_{L2}$[m] | $\theta_{err}$[deg] | $d_{L2}$[m] | Prismatic | Revolute |
| w/ DROID-SLAM poses | 14.67 | – | 18.12 | 0.10 | 0.71 | 0.94 |
| w/ Arti4D odometry | 14.54 | – | 17.14 | 0.07 | 0.68 | 0.98 |

### S.1.3 Evaluation Using Ground-Truth Interaction Segments

In addition to the results relying on predicted interaction segments using the windowed hand detection scheme, we provide an evaluation based on ground truth interaction segments that are labeled as part of Arti4D. In general, we would expect lower angular and translational errors given that the method is not affected by hand occlusions (under large opening angles), motion blur, or non-detected hands. However, we find that our articulation estimation is robust wrt. to the interaction segmentation as we observe smaller angular errors on prismatic objects but larger errors for revolute-jointed objects, thus not allowing a clear interpretation.

Table S.2: Evaluation under ground-truth interaction segments

| Method | Prismatic joints | | Revolute joints | | Type accuracy [%] | |
|---|---|---|---|---|---|---|
| | $\theta_{err}$[deg] | $d_{L2}$[m] | $\theta_{err}$[deg] | $d_{L2}$[m] | Prismatic | Revolute |
| w/ GT segments | **11.99** | – | 20.24 | 0.09 | **0.74** | 0.97 |
| w/ pred. segments | 14.54 | – | **17.14** | **0.07** | 0.68 | **0.98** |

### S.1.4 Additional Ablation Study

In the following sections, we share additional ablations on hyperparameters of ArtiPoint and scene-respective results.

#### S.1.4.1 Ablation on Hand Detection

Extracting the interaction intervals, as described in Sec. 3.1 and illustrated in Fig. S.1, is a critical component of the ArtiPoint pipeline as it directly affects the number of articulated objects detected. As such, it requires careful parameter tuning, as small values of $w_h$ or $T_{min}$ increase the number of false positives. To better understand the impact of these parameters on the final results, we conduct a detailed ablation study with its findings presented in Tab. S.3. Decreasing $T_{min}$ leads to a small degradation in both angular error ($\theta_{err}$) and joint type classification accuracy. Increasing $w_h$ leads to over-smoothing of the raw hand detection signal, causing the segments to contain elongated idle phases at the start and the end, thus inducing unfavorable noise, and resulting in a noticeable degradation in both angular error ($\theta_{err}$) and joint type classification accuracy for both joint types. Furthermore, we also list the effect of not filtering outlier tracks, not using backward tracking, using CoTracker2 over CoTracker3, and performing static trajectory using three-dimensional trajectory data instead of two-dimensional data that is less subject to noisy depth.

Table S.3: Ablations of key parameters for extracting the interaction intervals component Sec. 3.1 on the Arti4D dataset. As before, we report the same set of metrics as in Tab. 1. Default values: $T_{max} = 90, T_{min} = 30, w_h = 6$.

| Method | Prismatic joints | | Revolute joints | | Type accuracy [%] | |
|---|---|---|---|---|---|---|
| | $\theta_{err}$[deg] | $d_{L2}$[m] | $\theta_{err}$[deg] | $d_{L2}$[m] | Prismatic | Revolute |
| w/o filtering outlier tracks | 13.17 | – | 22.80 | 0.14 | 0.64 | 0.84 |
| w/o backward tracking | 13.92 | – | 20.67 | 0.13 | 0.71 | 0.94 |
| CoTracker2 | 15.78 | – | 21.04 | 0.10 | 0.66 | 0.94 |
| static traj. filter in 3D | 15.18 | – | 18.23 | 0.10 | 0.65 | 0.92 |
| $w_h = 12$ | 15.58 | – | 19.88 | 0.08 | 0.66 | 0.97 |
| $T_{max} = 120$ | 14.57 | – | 18.13 | 0.07 | 0.68 | 0.96 |
| $T_{min} = 15$ | 14.57 | – | 17.88 | 0.07 | 0.68 | 0.96 |
| ArtiPoint | 14.54 | – | 17.14 | 0.07 | 0.68 | 0.98 |

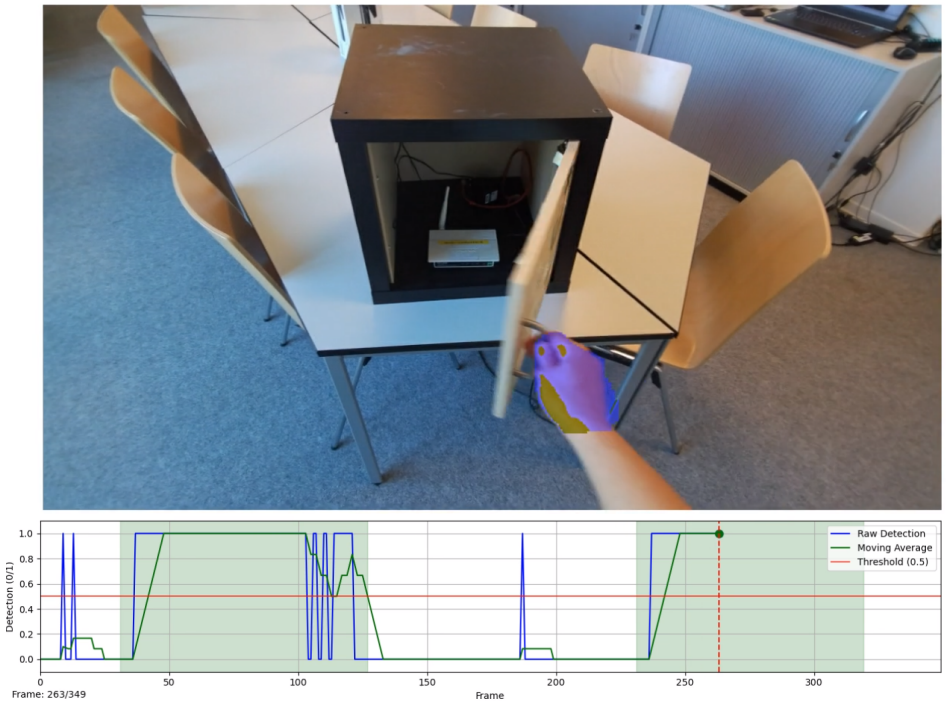
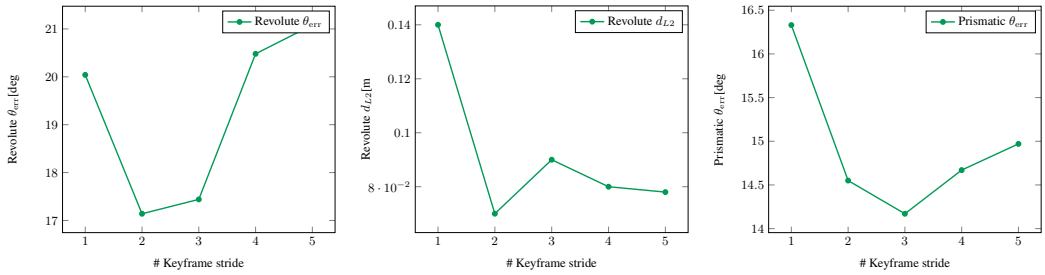



Figure S.1: Hand detection and interaction extraction: We visualize a live frame of an interaction, including a hand mask marked violet (top). In addition, we visualize the frequency of raw hand detections over time up to the live frame as well as its moving average (bottom). The horizontal red line indicates the threshold at which an interaction segment is created, given the moving average signal. The vertical dashed red line indicates the current frame.

### S.1.4.2 Ablation on Point Tracking

In this section, we present ablation results on the any-point tracking component as illustrated in Fig. S.2. In particular, we evaluate to which degree different keyframe strides used as input to the point tracking component affect the downstream axis prediction performance (see Sec. 3.2). Choosing the keyframe stride is vital hyperparameter of the point tracking stage. While a smaller stride leads to an increase in the number of detected points to be tracked by Cotracker3 [39], larger strides reduce the number of detected points to be tracked, thereby lowering the computational load. Retaining a sufficient number of points is necessary for estimating an object's 3D trajectory over time. Overall, we observe the lowest prediction errors for a keyframe stride of 2 to 3 based on the angular errors. However, in the case of the translational errors, we are not able to derive a distinct statement.




(a) Revolute $\theta_{\text{err}}$ (b) Revolute $d_{L2}$ (c) Prismatic $\theta_{\text{err}}$

Figure S.2: Impact of keyframe stride on tracking accuracy and error. (a) The revolute joint angular error $\theta_{\text{err}}$[deg] exhibits a minimum at stride 2. (b) The revolute joint positional error $d_{L2}$[m] is lowest at a stride of 2, with higher errors observed for both smaller and larger strides. (c) Similarly, the prismatic joint angular error $\theta_{\text{err}}$ exhibits its minimum at stride of 2. We conclude that a stride of 2 is optimal in terms of point density and computational efficiency.

### S.1.5 Scene-Respective Results

In addition, to the `EASY/HARD` differentiation evaluated in Tab. 2, we show scene-respective results in Tab. S.4. This involves averaging the predictions of all object interactions contained in a scene split. We list the number of sequences per scene as well as the number of labeled objects per sequence in Sec. S.2. As reported in Tab. S.4, `RH078` constitutes the most difficult split of the Arti4D dataset. In comparison, `DR080` and `RH201` seem to represent simpler environments.

We attribute worse results on `RH078` to a number of non-separable interactions as the hand is occasionally not fully retrieved between interactions. As a consequence, the proposed interaction extraction baseline potentially fails at differentiating two different interactions. We have mentioned the hand trigger limitation in Sec. 7 and leave improvements on that front to feature work. In addition to that, we observe that there are comparably more revolute joints in scene `RH078` whose associated objects are rather textureless and of metallic character, thus hindering consistent depth observations.

Table S.4: Scene-respective results: We report the scene-wise results using the established set of metrics. We find that `RH078` constitutes the most difficult split of the Arti4D dataset.

| Method | Prismatic joints | | Revolute joints | | Type accuracy | |
|---|---|---|---|---|---|---|
| | $\theta_{err}$[deg] | $d_{L2}$[m] | $\theta_{err}$[deg] | $d_{L2}$[m] | Prismatic | Revolute |
| `RH078` | 35.11 | – | 10.37 | 0.05 | 0.55 | 1.00 |
| `RR080` | 15.11 | – | 9.41 | 0.10 | 0.56 | 1.00 |
| `DR080` | 7.86 | – | 17.97 | 0.12 | 0.83 | 1.00 |
| `RH201` | 8.73 | – | 20.88 | 0.04 | 0.84 | 0.95 |
| Overall | **14.55** | – | **17.14** | **0.07** | **0.68** | **0.98** |

### S.1.6 Qualitative Results

In the following, we provide additional qualitative results. In Fig. S.3, we visualize the output of our proposed interaction extraction, any-point tracking and track filtering components. We observe a sufficient number of point trajectories even under partially missing depth measurements or feature-sparse textures. In addition to that, we visualize a full scene-level output of ArtiPoint on `DR080` scene in Fig. S.4. Overall, ArtiPoint detects the majority of interactions and produces reliable estimates considering the in-the-wild character of the recorded Arti4D sequences.

## S.2 Arti4D Dataset

In the following, we provide additional insight on the in-the-wild object articulation dataset Arti4D. We provide 45 sequences across four distinct environments as listed in Sec. S.2 and visualized in Fig. S.7. In addition to the sequence IDs and the recording names of the produced sequences, we

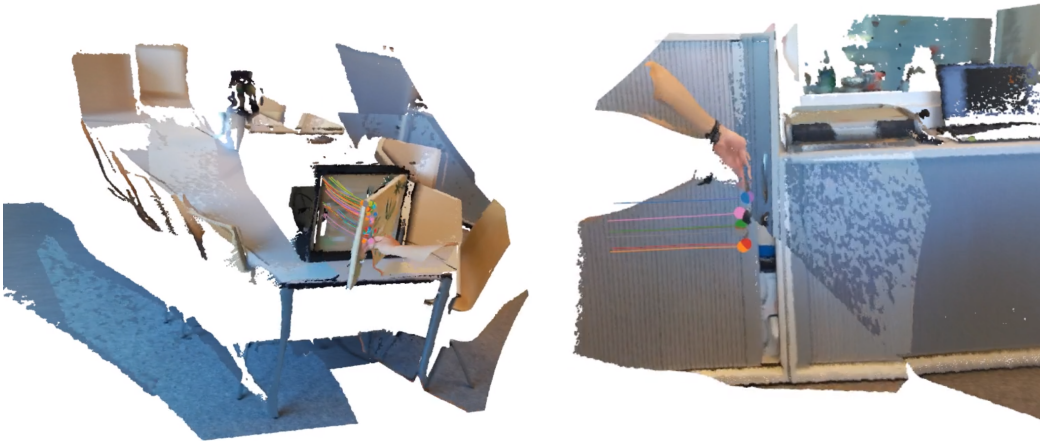


Figure S.3: Smoothed point trajectories: We visualize a cabinet (revolute) and its tracked keypoints (left) as well as a linear slider shelf (prismatic) on the right. Each keypoint trajectory is represented with a unique color. Both sets of point trajectories visualized constitute the output of our track filtering introduced in Sec. 3.3.

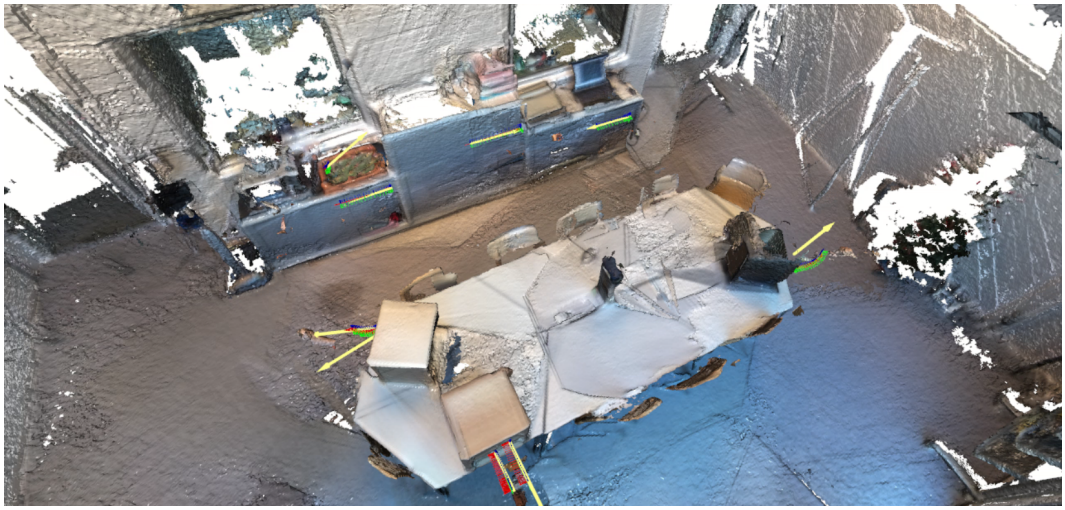


Figure S.4: Scene-level prediction: We depict a full scene-level output of the ArtiPoint framework on sequence `scene_2025-04-11-11-44-32` of the `DR080` scene. Yellow arrows denote axes of motion of predicted object interactions while coordinate frames represent the estimate part poses throughout articulation based on the proposed estimation framework (see Sec. 3.4).

report the number of labeled objects for each sequence, the ratio between prismatic and revolute joints as well as the ratio between easy and hard objects. First, note that the number of objects is not equal to the number of object interactions per sequence, as several sequences contain repeated interactions with the same single object instance. This constitutes a corner case in terms of articulated object interaction as it requires prediction methods to fuse, e.g., two predictions belonging to a single object. While most interactions are separable by detecting the absence of a hand mask, especially the `RH078` split which contains a number of hard-to-separate interactions. This is due to the fact that the interacting hand was not always fully retrieved in-between interactions of two distinct objects. Similar to the repeated interactions case mentioned before, this represents another corner case requiring advanced action recognition.

As part of the dataset, we make both rosbags and processed raw data public. While the rosbags include TF data at a higher frequency, the raw data includes aligned RGB, depth, and camera poses at 15 Hz. We employed an Azure Kinect RGB-D camera that was handheld throughout all interactions. In terms of ground truth camera pose retrieval, we employed external tracking using

HTC Vive trackers, which provide cm-level accuracy. In case of sudden odometry glitches induced by considerable occlusions or reflections on glass or metal, we have removed those sequences from the dataset. We found that running classical structure-from-motion approaches to reconstruct the underlying sequence fails as significant parts of the camera field of view cover articulations. In turn, the contained articulations break assumptions towards mostly *static* visual correspondence made in structure-from-motion methods. Thus, we leverage the ground truth camera poses and perform TSDF fusion to produce scene reconstructions. We depict four TSDF-reconstructed sequences stemming from each of the splits in Fig. S.5. The reconstructions reflect minimal ground truth odometry drift and enable precise anchoring of object axes.

The ground truth object axes were labeled based on the reconstructed sequences using Blender, exported as JSONs, and verified by a second reviewer. Furthermore, we provide metadata on difficulty levels and temporal interaction segment ground truth labels. We provide four exemplary depth-masked RGB frames covering an interaction of a microwave featuring semi-transparent glass and metal surfaces in Fig. S.6. As depicted, a considerable part of the object does not produce depth estimates. Objects of that kind are labeled using the `HARD` category.

Finally, we depict several RGB frames drawn from the four distinct scene splits in Fig. S.7, underlining the variety of objects and challenging conditions of the proposed dataset. We also make the sequence reconstructions in the form of meshes and point clouds available.

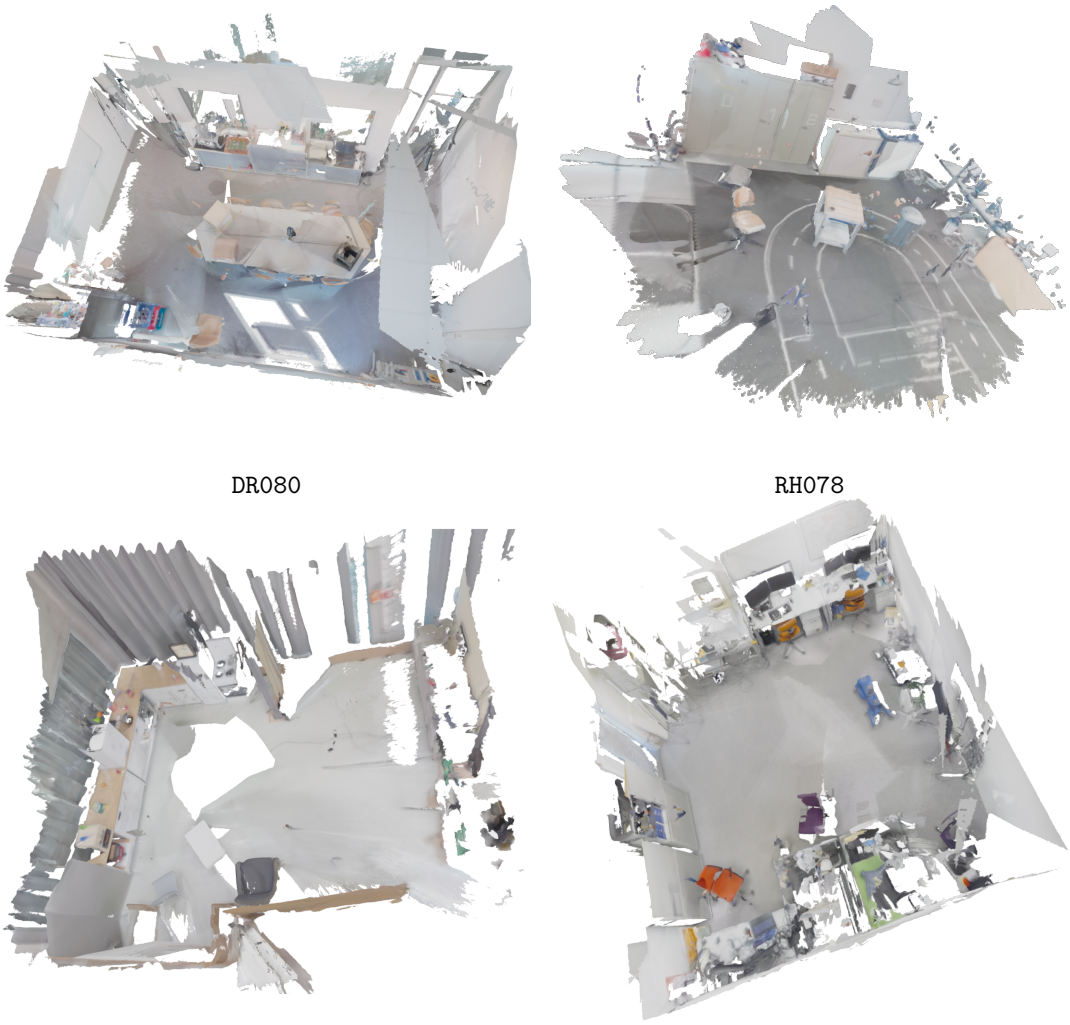



Figure S.5: We visualize the reconstructed scenes of the four Arti4D environments: DR080, RH078, RH201, and RR080.

Figure S.6: We visualize an exemplary interaction with the depth projected onto the RGB observation in red wherever a depth reading was unavailable for the particular image coordinate. As depicted, metallic, semi-transparent, or dark-colored objects do not produce reliable depth estimates, ultimately complicating the lifting of 3D point trajectories used to estimate the underlying articulation.

Table S.5: Overview of all Arti4D demonstration sequences: We assign sequence identifiers to each recording and list the number of objects interacted with per sequence as well as the distribution of prismatic vs. revolute joints (PRISM / REV) and objects classified as either EASY or HARD. The number of objects corresponds to the number of annotated axes per sequence, whereas the number of interactions denotes the number of performed articulations, thus including objects that are articulated multiple times per sequence.

| *Scene* | *Sequence ID* | *Recording* | *# Objects* | *# Interactions* | # PRISM / REV | # EASY / HARD |
|---|---|---|---|---|---|---|
| RH078 | RH078-00 | scene_2025-04-04-19-14-38 | 7 | 7 | 3 / 4 | 7 / 0 |
| | RH078-01 | scene_2025-04-04-19-18-54 | 6 | 7 | 2 / 4 | 4 / 2 |
| | RH078-02 | scene_2025-04-07-11-39-17 | 7 | 7 | 3 / 4 | 7 / 0 |
| | RH078-03 | scene_2025-04-07-11-41-52 | 8 | 9 | 3 / 5 | 6 / 2 |
| | RH078-04 | scene_2025-04-07-11-48-40 | 7 | 7 | 3 / 4 | 7 / 0 |
| | RH078-05 | scene_2025-04-09-10-30-11 | 8 | 8 | 5 / 3 | 3 / 5 |
| | RH078-06 | scene_2025-04-09-10-32-52 | 6 | 6 | 5 / 1 | 3 / 3 |
| | RH078-07 | scene_2025-04-09-10-35-47 | 7 | 7 | 6 / 1 | 2 / 5 |
| | RH078-08 | scene_2025-04-09-10-38-38 | 7 | 7 | 5 / 2 | 3 / 4 |
| | RH078-09 | scene_2025-04-09-10-46-48 | 7 | 8 | 7 / 0 | 4 / 3 |
| | RH078-10 | scene_2025-04-09-10-49-20 | 7 | 7 | 6 / 1 | 2 / 5 |
| RR080 | RR080-00 | scene_2025-04-10-13-11-16 | 17 | 17 | 14 / 3 | 9 / 8 |
| | RR080-01 | scene_2025-04-10-16-05-09 | 14 | 14 | 13 / 1 | 8 / 6 |
| | RR080-02 | scene_2025-04-17-15-25-14 | 11 | 12 | 11 / 0 | 7 / 4 |
| | RR080-03 | scene_2025-04-17-15-33-44 | 9 | 9 | 9 / 0 | 7 / 2 |
| | RR080-04 | scene_2025-04-22-09-53-49 | 8 | 8 | 8 / 0 | 5 / 3 |
| | RR080-05 | scene_2025-04-22-09-56-24 | 10 | 10 | 9 / 1 | 9 / 1 |
| | RR080-06 | scene_2025-04-22-09-58-49 | 7 | 8 | 7 / 0 | 4 / 3 |
| | RR080-07 | scene_2025-04-22-11-45-15 | 9 | 9 | 8 / 1 | 7 / 2 |
| | RR080-08 | scene_2025-04-22-11-48-01 | 9 | 9 | 8 / 1 | 7 / 2 |
| | RR080-09 | scene_2025-04-22-11-50-40 | 8 | 8 | 7 / 1 | 6 / 2 |
| DR080 | DR080-00 | scene_2025-04-11-11-44-32 | 11 | 11 | 7 / 4 | 5 / 6 |
| | DR080-01 | scene_2025-04-11-12-58-58 | 10 | 10 | 6 / 4 | 5 / 5 |
| | DR080-02 | scene_2025-04-11-13-01-59 | 9 | 9 | 5 / 4 | 4 / 5 |
| | DR080-03 | scene_2025-04-11-13-18-00 | 9 | 10 | 4 / 5 | 3 / 6 |
| | DR080-04 | scene_2025-04-11-13-43-03 | 11 | 11 | 7 / 4 | 5 / 6 |
| | DR080-05 | scene_2025-04-11-14-01-06 | 11 | 11 | 7 / 4 | 5 / 6 |
| | DR080-06 | scene_2025-04-11-15-43-24 | 11 | 12 | 7 / 4 | 5 / 6 |
| | DR080-07 | scene_2025-04-11-15-46-48 | 11 | 11 | 6 / 5 | 4 / 7 |
| RH201 | RH201-00 | scene_2025-04-24-17-52-21 | 11 | 11 | 5 / 6 | 6 / 5 |
| | RH201-01 | scene_2025-04-24-17-54-13 | 9 | 9 | 5 / 4 | 5 / 4 |
| | RH201-02 | scene_2025-04-24-19-18-42 | 11 | 11 | 5 / 6 | 7 / 4 |
| | RH201-03 | scene_2025-04-24-19-21-50 | 8 | 8 | 2 / 6 | 5 / 3 |
| | RH201-04 | scene_2025-04-24-19-24-09 | 8 | 8 | 5 / 3 | 5 / 3 |
| | RH201-05 | scene_2025-04-25-10-36-37 | 9 | 9 | 5 / 4 | 7 / 2 |
| | RH201-06 | scene_2025-04-25-10-53-40 | 8 | 8 | 4 / 4 | 3 / 5 |
| | RH201-07 | scene_2025-04-25-10-56-33 | 8 | 8 | 3 / 5 | 3 / 5 |
| | RH201-08 | scene_2025-04-25-11-11-47 | 15 | 16 | 6 / 9 | 6 / 9 |
| | RH201-09 | scene_2025-04-25-11-15-47 | 7 | 7 | 5 / 2 | 4 / 3 |
| | RH201-10 | scene_2025-04-25-14-58-42 | 9 | 9 | 6 / 3 | 4 / 5 |
| | RH201-11 | scene_2025-04-25-15-02-14 | 7 | 7 | 3 / 4 | 4 / 3 |
| | RH201-12 | scene_2025-04-25-15-04-48 | 7 | 7 | 4 / 3 | 3 / 4 |
| | RH201-13 | scene_2025-04-25-15-16-29 | 9 | 9 | 5 / 4 | 3 / 6 |
| | RH201-14 | scene_2025-04-25-15-19-22 | 10 | 10 | 5 / 5 | 5 / 5 |
| | RH201-15 | scene_2025-04-25-15-22-54 | 8 | 8 | 4 / 4 | 3 / 5 |

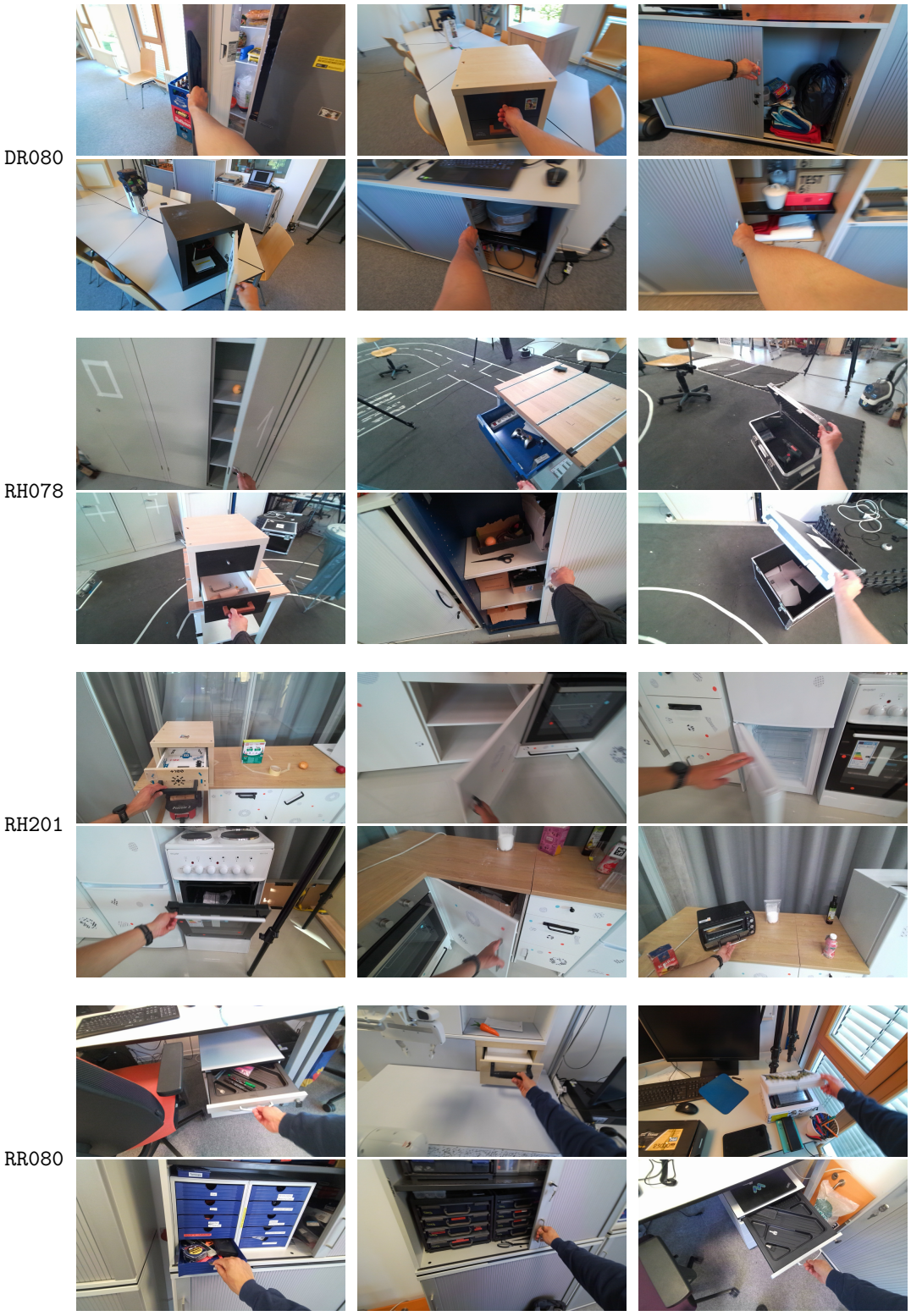



Figure S.7: We visualize several object interactions across the four different environments (DR080, RH078, RH201, RR080) captured as part of the Arti4D dataset.