This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

FUNREC
Reconstructing Functional 3D Scenes from Egocentric Interaction Videos

Alexandros Delitzas[1,2] Chenyangguang Zhang[1] Alexey Gavryushin[1]
Tommaso Di Mario[1] Boyang Sun[1] Rishabh Dabral[2] Leonidas Guibas[3]
Christian Theobalt[2] Marc Pollefeys[1,4] Francis Engelmann[3,5] Daniel Barath[1]
[1]ETH Zurich [2]Max Planck Institute for Informatics [3]Stanford University [4]Microsoft [5]USI Lugano

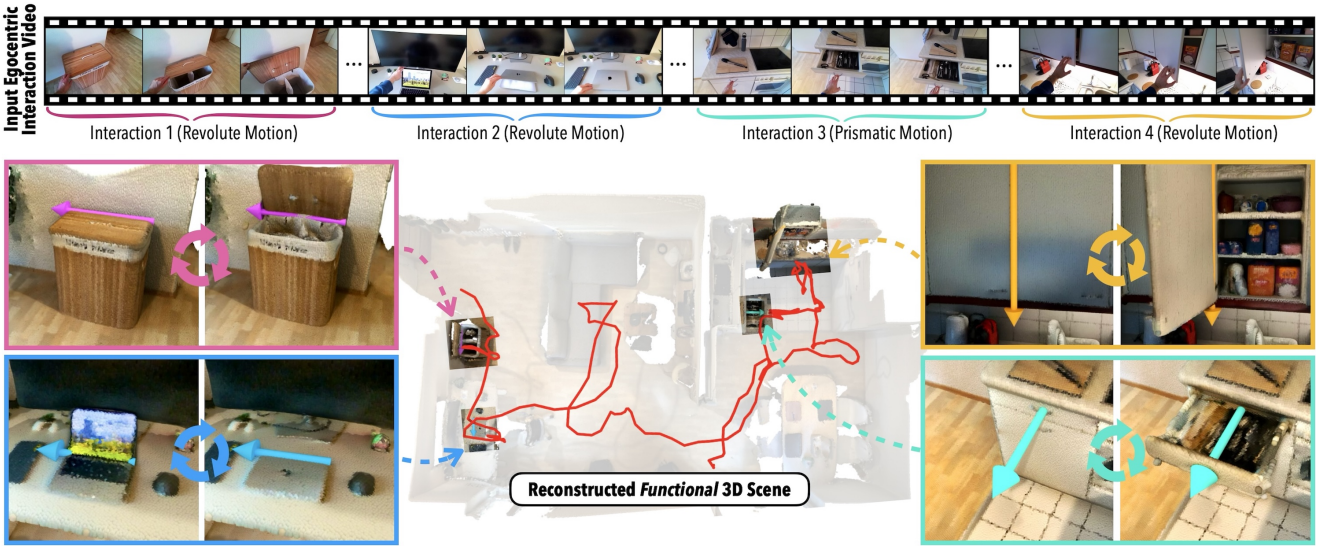


Figure 1. **Real-world functional digital twins.** FUNREC takes a single egocentric RGB-D interaction video *(top)* and reconstructs a functional 3D digital twin of the environment *(middle)*. The system automatically identifies articulated scene components, estimates their kinematic parameters along with per-timestep poses, and jointly reconstructs the static scene and each movable part, including interiors *(see left and right)*. The final output is a simulation-compatible 3D scene representation with fully interactable articulated elements.

Abstract

We present FunREC, a method for reconstructing functional 3D digital twins of indoor scenes directly from egocentric RGB-D interaction videos. Unlike existing methods on articulated reconstruction, which rely on controlled setups, multi-state captures, or CAD priors, FunREC operates directly on in-the-wild human interaction sequences to recover interactable 3D scenes. It automatically discovers articulated parts, estimates their kinematic parameters, tracks their 3D motion, and reconstructs static and moving geometry in canonical space, yielding simulation-compatible meshes. Across new real and simulated benchmarks, FunREC surpasses prior work by a large margin, achieving up to +50 mIoU improvement in part segmentation, 5-10× lower articulation and pose errors, and significantly higher reconstruction accuracy. We further demonstrate applications on URDF/USD export for simulation, hand-guided affordance mapping and robot-scene interaction. Our project page is: functionalscenes.github.io.

1. Introduction

Humans make sense of the world not merely through observation, but through *interacting* with it. Reconstructing *functional* 3D environments, capturing not only static geometry but also how objects move and articulate, is a central goal in computer vision, robotics, and embodied AI. While recent progress in 3D reconstruction and large-scale RGB-D datasets [3, 6, 7, 24, 57, 58, 61, 73, 76] has advanced static scene understanding [43, 45], these datasets represent only a *single state* of each environment. They fail to capture how scenes change under human interaction (*e.g.*, doors opening, drawers sliding, fridges revealing interiors) which is essential for agents that must perceive, plan, and act in the physical world. This missing notion of *functionality* remains a key limitation in 3D scene reconstruction.

Recent works have taken steps toward addressing this challenge, but notable limitations persist. MultiScan [44] manually aligns multiple scans of the same room in different states (*e.g.*, cabinet open/closed) to annotate articulated

parts. SceneFun3D [11] and Articulate3D [18] enrich static LiDAR scans with fine-grained functional and affordance annotations. Parallel work on "digital cousins" [8, 22, 77] retrieves CAD proxies resembling static reconstructions to obtain simulated, interactive counterparts. While these efforts produce interactive scenes, they rely on labor-intensive multi-state captures, manual annotations, or proxy reconstruction, only weakly tied to the actual 3D geometry. At the object level, several works [16, 20, 26, 29, 37, 39, 42, 69, 71] recover articulated objects from demonstrations, but typically assume controlled setups, fixed cameras, or known CAD models. They remain limited to object-centric or synthetic scenarios, and none provides an end-to-end solution for reconstructing *scene-scale, physically grounded* digital twins directly from real, egocentric interaction videos.

Our motivation is that human interaction provides the most direct and rich supervision for functional scene reconstruction. As people move and manipulate their surroundings, egocentric observations naturally reveal which parts articulate, around what joints, what volumes are exposed, and the associated affordances. Building on this insight, we introduce FunREC, a system that reconstructs a coherent, articulated 3D digital twin of a real environment directly from a single egocentric RGB-D interaction video.

FunREC automatically detects articulated parts, estimates their articulation parameters, tracks scene and part motion jointly, and reconstructs both static and moving geometry, including occluded interiors. The result is a physically consistent, interactable 3D scene in which articulated components can be continuously manipulated along their inferred motion axes. At its core, FunREC is a training-free, optimization-based pipeline that integrates geometric reasoning with semantic and motion priors from foundation models. It decomposes the input video into short fragments, identifies interactions via a video-language model, clusters motion trajectories into articulated components, and derives pixel-accurate interacted-part masks. We then jointly optimize part poses and articulation parameters, and fuse TSDF reconstructions using the estimated camera and part poses to obtain globally aligned functional digital twins.

We evaluate FunREC on HOI4D [40], which is recorded in the lab showing a single object interaction, and we introduce two new egocentric interaction datasets that show a more realistic setup in real-world scenes: RealFun4D, containing 351 real interaction videos from indoor spaces across 60 apartments in four countries, and OmniFun4D, comprising 127 photorealistic sequences rendered in 12 OmniGibson scenes [13, 56]. Across all three datasets, FunREC outperforms state-of-the-art baselines in motion estimation, segmentation, and reconstruction quality. Finally, we also demonstrate the practical applications of our method for URDF/USD export for simulation, affordance mapping, and robot-scene interaction.

2. Related Work

Static and Interactive 3D Scene Understanding. Over recent years, substantial progress has been made across a range of 3D scene understanding tasks, such as segmentation [31, 52], or object detection [33, 53]. This progress has been enabled in large part by the availability of large-scale 3D datasets [3, 6, 7, 76]. While these datasets rely on *static* scans, some have begun to include changes over time, *e.g.*, furniture rearrangements in RIO [61] or structural evolution on construction sites in NSS [59]. More recently, the field has turned towards modeling *articulations* and *functionalities* in 3D scenes [11, 18, 44]. MultiScan [44] annotates articulated object parts (*e.g.*, cabinet doors and drawers) by capturing each room twice, once with objects closed and once opened, and manually matching parts across states. SceneFun3D [11] and Articulate3D [18] focus on understanding functionality and affordances, with detailed annotations on high-resolution LiDAR scans that capture fine-grained interactive elements such as knobs and switches. However, because these scenes are static, the causal and kinematic properties of articulated objects cannot directly be observed. A related direction seeks to reconstruct *digital cousins* of real environments [8, 22, 77] by retrieving synthetic proxies that resemble the observed scenes. These approaches aim to provide functional digital replicas but rely on similarity-based substitution rather than observed physical interaction. In this work, we introduce a complementary paradigm for capturing functional, *causal* 3D scene replicas. Instead of inferring articulation from static geometry, we reconstruct the scene and recover functionalities and kinematic properties directly from observed interactions.

3D Articulated Object Reconstruction. Modeling interactive scenes begins with understanding their fundamental subcomponents: the objects themselves. A substantial body of work has focused on articulated 3D objects [38]. Most approaches [26, 37, 42, 69, 70] recover articulation by capturing multi-view RGB(-D) observations across discrete object states and leveraging these to infer articulation parameters and reconstruct part-level geometry. However, the reliance on multi-view capture makes the data acquisition complex and limits practicality. Another line of research integrates vision and language models to infer articulation by generating executable code [82] or retrieving CAD models from visual input [34], while others [16, 20, 23, 25, 39, 60, 72] operate from single observations. Nevertheless, the generalization of these works to in-the-wild settings is hindered by inaccurate CAD model retrievals or the reliance on 2D articulation estimators trained on small-scale datasets. More recently, a new family of methods has emerged that jointly capture geometry and motion from a human demonstration video [29, 71]. However, these approaches rely on demonstrations recorded with a

static camera and require pre-scanned 3D object models, which restricts their applicability in unconstrained, real-world settings. Concurrently with our work, iTACO [50] and VideoArtGS [41] reconstruct articulated objects from dynamic videos, but operate in object-centric setups and do not address scene-scale reconstruction in the wild.

Tracking Interactions in Dynamic 3D Scenes. While several impactful datasets have been proposed for dynamic object interactions ranging from rigid [1, 32, 40, 62] to articulated objects [14, 30], they primarily operate in a controlled, table-top setting with objects artificially scattered in the scene. Consequently, their applicability to in-the-wild scenes, *e.g.*, real apartments, remains limited. Recent progress in 6D object pose tracking [66–68] has enabled several works that track objects and scene changes from egocentric videos [4, 17, 80] in real scenes. Yet, they either only track rigid changes, or assume a given 3D scene scan pre-labeled with part-wise motions. Several recent works have focused on monocular 4D reconstruction [35, 36, 64, 65, 78, 81]. Typically trained to reconstruct the dynamic point maps and camera trajectories in a feed-forward fashion, such methods do not understand the scene and object semantics, let alone the articulation parameters of the individual objects. Furthermore, they often operate on short video sequences and struggle with depth inconsistencies and flickering artifacts in the presence of occlusions, which are routinely encountered during in-the-wild egocentric scene interactions. While most methods do not predict 3D motion tracks, the ones that do [15, 75] struggle in the presence of occlusions due to interactions. Meanwhile, 2D and 3D point trackers [19, 27, 28, 46, 54, 63, 74, 79] provide useful motion priors but can be noisy. How to leverage them to robustly track the evolution of scene states in real-world articulated 3D environments remains an open question. We propose a training-free approach that leverages the semantic and motion priors of foundational models to robustly *discover* functionalities and track scene states from casually captured egocentric interaction videos.

3. Proposed Method

Problem formulation. Let $\mathcal{V} = \{(I_i, D_i)\}_{i=1}^{N}$ denote an RGB-D video consisting of N frames, where I_i and D_i represent the RGB image and the depth map at time i, respectively, and camera intrinsics are known. The scene is composed of a static background geometry $\mathcal{P}^s$ and a set of K articulated parts $\{\mathcal{P}^{m_k}\}_{k=1}^{K}$. Each part m_k is parameterized by a set of articulation parameters ϕ_k and a sequence of rigid transformations $\{T_i^{m_k}\}_{i=1}^{N}$, where each $T_i^{m_k} \in \mathrm{SE}(3)$ maps the part from its canonical coordinate frame to the world frame at time i. We model two articulation types:

Prismatic joints. For sliding object parts such as drawers, we define a unit translation axis $\mathbf{a} \in \mathbb{S}^2$ and a scalar displacement $\lambda_i \in \mathbb{R}$ at frame i, where $\mathbb{S}^2$ is the unit sphere. The transformation at frame i is then given by:

$$\mathcal{T}_{\text{pris}}(\mathbf{a}, \lambda_i) = \begin{bmatrix} I & \lambda_i \mathbf{a} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \tag{1}$$

where I is the 3×3 identity matrix.

Revolute joints. For rotating object parts such as doors, we define a unit rotation axis $\mathbf{a} \in \mathbb{S}^2$, a pivot point $\mathbf{p} \in \mathbb{R}^3$ lying on the axis and closest to the scene origin (*i.e.*, $\mathbf{a}^\top \mathbf{p} = 0$), and a rotation angle $\theta_i \in \mathbb{S}^1$ at frame i, where $\mathbb{S}^1$ is the unit circle. The corresponding transformation is:

$$\mathcal{T}_{\text{rev}}(\mathbf{a}, \mathbf{p}, \theta_i) = \begin{bmatrix} R(\mathbf{a}, \theta_i) & (I - R(\mathbf{a}, \theta_i))\,\mathbf{p} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \tag{2}$$

where $R(\mathbf{a}, \theta)$ is the 3×3 rotation matrix around axis $\mathbf{a}$ by angle θ.

Thus, the pose of part m_k at time i is:

$$T_i^{m_k} = \begin{cases} \mathcal{T}_{\text{pris}}(\mathbf{a}_k, \lambda_{i,k}), & \text{if } m_k \text{ is prismatic,} \\ \mathcal{T}_{\text{rev}}(\mathbf{a}_k, \mathbf{p}_k, \theta_{i,k}), & \text{if } m_k \text{ is revolute.} \end{cases}$$

3.1. Fragment Construction

The input video $\mathcal{V}$ is divided into temporally contiguous fragments $\mathcal{V}_k$ (Fig. 2), such that $\mathcal{V} = \bigcup_{k=1}^{K} \mathcal{V}_k$. Each fragment is automatically classified as either *static* (no interaction) or *dynamic* (interaction with an articulated part) using a video-language model (VLM) [10]. For each dynamic fragment, the VLM also predicts the articulation type $\sigma_k \in \{\text{prismatic}, \text{revolute}\}$. Each fragment is processed independently as described below.

3.2. Dynamic Fragment Reconstruction

Camera pose estimation. Dense point correspondences are extracted between consecutive RGB-D frames using RoMA [12]. We filter these matches using per-frame hand and interacted-object masks from VISOR [9]: points whose pixels lie inside the hand mask $\{M_i^h\}$ are discarded, and those inside the interacted-object mask $\{M_i^{obj}\}$ are down-weighted by scaling their RoMA confidence scores. Each surviving 2D correspondence is lifted into 3D, and the relative camera motion between frames is estimated using SupeRANSAC [2]. A fragment-level pose graph optimization ensures globally consistent camera poses $\{T_i^c\}$.

Sparse 3D trajectories. We obtain sparse 3D point trajectories directly in the world frame using TAPIP3D [79]. This produces track positions $\tau \in \mathbb{R}^{T\times N\times 3}$ and per-frame visibility scores $o \in \mathbb{R}^{T\times N}$, where T is the number of tracked points. New tracks are periodically initialized on a uniform 2D grid to capture surfaces revealed during interaction. Having established sparse 3D trajectories, next we assign tracks to the interacted articulated part.

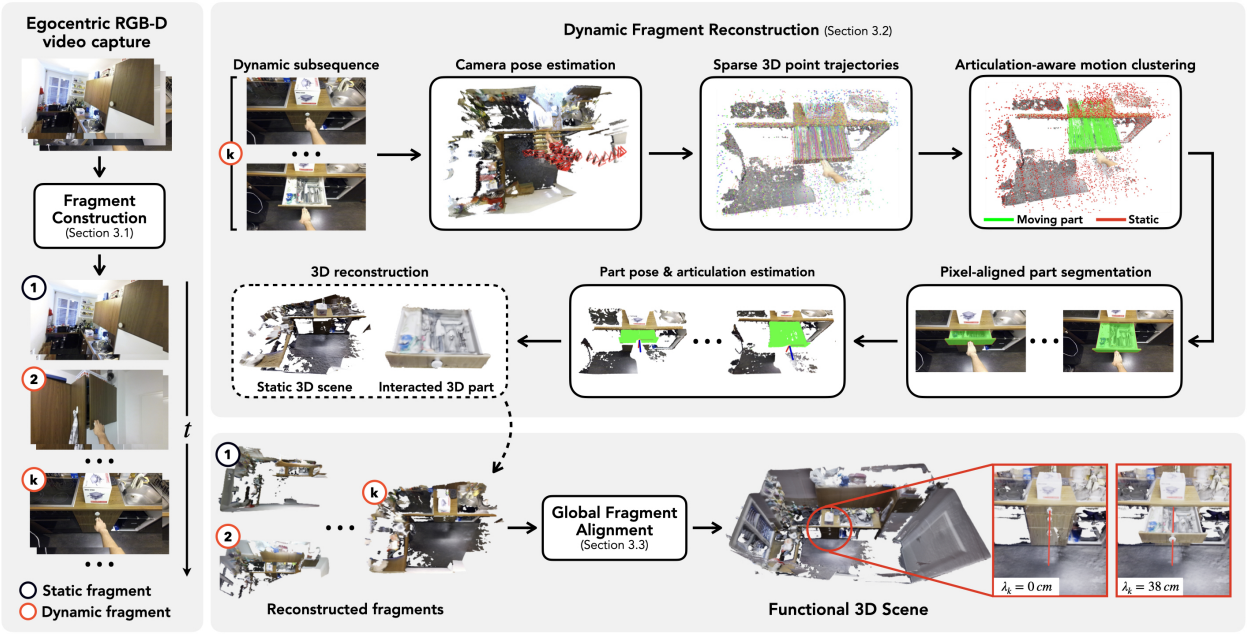


Figure 2. **Method overview**. Given an egocentric RGB-D interaction video, FunREC first divides it into static and dynamic fragments. For each dynamic fragment, it estimates camera poses, computes sparse 3D point trajectories, and clusters them into articulated components via articulation-aware motion modeling. The interacting part is then segmented to obtain dense masks and reconstructed together with the static scene. Part pose and articulation parameters are optimized jointly to yield consistent motion across frames. Finally, all reconstructed fragments are globally aligned to produce a coherent, functional 3D digital twin where articulated parts can be interactively manipulated.

Articulation-aware motion clustering. Let $\tau_l = \{\tau_{l,i}\}_{i=1}^{N}$ denote the 3D coordinates of the l-th track. Tracks with negligible motion are discarded by thresholding their maximal displacement with ϵ_s. For each remaining track, we estimate a per-track articulation hypothesis $\hat{\phi}_l$ consistent with the fragment-wide articulation type σ_k as:

$$\hat{\phi}_l = \begin{cases} (\hat{\mathbf{a}}_l, \{\hat{\lambda}_{l,i}\}), & \text{if } \sigma_k = \text{prismatic}, \\ (\hat{\mathbf{a}}_l, \hat{\mathbf{p}}_l, \{\hat{\theta}_{l,i}\}), & \text{if } \sigma_k = \text{revolute}. \end{cases}$$

The motion of each track is fitted by a 3D line (prismatic) or circle (revolute). We compute predicted track positions $\hat{\tau}_{l,i}$ by transforming the initial point $\tau_{l,1}$ with the estimated motion parameters as follows:

$$\hat{\tau}_{l,i} = \begin{cases} \mathcal{T}_{\text{pris}}(\hat{\mathbf{a}}_l, \hat{\lambda}_{l,i})\tau_{l,1}, & \text{if prismatic}, \\ \mathcal{T}_{\text{rev}}(\hat{\mathbf{a}}_l, \hat{\mathbf{p}}_l, \hat{\theta}_{l,i})\tau_{l,1}, & \text{if revolute}. \end{cases}$$

A track is retained if its average fitting error satisfies

$$\frac{1}{|\tau_l|} \sum_{\tau_{l,i} \in \tau_l} \|\tau_{l,i} - \hat{\tau}_{l,i}\|_2 < \epsilon_f.$$

Tracks passing this filter are clustered using HDBSCAN [5] according to the similarity of their fitted joint parameters (axis, pivot, and motion pattern), forming motion clusters that represent independently moving parts.

To identify the cluster corresponding to the manipulated part, we compare clusters against 2D interaction evidence. Let $\mathcal{C}_i^{obj}$ be the confidence associated with the interacted-object mask $\mathcal{M}_i^{obj}$ at frame i. For each cluster γ, we compute a consistency score as:

$$s_\gamma = \sum_{l \in \gamma} \sum_i o_{l,i} \cdot \mathcal{C}_i^{obj} \cdot \mathbb{I}\left[\pi(\tau_{l,i}) \in \mathcal{M}_i^{obj}\right], \tag{3}$$

where $\pi(\cdot)$ denotes projection to the image plane. The cluster with the highest score s_γ is selected as the interacted part, yielding the moving track set τ^m and the static set τ^s.

Pixel-aligned part segmentation. Given the moving track set τ^m, our goal is to obtain a dense, pixel-aligned segmentation mask of the articulated part for each frame. Since τ^m is sparse and may contain outliers, directly projecting tracks or prompting a segmentation model is unreliable. To obtain a robust segmentation, we combine geometric evidence from the tracks with image-level semantic grouping.

We first select a set of keyframes $\{I_q\}_{q=1}^{Q}$ uniformly across the fragment. On each keyframe I_q, we apply SAM's automatic mask generator to produce an over-segmentation $G_q : \{1, \dots, H\} \times \{1, \dots, W\} \to \mathbb{N}$, where $G_q(p)$ assigns each pixel p to a semantic region (mask) hypothesis, and H, W denote the image height and width. The static and moving 3D tracks are projected into the keyframe using the corresponding camera pose T_q^c as $\pi(\tau_q^m)$ and $\pi(\tau_q^s)$, where $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ is the projection function.

For each semantic region r in G_q, we count the number of projected moving and static tracks contained within it:

$$n_r^m = \sum_{p \in r} \mathbb{I}\left[p \in \pi(\tau_q^m)\right], \quad n_r^s = \sum_{p \in r} \mathbb{I}\left[p \in \pi(\tau_q^s)\right], \tag{4}$$

where $\mathbb{I}$ is an indicator. We compute a motion ratio, quantifying if the current region contains the moving part, as $\gamma_r = n_r^m/(n_r^m + n_r^s + \epsilon)$, where ϵ prevents division by 0. Regions with $\gamma_r > \eta_m$ are classified as belonging to the moving articulated part, where η_m is a threshold. This produces a motion mask $\mathcal{M}_q^{sm}$ on each selected keyframe q as:

$$\mathcal{M}_q^{sm}(p) = \begin{cases} 1, & \text{if } G_q(p) \in \{r \mid \gamma_r > \eta_m\}, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The keyframe motion masks $\{\mathcal{M}_q^{sm}\}$ serve as prompts for SAM2's video propagation module [55], producing a temporally consistent sequence of articulated part masks $\{\mathcal{M}_i^m\}_{i=1}^N$. These masks accurately delineate the articulated part across the entire fragment, providing the pixel-level support necessary for dense part reconstruction in subsequent steps.

Part pose and articulation estimation. Given the moving track set $\tau^m = \{\tau_{l,i}^m\}$ and corresponding visibility scores $o^m = \{o_{l,i}^m\}$, our goal is to recover the globally consistent part poses $\{T_i^m\}_{i=1}^N$ and articulation parameters ϕ^m describing the motion of the part across all frames. Each pose $T_i^m \in \mathrm{SE}(3)$ maps the canonical part coordinate frame to the world frame at time i. The articulation parameters ϕ^m encode the joint model, defined as $(\mathbf{a}, \{\lambda_i\})$ for prismatic motion or $(\mathbf{a}, \mathbf{p}, \{\theta_i\})$ for revolute motion.

For each pair of frames (i, j) within a fragment, we construct 3D-3D correspondences between visible points in τ^m and estimate the relative transformation of the part, denoted $T_{i\to j}^m \in \mathrm{SE}(3)$, using SupeRANSAC [2]. Each correspondence is weighted by the product of its visibility scores $o_{l,i}^m \cdot o_{l,j}^m$. The resulting set of relative transforms forms a pose graph connecting all part poses $\{T_i^m\}$ in the fragment.

To jointly recover the absolute part poses and articulation parameters, we minimize the following objective:

$$\mathcal{L}(T^m, L^m, \phi^m) = \sum_i f(T_i^m, T_{i+1}^m, T_{i\to i+1}^m) + \sum_{i,j} l_{ij}^m \, f(T_i^m, T_j^m, T_{i\to j}^m) + \mu \sum_{i,j} (\sqrt{l_{ij}^m} - 1)^2, \tag{6}$$

where $l_{ij}^m \in [0, 1]$ are optimized loop-closure confidences, and μ controls their regularization. The term

$$f(T_i^m, T_j^m, T_{i\to j}^m) = e_{ij}^\top \Omega_{ij} e_{ij},$$

where $e_{ij} = \log_{\mathrm{SE}(3)}\left((T_i^m)^{-1} T_j^m T_{i\to j}^m\right)$ measures the discrepancy of the estimated relative transformation $(T_i^m)^{-1} T_j^m$ and the observed transformation $T_{i\to j}^m$, weighted by the information matrix Ω_{ij}. The logarithm $\log_{\mathrm{SE}(3)}(\cdot)$ maps the residual to the tangent space of $\mathrm{SE}(3)$.

The articulation parameters ϕ^m are initialized via least-squares fitting to the observed tracks τ^m, providing an initial estimate of the joint axis and motion states. Both $\{T_i^m\}$ and ϕ^m are then jointly refined through non-linear optimization using Ceres Solver, employing manifold optimization to ensure the articulation parameters remain on their respective manifolds as defined in our problem formulation Sec. 3.

Reconstruction and interactive manipulation. Given the per-pixel segmentation masks and estimated poses, we reconstruct the geometry using two separate truncated signed distance function (TSDF) volumes: one for the static background and one for the articulated part. The static TSDF volume is integrated in the world coordinate frame using the estimated camera poses $\{T_i^c\}$, while excluding dynamic regions corresponding to hand and moving-part pixels, as indicated by binary masks $\mathcal{M}_i^h$ and $\mathcal{M}_i^m$, respectively. The articulated part is reconstructed in its canonical coordinate frame. For each frame i, the depth map is first transformed from the camera frame into the part frame using $(T_i^m)^{-1} T_i^c$, and only pixels belonging to the moving-part mask $\mathcal{M}_i^m$ are fused. This yields a clean canonical 3D model $\mathcal{P}^m$, free from camera motion and occlusions. After both TSDF volumes are fused, meshes are extracted for the static scene $\mathcal{P}^s$ and articulated part $\mathcal{P}^m$. The complete scene at time i is obtained as: $\mathcal{P}_i = \mathcal{P}^s \cup T_i^m(\mathcal{P}^m)$, where $T_i^m(\mathcal{P}^m)$ applies the estimated pose of the part to its canonical geometry, placing it in the global frame. To support interactive manipulation, we estimate the feasible motion range of the articulation parameters from the tracked motion trajectory: $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ or $\theta \in [\theta_{\min}, \theta_{\max}]$, allowing the reconstructed scene to be rendered or physically simulated at any intermediate state consistent with the articulation ϕ^m.

3.3. Global Fragment Alignment

Each dynamic fragment yields a local submap $\mathcal{S}_k$ containing its reconstructed static geometry and articulated parts. To produce static fragment submaps, we perform only camera pose estimation, since the scene remains static. To form a globally consistent scene, we align all submaps $\{\mathcal{S}_k\}_{k=1}^K$ within a shared coordinate frame. For each submap pair $(\mathcal{S}_k, \mathcal{S}_{k'})$, geometric correspondences are extracted, and a relative rigid transformation is estimated using PREDATOR [21]. Loop closures are accepted only if the point alignment root-mean-square error (RMSE) falls below a predefined threshold. These relative transformations define a scene-level pose graph over the submaps, which we optimize to obtain globally consistent submap poses. The aligned submaps are finally fused into a unified TSDF volume, from which we extract the complete static scene $\mathcal{P}^s$ and the set of all articulated part meshes $\{\mathcal{P}^{m_k}\}$.

4. Data Collection

HOI4D [40] is the only prior dataset for our task but offers low-motion, single-object interactions and no full scenes. Thus, we collect two new egocentric datasets with realistic, diverse interactions in real and simulated scenes (Fig. 3).

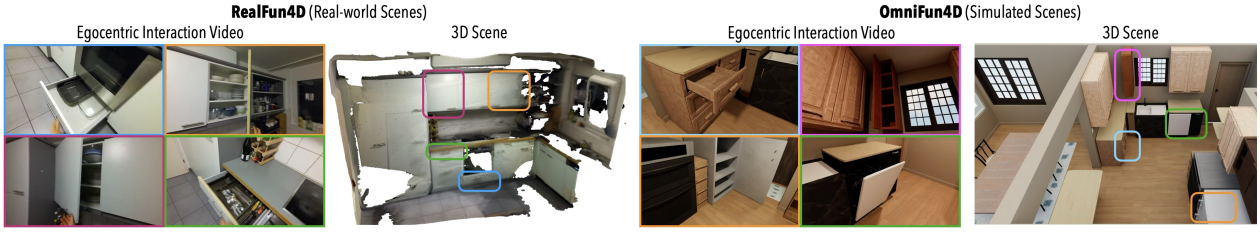


Figure 3. **Proposed Datasets.** We introduce two datasets for functional 3D scene reconstruction and evaluation: RealFun4D *(left)*, capturing egocentric interactions in real scenes, and OmniFun4D *(right)*, providing photorealistic simulated interactions in synthetic scenes.

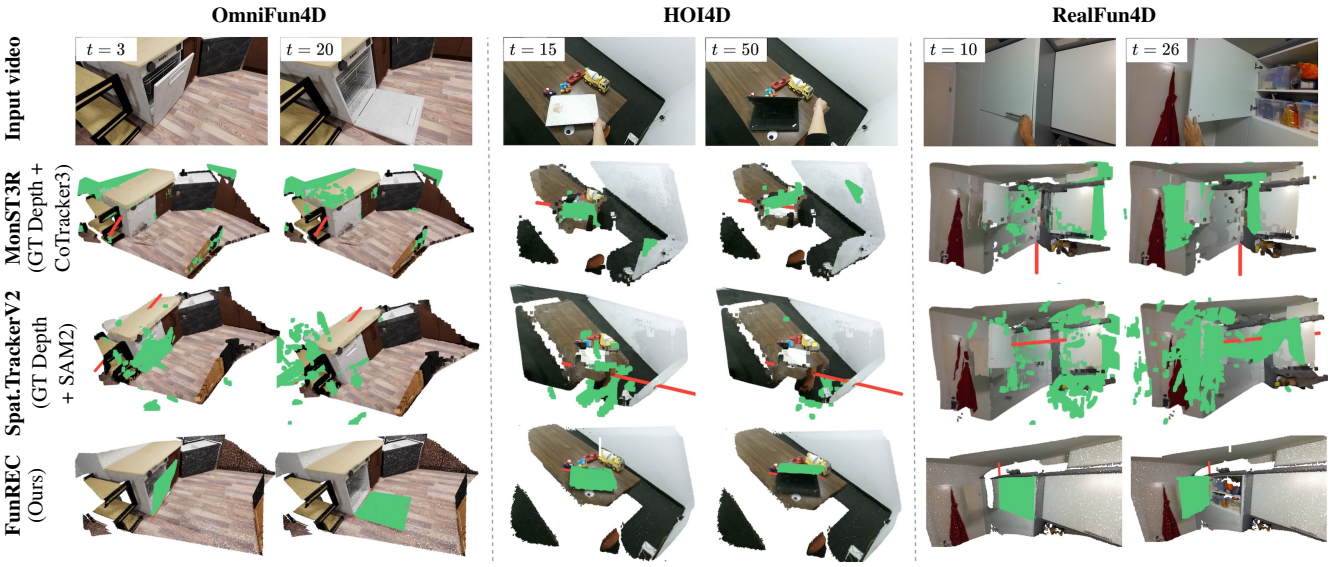


Figure 4. **Qualitative comparisons**. We show qualitative comparisons between baselines and our method. For each method, we accumulate the reconstructed point clouds of both the articulated part and the static scene across all timesteps, and visualize them under two selected scene states. Green indicates the articulated part, and red lines denote the estimated articulation axes.

Real-world scenes: RealFun4D. Our real dataset contains 351 *in-the-wild* human-scene interactions recorded across 60 apartments in four countries. Each sequence is captured with a head-mounted Azure Kinect DK (1920 × 1080, 15 FPS), providing synchronized RGB and depth. We annotate interaction intervals and textual descriptions, label per-frame 2D hand and part masks for dynamic-region filtering and static-scene reconstruction [40], and mark articulation joints along with per-frame 2D part tracks to enable alignment and full 3D part reconstruction.

Simulated scenes: OmniFun4D. We further record 127 interactions in 12 OmniGibson [13] scenes derived from iGibson [56]. A human operator navigates the environment and triggers scripted interactions; camera poses and events are logged and replayed offline to render high-quality RGB-D and masks using NVIDIA RTX Path Tracing [48]. We add stochastic Gaussian perturbations to the camera poses to emulate natural head motion. Further details are provided in the supplementary material.

5. Experiments

Datasets. We evaluate our method on three datasets. The first two, RealFun4D and OmniFun4D, are newly collected as described in Sec. 4. To enable consistent benchmarking with the baselines, we construct a diverse evaluation set of 60 interaction sequences from these datasets, using the annotated interaction intervals. Additionally, we use the HOI4D dataset [40], which contains short, single-object egocentric RGB-D interaction videos. To adapt it to our setting, we extract 30 interactions involving four articulated object categories ("laptop", "cabinet", "safe" and "trash can") and post-process the provided 6D part poses to obtain ground-truth articulation parameters.

Evaluation tasks and metrics. We consider four tasks: articulated motion estimation, 6D part pose estimation, part segmentation, and 3D reconstruction. For articulated motion estimation, we report axis direction and position errors, and per-timestep joint state error with separate results for revolute and prismatic joints. Additionally, we include the

Methods	OmniFun4D				HOI4D				RealFun4D			
	Axis (°)	Pos (m)	State (°/m)	Fail (%)	Axis (°)	Pos (m)	State (°/m)	Fail (%)	Axis (°)	Pos (m)	State (°/m)	Fail (%)
MonST3R [81] (ICP)	52.7/61.5	1.14	66.2/0.27	11.7	67.3/66.1	0.45	60.8/0.06	**0.0**	54.6/59.1	0.68	65.9/0.21	**0.0**
MonST3R (CoTr3)	46.8/58.9	1.20	45.3/0.18	11.7	54.7/51.5	0.44	44.8/0.07	**0.0**	56.8/52.4	0.70	57.4/0.22	**0.0**
MonST3R (GT depth+CoTr3)	57.6/63.5	1.10	46.5/0.18	11.7	48.6/30.1	0.41	37.6/0.06	**0.0**	51.4/56.2	0.55	54.3/0.21	**0.0**
SpatialTrackerV2 [75]	59.8/69.1	1.22	48.6/0.19	30.0	57.9/44.6	0.46	41.9/0.07	**0.0**	59.4/49.8	0.49	57.5/0.22	20.0
SpatialTrackerV2 (GT depth)	48.7/69.3	1.10	46.8/0.19	30.0	54.3/38.8	0.43	38.7/0.07	**0.0**	60.8/42.6	0.58	55.6/0.22	20.0
BundleSDF [67] (GT mask)	38.2/55.9	0.95	23.4/0.20	55.0	26.3/24.5	0.24	9.1/0.07	43.3	32.0/52.1	0.58	15.6/0.22	36.7
ArtGS [42]	64.3/23.4	1.69	–/–	95.0	60.5/53.2	0.35	–/–	66.7	70.1/18.2	0.34	–/–	66.7
FunREC (Ours)	**5.3/ 5.4**	**0.03**	**5.0/0.02**	**1.7**	**12.4/ 1.3**	**0.06**	**9.1/0.02**	**0.0**	**5.6/ 5.5**	**0.05**	**8.4/0.03**	**0.0**

Table 1. **Articulated motion estimation.** We report articulation axis direction error (°), axis position error (meters; applicable only to revolute joints), joint state error (° for revolute and meters for prismatic joints), and failure rate (%). Values are shown as "XX / YY" for revolute (XX) and prismatic (YY) joints. **Best**, second-best, and third-best results are highlighted. All metrics are lower-is-better.

Methods	OmniFun4D			HOI4D			RealFun4D		
	6D part pose		Reconstruction	6D part pose		Reconstruction	6D part pose		Reconstruction
	ADD-S (%) ↑	ADD (%) ↑	CD (cm) ↓	ADD-S (%) ↑	ADD (%) ↑	CD (cm) ↓	ADD-S (%) ↑	ADD (%) ↑	CD (cm) ↓
MonST3R [81] (ICP)	14.36	9.62	29.8	29.41	19.12	6.4	13.21	9.11	13.4
MonST3R [81] (CoTr3)	33.58	21.94	23.6	44.72	25.98	2.1	18.64	11.35	17.6
MonST3R [81] (GT depth+CoTr3)	37.12	30.48	13.9	54.83	37.11	1.3	20.11	14.62	13.2
SpatialTrackerV2 [75]	26.94	13.11	34.7	61.25	41.02	0.9	19.77	13.92	18.5
SpatialTrackerV2 [75] (GT depth)	29.71	15.43	9.88	60.98	40.21	0.8	22.94	15.22	13.5
BundleSDF [67] (GT mask)	22.84	12.37	17.1	53.12	35.22	1.4	20.88	12.91	10.6
ArtGS [42]	–	–	20.3	–	–	6.8	–	–	20.6
FunREC (Ours)	**78.96**	**71.28**	**3.2**	**79.43**	**69.85**	**0.7**	**75.62**	**68.11**	**6.1**

Table 2. **6D part pose estimation and reconstruction.** We report 6D part pose accuracy using ADD-S and ADD (higher is better) and surface reconstruction quality using Chamfer Distance (CD, lower is better) across the OmniFun4D, HOI4D, and RealFun4D datasets.

Methods	OmniFun4D	HOI4D	RealFun4D
	mIoU ↑	mIoU ↑	mIoU ↑
MonST3R [81]	23.6	26.8	23.7
SpatialTrackerV2 [75] (SAM2)	6.2	5.8	13.4
FunREC (Ours)	**77.9**	**76.4**	**74.8**

Table 3. **Moving part segmentation.** Mean Intersection-over-Union (mIoU) for moving-part segmentation.

failure rate, defined as the proportion of videos where the method either failed to process or predicted an incorrect joint type. For 6D part pose estimation, we follow standard evaluation using ADD-S and ADD metrics [66, 67], while 3D surface reconstruction quality is measured by Chamfer Distance (CD). For segmentation of moving parts we report mean Intersection-over-Union (mIoU).

Baselines. We compare against three representative categories: (Type 1) 4D reconstruction pipelines (MonST3R [81], SpatialTrackerV2 [75]), (Type 2) 6D object pose tracking (BundleSDF [67]), and (Type 3) articulated object reconstruction (ArtGS [42]). For MonST3R, we integrate ICP and CoTracker3 [28] to enable moving-part pose tracking. SpatialTrackerV2 is augmented with SAM2 to obtain part segmentations. For both Type 1 baselines, we lift dynamic regions or tracks into 3D and estimate part poses via RANSAC-based model fitting, optionally using ground-truth depth for fair comparison. BundleSDF is given ground-truth camera poses and segmentation masks to recover canonical part geometry and per-frame part transformations. ArtGS operates on two static articulation states and cannot track motion in video. Thus, we report its static reconstruction quality and estimated articulation parameters using ground-truth camera poses. More details are provided in the Supp. Mat.

Articulated motion estimation. Results are reported in Tab. 1, with visual examples in Fig. 4. FunREC achieves the lowest errors across all datasets and motion types. On OmniFun4D, it estimates articulation axis direction within $5.3°$ and position within $0.03\,\mathrm{m}$, outperforming BundleSDF [67] by over $30°$ and an order of magnitude in distance. The same trend holds for HOI4D [40] and RealFun4D, where FunREC reduces the state prediction error to $9.1°/0.02\,\mathrm{m}$ and $8.4°/0.03\,\mathrm{m}$, respectively, with a failure rate of 0%. In contrast, existing dynamic reconstruction pipelines [75, 81] and articulated modeling methods [42] exhibit large deviations and frequent failures. These results highlight the reliability of FunREC in recovering physically consistent joint parameters directly from egocentric videos.

Moving part segmentation. Results are presented in Tab. 3. Across all datasets, FunREC achieves the highest segmentation accuracy, with mIoU scores of 77.9 on OmniFun4D, 76.4 on HOI4D, and 74.8 on RealFun4D. These results represent a substantial improvement over prior methods, including MonST3R [81] (23.6-26.8 mIoU) and SpatialTrackerV2 [75] (5.8-13.4 mIoU), demonstrating consistently accurate delineation of moving parts across synthetic, controlled, and real-world scenes.

6D part pose estimation. Tab. 2 summarizes 6D part pose accuracy, measured via ADD-S and ADD. Across all datasets, FunREC consistently achieves the best results: 79.0% ADD-S / 71.3% ADD on OmniFun4D, 79.4%

/ 69.9% on HOI4D, and 75.6% / 68.1% on RealFun4D. These represent more than a twofold improvement over BundleSDF [67] and a large margin over dynamic trackers [75, 81]. The improvements indicate that the proposed joint optimization of part pose and articulation parameters produces stable, temporally consistent motion trajectories.

Articulated reconstruction. Finally, FunREC delivers high-fidelity articulated 3D reconstructions, as reported by Chamfer Distance (CD) in Tab. 2. It achieves 3.2 cm CD on OmniFun4D, 0.7 cm on HOI4D, and 6.1 cm on RealFun4D, substantially improving over *all* baselines. Examples (Fig. 4) show that FunREC reconstructs both static geometry and articulated parts with accurate motion ranges, enabling continuous re-simulation of interactions in 3D. Overall, these results confirm that our training-free, optimization-based approach generalizes across synthetic, controlled, and real-world egocentric recordings, producing physically consistent and interactable digital twins.

6. Applications

Simulation-ready export. Using the reconstructed geometry and articulation information, we generate files that enable interactive digital replicas of real-world scenes for physical simulation. Specifically, we use the articulation parameters to define revolute and prismatic joints between static structures and movable parts, and export them in URDF or USD format. Physical properties such as mass and inertia can be inferred from RGB images by querying a vision-language model (*e.g.*, GPT-5). This process allows the reconstructed scene to be directly loaded into physics simulators, enabling a wide range of downstream tasks such as robot-scene interaction. Fig. 5 illustrates an example in Isaac Sim [47] where a simulated robot arm interacts with a scene reconstructed by FunREC from real-world scans.

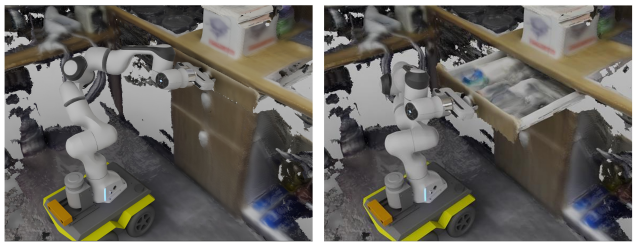


Figure 5. **Isaac Sim deployment.** A mobile manipulator interacts with a drawer reconstructed from a real-world scan.

Hand-guided affordance mapping. As shown in Fig. 6, our framework naturally extends to incorporate hand-centric affordance information. By extracting 3D hand meshes from off-the-shelf estimators [49, 51] and aligning them within our functional digital-twin space, we can localize the hand in 3D and recover its contact regions on the object. This enables joint reasoning over the hand's motion and the motion of the interacted object part.



Figure 6. **Hand-scene interaction.** Estimated 3D hand mesh (*left*) and inferred affordance map (*right*). Integrating the hand pose into our functional reconstruction enables finding contact regions and consistent reasoning over the associated scene-part motion.

Robot-scene interaction from human demonstration. The functional scene model can be directly transferred to a mobile manipulator, enabling robot-scene interaction from human demonstrations. Fig. 7 shows a Boston Dynamics Spot with Arm interacting with articulated objects in the real world. Given the inferred contact points, articulation parameters, and interaction trajectories, the robot can reproduce the same interactions reliably and stably.



Figure 7. **Robot-scene interaction**. *Left:* Human demonstration of opening a cabinet. *Center:* The articulation trajectory derived from the functional scene model. *Right:* The robot leverages the functional information to reliably reproduce the same interaction.

7. Conclusion

We present FunREC, a training-free method to reconstruct functional, articulated 3D digital twins of real environments from a single egocentric RGB-D interaction video. By combining geometric reasoning with semantic and motion priors from foundation models, FunREC jointly estimates camera motion, part articulation, and scene geometry. Our two new egocentric datasets, RealFun4D and OmniFun4D, enable quantitative evaluation and future research on functional scene understanding. Experiments across real and simulated settings demonstrate that FunREC substantially outperforms all baselines in articulation estimation, pose estimation, segmentation, and reconstruction quality, and produces digital twins that can be directly used for simulation, affordance reasoning, and robot-scene interaction.

Acknowledgements. This work was supported by the Swiss National Science Foundation Advanced Grant 216260: "Beyond Frozen Worlds: Capturing Functional 3D Digital Twins from the Real World". The authors also acknowledge the support from a SwissAI Grant for Small Projects and an Academic Grant from NVIDIA. Alexandros Delitzas is also supported by the Max Planck ETH Center for Learning Systems (CLS).

References

[1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[2] Daniel Barath. Superansac: One ransac to rule them all. *arXiv preprint arXiv:2506.04803*, 2025. 3, 5

[3] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. ARKitScenes: A Diverse Real-world Dataset for 3D Indoor Scene Understanding Using Mobile RGB-D Data. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2

[4] Tjark Behrens, René Zurbrügg, Marc Pollefeys, Zuria Bauer, and Hermann Blum. Lost & Found: Tracking Changes from Egocentric Observations in 3D Dynamic Scene Graphs. *IEEE Robotics and Automation Letters (RA-L)*, 2025. 3

[5] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*, 2013. 4

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *International Conference on 3d Vision (3dV)*, 2017. 1, 2

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[8] Tianyuan Dai, Josiah Wong, Yunfan Jiang, Chen Wang, Cem Gokmen, Ruohan Zhang, Jiajun Wu, and Li Fei-Fei. Automated Creation of Digital Cousins for Robust Policy Learning. In *Conference on Robot Learning (CoRL)*, 2024. 2

[9] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3

[10] Google Deepmind. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3

[11] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[12] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[13] Chengshu Li et al. BEHAVIOR-1K: A Human-Centered, Embodied AI Benchmark with 1,000 Everyday Activities and Realistic Simulation. *arXiv preprint arXiv:2403.09227*, 2024. 2, 6

[14] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[15] Haiwen Feng, Junyi Zhang, Qianqian Wang, Yufei Ye, Pengcheng Yu, Michael J. Black, Trevor Darrell, and Angjoo Kanazawa. St4RTrack: Simultaneous 4D Reconstruction and Tracking in the World. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[16] Pradyumn Goyal, Dmitry Petrov, Sheldon Andrews, Yizhak Ben-Shabat, Hsueh-Ti Derek Liu, and Evangelos Kalogerakis. GEOPARD: Geometric Pretraining for Articulation Prediction in 3D Shapes. In *International Conference on Computer Vision (ICCV)*, 2025. 2

[17] Vladimir Guzov, Julian Chibane, Riccardo Marin, Yannan He, Yunus Saracoglu, Torsten Sattler, and Gerard Pons-Moll. Interaction Replica: Tracking human–object interaction and scene changes from human motion. In *International Conference on 3d Vision (3dV)*, 2024. 3

[18] Anna-Maria Halacheva, Yang Miao, Jan-Nico Zaech, Xi Wang, Luc Van Gool, and Danda Pani Paudel. Holistic Understanding of 3D Scenes as Universal Scene Description. In *International Conference on Computer Vision (ICCV)*, 2025. 2

[19] Adam W. Harley, Yang You, Xinglong Sun, Yang Zheng, Nikhil Raghuraman, Yunqi Gu, Sheldon Liang, Wen-Hsuan Chu, Achal Dave, Pavel Tokmakov, Suya You, Rares Ambrus, Katerina Fragkiadaki, and Leonidas J. Guibas. AllTracker: Efficient Dense Point Tracking at High Resolution. In *International Conference on Computer Vision (ICCV)*, 2025. 3

[20] Nick Heppert, Muhammad Zubair Irshad, Sergey Zakharov, Katherine Liu, Rares Andrei Ambrus, Jeannette Bohg, Abhinav Valada, and Thomas Kollar. CARTO: Category and Joint Agnostic Reconstruction of ARTiculated Objects. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[21] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, and Konrad Schindler Andreas Wieser. PREDATOR: Registration of 3D Point Clouds with Low Overlap. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[22] Zhening Huang, Xiaoyang Wu, Fangcheng Zhong, Hengshuang Zhao, Matthias Nießner, and Joan Lasenby. LiteReality: Graphics-Ready 3D Scene Reconstruction from RGB-D Scans. *arXiv preprint arXiv:2507.02861*, 2025. 2

[23] Zhao Huang, Boyang Sun, Alexandros Delitzas, Jiaqi Chen, and Marc Pollefeys. React3d: Recovering articulations for interactive physical 3d scenes. *IEEE Robotics and Automation Letters (RA-L)*, 2026. 2

[24] Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T. Barron, Mario Fritz, Kate Saenko, and Trevor Darrell. A Category-level 3D Object Dataset: Putting the Kinect to Work. In *International Conference on Computer Vision (ICCV) Workshops*, 2011. 1

[25] Hanxiao Jiang, Yongsen Mao, Manolis Savva, and Angel X Chang. OPD: Single-view 3D Openable Part Detection. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[26] Zhenyu Jiang, Cheng-Chun Hsu, and Yuke Zhu. Ditto: Building Digital Twins of Articulated Objects from Interaction. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[27] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is Better to Track Together. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[28] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos. In *International Conference on Computer Vision (ICCV)*, 2025. 3, 7

[29] Justin Kerr, Chung Min Kim, Mingxuan Wu, Brent Yi, Qianqian Wang, Ken Goldberg, and Angjoo Kanazawa. Robot See Robot Do: Imitating Articulated Object Manipulation with Monocular 4D Reconstruction. In *Conference on Robot Learning (CoRL)*, 2024. 2

[30] Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[31] Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. OneFormer3D: One transformer for Unified Point Cloud Segmentation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[32] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two Hands Manipulating Objects for First Person Interaction Recognition. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[33] Justin Lazarow, David Griffiths, Gefen Kohavi, Francisco Crespo, and Afshin Dehghan. Cubify Anything: Scaling Indoor 3D Object Detection. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[34] Long Le, Jason Xie, William Liang, Hung-Ju Wang, Yue Yang, Yecheng Jason Ma, Kyle Vedder, Arjun Krishna, Dinesh Jayaraman, and Eric Eaton. Articulate-anything: Automatic modeling of articulated objects via a vision-language foundation model. In *International Conference on Learning Representations (ICLR)*, 2025. 2

[35] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. MoSca: Dynamic Gaussian Fusion from Casual Videos via 4D Motion Scaffolds. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[36] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, Fast and Robust Structure and Motion from Casual Dynamic Videos. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[37] Jiayi Liu, Ali Mahdavi-Amiri, and Manolis Savva. PARIS: Part-level Reconstruction and Motion Analysis for Articulated Objects. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[38] Jiayi Liu, Manolis Savva, and Ali Mahdavi-Amiri. Survey on Modeling of Human-made Articulated Objects. *arXiv preprint arXiv:2403.14937*, 2025. 2

[39] Xueyi Liu, Ji Zhang, Ruizhen Hu, Haibin Huang, He Wang, and Li Yi. Self-Supervised Category-Level Articulated Object Pose Estimation with Part-Level SE(3) Equivariance. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[40] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 5, 6, 7

[41] Yu Liu, Baoxiong Jia, Ruijie Lu, Chuyue Gan, Huayu Chen, Junfeng Ni, Song-Chun Zhu, and Siyuan Huang. Videoartgs: Building digital twins of articulated objects from monocular video. *arXiv preprint arXiv:2509.17647*, 2025. 3

[42] 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 *International Conference on Learning Representations (ICLR)*, 2025. 2, 7

[43] Xianzheng Ma, Yash Bhalgat, Brandon Smart, Shuai Chen, Xinghui Li, Jian Ding, Jindong Gu, Dave Zhenyu Chen, Songyou Peng, Jia-Wang Bian, Philip H Torr, Marc Pollefeys, Matthias Nießner, Ian D Reid, Angel X. Chang, Iro Laina, and Victor Adrian Prisacariu. When LLMs step into the 3D World: A Survey and Meta-Analysis of 3D Tasks via Multi-modal Large Language Models, 2024. 1

[44] Yongsen Mao, Yiming Zhang, Hanxiao Jiang, Angel X Chang, and Manolis Savva. MultiScan: Scalable RGBD Scanning for 3D Environments With Articulated Objects. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2

[45] Muzammal Naseer, Salman Khan, and Fatih Porikli. Indoor Scene Understanding in 2.5/3D for Autonomous Agents: A Survey. *IEEE Access*, 2018. 1

[46] Tuan Ngo, Peiye Zhuang, Evangelos Kalogerakis, Chuang Gan, Sergey Tulyakov, Hsin-Ying Lee, and Chaoyang Wang. DELTA: Dense Efficient Long-Range 3D Tracking for Any Video. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[47] NVIDIA. Isaac Sim. 8

[48] NVIDIA. NVIDIA® RTX Path Tracing, 2023. 6

[49] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing Hands in 3D With Transformers. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 8

[50] Weikun Peng, Jun Lv, Cewu Lu, and Manolis Savva. iTACO: Interactable Digital Twins of Articulated Objects from Casually Captured RGBD Videos. In *International Conference on 3d Vision (3dV)*, 2026. 3

[51] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 8

[52] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[53] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep Hugh Voting for 3D Object Detection in Point Clouds. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[54] Frano Rajič, Haofei Xu, Marko Mihajlovic, Siyuan Li, Irem Demir, Emircan Gündoğdu, Lei Ke, Sergey Prokudin, Marc Pollefeys, and Siyu Tang. Multi-view 3d point tracking. In *International Conference on Computer Vision (ICCV)*, 2025. 3

[55] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2510.11340*, 2024. 5

[56] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne P. Tchapmi, Micael E. Tchapmi, Kent Vainio, Josiah Wong, Li Fei-Fei, and Silvio Savarese. iGibson 1.0: A Simulation Environment for Interactive Tasks in Large Realistic Scenes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021. 2, 6

[57] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *European Conference on Computer Vision (ECCV)*, 2012. 1

[58] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[59] Tao Sun, Yan Hao, Shengyu Huang, Silvio Savarese, Konrad Schindler, Marc Pollefeys, and Iro Armeni. Nothing Stands Still: A Spatiotemporal Benchmark on 3D Point Cloud Registration Under Large Geometric and Temporal Change. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2025. 2

[60] Xiaohao Sun, Hanxiao Jiang, Manolis Savva, and Angel Chang. OPDMulti: Openable Part Detection for Multiple Objects. In *International Conference on 3d Vision (3dV)*, 2024. 2

[61] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Niessner. RIO: 3D Object Instance Re-Localization in Changing Indoor Environments. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[62] Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. HO-Cap: A Capture System and Dataset for 3D Reconstruction and Pose Tracking of Hand-Object Interaction. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2025. 3

[63] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking Everything Everywhere All at Once. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[64] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of Motion: 4D Reconstruction from a Single Video. In *International Conference on Computer Vision (ICCV)*, 2025. 3

[65] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D Perception Model with Persistent State. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[66] B Wen and Kostas E Bekris. BundleTrack: 6D Pose Tracking for Novel Objects without Instance or Category-Level 3D Models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021. 3, 7

[67] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Muller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. BundleSDF: Neural 6-DoF Tracking and 3D Reconstruction of Unknown Objects. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 7, 8

[68] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[69] Yijia Weng, Bowen Wen, Jonathan Tremblay, Valts Blukis, Dieter Fox, Leonidas Guibas, and Stan Birchfield. Neural Implicit Representation for Building Digital Twins of Unknown Articulated Objects. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[70] 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 *International Conference on Neural Information Processing Systems (NeurIPS)*, 2025. 2

[71] Mingxuan Wu, Huang Huang, Justin Kerr, Chung Min Kim, Anthony Zhang, Brent Yi, and Angjoo Kanazawa. Predict-Optimize-Distill: A Self-Improving Cycle for 4D Object Understanding. In *International Conference on Computer Vision (ICCV)*, 2025. 2

[72] Hongchi Xia, Entong Su, Marius Memmel, Arhan Jain, Raymond Yu, Numfor Mbiziwo-Tiapo, Ali Farhadi, Abhishek Gupta, Shenlong Wang, and Wei-Chiu Ma. Drawer: Digital Reconstruction and Articulation with Environment Realism. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[73] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. SUN3D: A Database of Big Spaces Reconstructed Using SfM and Object Labels. In *International Conference on Computer Vision (ICCV)*, 2013. 1

[74] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. SpatialTracker: Tracking Any 2D Pixels in 3D Space. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[75] Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. Spatialtrackerv2: Advancing 3d point tracking with explicit camera motion. In *International Conference on Computer Vision (ICCV)*, 2025. 3, 7, 8

[76] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A High-fidelity Dataset of 3D Indoor Scenes. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 2

[77] Huangyue Yu, Baoxiong Jia, Yixin Chen, Yandan Yang, Puhao Li, Rongpeng Su, Jiaxin Li, Qing Li, Wei Liang, Zhu Song-Chun, Tengyu Liu, and Siyuan Huang. METASCENES: Towards Automated Replica Creation for Real-world 3D Scans. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[78] Chengbo Yuan, Geng Chen, Li Yi, and Yang Gao. Self-Supervised Monocular 4D Scene Reconstruction for Egocentric Videos. In *International Conference on Computer Vision (ICCV)*, 2025. 3

[79] Bowei Zhang, Lei Ke, Adam W Harley, and Katerina Fragkiadaki. TAPIP3D: Tracking Any Point in Persistent 3D Geometry. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2025. 3

[80] Daiwei Zhang, Gengyan Li, Jiajie Li, Mickaël Bressieux, Otmar Hilliges, Marc Pollefeys, Luc Van Gool, and Xi Wang. EgoGaussian: Dynamic Scene Understanding from Egocentric Video with 3D Gaussian Splatting. In *International Conference on 3d Vision (3dV)*, 2025. 3

[81] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A Simple Approach for Estimating Geometry in the Presence of Motion. In *International Conference on Learning Representations (ICLR)*, 2025. 3, 7, 8

[82] Mandi Zhao, Yijia Weng, Dominik Bauer, and Shuran Song. Real2Code: Reconstruct Articulated Objects via Code Generation. In *International Conference on Learning Representations (ICLR)*, 2025. 2