Articulat3D: Reconstructing Articulated Digital Twins From Monocular Videos with Geometric and Motion Constraints

Lijun Guo*[1], Haoyu Zhao*[2,3], Xingyue Zhao[4], Rong Fu[5], Linghao Zhuang[1], Siteng Huang[6], Zhongyu Li[2,3], and Hua Zou†[1]

[1] Wuhan University, [2] Hong Kong Embodied AI Lab,
[3] The Chinese University of Hong Kong, [4] Peking Union Medical College,
[5] University of Macau, [6] Zhejiang University

Abstract. Building high-fidelity digital twins of articulated objects from visual data remains a central challenge. Existing approaches depend on multi-view captures of the object in discrete, static states, which severely constrains their real-world scalability. In this paper, we introduce Articulat3D, a novel framework that constructs such digital twins from casually captured monocular videos by jointly enforcing explicit 3D geometric and motion constraints. We first propose Motion Prior–Driven Initialization, which leverages 3D point tracks to exploit the low-dimensional structure of articulated motion. By modeling scene dynamics with a compact set of motion bases, we facilitate soft decomposition of the scene into multiple rigidly-moving groups. Building on this initialization, we introduce Geometric and Motion Constraints Refinement, which enforces physically plausible articulation through learnable kinematic primitives parameterized by a joint axis, a pivot point, and per-frame motion scalars, yielding reconstructions that are both geometrically accurate and temporally coherent. Extensive experiments demonstrate that Articulat3D achieves state-of-the-art performance on synthetic benchmarks and real-world casually captured monocular videos, significantly advancing the feasibility of digital twin creation under uncontrolled real-world conditions. Here is our project page.

1 Introduction

Articulated objects, prevalent in our daily life, are a major focus in computer vision and robotics [12, 16, 26, 33, 43]. Reconstructing an interactable digital twin of an articulated object—capturing its part geometry, visual appearance, and articulation parameters—enables a wide range of downstream applications, including augmented reality [39], robotics simulation [41], and interactive scene understanding [13]. Among possible sensing modalities, monocular videos are particularly attractive because they are cheap to acquire at scale and readily

* Equal contributions.
†Corresponding Author.

no multi-state capture	explicit articulation model	physically-grounded geometry	monocular video input	no separate static scan	joint optimization of frames	
✓	✗	✗	✓	✓	✗	Shape of Motion [22]
✓	✓	✗	✗	✓	✗	Articulate Anything [9]
✗	✓	✗	✗	✗	✗	PARIS [11]
✗	✓	✓	✗	✗	✗	ArtGS [14]
✗	✓	✓	✗	✗	✗	ArticulatedGS [4]
✓	✓	✓	✗	✗	✗	iTACO [18]
✓	✓	✓	✓	✗	✗	RSRD [8]
✓	✓	✓	✓	✗	✗	VideoArtGS [13]
✓	✓	✓	✓	✓	✓	**Ours**

Table 1: Comparison to Concurrent Works. Articulat3D uniquely enables the reconstruction of explicit articulation models from casual monocular videos by enforcing explicit physical motion and rigorous geometric joint parameter constraints.

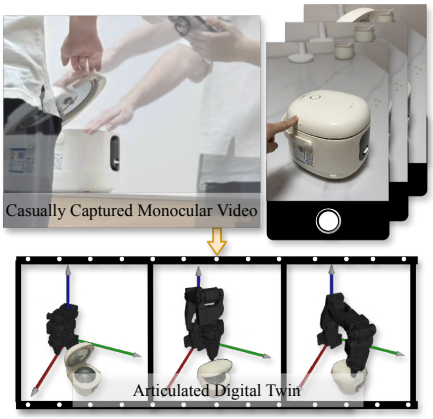

Fig. 1: From casual monocular video to interactive digital twin. Given a video of an object's articulation, Articulat3D produces physically-consistent models that are ready for interaction and manipulation in simulation environments.

available from casual capture and internet content, making it feasible for robotic agents to model objects directly from onboard observations. Achieving accurate articulated digital twins from such monocular inputs is therefore a key step toward scalable scene and object modeling and toward reducing the sim-to-real gap in robotic manipulation and interaction [8, 21].

Recent approaches to reconstructing articulated objects can be broadly categorized into three families based on the way to estimate articulation parameters. One line of work uses feed-forward predictors to infer articulation directly from images or videos [7, 9, 17]. While efficient at test time, these methods often depend on an external library for object retrieval or require data for fine-tuning, which limits scalability and robustness in real-world deployment. A second, more common family reconstructs objects by explicitly estimating joint parameters from multi-view images of the object in two or more discrete states [11, 14, 26]. While benefiting from geometric constraints, they necessitate controlled data capture setups and, crucially, require precise camera intrinsics and extrinsics for reconstruction, which is impractical for spontaneous, in-the-wild scenarios. A far more practical and scalable paradigm is reconstructing articulated objects from casually captured monocular videos which enables the ability to learn from internet videos and allows robotic agents to model objects directly from their visual observations [8, 13, 18]. While seemingly more scalable, these methods still implicitly rely on an initial static scan of the object, typically requiring the video to start with a segment where the camera moves around the object.

More fundamentally, a pervasive but overlooked limitation shared by these paradigms is their reliance on optimizing motion parameters (*e.g.*, dual quaternions) for discrete states or individual frames independently. This frame-wise optimization is inherently disjointed; the model focuses on fitting the visual appearance at specific timestamps rather than learning the actual evolutionary process of the motion. In reality, articulated motion, such as a door transitioning from closed to open, follows a continuous physical trajectory. By treating each frame as an isolated optimization target, existing methods fail to capture this temporal coherence, leading to reconstructions that lack physical plausibility during the transition phases, which limits their use on in-the-wild videos that start with immediate motion. As summarized in Tab. 1, existing methods often lack the ability to jointly optimize all frames under explicit articulation constraints without a separate static scan. We therefore ask this question: *how can we reconstruct articulated digital twins from just casually captured monocular videos by imbuing the model with an understanding of the fundamental geometric constraints that govern articulated motion*?

To this end, we present **Articulat3D**, a novel framework for reconstructing high-fidelity, interactable articulated digital twins from a single monocular video by modeling motion as a *continuous, constrained trajectory* rather than a collection of independent per-frame states, as shown in Fig. 1. Articulat3D follows a two-stage optimization strategy. First, in *Motion Prior-Driven Initialization*, instead of optimizing per-frame displacements, we exploit the low-dimensional structure of articulated movement by modeling the scene's dynamics through a compact set of motion bases. By representing each trajectory as a linear combination of these bases, we effectively regularize the motion space, ensuring global temporal consistency from the outset. Second, in *Geometric and Motion Constraints Refinement*, we explicitly parameterize articulation with learnable kinematic primitives—joint axes, pivot points, and per-frame motion scalars—and enforce these constraints in a differentiable optimization loop, encouraging physically plausible motion throughout the entire sequence, including transition phases. Extensive experiments on both synthetic benchmarks and real-world casual captures demonstrate that our method significantly outperforms existing approaches, achieving state-of-the-art accuracy in both geometry reconstruction and motion estimation. In summary, our work makes the following contributions:

- We present **Articulat3D**, a novel framework for high-fidelity, interactable digital twin reconstruction from monocular videos.
- We introduce a two-stage optimization strategy: Motion Prior-Driven Initialization exploits the low-dimensional structure of 3D point tracks via motion bases, and Geometric and Motion Constraints Refinement enforces strict physical plausibility through learnable kinematic primitives.
- We introduce two challenging benchmarks featuring multi-part kinematics and uncontrolled captures. Experiments demonstrate that our method achieves state-of-the-art performance in all metrics.

2 Related Work

2.1 Dynamic Scene Reconstruction

Dynamic scene reconstruction is a long-standing challenge. One significant line of research focuses on jointly estimating camera poses and scene geometry. Pioneering optimization-based methods [36] like DROID-SLAM [20], and MegaSaM [10] established robust frameworks for this task. More recently, foundation models have provided a powerful basis for 3D reconstruction, with DUSt3R [24] being a prominent example. Subsequent works extend these models to better handle dynamic content, such as MonST3R [37], which fine-tunes DUSt3R on dynamic datasets, and CUT3R [23], which proposes a continuous model for online reconstruction. While these methods effectively recover camera parameters and static geometry, modeling scene dynamics remains a formidable challenge.

A parallel and highly successful approach represents dynamic scenes with a deformation field, evolving from signed distance functions (SDFs) to modern neural representations. In particular, 4D extensions of 3D Gaussian Splatting [2, 15, 22, 30, 34, 38–40, 42] achieve state-of-the-art results in shape reconstruction and novel-view rendering. However, a critical limitation is that implicit deformation field is difficult to convert into an explicit motion representation, such as a rigid transformation matrix or an articulated structure. While concurrent works like Shape of Motion [22] attempt to extract motion from gaussians, they still treat motion as generic displacements rather than constrained articulated trajectories, a gap our work specifically addresses. Articulat3D integrates an explicit, constraint-driven articulation model into the Gaussian framework. By regularizing motion with kinematic primitives, our method produces a digital twin that is both visually accurate and physically controllable for downstream simulation across diverse articulated-object configurations and interaction scenarios.

2.2 Articulated Object Reconstruction

Reconstruction of articulated objects is a long-standing task in computer vision and robotics, which presents a dual challenge: one must solve for both the part-level geometry and the underlying articulation parameters. One family of methods employs large language and multi-modal models to predict both part segmentation and joint parameters [3, 5, 17, 25, 31], while some similar methods only predict articulation parameters [6, 35]. Leveraging vision-language models, Articulate Anything [9] take image or video as input to generate a code representation for articulated objects. Their fundamental limitation, however, is a reliance on large, existing mesh libraries, which prevents them from generalizing to unseen object categories in real-world setting.

Another line of work attempts to infer joint parameters and build part-level geometry of articulated objects simultaneously [4, 7, 11, 14, 26, 28, 29]. They rely on multi-view observations at discrete multi-state. These methods leverage strong geometric constraints, which simplify the problem but require impractical and controlled data capture setups. For example, ArtGS [14] and ArticulatedGS [4]

must be provided with two static 3D reconstructions of the object at different articulation states and assume all observations are aligned to the same coordinates. While this simplifies the correspondence problem, the underlying data acquisition process remains impractical for real-world use.

A more practical but far more challenging setting is reconstruction from just a monocular video. Existing methods are typically limited to simple objects [18, 19, 44] or implicitly rely on an initial static scan of the object [8, 13, 18]. For example, RSRD [8] and iTACO [18] require a pre-scanned 3D model before analyzing the dynamic video. Even methods designed for monocular input, such as VideoArtGS [13], still needs an initial static sequence for geometric initialization. This prerequisite fundamentally limits their applicability to in-the-wild videos that often begin with immediate motion. In contrast, our Articulat3D bypasses the need for static scans or templates by leveraging low-dimensional motion bases and explicit kinematic constraints, enabling robust joint optimization of the entire motion trajectory from the very first frame.

3 Our Methodology

As illustrated in Fig. 2, **Articulat3D** reconstructs a high-fidelity digital twin from a casual monocular video sequence $X = \{x_i\}_{i=1}^{N}$ through a two-stage optimization pipeline. We first preprocess the input frames by segmenting the object using SAM3 [1], filtering occluding hands via Qwen-Image [27], and extracting initial 3D motion trajectories with TAPIP3D [36] to serve as a foundational motion prior. In the first stage, we introduce **Motion Prior-Driven Initialization** to exploit the low-dimensional structure of articulated motion by modeling scene dynamics with a compact set of motion bases, where each noisy trajectory is fused into a coherent 3D Gaussian Splatting (3DGS) representation as a linear combination of these bases. Building upon this, the second stage, **Geometric and Motion Constraints Refinement**, enforces physical plausibility by introducing learnable kinematic primitives. By parameterizing component movements as either *revolute* or *prismatic* joints—defined by a joint axis, a pivot point, and per-frame motion scalars—we constrain each part to adhere strictly to rigid body transformations. This joint enforcement of geometric and kinematic constraints yields a reconstructed digital twin that is not only visually accurate and temporally coherent but also inherently controllable for downstream physical simulations.

3.1 Motion Prior-Driven Initialization

Prior Extraction. Casual captures of articulated objects often suffer from significant hand occlusions, leading to incomplete visual observations and distorted 3D reconstructions. To address this, we first employ Qwen-Image [27] to perform inpainting, removing occluding hands and recovering the holistic appearance of the object. This process yields a sequence of de-occluded images with corresponding complete segmentation masks $M = \{m_i\}_{i=1}^{N}$ generated by SAM3 [1]. We also utilize TAPIP3D [36] to obtain sparse 3D point trajectories with visibility. Each

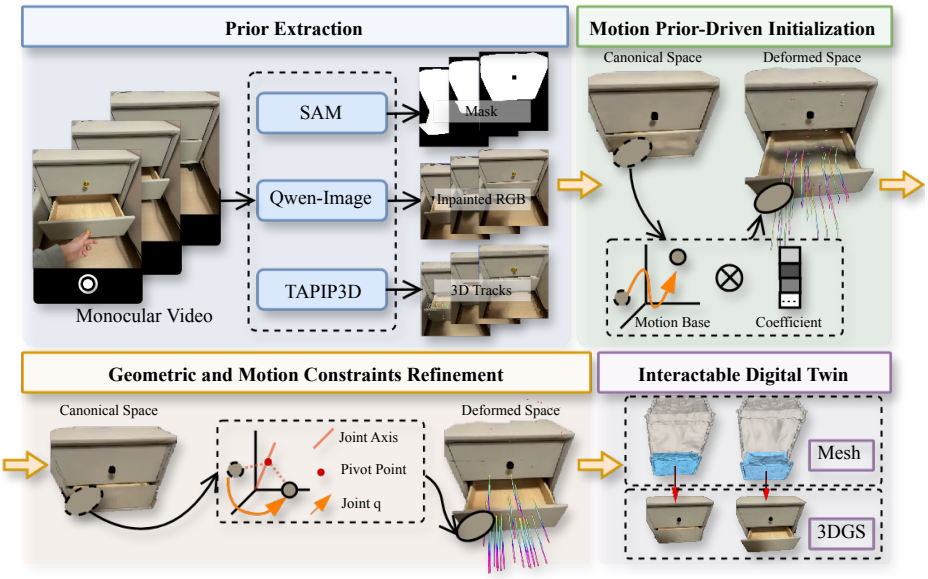

Fig. 2: Overview of Articulat3D . Given a monocular video, Articulat3D reconstructs a high-fidelity articulated digital twin via two stages. (1) *Motion Prior–Driven Initialization* exploits the low-dimensional structure of 3D point tracks by modeling them as a linear combination of shared motion bases. (2) *Geometric and Motion Constraints Refinement* then enforces physical plausibility by optimizing learnable kinematic primitives, including joint axes and pivot points. This joint optimization ensures that the final reconstruction is both geometrically accurate and temporally coherent.

trajectory is independently classified as static, prismatic, revolute, or invalid by fitting simple geometric motion primitives. Static points are detected by bounded displacement around their visible mean. Prismatic candidates are obtained by robust 3D line fitting, while revolute candidates are obtained by fitting a plane and a circle to the projected trajectory, yielding an axis direction and pivot. We then cluster trajectories within each motion type using KMeans over primitive-aware features, including initial position, mean position, motion direction, pivot for revolute motion, and temporal displacement or angular signatures. The number of prismatic and revolute clusters is determined by the predicted joint-type prior. This produces a per-track part label, where static points share label 0 and each dynamic cluster corresponds to one candidate movable part.

Motion Bases-based Modeling. Directly optimizing per-Gaussian trajectories from noisy sparse tracks is ill-posed. Instead, we leverage the fact that articulated motion resides in a low-dimensional subspace. Inspired by Shape of Motion [22], we model the scene dynamics using a compact set of B shared, learnable $\mathbb{SE}(3)$ motion bases. Unlike Shape of Motion, which represents object motion with generic free motion bases and soft per-Gaussian mixture weights, our motion prior-driven initialization is designed to produce a part-aware motion

representation from the beginning. Instead of allowing each Gaussian to freely combine multiple redundant bases, we bias the optimization toward compact, sparse, and near-discrete motion assignments that align with articulated parts. We also decouple the true articulation interval from static camera-orbiting frames, so temporal regularization is applied only over physically meaningful motion segments. As a result, motion prior-driven initialization provides a more coherent and interpretable part-level motion initialization, making the subsequent joint-axis and joint-value estimation more stable.

Given the TAPIP3D trajectories from the prior extraction step, we cluster them into B part-level motion groups using spatiotemporal K-means (see Supp. Mat. for details), where each recovered cluster corresponds to one motion basis, including an identity basis for the static part. For each dynamic cluster b, we initialize its per-frame SE(3) basis $\mathbf{T}_{0\to t}^{(b)}$ by solving a weighted Procrustes alignment between the canonical 3D tracks at frame t_0 and their observations at each target frame t. Each Gaussian is then assigned a strong one-hot motion coefficient according to its associated trajectory cluster. These coefficients remain trainable during Stage-I optimization, but are regularized to stay sparse and near-discrete, encouraging compact part-level motion instead of arbitrary soft combinations of redundant bases. Specifically, for a 3D Gaussian at time t, its position $\boldsymbol{\mu}_t$ and orientation $\mathbf{R}_t$ are transformed from the canonical state $(\boldsymbol{\mu}_0, \mathbf{R}_0)$ via a rigid transformation $\mathbf{T}_{0\to t} = [\mathbf{R}_{0\to t} \mid \mathbf{t}_{0\to t}] \in \mathbb{SE}(3)$:

$$\boldsymbol{\mu}_t = \mathbf{R}_{0\to t}\mu_0 + \mathbf{t}_{0\to t}, \quad \mathbf{R}_t = \mathbf{R}_{0\to t}\mathbf{R}_0. \tag{1}$$

To represent complex articulated motion, each Gaussian is assigned a set of learnable coefficients $\{\mathbf{w}^{(b)}\}_{b=1}^{B}$, normalized via a softmax function such that $\sum_b \mathbf{w}^{(b)} = 1$. The aggregate transformation $\mathbf{T}_{0\to t}$ is formulated as a weighted combination of the B shared bases:

$$\mathbf{T}_{0\to t} = \text{Normalize}\left(\sum_{b=1}^{B} \mathbf{w}^{(b)}\ \mathbf{T}_{0\to t}^{(b)}\right), \tag{2}$$

where the normalization ensures the resulting transformation remains a valid $\mathbb{SE}(3)$ member (e.g., via SVD-based orthogonalization of the rotation component). This formulation regularizes the motion space, forcing Gaussians with similar semantic properties to share consistent, low-dimensional trajectories, thereby ensuring spatio-temporal coherence even in regions with sparse tracking data.

3.2 Geometric and Motion Constraints Refinement

While the motion bases in Sec. 3.1 provide a flexible initialization, they lack the physical rigour required for true digital twins, as they do not guarantee axis-aligned rigid-body motions. To enforce physical plausibility, we transition from unconstrained $\mathbb{SE}(3)$ bases to explicit kinematic primitives, specifically *revolute* and *prismatic* joints forcing each component to adhere to strict kinematic laws.

Kinematic Parameterization. We parameterize the motion of each object part k using a normalized joint axis $\mathbf{a}_k \in \mathbb{S}^2$, a pivot point $\mathbf{c}_k \in \mathbb{R}^3$, and a per-frame

scalar $q_k(t)$ representing displacement or rotation. These parameters define a per-part rigid transformation $\mathbf{T}_k(t) = [\mathbf{R}_k(t) \mid \mathbf{t}_k(t)] \in \mathbb{SE}(3)$.

For a **revolute joint** with rotation angle $\theta_k(t) = q_k(t)$, the transformation follows Rodrigues' formula around pivot $\mathbf{c}_k$:

$$\mathbf{R}_k(t) = \exp(\theta_k(t)[\mathbf{a}_k]_\times), \quad \mathbf{t}_k(t) = (\mathbf{I} - \mathbf{R}_k(t))\mathbf{c}_k, \tag{3}$$

where $[\mathbf{a}_k]_\times$ denotes the skew-symmetric matrix of $\mathbf{a}_k$. For a **prismatic joint** with displacement $d_k(t) = q_k(t)$, the motion is a pure translation along the axis:

$$\mathbf{R}_k(t) = \mathbf{I}, \quad \mathbf{t}_k(t) = d_k(t)\mathbf{a}_k. \tag{4}$$

We provide more details about kinematic parameterization in our Supp. Mat.

Robust Joint Initialization To bridge the gap between soft bases and rigid joints, we first perform a hard assignment where each Gaussian i is assigned to part $k = \arg\max_b w_i^{(b)}$ based on its learned coefficients. We then initialize the kinematic parameters by analyzing the aggregate motion: 1) *Robust Trajectory Synthesis*: For each part k, we compute a robust center trajectory $\mathbf{c}_k(t)$ using a trimmed mean of the positions of its assigned Gaussians to filter out tracking noise. 2) *Kinematic Discovery via PCA*: We apply Principal Component Analysis (PCA) to the centered trajectory $\{\mathbf{c}_k(t) - \bar{\mathbf{c}}_k\}$. The joint type is determined by the eigenvalue distribution: a single dominant eigenvalue indicates a **prismatic** joint (linear motion), while two significant eigenvalues suggest a **revolute** joint (planar arc). In the latter case, the joint axis $\mathbf{a}_k$ is initialized as the normal to the motion plane (the third principal component). We provide more details about robust joint initialization in our Supp. Mat.

Joint Refinement of Part Assignment. Initial assignments based on motion bases may be noisy near kinematic boundaries. We refine these assignments jointly with the joint parameters by associating each Gaussian with a learnable latent vector $\mathbf{z}_i \in \mathbb{R}^K$ (initialized from $\mathbf{w}_i$), as shown in Fig. 3. Soft assignment probabilities $p_{i,k}$ are obtained via a softmax with temperature τ.

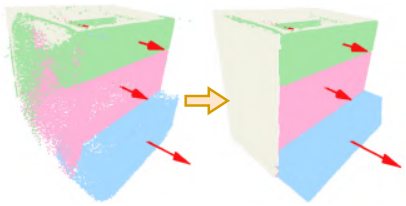

Fig. 3: Joint Refinement of Part Assignment. Kinematic consistency corrects initial misalignments (left) into physically-grounded segments (right).

To maintain strict rigidity in the forward pass, we apply the transformation of the most probable joint: $\mathbf{T}_i = \mathbf{T}_{k^*}$ where $k^* = \arg\max_k p_{i,k}$. For end-to-end optimization, we utilize a Straight-Through Estimator (STE) in the backward pass. The gradient of the loss $\mathcal{L}$ with respect to $\mathbf{z}_i$ is approximated by treating the transformation as a soft mixture $\sum_j p_{i,j}\mathbf{T}_j$:

$$\frac{\partial \mathcal{L}}{\partial z_{i,k}} \approx p_{i,k} \sum_{j=1}^{K} \left\langle \frac{\partial \mathcal{L}}{\partial \mathbf{T}_i}, \mathbf{T}_j \right\rangle (\delta_{jk} - p_{i,j}), \tag{5}$$

Table 2: Comparison of Articulat3D with state-of-the-art baselines. We evaluate kinematic accuracy, reconstruction fidelity, 3D tracking, and view synthesis across synthetic and real-world datasets.

Method	Joint Estimation		Reconstruction			Tracking	View Synthesis		
	Axis Err ↓	Pos Err ↓	CD-w ↓	CD-m ↓	CD-s ↓	EPE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Video2Articulation-S									
RSRD [8]	68.49	203.00	339.00	82.00	14.00	N/A	24.78	0.77	0.16
Articulate Anything [9]	49.85	81.00	11.00	59.00	7.00	N/A	N/A	N/A	N/A
iTACO [18]	16.05	13.00	1.00	13.00	6.00	N/A	N/A	N/A	N/A
Shape of Motion [22]	N/A	N/A	N/A	N/A	N/A	N/A	31.88	0.96	0.05
Articulat3D	**1.60**	**1.83**	**0.82**	**1.12**	**1.82**	N/A	**35.91**	**0.98**	**0.05**
Articulat3D-Sim									
RSRD [8]	55.75	91.52	69.16	55.41	10.05	3.14	28.19	0.94	0.06
Articulate Anything [9]	44.65	129.00	16.10	17.74	16.36	2.02	N/A	N/A	N/A
iTACO [18]	48.50	38.00	3.13	1.63	3.72	0.14	N/A	N/A	N/A
Shape of Motion [22]	N/A	N/A	N/A	N/A	N/A	0.08	33.37	0.97	0.04
Articulat3D	**0.53**	**0.65**	**0.76**	**0.84**	**0.85**	**0.04**	**37.80**	**0.98**	**0.03**
Articulat3D-Real									
RSRD [8]	N/A	N/A	N/A	N/A	N/A	N/A	16.12	0.78	0.34
Shape of Motion [22]	N/A	N/A	N/A	N/A	N/A	N/A	24.13	0.91	0.15
Articulat3D	N/A	N/A	N/A	N/A	N/A	N/A	**26.73**	**0.93**	**0.14**

where δ_{jk} is the Kronecker delta and $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product. This "re-skinning" mechanism allows model to correct misassignments, ensuring each Gaussian aligns with the most physically consistent kinematic structure.

3.3 Optimization

To reconstruct high-fidelity geometry and physically plausible motion, we jointly optimize the canonical Gaussian parameters $\mathcal{G}^c$, the latent assignments $\mathbf{z}_i$, and the kinematic articulation parameters $\Theta = \{\mathbf{a}_k, \mathbf{c}_k, q_k(t)\}$. The total objective function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{render}} + \lambda_{\text{acc}}\mathcal{L}_{\text{acc}} + \lambda_z\mathcal{L}_z \tag{6}$$

where λ_{acc} and λ_z are hyper-parameters balancing rendering accuracy, motion smoothness, and assignment consistency.

Rendering Loss. Following ArtGS [14], $\mathcal{L}_{\text{render}}$ combines photometric accuracy with geometric supervision. Specifically, $\mathcal{L}_{\text{render}} = (1-\lambda_{\text{SSIM}})\mathcal{L}_1 + \lambda_{\text{SSIM}}\mathcal{L}_{\text{D-SSIM}} + \lambda_D\mathcal{L}_D$. For the depth supervision, we utilize monocular depth estimates $\bar{\boldsymbol{D}}$ (e.g., from [10]) and apply a robust scale-invariant loss: $\mathcal{L}_D = \log(1 + \|\boldsymbol{D} - \bar{\boldsymbol{D}}\|_1)$, where $\boldsymbol{D}$ is the rendered depth map. To ensure temporal smoothness and mitigate monocular depth ambiguities, we employ an Acceleration Loss ($\mathcal{L}_{acc}$) and a Depth-Stability Loss ($\mathcal{L}_z$). A detailed mathematical formulation and a discussion of their physical significance are provided in our Supp. Mat.

4 Experiments

4.1 Datasets

We conduct a comprehensive evaluation on three distinct datasets to assess the performance of existing methods on objects with varying articulation complexity.

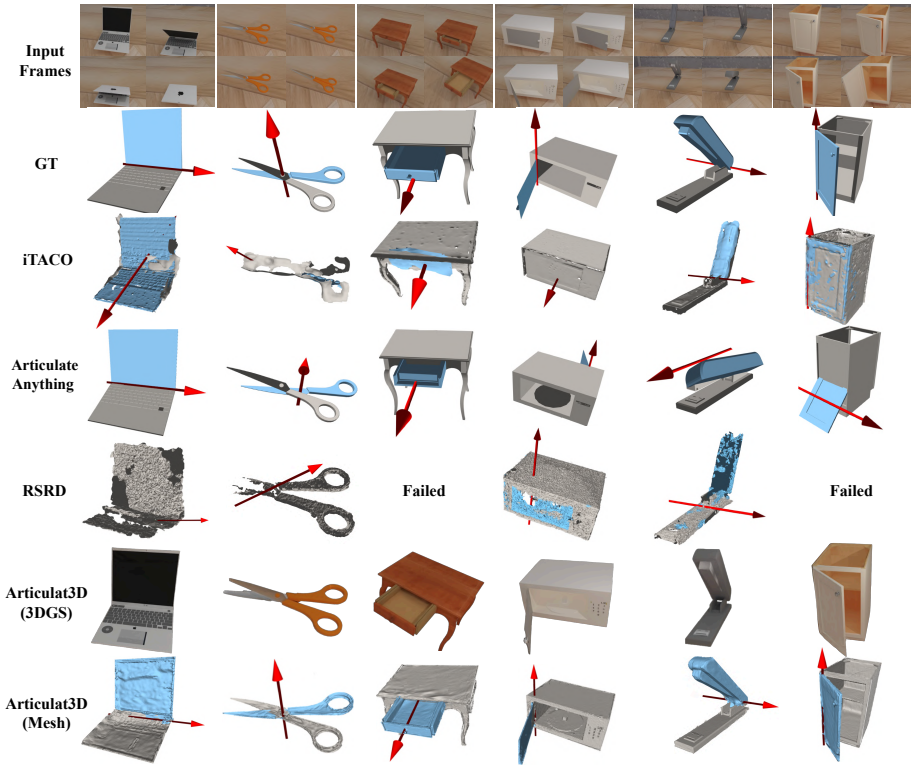

Fig. 4: Visual results on Video2Articulation-S dataset [18].

Video2Articulation-S. It is a dataset proposed by [18], which serves as our benchmark for simple articulated objects. It consists of 73 test videos across 11 categories of synthetic objects, where each object has only a single movable part.
Articulat3D-Sim. We introduce this dataset of 20 videos across 17 categories from PartNet-Mobility dataset [32] to evaluate complex multi-part kinematics (ranging from 2 to 7 parts). Each sequence features natural interaction, combining synchronized articulations with significant camera motion.
Articulat3D-Real. To evaluate performance in real world, we introduce this dataset, which consists of 12 sequences of 9 distinct categories (*e.g.*, cabinets, and doors). Captured using an iPhone 16 Pro Max, these sequences feature diverse lighting conditions and complex background clutter, providing a benchmark for assessing the generalizability of articulated reconstruction in the wild.

4.2 Evaluation Metrics

We evaluate Articulat3D across four dimensions: (1) **Joint Estimation**: Axis Err (°) and Position Err (cm) for kinematic accuracy; (2) **Reconstruction**: Chamfer Distance (cm) for the whole object (CD-w), movable parts (CD-m), and

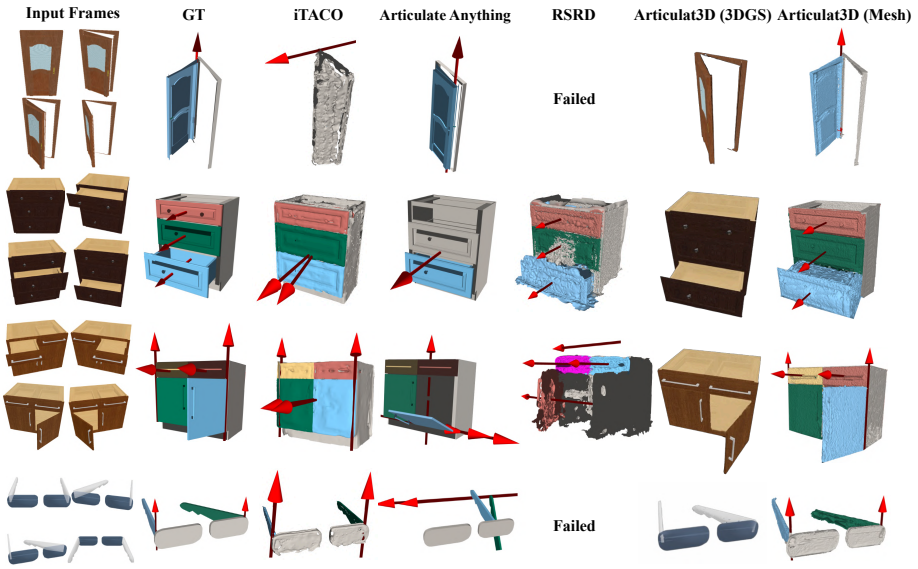

Fig. 5: Visual results on Articulat3D-Sim dataset.

static base (CD-s); (3) **Tracking**: End-Point Error (EPE) for temporal motion coherence; (4) **View Synthesis**: PSNR, SSIM, and LPIPS for photometric quality. For qualitative results, we visualize reconstruction results as meshes, using grey for static components, various colors for dynamic parts, and red arrows for joint axes. Detailed definitions are provided in our Supp. Mat.

4.3 Results

We compare Articulat3D against four state-of-the-art baselines: (1) **RSRD** [8], which optimizes part-level $SE(3)$ transformations; (2) **Articulate Anything** [9], a foundation model that retrieves joint parameters from a mesh library; (3) **iTACO** [18], which estimates articulation via point cloud motion analysis; (4) **Shape of Motion** [22], a 4D-GS based method representing dynamics with generic motion bases. For a fair comparison, RSRD [8] is adapted with NKSR mesh reconstruction, Articulate Anything [9] is provided with the PartNet-Mobility library, and iTACO [18] is given ground-truth depth. We provide exhaustive implementation details and configuration settings for all baselines in Supp. Mat.

For quantitative results, N/A entries in Tab. 2 indicate inapplicable metrics: EPE is omitted for datasets lacking ground-truth trajectories; view synthesis metrics are excluded for Articulate Anything [9] and iTACO [18] as they do not support rendering. For qualitative results, we visualize reconstruction results as meshes, using grey for static components, various colors for dynamic parts, and red arrows for joint axes. Failed indicates cases where RSRD [8] failed to reconstruct geometry. Since baseline methods such as Articulate Anything [9]

Table 3: Ablation Study on Articulat3D-Sim dataset.

Method	Joint Estimation		Reconstruction			Tracking	View Synthesis		
	Axis Err ↓	Pos Err ↓	CD-w ↓	CD-m ↓	CD-s ↓	EPE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Articulat3D w/o motion prior	43.42	31.26	18.78	92.45	17.73	28.29	20.78	0.95	0.07
Articulat3D w/noise	0.56	0.67	0.78	0.88	0.86	0.05	37.62	0.98	0.03
Articulat3D w/o prior init	1.36	3.42	1.13	1.67	0.96	0.06	30.89	0.98	0.04
Articulat3D w/o kinem. const.	0.86	4.98	5.19	7.71	2.76	0.09	35.45	0.97	0.02
Articulat3D w less data	0.61	0.68	0.83	0.92	0.86	0.07	34.60	0.97	0.03
Articulat3D	**0.53**	**0.65**	**0.76**	**0.84**	**0.85**	**0.04**	**37.80**	**0.98**	**0.03**

and iTACO [18] only provide static meshes rather than per-frame poses, their results are shown in fixed configurations that may not match the GT states.

Quantitative results. As shown in Tab. 2, Articulat3D consistently outperforms all baselines. Regarding **Joint Estimation**, Articulat3D achieves best precision with an average axis error of 1.60° on Video2Articulation-S and 0.53° on Articulat3D-Sim. In contrast, iTACO [18] fails to handle multi-part objects (48.50° axis error), highlighting its limitations in complex kinematic scenarios. Regarding **Reconstruction Fidelity**, Articulate Anything [9] and RSRD [8] suffer from catastrophic position errors (> 90 cm) due to template mismatch or structural collapse. Conversely, Articulat3D still maintains sub-centimeter accuracy and high-fidelity synthesis. While Shape of Motion [22] is a dedicated dynamic renderer, its motion bases lead to significant geometric drift. In contrast, our kinematic refinement enforces strict rigid-body laws, ensuring both high-fidelity synthesis (37.80 PSNR) and physically plausible 3D trajectories.

Qualitative Results. Visual comparisons in Fig. 4, 5, and 6 confirm our quantitative gains. In **Video2Articulation-S** (Fig. 4), Articulat3D preserves structural rigidity by filtering tracking noise. On more challenging **Articulat3D-Sim** dataset (Fig. 5), our framework effectively reconstructs complex multi-part kinematics. While concurrent methods often suffer from ghosting or part-intersections, our two-stage optimization ensures clean, axis-aligned trajectories. Finally, results on real world dataset, **Articulat3D-Real** (Fig. 6) demonstrate our robustness to in-the-wild challenges, including handheld jitter, cluttered backgrounds, and significant hand occlusions. Articulat3D still produces high-fidelity digital twins that are inherently ready for physical interaction.

4.4 Ablation Study

Motion Prior Guidance. Removing 3D track guidance leads to performance drop (Articulat3D w/o motion prior) in Tab. 3 and visible artifacts, as shown in Fig. 7. This prior is essential for resolving the inherent ambiguity between geometry and kinematics and decoupling local geometry from global motion dynamics. Our full model successfully resolves this coupling, yielding stable and accurate digital twins.

Robustness to noise in Motion Priors. By injecting Gaussian noise into raw 3D trajectories (Articulat3D w/ noise), we demonstrate that Articulat3D remains robust, as shown in Tab. 3 and Fig. 9. Our explicit axis-angle parameterization acts as a rigid kinematic bottleneck that filters non-physical noise. Instead of

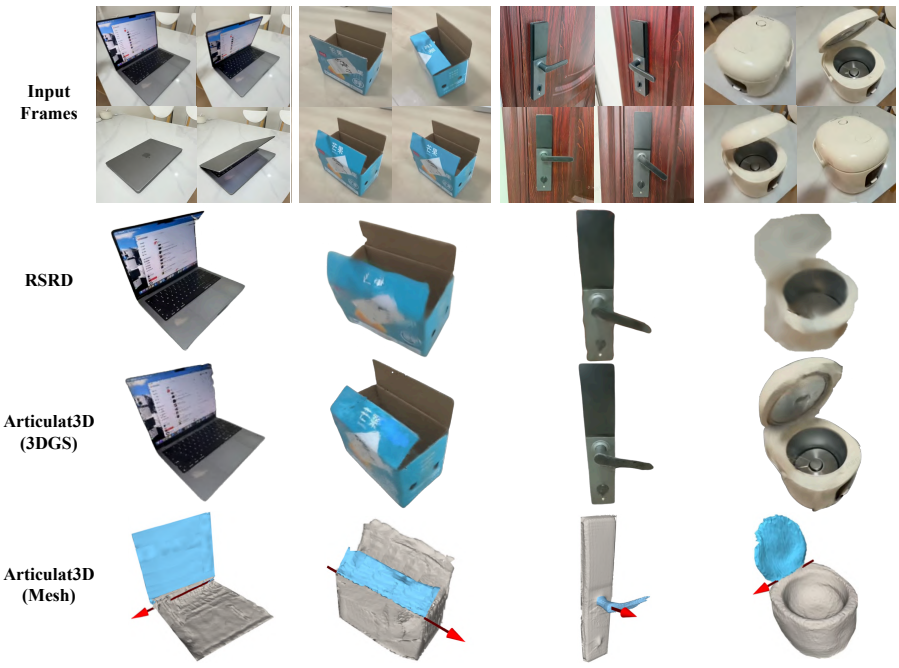

Fig. 6: Visual results on Articulat3D-Real dataset.

directly supervising on inconsistent raw coordinates, this formulation forces motion onto a valid kinematic manifold, effectively refining and correcting the upstream trajectories in a geometrically consistent manner.

Low-Dimensional Motion Bases Initialization. We evaluate the necessity of Motion Prior–Driven Initialization by attempting to initialize kinematic parameters directly from raw 3D tracks (Articulat3D w/o prior init). While explicit kinematic primitives provide strong physical grounding, their optimization is inherently non-convex and highly sensitive to initial values. As shown in Tab. 3 and Fig. 8, bypassing this stage causes the model to converge to poor local minima, characterized by misaligned part segmentation and inaccurate joint axes. This initialization stage exploits the low-dimensional structure of articulated motion to transform an ill-posed monocular tracking problem into a well-posed kinematic refinement task, ensuring temporal coherence and geometric accuracy.

Novel-view synthesis on synthetic data. We additionally render the reconstructed articulated objects from viewpoints that are not observed during training. As shown in Fig 10, Articulat3D produces visually consistent novel-view renderings while preserving object appearance, geometric structure, and articulation details. This indicates that the learned representation captures a coherent 3D articulated object rather than merely reproducing the observed training views from the input sequence directly.

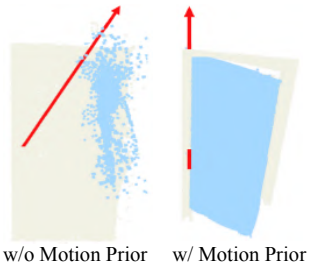

Fig. 7: Visual impact of motion prior. Without 3D motion guidance, motion-geometry ambiguity leads to distorted structures.

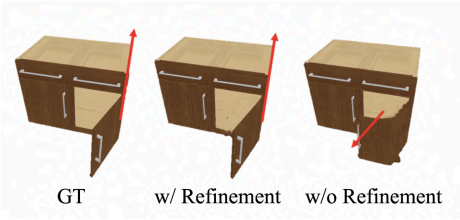

Fig. 8: Visual impact of refinement. Without our initialization, the model converges to local minima with misaligned joint axes and distorted geometry.

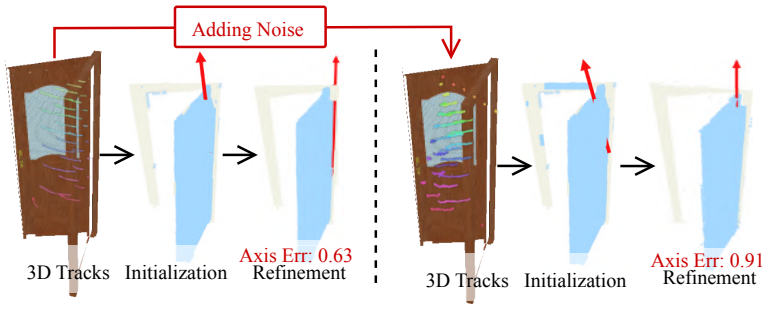

Fig. 9: Robustness to noisy motion priors. Our kinematic refinement stage recovers accurate part segmentation and joint parameters even when initialized with highly inconsistent 3D tracks.

Geometric and Motion Constraints Refinement. Removing the geometric and motion refinement stage leads to a marked degradation in joint accuracy and structural fidelity (Articulat3D w/o kinem. const.), as shown in Tab. 3. Without the explicit kinematic bottleneck, the model fails to enforce rigid-body physical laws, resulting in "floating" components and geometric drift, as shown in Fig. 8. This is evidenced by the sharp increase in Position Error (from 0.65 to 4.98) and CD-m (from 0.84 to 7.71), which confirm that our refinement stage is indispensable for transforming initial motion estimates into physically plausible and temporally coherent digital twins.

Robustness to data sparsity. To evaluate the efficiency and robustness of our under limited visual observations, we conduct an experiment by uniformly subsampling the input video to use only 1/3 of the original frames. As reported in Tab. 3 and Fig. 11, while the rendering quality experiences a moderate decline (PSNR decreases from 37.80 to 34.60) due to fewer photometric constraints for Gaussian optimization, the kinematic accuracy remains remarkably stable. Specifically, the Axis Error and Position Error only show marginal increases ($0.53^\circ \rightarrow 0.61^\circ$ and $0.65 \rightarrow 0.68$ cm, respectively). This demonstrates that our Motion Prior-Driven

Fig. 10: Novel-view synthesis results on the Articulat3D-Sim dataset. The displayed viewpoints are not observed during training.

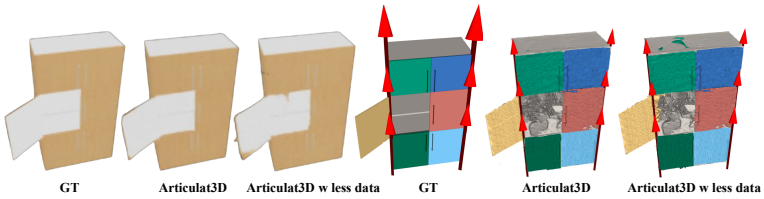

Fig. 11: Robustness to data sparsity. Comparison the results between GT, Articulat3D trained with full frames, and Articulat3D trained with 1/3 sub-sampled frames.

Initialization effectively captures the low-dimensional structure of articulation even from temporally sparse data, and the Geometric Constraints Refinement successfully maintains structural integrity despite the reduced supervision.

5 Conclusion

We present **Articulat3D**, a framework for reconstructing physically plausible, interactable digital twins of articulated objects from a single casually captured monocular video, even when motion begins immediately. A motion-prior-driven initialization combined with rigorous geometric constraint refinement successfully regularizes the ill-posed monocular reconstruction problem into a structured optimization of kinematic primitives. Without requiring static scans or controlled environments, Articulat3D offers a scalable, robust solution for modeling the articulated world and enabling advanced robotic manipulation applications.
Limitation and future work. Articulat3D 's performance may degrade on texture-less surfaces where sparse 3D tracks become unreliable. Moreover, our current implementation assumes independent joints rather than hierarchical kinematic structures, which limits the modeling of nested assemblies. Future work will address these issues by integrating multi-view geometric priors and exploring the estimation of dynamic physical parameters (*e.g.*, mass and friction).

Acknowledgements

This work was partly supported by the Industry-University-Research Innovation Fund of the Ministry of Education (Grant No. 2025XJS023), the Central Guidance for Local Science and Technology Development Fund(Grant No. ZYYD2025QY19).

Bibliography

[1] Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025) 5

[2] Duan, Y., Wei, F., Dai, Q., He, Y., Chen, W., Chen, B.: 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024) 4

[3] Goyal, P., Petrov, D., Andrews, S., Ben-Shabat, Y., Liu, H.T.D., Kalogerakis, E.: GEOPARD: Geometric pretraining for articulation prediction in 3d shapes. arXiv preprint arXiv:2504.02747 (2025) 4

[4] Guo, J., Xin, Y., Liu, G., Xu, K., Liu, L., Hu, R.: ArticulatedGS: Self-supervised digital twin modeling of articulated objects using 3d gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27144–27153 (2025) 2, 4

[5] Heppert, N., Irshad, M.Z., Zakharov, S., Liu, K., Ambrus, R.A., Bohg, J., Valada, A., Kollar, T.: Carto: Category and joint agnostic reconstruction of articulated objects. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. pp. 21201–21210 (2023) 4

[6] Hu, R., Li, W., Van Kaick, O., Shamir, A., Zhang, H., Huang, H.: Learning to predict part mobility from a single static snapshot. ACM Transactions On Graphics (TOG) **36**(6), 1–13 (2017) 4

[7] Jiang, Z., Hsu, C.C., Zhu, Y.: Ditto: Building digital twins of articulated objects from interaction. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. pp. 5616–5626 (2022) 2, 4

[8] Kerr, J., Kim, C.M., Wu, M., Yi, B., Wang, Q., Goldberg, K., Kanazawa, A.: Robot see robot do: Imitating articulated object manipulation with monocular 4d reconstruction. arXiv preprint arXiv:2409.18121 (2024) 2, 5, 9, 11, 12

[9] Le, L., Xie, J., Liang, W., Wang, H.J., Yang, Y., Ma, Y.J., Vedder, K., Krishna, A., Jayaraman, D., Eaton, E.: Articulate-anything: Automatic modeling of articulated objects via a vision-language foundation model. Proc. of International Conference on Learning Representations (2024) 2, 4, 9, 11, 12

[10] Li, Z., Tucker, R., Cole, F., Wang, Q., Jin, L., Ye, V., Kanazawa, A., Holynski, A., Snavely, N.: MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. pp. 10486–10496 (2025) 4, 9

[11] Liu, J., Mahdavi-Amiri, A., Savva, M.: Paris: Part-level reconstruction and motion analysis for articulated objects. In: Proc. of IEEE Intl. Conf. on Computer Vision. pp. 352–363 (2023) 2, 4

[12] Liu, J., Tam, H.I.I., Mahdavi-Amiri, A., Savva, M.: Cage: Controllable articulation generation. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. pp. 17880–17889 (2024) 1

[13] Liu, Y., Jia, B., Lu, R., Gan, C., Chen, H., Ni, J., Zhu, S.C., Huang, S.: Videoartgs: Building digital twins of articulated objects from monocular video. arXiv preprint arXiv:2509.17647 (2025) 1, 2, 5

[14] Liu, Y., Jia, B., Lu, R., Ni, J., Zhu, S.C., Huang, S.: Building interactable replicas of complex articulated objects via gaussian splatting. In: Proc. of International Conference on Learning Representations (2025) 2, 4, 9

[15] Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In: Proc. of Intl. Conf. on 3D Vision. pp. 800–809 (2024) 4

[16] Luo, R., Geng, H., Deng, C., Li, P., Wang, Z., Jia, B., Guibas, L., Huang, S.: Physpart: Physically plausible part completion for interactable objects. In: IEEE International Conference on Robotics and Automation. pp. 12386–12393. IEEE (2025) 1

[17] Mandi, Z., Weng, Y., Bauer, D., Song, S.: Real2Code: Reconstruct articulated objects via code generation. arXiv preprint arXiv:2406.08474 (2024) 2, 4

[18] Peng, W., Lv, J., Lu, C., Savva, M.: iTACO: Interactable Digital Twins of Articulated Objects from Casually Captured RGBD Videos. In: Proc. of Intl. Conf. on 3D Vision (2025) 2, 5, 9, 10, 11, 12

[19] Song, C., Wei, J., Foo, C.S., Lin, G., Liu, F.: Reacto: Reconstructing articulated objects from a single video. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. pp. 5384–5395 (2024) 5

[20] Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. Proc. of Advances in Neural Information Processing Systems **34**, 16558–16569 (2021) 4

[21] Torne, M., Simeonov, A., Li, Z., Chan, A., Chen, T., Gupta, A., Agrawal, P.: Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. arXiv preprint arXiv:2403.03949 (2024) 2

[22] Wang, Q., Ye, V., Gao, H., Zeng, W., Austin, J., Li, Z., Kanazawa, A.: Shape of motion: 4d reconstruction from a single video. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. pp. 9660–9672 (2025) 2, 4, 6, 9, 11, 12

[23] Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. pp. 10510–10522 (2025) 4

[24] Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024) 4

[25] Wei, F., Chabra, R., Ma, L., Lassner, C., Zollhöfer, M., Rusinkiewicz, S., Sweeney, C., Newcombe, R., Slavcheva, M.: Self-supervised neural articulated shape and appearance models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15816–15826 (2022) 4

[26] Weng, Y., Wen, B., Tremblay, J., Blukis, V., Fox, D., Guibas, L., Birchfield, S.: Neural implicit representation for building digital twins of unknown articulated objects. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. pp. 3141–3150 (2024) 1, 2, 4

[27] Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025) 5
[28] Wu, D., Liu, L., Huang, A., Liu, Y., Yu, Q., Liu, S., Song, L., Lu, C.: REArtGS++: Generalizable articulation reconstruction with temporal geometry constraint via planar gaussian splatting. arXiv preprint arXiv:2511.17059 (2025) 4
[29] Wu, D., Liu, L., Linli, Z., Huang, A., Song, L., Yu, Q., Wu, Q., Lu, C.: REArtGS: Reconstructing and generating articulated objects via 3d gaussian splatting with geometric and motion constraints. arXiv preprint arXiv:2503.06677 (2025) 4
[30] Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. pp. 20310–20320 (2024) 4
[31] Xia, H., Su, E., Memmel, M., Jain, A., Yu, R., Mbiziwo-Tiapo, N., Farhadi, A., Gupta, A., Wang, S., Ma, W.C.: Drawer: Digital reconstruction and articulation with environment realism. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21771–21782 (2025) 4
[32] Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Jiang, H., Yuan, Y., Wang, H., et al.: Sapien: A simulated part-based interactive environment. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. pp. 11097–11107 (2020) 10
[33] Yang, G., Wang, C., Reddy, N.D., Ramanan, D.: Reconstructing animatable categories from videos. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. pp. 16995–17005 (2023) 1
[34] Yang, Z., Yang, H., Pan, Z., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. Proc. of International Conference on Learning Representations (2024) 4
[35] Yi, L., Huang, H., Liu, D., Kalogerakis, E., Su, H., Guibas, L.: Deep part induction from articulated object pairs. arXiv preprint arXiv:1809.07417 (2018) 4
[36] Zhang, B., Ke, L., Harley, A.W., Fragkiadaki, K.: TAPIP3D: Tracking any point in persistent 3d geometry. arXiv preprint arXiv:2504.14717 (2025) 4, 5
[37] Zhang, J., Herrmann, C., Hur, J., Jampani, V., Darrell, T., Cole, F., Sun, D., Yang, M.H.: Monst3r: A simple approach for estimating geometry in the presence of motion. Proc. of International Conference on Learning Representations (2025) 4
[38] Zhao, H., Wang, H., Yang, C., Shen, W.: 3d-consistent human avatars with sparse inputs via gaussian splatting and contrastive learning. arXiv preprint arXiv:2408.09663 (2024) 4
[39] Zhao, H., Wang, H., Zhao, X., Fei, H., Wang, H., Long, C., Zou, H.: Physsplat: Efficient physics simulation for 3d scenes via mllm-guided gaussian splatting. In: Proc. of IEEE Intl. Conf. on Computer Vision. pp. 5242–5252 (2025) 1
[40] Zhao, H., Yang, C., Wang, H., Zhao, X., Shen, W.: Sg-gs: Photo-realistic animatable human avatars with semantically-guided gaussian splatting. arXiv preprint arXiv:2408.09665 **2**(3) (2024) 4

[41] Zhao, H., Zeng, C., Zhuang, L., Zhao, Y., Xue, S., Wang, H., Zhao, X., Li, Z., Li, K., Huang, S., et al.: High-fidelity simulated data generation for real-world zero-shot robotic manipulation learning with gaussian splatting. arXiv preprint arXiv:2510.10637 (2025) 1
[42] Zhao, H., Zhao, X., Zhu, L., Zheng, W., Xu, Y.: Hfgs: 4d gaussian splatting with emphasis on spatial and temporal high-frequency components for endoscopic scene reconstruction. arXiv preprint arXiv:2405.17872 (2024) 4
[43] Zhao, H., Zhuang, L., Zhao, X., Zeng, C., Xu, H., Jiang, Y., Cen, J., Wang, K., Guo, J., Huang, S., et al.: Towards affordance-aware robotic dexterous grasping with human-like priors. arXiv preprint arXiv:2508.08896 (2025) 1
[44] Zhou, H., Guo, Y., Wang, X., Xu, K.: Monomobility: Zero-shot 3d mobility analysis from monocular videos. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. pp. 8800–8809 (2025) 5