SIAM: Towards Generalizable Articulated Object Modeling via Single Robot-Object Interaction

Yuyan Liu[1*], Li Zhang[2*], Di Wu[2], Yan Zhang[1], Anran Huang[1], Zhi Wang[1], Liu Liu[1†], Dan Guo[1]

[1]Hefei University of Technology, Hefei, China
[2]University of Science and Technology of China, Hefei, China
2024110532@mail.hfut.edu.cn

Abstract

Articulated object modeling, which represents interconnected rigid bodies with their geometry, part segmentation, articulation tree, and physical properties, is crucial for robotic perception and manipulation. Recently existing methods like SAGCI leverage Interactive Perception (IP) to refine models through robot interaction. However, SAGCI suffers from prior-dependency (requiring initialization), neglects kinematic/dynamic constraints, and generates non-watertight meshes. To overcome these limitations, we propose SIAM, a novel framework for efficient and generalizable **S**ingle-**I**nteraction **A**rticulated **M**odeling. Given an initial point cloud, SIAM first enables minimal robot interaction to trigger object motion. It then precisely segments parts by analyzing point cloud differences pre- and post-interaction. For joint parameter estimation, we introduce an optimization incorporating novel kinematic energy constraints, enhancing physical consistency. Finally, we reconstruct a high-quality, topologically watertight mesh by learning 3D Gaussian Primitives from multi-view RGB-D observations under deformation. Extensive experiments on the PartNet-Mobility benchmark demonstrate state-of-the-art articulation modeling performance. Successful real-world deployment with an xArm robot further validates the framework's practicality and transferability. SIAM achieves accurate, prior-free modeling with significantly reduced interaction cost.

Introduction

Articulated object modeling describes the structured representation of 3D objects composed of interconnected rigid bodies, characterized by annotated full geometry, part segmentation, and an articulation tree encoding kinematic information (Anguelov et al. 2012). This model specifies each part's type, connectivity, degrees of freedom, and motion constraints, unifying visual appearance, topology, semantics, and physics—crucial for robotic perception (Premebida, Ambrus, and Marton 2018; Jiang et al. 2023; Li et al. 2023), grasp planning (Miller et al. 2003; Berenson et al. 2007), and manipulation (Liu, Savva, and Mahdavi-Amiri 2025). It offers agents richer, actionable object representations. However, existing CAD or synthetic datasets (e.g., PartNet-Mobility (Xiang et al. 2020)) lack realistic textures, fine geometry, and accurate physics, limiting real-world applicability. Real-scanned datasets like AKB-48 (Liu et al. 2022) provide more faithful appearance and physical traits but remain time-consuming, labor-intensive, and costly to construct.

The SAGCI system (Lv et al. 2022) introduces Interactive Perception (IP), where robots actively manipulate objects and observe state transitions to refine internal models. This enables automatic correction of articulation attributes such as joint types, parameters, and physical properties, yielding more accurate representations. Unlike static or manually annotated methods, IP lets robots iteratively learn structural and kinematic properties through interaction. Moreover, integrating IP with differentiable simulation and model-driven learning improves sample efficiency and generalization across simulation and real environments, enhancing performance in articulated object manipulation and physical reasoning. Despite these promising results, several challenges still remain.

Despite SAGCI's pioneering contributions to robotic perception, several limitations remain: (1) The system relies on human interventions for initialization, making it prior-dependent and incapable of fully autonomous, closed-loop modeling. (2) For kinematic modeling, SAGCI requires an extra network to learn parameters, incurring additional training costs. Furthermore, it lacks kinematic constraints like energy conservation, limiting physical consistency during manipulation. (3) The generated models build the whole object mesh rather than per-part generation, often exhibiting non-watertight surfaces that hinder accurate simulation.

To address these, we propose **SIAM** for efficient generalizable **A**rticulated object **M**odeling within a **S**ingle **I**nteraction. Specifically, SIAM introduces coarse actionable part perception to enable slight robot interaction, reducing reliance on human intervention for a truly prior-free pipeline. Next, we input pre- and post-interaction point clouds to obtain precise segmentation by mining object motions. For joint parameter estimation, we augment training-free optimization with a novel energy function considering geometric and kinematic constraints, thereby enhancing the structural and physical consistency as well as the interpretability of the estimated results. Finally, we utilize part segments to learn per-part 3D Gaussian Primitives from

*These authors contributed equally.
†Corresponding author.

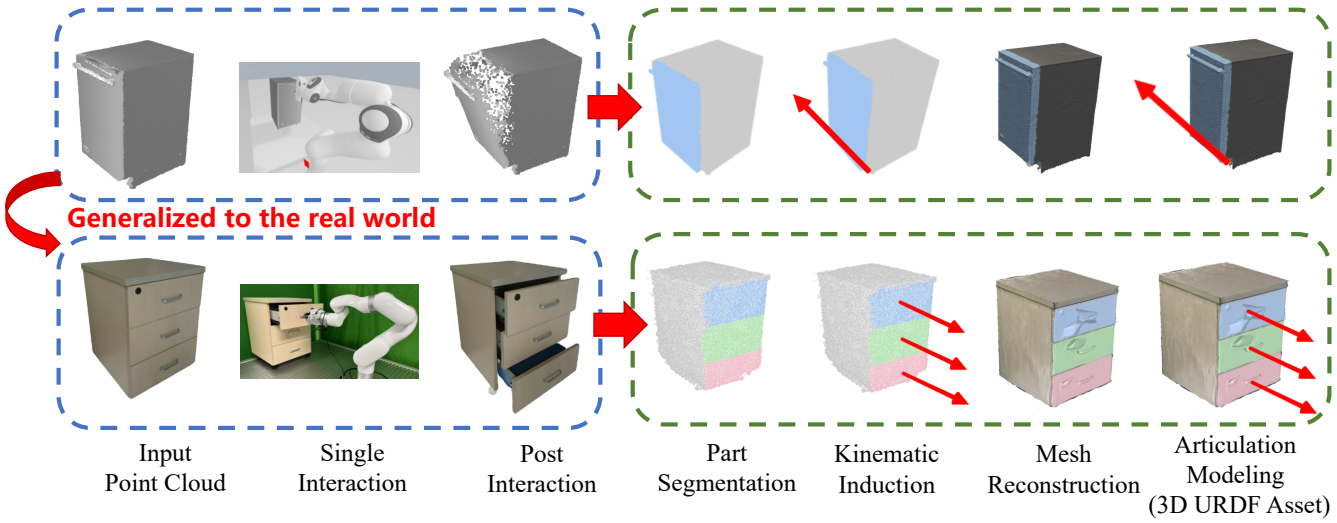

Figure 1: Given an initial point cloud observation and a single robot-object interaction, our SIAM achieves high-precise part segmentation, kinematic induction, and mesh reconstruction tasks, and outputs an articulated object modeling represented as a 3D URDF asset.

multi-view RGB-D images, generating topologically watertight meshes that largely improve simulation compatibility. Experiments on PartNet-Mobility demonstrate superior state-of-the-art performance, with successful transfer to real-world interaction using a xArm robot.

The contributions can be summarized as follows:

- We propose the SIAM, which aims at generalizable articulated object modeling via single robot-object interaction.
- To cope with part segmentation, we introduce a motion-geometry-guided network that leverages scene flow for robust part discovery. To address joint estimation, we design an energy-based optimization framework with kinematic constraints for physically consistent modeling.
- Extensive experiments in both simulation and real-world scenarios demonstrate the effectiveness of our method and its strong generalization to unseen articulated objects, paving the way for generalizable part-level robotic manipulation.

Related Work

Part Instance Segmentation from Point Clouds

3D part instance segmentation (Papandreou et al. 2018) has been extensively studied with the development of large-scale 3D datasets such as ShapeNet (Chang et al. 2015) and PartNet (Mo et al. 2019), which enable learning part-level semantic understanding from point clouds. Prior works (Li et al. 2020; Zhang et al. 2022; Wang et al. 2024) have advanced unified point cloud learning architectures, supervised segmentation networks, and unsupervised part discovery. However, most of these methods focus on category-specific segmentation, where part definitions and structures are limited within the same object class. In contrast, GAPartNet (Geng et al. 2023) introduces a cross-category part segmentation benchmark that encourages learning geometry- and function-based part concepts shared across different object categories. Building upon this, our work enhances part segmentation by introducing interaction-induced dynamic cues, which provide additional geometric signals to resolve ambiguities that static observations cannot address.

Kinematic Structure Reconstruction

Reconstructing the kinematic structure of articulated objects from point clouds is crucial for robotic manipulation (Liu, Mahdavi-Amiri, and Savva 2023; Taylor 2000; Heppert et al. 2023). Existing methods such as RPMNet (Yan et al. 2020) and Shape2Motion (Wang et al. 2019) utilize point-wise motion prediction to separate articulated parts in unknown objects, but they do not generate URDF models suitable for downstream embodied tasks. Similarly, URDFormer (Chen et al. 2024) employs a Transformer architecture to predict part connectivity, yet its performance is limited by reliance on single-view inputs and lacks precise articulation parameters. In contrast, our approach processes paired point clouds captured before and after a single interaction, enabling unsupervised and accurate generation of URDF models including detailed joint parameters. By integrating segmentation, mesh reconstruction, and kinematic modeling in a unified pipeline, our method produces physically accurate and simulation-ready models, offering a practical and generalizable solution for real-world articulated object reconstruction.

Cross-Category Generalization

Generalizing 3D part perception and manipulation to unseen object categories remains challenging (Mateo, Gil, and

Torres 2016; Stückler et al. 2011). Prior works explore category-agnostic part discovery through interaction (Qian and et al. 2023) or bottom-up learning (Luo 2023), but their performance on complex objects is limited. Some methods reconstruct unknown articulated shapes via latent encodings (Heppert et al. 2023), while others improve manipulation generalization using low-level geometric cues (Mo et al. 2021; Wu et al. 2022). Interactive pipelines like (Gadre, Ehsani, and Song 2021) segment movable parts but handle only simple objects with few parts and lack use of consistent geometric priors. Rigid object manipulation methods (Breyer et al. 2020; Fang et al. 2020) provide foundations but do not address articulated parts. Our method improves generalization by leveraging temporal geometric cues from interaction and part priors, enabling robust perception and kinematic reasoning across diverse unseen articulated objects.

Problem Statement

This work focuses on articulated object modeling from a single robot-object interaction in previously unseen environments. The input to our system is a single-view point cloud $P \in \mathbb{R}^{N\times 3}$, by an eye-in-hand RGB-D camera, depicting a previously unknown articulated object. The output is a complete and simulation-ready asset that includes: (1) fine-grained part segmentation, (2) estimated joint parameters with explicit kinematic types (e.g., revolute or prismatic), and (3) a watertight mesh representation annotated with URDF-compatible structure. Without relying on prior CAD models or category labels, our goal is to recover the object's part-level semantics and motion configuration by leveraging physical cues collected during a minimal, autonomous interaction.

Given the initial observation P, we first use a pretrained part prior network to generate coarse interaction proposals, guiding the robot to execute a manipulation trajectory T. Post-interaction, a second point cloud P' is captured, and a scene flow field F is estimated between P and P' using our DifFlow3D network. These dynamic cues, combined with spatial features, are processed by a SparseUNet-based segmentation backbone to generate instance proposals $\{S_k\}_{k=1}^{K}$. For each segmented part, we infer its motion type and parameters—revolute parts are parameterized by axis point p, direction n, and angle $\Delta\theta$; prismatic parts by axis direction n and displacement Δd. This inference is formulated as a physically grounded optimization problem minimizing geometric misalignment E_G and motion consistency error E_K. Concurrently, multi-view RGB-D observations captured along T are used to reconstruct a high-fidelity 3D mesh using planar-guided Gaussian Splatting. Finally, mesh parts are aligned with point cloud proposals to produce a unified, articulated representation suitable for downstream robotic applications.

Method

Initial Perception for Single Interaction

Given a single-view point cloud as input, we begin by employing GAPartNet to generate an initial coarse part segmentation and to estimate a candidate grasp pose. This preliminary prediction acts as a weak prior, guiding the subsequent interaction process.

Based on the predicted pose, we construct a heuristic manipulation trajectory T that directs the robotic arm to interact with the object. During this interaction, the robot executes a predefined motion path while actively collecting sparse multi-view RGB observations from varying viewpoints. This interaction not only facilitates physical contact but also introduces necessary viewpoint diversity, which is critical for comprehensive perception.

Upon completing the interaction, the robot returns to its initial observation pose and captures a post-interaction point cloud that reflects the object's deformed or displaced geometry. To extract motion cues, we estimate the scene flow between the pre- and post-interaction point clouds using DifFlow3D. The resulting motion field reveals fine-grained part-level displacements, offering dynamic information that complements static shape observations and highlights articulated components.

The post-interaction point clouds and estimated scene flow are then input into our motion-geometry-guided part segmentation network. By integrating both geometric structure and observed dynamics, the network refines the initial segmentation and enhances structural understanding. Compared to purely geometry-based methods, our approach demonstrates improved robustness in cluttered scenes and on previously unseen object categories by grounding part-level predictions in physically observed motion patterns.

Motion-Geometry-Guided Part Segmentation

To effectively integrate both geometric structure and dynamic motion cues, we adopt a SparseUNet backbone that jointly processes spatial coordinates (x, y, z) and per-point scene flow vectors $\mathcal{F} = (f_x, f_y, f_z)$. The SparseUNet architecture is well-suited for this task due to its efficient sparse tensor operations and its ability to preserve fine-scale geometric details in irregular 3D point clouds. Notably, we exclude RGB inputs, as color information offers limited utility in distinguishing kinematic relationships such as revolute versus prismatic joints.

Given the input point cloud $\mathcal{P}'$ and the associated scene flow $\mathcal{F}$, the network extracts point-wise features $\mathbf{F}_p$, which are subsequently processed by two parallel MLP heads. The first head predicts motion-based part segmentation, while the second estimates per-point offsets to guide kinematic parameter inference. This dual-head design enables joint learning of semantic part decomposition and motion-aware structural attributes, supporting a more complete understanding of the object's articulation. The network is trained using two complementary loss functions: a segmentation loss that encourages accurate motion part prediction, and an offset loss that supervises the regression of spatial displacements used in downstream part proposal via a ball-query clustering mechanism. Together, these objectives facilitate robust and physically grounded part-level understanding from sparse, interaction-driven observations.

Segmentation loss. The segmentation head is optimized using a composite loss L_{seg} combining two complemen-

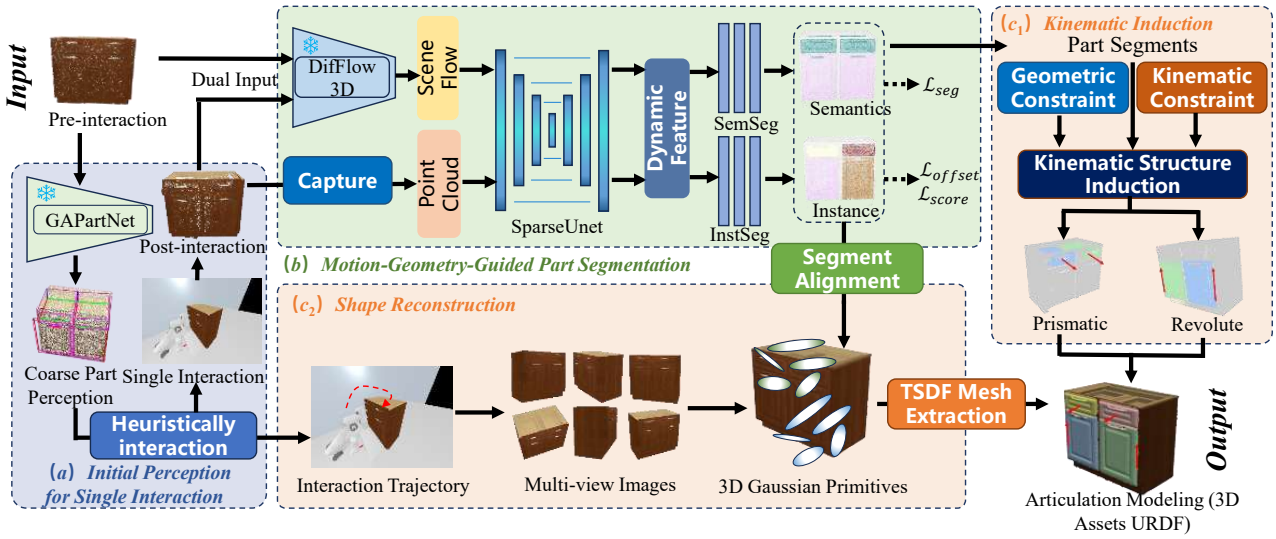

Figure 2: The Overview of the Proposed SIAM. We propose SIAM, a unified framework for generalizable articulated object modeling from a single robot-object interaction. Given an initial point cloud observation of an unknown articulated object, SIAM performs motion-guided part segmentation, kinematic structure induction, and watertight shape reconstruction to produce URDF-compatible, simulation-ready assets.

tary objectives: $\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{focal}} + \mathcal{L}_{\text{Lovász}}$ where Focal Loss $\mathcal{L}_{\text{focal}}$ addresses class imbalance by down-weighting well-classified examples, particularly effective for our task where part categories have uneven distributions.

Given the dense spatial distribution of revolute and prismatic parts in our task, semantic boundaries between adjacent instances are often ambiguous. To address this, we adopt the Lovász-Softmax loss (Berman, Triki, and Blaschko 2018), which directly optimizes mean IoU (mIoU) and improves segmentation quality in such challenging regions. Specifically, for semantic logits $\mathbf{s} \in \mathbb{R}^{C\times N}$ over N points and C classes, we compute softmax probabilities $\mathbf{p} = \text{softmax}(\mathbf{s}, \text{dim} = 1)$ and ground-truth labels $\mathbf{y} \in \{0, \dots, C-1\}^N$. The loss is given by

$$\mathcal{L}_{\text{Lovász}}(\mathbf{p}, \mathbf{y}) = \frac{1}{C}\sum_{c=1}^{C} \text{LovászHinge}(m^{(c)}) \tag{1}$$

where $m_i^{(c)} = 1 - \mathbf{p}_{c,i}$ if $y_i = c$, and $m_i^{(c)} = \mathbf{p}_{c,i}$ otherwise.

Offset Loss. The offset loss supervises the prediction of per-point displacement vectors from each point in the input point cloud $\mathcal{P}'$ to its corresponding instance center. For point i in the valid instance point set $\mathcal{V} \subset \mathcal{P}'$, let $\mathbf{o}_i^*$ denote the ground truth offset vector and $\mathbf{o}_i$ denote the predicted offset. The loss enforces both magnitude accuracy and directional consistency through:

$$\mathcal{L}_{\text{offset}} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \left(\|\mathbf{o}_i - \mathbf{o}_i^*\|_1 + \left(1 - \frac{\mathbf{o}_i \cdot \mathbf{o}_i^*}{\|\mathbf{o}_i\| \cdot \|\mathbf{o}_i^*\|}\right)\right) \tag{2}$$

Given the predicted part semantics and per-point offset vectors, we generate instance proposals $\{\mathcal{S}_k\}_{k=1}^K$ through a clustering process. Specifically, for each point, we perform a ball query in 3D space to group nearby points with the same predicted semantic label. To enhance the quality of proposals, we further incorporate the estimated scene flow $\mathcal{F}$: for each queried neighbor, we compute the angular and magnitude differences between its flow vector and that of the query center, retaining only neighbors within predefined thresholds. In our implementation, we set the angular threshold to $40°$ to avoid grouping points with significantly different flow directions, and the magnitude threshold to 0.02 to exclude neighbors with inconsistent motion amplitudes. This motion-guided filtering helps suppress noisy or irrelevant points, especially in cluttered or ambiguous regions.

Proposal Score Loss. This loss evaluates the quality of each generated proposal $\mathcal{S}_i$ using binary classification. Let $s_i^* \in [0, 1]$ denote the target quality score derived from the IoU between proposal $\mathcal{S}_i$ and the corresponding ground-truth segment, and let $s_i \in [0, 1]$ represent the predicted score by the network. The loss is defined as:

$$\mathcal{L}_{\text{score}} = \frac{1}{|\mathcal{S}|}\sum_{i=1}^{|\mathcal{S}|} \text{BCE}(s_i, s_i^*) \tag{3}$$

where $|\mathcal{S}|$ is the total number of generated proposals, and $\text{BCE}(\cdot)$ is the binary cross-entropy function. The IoU-derived s_i^* provides continuous supervision to guide proposal quality prediction.

To address the imbalance between multiple task losses with different magnitudes and optimization difficulties, we adopt an uncertainty-based weighting strategy (Kendall, Gal, and Cipolla 2018). The total loss is formulated as:

$$\mathcal{L}_{\text{total}} = \sum \frac{1}{2\sigma_i^2}\mathcal{L}_i + \log \sigma_i \tag{4}$$

where $\mathcal{L}_i$ is the loss of the i-th task, and σ_i is a learnable parameter (implemented by Kaiming initialization) representing its uncertainty. This formulation allows the model to dynamically adjust the contribution of each task during training.

This architecture effectively integrates scene geometry and motion to segment articulated parts, especially when geometric cues alone are insufficient to distinguish functional categories like sliders and hinges.

Kinematic Induction and Shape Reconstruction

Given the instance segmentation results, we perform kinematic structure induction for each part proposal using the predicted scene flow. For each proposal, we first classify its kinematic type as either *revolute* or *prismatic*, based on its motion characteristics.

For a **revolute** part, the objective is to estimate the rotation axis, which is parameterized by a point $\mathbf{p} \in \mathbb{R}^3$ on the axis and a unit direction vector $\mathbf{n} \in \mathbb{R}^3$. The rotation angle $\Delta\theta$ is assumed to be known or estimated beforehand and is used to guide the optimization of the axis parameters $(\mathbf{p}, \mathbf{n})$. The input consists of the part's point cloud $\{P_i\}_{i=1}^N$ at frame m_1, along with the predicted scene flow vectors $\{\mathbf{F}_i^{m_{12}}\}_{i=1}^N$ from frame m_1 to frame m_2.

To infer the axis parameters, we define a geometric constraint energy composed of two terms:

$$E_G^{(r)} = E_V^{(r)} + E_C^{(r)} \tag{5}$$

where the first term

$$E_V^{(r)} = \sum_{i=1}^{N} w_i \left| \frac{\mathbf{n} \cdot \mathbf{F}_i}{\|\mathbf{n}\|\|\mathbf{F}_i\|} \right| \tag{6}$$

encourages the flow vectors to be perpendicular to the rotation axis, reflecting the tangential nature of rotational motion. We introduce a normalized flow magnitude weight w_i (Eq. 7) in $E_V^{(r)}$ to suppress the influence of noisy or nearly static points, allowing the estimation to focus on dynamic regions with more informative motion cues.

$$w_i = \frac{\|\mathbf{F}_i\|}{\sum_{k=1}^{N} \|\mathbf{F}_k\|} \tag{7}$$

The second term

$$E_C^{(r)} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left| \frac{D(P_i, \mathbf{p}, \mathbf{n})}{D(P_j, \mathbf{p}, \mathbf{n})} - \frac{\|\mathbf{F}_i\|}{\|\mathbf{F}_j\|} \right| \tag{8}$$

encourages consistency between the relative distances from points to the axis and the relative magnitudes of their motion. Here, $D(P_i, \mathbf{p}, \mathbf{n})$ denotes the perpendicular distance from point P_i to the axis $(\mathbf{p}, \mathbf{n})$. This energy formulation leverages the spatial pattern of rotational motion to robustly estimate the underlying kinematic axis.

$E_V^{(r)}$ and $E_C^{(r)}$ in Eq 5 enforcing flow-axis alignment and distance-ratio consistency, respectively. Here, w_i is the normalized flow magnitude weight, and $D(P_i, \mathbf{p}, \mathbf{n})$ denotes the perpendicular distance from P_i to the axis.

To further constrain the axis parameters, we introduce a motion consistency term:

$$E_K^{(r)} = \frac{1}{N} \sum_i \|(P_i^{m_1} + \mathbf{F}_i^{m_{12}}) - T_r(P_i^{m_1}, \mathbf{p}, \mathbf{n}, \Delta\theta)\|_2 \tag{9}$$

where T_r is a rotation transformation function that outputs the post-rotated $P_i^{m_1}$ using joint $(\mathbf{p}, \mathbf{n})$ with angle $\Delta\theta$. Eq. 9 encourages the transformed points to align with the predicted flow vectors under the estimated rotation.

For a **prismatic** part, we only estimate the translation axis direction $\mathbf{n} \in \mathbb{R}^3$ and the displacement Δd. The input is the same: the point cloud and predicted flow. The geometric constraint becomes:

$$E_G^{(p)} = \sum_i |\mathbf{n} \cdot \mathbf{F}_i^{m_{12}} - \|\mathbf{n}\|\|\mathbf{F}_i^{m_{12}}\|| \tag{10}$$

and the motion consistency term is defined as:

$$E_K^{(p)} = \frac{1}{N} \sum_i \|(P_i^{m_1} + \mathbf{F}_i^{m_{12}}) - (P_i^{m_1} + \mathbf{n} \cdot \Delta d)\|_2 \tag{11}$$

The final energy to optimize for both motion types is:

$$E = E_G + E_K \tag{12}$$

This formulation allows us to infer the axis parameters and motion type for each moving part. Once the kinematic structure is recovered on the object point cloud, we proceed to *Shape Reconstruction*, and finally combine both components to generate complete 3D assets in URDF format.

To reconstruct 3D geometry, we use RGB images captured during heuristic manipulation, which offer both interaction cues and diverse viewpoints.

After optimizing the unified 3D Gaussians, we fuse them via TSDF to obtain a watertight mesh. This mesh is then globally aligned to the pre-interaction point cloud P using saved camera poses, ensuring consistency with the original coordinate frame.

Given the aligned mesh and the set of part proposals $\{S_k\}_{k=1}^K$ derived from motion-guided segmentation: due to the downsampled nature of our point cloud, we use a KDTree to locate the 50 nearest vertices in the global mesh for each point in a proposal, and assign the corresponding part label via majority voting among these neighbors, ensuring a more robust association. This approach ensures more robust nearest-neighbor association: for each point in a proposal, we find its 50 closest vertices in the global mesh and inherit the corresponding part label by majority voting among these neighbors. This process yields a collection of per-part meshes that are both topologically watertight and kinematically consistent with the recovered joint parameters. By performing segmentation in the reconstructed 3D space rather than per-view masking, our pipeline avoids redundancy, preserves surface continuity, and ensures that the final URDF-compatible asset faithfully reflects the estimated articulation structure for downstream manipulation and simulation tasks. Finally, we construct kinematic structures and export the object as URDF-compatible 3D assets, suitable for downstream physics-based simulation and manipulation.

Method	Seen Category					Unseen Category				
	Safe	Box	Dishwasher	Laptop	Storage	Door	Kitchenpot	Refrig.	Table	Oven
Part Segmentation (mean AP for Revolute part and Prismatic part)										
SAGCI (Lv et al. 2022)	28.93	16.24	23.34	16.08	18.17	20.86	16.66	25.42	26.87	17.98
Vanilla PartNet (Mo et al. 2019)	32.16	15.99	22.59	27.54	18.63	23.65	19.05	23.69	22.35	13.65
GAPartNet (Geng et al. 2023)	30.12	32.18	30.85	32.66	41.77	36.02	31.07	30.26	28.82	29.21
SIAM (Ours)	**49.68**	**77.50**	**68.99**	**88.84**	**74.84**	**89.47**	**95.14**	**83.18**	**68.88**	**33.7**
Kinematic Induction (average Axis Angle Error and Distance Error)										
SAGCI (Lv et al. 2022)	11.0°, 0.15m	9.8°, 0.16m	10.2°, 0.16m	6.9°, 0.09m	5.8°, 0.09m	9.6°, 0.18m	6.3°, 0.08m	12.7°, 0.20m	8.9°, 0.07m	8.5°, 0.07m
U-COPE (Zhang et al. 2024)	3.0°, **0.04m**	3.8°, 0.05m	7.5°, 0.10m	5.1°, 0.06m	3.7°, 0.04m	6.2°, 0.07m	3.6°, 0.06m	9.3°, 0.13m	6.0°, 0.06m	4.9°, 0.04m
EfficientCAPER (Yu et al. 2024)	2.2°, 0.05m	**3.6°**, **0.03m**	6.0°, 0.05m	3.1°, 0.04m	3.2°, 0.02m	4.3°, 0.05m	3.3°, 0.07m	7.0°, 0.09m	4.1°, 0.05m	5.2°, 0.07m
SIAM (Ours)	**2.1°**, 0.06m	3.9°, 0.10m	**2.8°**, **0.04m**	**3.6°**, **0.03m**	**3.0°**, **0.02m**	**2.5°**, **0.05m**	**2.7°**, **0.06m**	**5.9°**, **0.07m**	**3.8°**, **0.03m**	**4.8°**, **0.04m**
Shape Reconstruction (Chamfer Distance (CD) for Part-level Shape and Whole Shape)										
SAGCI (Lv et al. 2022)	18.64, 16.58	17.06, 16.35	23.64, 14.98	17.40, 17.53	26.14, 23.61	12.84, 13.13	8.65, 6.97	28.09, 27.68	24.31, 22.07	20.08, 21.43
A-SDF (Mu et al. 2021)	4.89, 3.06	9.48, 6.25	11.06, 10.13	4.11, 2.51	11.87, 8.63	9.26, 5.38	6.18, 4.05	13.62, 7.97	10.18, 7.90	10.14, 6.58
CARTO (Heppert et al. 2023)	**4.11**, 2.99	7.34, 4.08	8.03, 6.15	**2.09**, 1.58	10.06, 8.31	5.17, 2.69	5.01, 3.26	12.39, 7.01	7.69, **5.36**	4.15, **2.03**
SIAM (Ours)	4.90, **2.65**	**7.38**, **5.10**	**7.34**, **5.74**	2.23, **1.31**	**7.68**, **6.92**	**1.95**, **1.62**	**3.36**, **2.52**	**5.39**, **4.20**	**7.25**, 5.56	**4.01**, 3.14

Table 1: Quantitative Results on Datasets from PartNet-Mobility.

Experiments

Experimental Setup

Datasets. We use the articulated objects from PartNet-Mobility (Xiang et al. 2020) to render RGB-D images with pre-interaction and post-interaction states. The motion magnitude is constrained to be within 40% of the part's maximal articulation range, simulating a realistic one-time interaction. All the objects are split into seen and unseen sets to validate the generalization capacity of our SIAM. The part segmentation and shape reconstruction modules are trained separately in our training protocol and inferred together, since kinematic induction is a training-free method. The input point clouds are sampled into 2,048 points. All experiments are implemented using PyTorch and trained on a single NVIDIA RTX 4090 GPU with 48GB of memory.

Baselines and Metrics. We evaluate our SIAM by both systematical and individual manner. For systematical comparison, we employ SAGCI system (Lv et al. 2022) as a baseline since it holds the similar problem setting, and use the raw observed point cloud input to compare each component's performance. For individual comparison, we select different SOTA methods that designed for specific articulation-related tasks, e.g. GAPartNet for part segmentation, EfficientCAPER for joint estimation and CARTO for shape reconstruction. We feed the ground truth results from former components into these individual methods and compare their articulation modeling performance with our SIAM. In terms of metrics, we use Average Precision at Interaction-over-Union (IoU) 0.5 for part segmentation, angle error measured by degree and distance error measured by meter for joint estimation. As for shape reconstruction, we use Chamfer Distance (CD) for the whole object shape and per-part shape as metrics.

Comparison with the SOTA Methods

We report articulated object modeling performance evaluated on seen and unseen categories whose results are illustrated in Table 1. As it is shown, compared with current articulated object modeling approach SAGCI, our SIAM outperforms it with a large margin among part segmentation, joint estimation and shape reconstruction tasks, which achieve averagely around **74.07** AP50, **3.94°**, **0.05m** and over **3.90** Chamfer Distance performance on unseen categories. This proves that SIAM works well for generalizable articulation modeling with only single robot-object interaction. In terms of individual comparison, within the assistance of object interaction, our SIAM also obtains the state-of-the-art performance on these tasks. It is worthy noting that on the unseen categories, the joint parameters can still estimated with the similar performance comparing with them from seen categories. This can be explained by the fact that the kinematic structure induction process is designed as a training-free manner so it is not sensitive to the categories. Therefore, we can conclude that SIAM is an effective and novel generalizable articulated object modeling framework. Qualitative results are shown in Fig. 3.

Ablation Study

Effect of initial interaction prediction. Our method is sensitive to the predicted initial interaction. In Table 2, we compare our SIAM with different types of grasp prediction methods, such as deep learning approaches Where2act, foundation model AnyGrasp and reinforcement learning method AKBNet. As it can be seen, random grasp generation manner achieves the worst performance since numerous grasp poses cannot enable articulated object to move its parts with only 13.81 AP50, 10.0°, 0.15m joint estimation results and 10.32, 8.69 Chamfer Distance. One the other hand, these learning based approaches provide a good grasp initialization so achieve higher points. However, thanks for the specific architecture design of our SIAM, a slight part movement is enough for obtain the best articulation modeling result, with **86.47** AP50, **3.1°, 0.06m** joint parameter error and **3.51, 2.18** reconstruction results.

Effect of energy function for kinematic induction. To study the contribution of each energy function defined for the joint estimation task, ablated investigations are shown in Table 3. Generally, within both proposed kinematic and geometry constrained optimization, SIAM could better perceive joint parameters with only **4.0°** and **0.06m** angle error and distance error for revolute joint, as well as **2.8°** angle er-

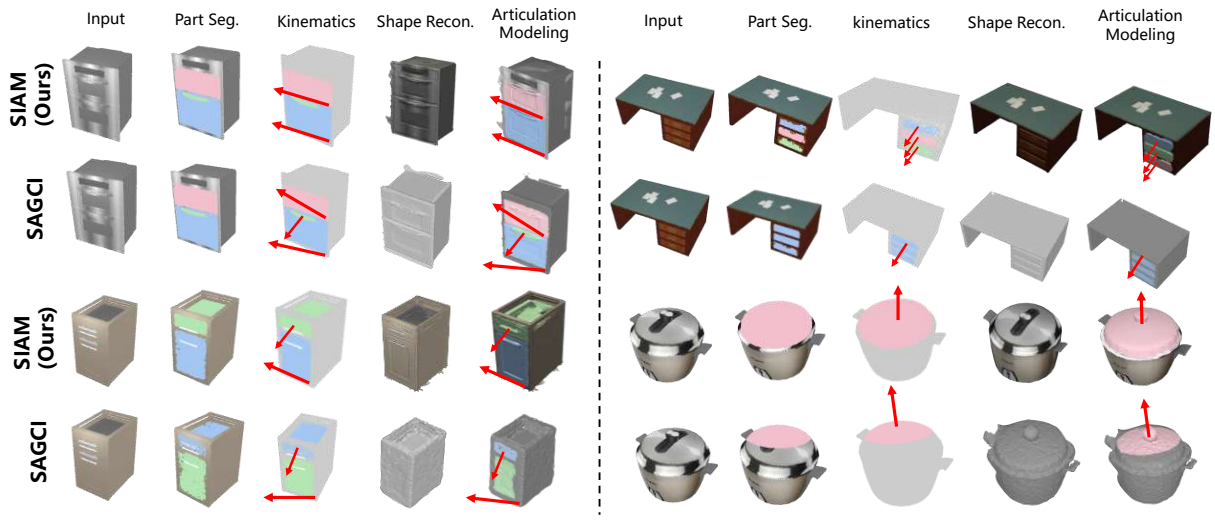

Figure 3: Qualitative results comparison between our SIAM and SAGCI. We show the results from part segmentation, kinematic induction, shape reconstruction and the final articulation modeling.

Initial Grasp	Part Seg.	Joint Est.	Shape Recon.
Random	13.81	20.6°, 0.23m	19.68, 17.49
Where2act (Mo et al. 2021)	23.07	10.0°, 0.15m	10.32, 8.69
AnyGrasp (Fang et al. 2023)	26.98	9.8°, 0.13m	11.07, 8.83
AKBNet (Liu et al. 2022)	32.16	6.2°, 0.10m	6.35, 4.02
Ours	**86.47**	**3.1°, 0.06m**	**3.51, 2.18**

Table 2: Articulation modeling comparison among different grasp initialization approaches

ror for prismatic joint. Generally, comparing the E_G and E_K energy functions, they contribute differently for these two types of joints. Observed from Table 3, we can conclude that prismatic motion is more sensitive to kinematic constraint and revolute motion relies more on geometric constraint.

Energy Func	Revolute Joint		Prismatic Joint
	Angle Error	Dist. Error	Angle Error
E_K-only	5.6°	0.10m	3.0°
E_G w/o E_V	4.8°	0.11m	3.9°
E_G w/o E_C	4.5°	0.09m	3.8°
E_G-only	4.2°	0.07m	3.6°
E_K + E_G	**4.0°**	**0.06m**	**2.8°**

Table 3: Ablation study of energy function

Generalization Capacity on Real-World Scenarios

To verify the generalization capacity of the proposed SIAM, we conduct experiments on real-world scenarios. We use the xArm robot arm as the interaction agent and deploy our SIAM framework in the robot. Fig. 4 shows the demonstration of using SIAM and the real-world robot to achieve articulated object modeling. The outputted URDF asset can support the simulation training for embodied AI.

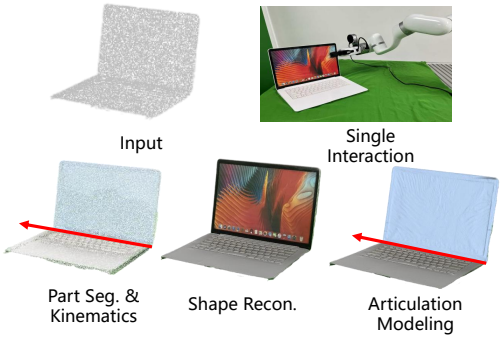

Figure 4: Articulation modeling in the real world

Conclusion

In this work, we present a unified framework for articulated object modeling through a single robot-object interaction, enabling generalizable part segmentation and kinematic reconstruction across diverse categories. By leveraging dynamic scene flow estimation and motion-guided segmentation, our method accurately identifies articulated parts and estimates their joint parameters in a physically consistent manner. Furthermore, we integrate multi-view RGB-D observations to reconstruct high-fidelity, watertight meshes aligned with the recovered kinematic structure. Extensive evaluations on both synthetic and real-world settings demonstrate the framework's robustness, generalization capability, and practical utility for downstream manipulation tasks. This approach paves the way for scalable, autonomous generation of simulation-ready articulated assets without reliance on category priors or predefined models.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62302143, National Key R&D Program of China (NO.2024YFB3311602), Natural Science Foundation of China (62272144), the Anhui Provincial Natural Science Foundation (2408085J040), and the Major Project of Anhui Provincial Science and Technology Breakthrough Program (202423k09020001).

References

Anguelov, D.; Koller, D.; Pang, H.-C.; Srinivasan, P.; and Thrun, S. 2012. Recovering articulated object models from 3d range data. *arXiv preprint arXiv:1207.4129*.

Berenson, D.; Diankov, R.; Nishiwaki, K.; Kagami, S.; and Kuffner, J. 2007. Grasp planning in complex scenes. In *2007 7th IEEE-RAS International Conference on Humanoid Robots*, 42–48. IEEE.

Berman, M.; Triki, A. R.; and Blaschko, M. B. 2018. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4413–4421.

Breyer, M.; Chung, J.; Ott, L.; Roland, S.; and Nieto, J. 2020. Volumetric Grasping Network: Real-Time 6 DOF Grasp Detection in Clutter. In *Conference on Robot Learning (CoRL)*, 3.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Chen, Z.; Walsman, A.; Memmel, M.; Mo, K.; Fang, A.; Vemuri, K.; Wu, A.; Fox, D.; and Gupta, A. 2024. Urdformer: A pipeline for constructing articulated simulation environments from real-world images. *arXiv preprint arXiv:2405.11656*.

Fang, H.-S.; Wang, C.; Fang, H.; Gou, M.; Liu, J.; Yan, H.; Liu, W.; Xie, Y.; and Lu, C. 2023. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 39(5): 3929–3945.

Fang, H.-S.; Wang, C.; Gou, M.; and Lu, C. 2020. GraspNet-1Billion: A Large-Scale Benchmark for General Object Grasping. In *CVPR*, 11444–11453.

Gadre, S.; Ehsani, K.; and Song, S. 2021. Act the Part: Learning Interaction Strategies for Articulated Object Part Discovery. In *ICCV*, 15752–15761.

Geng, H.; Xu, H.; Zhao, C.; Xu, C.; Yi, L.; Huang, S.; and Wang, H. 2023. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7081–7091.

Heppert, N.; Irshad, M. Z.; Zakharov, S.; Liu, K.; Ambrus, R. A.; Bohg, J.; Valada, A.; and Kollar, T. 2023. Carto: Category and joint agnostic reconstruction of articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21201–21210.

Jiang, J.; Cao, G.; Deng, J.; Do, T.-T.; and Luo, S. 2023. Robotic perception of transparent objects: A review. *IEEE Transactions on Artificial Intelligence*, 5(6): 2547–2567.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. arXiv:1705.07115.

Li, W.; Wang, Z.; Mai, R.; and et al. 2023. Modular design automation of the morphologies, controllers, and vision systems for intelligent robots: a survey. *Vis. Intell.*, 1: 2.

Li, Y.; Ma, L.; Zhong, Z.; Liu, F.; Chapman, M. A.; Cao, D.; and Li, J. 2020. Deep learning for lidar point clouds in autonomous driving: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8): 3412–3432.

Liu, J.; Mahdavi-Amiri, A.; and Savva, M. 2023. Paris: Part-level reconstruction and motion analysis for articulated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 352–363.

Liu, J.; Savva, M.; and Mahdavi-Amiri, A. 2025. Survey on Modeling of Human-made Articulated Objects. In *Computer Graphics Forum*, e70092. Wiley Online Library.

Liu, L.; Xu, W.; Fu, H.; Qian, S.; Yu, Q.; Han, Y.; and Lu, C. 2022. Akb-48: A real-world articulated object knowledge base. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14809–14818.

Luo, e. a. 2023. Learning to Group: A Bottom-Up Framework for 3D Part Discovery in Unseen Categories. In *ICLR*.

Lv, J.; Yu, Q.; Shao, L.; Liu, W.; Xu, W.; and Lu, C. 2022. Sagci-system: Towards sample-efficient, generalizable, compositional, and incremental robot learning. In *2022 International Conference on Robotics and Automation (ICRA)*, 98–105. IEEE.

Mateo, C.; Gil, P.; and Torres, F. 2016. Visual perception for the 3D recognition of geometric pieces in robotic manipulation. *The International Journal of Advanced Manufacturing Technology*, 83(9): 1999–2013.

Miller, A. T.; Knoop, S.; Christensen, H. I.; and Allen, P. K. 2003. Automatic grasp planning using shape primitives. In *2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422)*, volume 2, 1824–1829. IEEE.

Mo, K.; Guibas, L. J.; Mukadam, M.; Gupta, A.; and Tulsiani, S. 2021. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6813–6823.

Mo, K.; Zhu, S.; Chang, A. X.; Yi, L.; Tripathi, S.; Guibas, L. J.; and Su, H. 2019. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 909–918.

Mu, J.; Qiu, W.; Kortylewski, A.; Yuille, A.; Vasconcelos, N.; and Wang, X. 2021. A-sdf: Learning disentangled signed distance functions for articulated shape representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13001–13011.

Papandreou, G.; Zhu, T.; Chen, L.-C.; Gidaris, S.; Tompson, J.; and Murphy, K. 2018. Personlab: Person pose estimation

and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European conference on computer vision (ECCV)*, 269–286.

Premebida, C.; Ambrus, R.; and Marton, Z.-C. 2018. Intelligent robotic perception systems. *Applications of mobile robots*, 111–127.

Qian, Y.; and et al. 2023. Interactive Fine-Grained Part Segmentation via User Interaction. In *International Conference on Computer Vision (ICCV)*.

Stückler, J.; Steffens, R.; Holz, D.; and Behnke, S. 2011. Real-Time 3D Perception and Efficient Grasp Planning for Everyday Manipulation Tasks. In *ECMR*, 177–182.

Taylor, C. J. 2000. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding*, 80(3): 349–363.

Wang, X.; Zhou, B.; Shi, Y.; Chen, X.; Zhao, Q.; and Xu, K. 2019. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8876–8884.

Wang, Y.; Wang, L.; Hu, Q.; and et al. 2024. Panoptic segmentation of 3D point clouds with Gaussian mixture model in outdoor scenes. *Vis. Intell.*, 2: 10.

Wu, R.; Zhao, Y.; Mo, K.; Guo, Z.; Wang, Y.; Wu, T.; Fan, Q.; Chen, X.; Guibas, L.; and Dong, H. 2022. VAT-MART: Learning Visual Action Trajectory Proposals for Manipulating 3D Articulated Objects. In *International Conference on Learning Representations (ICLR)*.

Xiang, F.; Qin, Y.; Mo, K.; Xia, Y.; Zhu, H.; Liu, F.; Liu, M.; Jiang, H.; Yuan, Y.; Wang, H.; et al. 2020. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11097–11107.

Yan, Z.; Hu, R.; Yan, X.; Chen, L.; Van Kaick, O.; Zhang, H.; and Huang, H. 2020. RPM-Net: recurrent prediction of motion and parts from point cloud. *arXiv preprint arXiv:2006.14865*.

Yu, X.; Jiang, H.; Zhang, L.; Wu, L. Y.; Ou, L.; and Liu, L. 2024. EfficientCAPER: An End-to-End Framework for Fast and Robust Category-Level Articulated Object Pose Estimation. *Advances in Neural Information Processing Systems*, 37: 31968–31989.

Zhang, C.; Wan, H.; Shen, X.; and Wu, Z. 2022. PVT: Point-voxel transformer for point cloud learning. *International Journal of Intelligent Systems*, 37(12): 11985–12008.

Zhang, L.; Meng, W.; Zhong, Y.; Kong, B.; Xu, M.; Du, J.; Wang, X.; Wang, R.; and Liu, L. 2024. U-cope: Taking a further step to universal 9d category-level object pose estimation. In *European Conference on Computer Vision*, 254–270. Springer.