MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

ORIGAMI: Object Representation Inferred Geometrically for Articulated ManIpulation

Deng, Yunfu; Nikovski, Daniel N.

TR2026-141 September 29, 2026

Abstract

We present ORIGAMI, a geometric pipeline that discovers the kinematic structure of unknown articulated mechanisms from a single RGB-D demonstration and maps each subsequent high-dimensional visual observation into a compact joint-position vector to be used for reinforcement learning. Existing approaches to visual articulated manipulation either learn policies directly from pixels, suffering from the high dimensionality of the observation space, or depend on pretrained visual representations that require large-scale data collection and offer no structural guarantee about the resulting features. ORIGAMI sidesteps both limitations: from tracked 3D keypoints, it recovers link segmentation, joint types, axis parameters, and kinematic topology through purely geometric reasoning, without any task-specific training or category-level priors. A formal consistency analysis shows that the systematic estimation errors of the geometric estimator cancel to first order in the goal-conditioned displacement observed by the policy, eliminating the need for learned correction. Experiments on revolute and prismatic mechanisms in SAPIEN simulation and on real hardware demonstrate that the discovered low-dimensional representation substantially narrows the gap between pixelbased reinforcement learning and privileged ground-truth state, while outperforming methods built on pretrained visual representations.

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2026

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

ORIGAMI: Object Representation Inferred Geometrically for Articulated Manipulation

Yunfu Deng and Daniel N. Nikovski

***Abstract*— We present ORIGAMI, a geometric pipeline that discovers the kinematic structure of unknown articulated mechanisms from a single RGB-D demonstration and maps each subsequent high-dimensional visual observation into a compact joint-position vector to be used for reinforcement learning. Existing approaches to visual articulated manipulation either learn policies directly from pixels, suffering from the high dimensionality of the observation space, or depend on pretrained visual representations that require large-scale data collection and offer no structural guarantee about the resulting features. ORIGAMI sidesteps both limitations: from tracked 3D keypoints, it recovers link segmentation, joint types, axis parameters, and kinematic topology through purely geometric reasoning, without any task-specific training or category-level priors. A formal consistency analysis shows that the systematic estimation errors of the geometric estimator cancel to first order in the goal-conditioned displacement observed by the policy, eliminating the need for learned correction. Experiments on revolute and prismatic mechanisms in SAPIEN simulation and on real hardware demonstrate that the discovered low-dimensional representation substantially narrows the gap between pixel-based reinforcement learning and privileged ground-truth state, while outperforming methods built on pretrained visual representations.**

I. Introduction

Reinforcement learning from compact state representations – joint positions (angles and displacements), end-effector poses, and contact flags – usually converges orders of magnitude faster than learning from raw pixels on continuous control benchmarks [1], [2]. Data augmentation [3] and pretrained visual encoders [4], [5], [6] narrow this gap but do not close it, particularly on contact-rich manipulation tasks where the relevant degrees of freedom are a small subset of the visual observation [7], [8]. The gap is especially pronounced for articulated mechanisms, such as folding rulers, desk lamps, and pliers, whose configuration spaces are low-dimensional (a handful of joint angles), and yet their visual appearance changes in complex, nonlinear ways as joints rotate. For such mechanisms, the visual-state efficiency gap is fundamentally a representation problem: the policy must implicitly discover the object's kinematic structure from high-dimensional images, a task that consumes the majority of the sample budget.

Existing approaches to bridging this gap fall into two categories, both requiring training. General-purpose representation learning methods either jointly optimize an encoder alongside the policy [9], [10], [11] or rely on large-scale visual pretraining [4], [5], [6]. Category-level articulated object estimators [12], [13], [14] provide structured joint-angle predictions but depend on category-specific training data and learned reconstruction networks, limiting their applicability to object instances covered by the training distribution. In both cases, the representation must be learned – either from scratch for each task or transferred from a large external dataset – before the policy can benefit from it.

This paper takes a different approach: instead of fitting a representation to data, ORIGAMI constructs one through geometric analysis of a single demonstration video, exploiting the rigid-body structure of articulated mechanisms as an inductive prior. The proposed method, ORIGAMI (**O**bject **R**epresentation **I**nferred **G**eometrically for **A**rticulated **M**ani**i**pulation), builds on a key observation: the kinematic structure of an articulated mechanism is fully determined by the pairwise distance rigidity constraint – two points on the same rigid link maintain constant mutual distance under any configuration change. Given 3D keypoint trajectories extracted from an RGB-D demonstration using off-the-shelf pretrained perception (SAM [15] for segmentation, CoTracker3 [16] for tracking), ORIGAMI applies spectral clustering on distance-invariance statistics to segment rigid bodies, Kabsch-Umeyama alignment [17] with screw-theoretic decomposition to estimate joint parameters, and minimum spanning tree construction to recover the kinematic topology. The entire structure discovery process requires no task-specific training, no category-level priors, and no learned components beyond the perception front-end. At deployment time, the discovered model maps each new RGB-D frame to a compact joint-position vector through the same geometric procedure, producing a low-dimensional state representation that a standard RL algorithm can exploit directly. Because the demonstration can be generated either by a human manually articulating the object or in simulation by commanding the object's joints, the only requirement is that each joint is exercised over a sufficient range of motion. Furthermore, for goal-conditioned policies, systematic estimation errors arising from tracking noise and depth quantization cancel to first order in the configuration displacement between current and goal states, making the raw geometric estimates directly usable without calibration or learned correction.

The contributions of this work are three-fold. First, ORIGAMI constructs a compact joint-angle state representation for articulated mechanisms from a single RGB-D

This work was conducted while Yunfu Deng was an intern at Mitsubishi Electric Research Laboratories (MERL).

Yunfu Deng is with the Department of Computer Sciences, University of Wisconsin–Madison, Madison, WI 53706 USA (Email: yunfu.deng@wisc.edu).

Daniel N. Nikovski is with Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139 USA (Email: nikovski@merl.com).

demonstration via purely geometric kinematic discovery, requiring no task-specific or category-specific training. Second, the complete pipeline – from video observation to real-time state estimation – is presented, integrating distance-invariance segmentation, screw-theoretic joint estimation, and topology inference into a unified geometric framework. Third, experiments on articulated manipulation tasks in both SAPIEN simulation and real-world settings demonstrate that the discovered representation enables policy learning with sample efficiency approaching that of privileged ground-truth state access, substantially outperforming both pixel-based and learned-representation baselines.

II. Related Work

A. Visual Reinforcement Learning for Manipulation

Learning manipulation policies directly from visual observations remains sample-inefficient compared to learning from compact state representations. Yarats et al. [1] quantify this gap by showing that RL using the Soft Actor-Critic (SAC) algorithm from pixels requires substantially more environment steps than SAC from proprioceptive state on continuous control tasks. Data augmentation techniques [3], [2] partially close this gap but still require millions of interactions for complex tasks. Standard manipulation benchmarks [18], [8] further expose the difficulty: even state-of-the-art visual RL methods struggle on tasks that are straightforward with low-dimensional state access. Mu et al. [7] provide the most recent empirical comparison across 16 manipulation tasks, confirming that the performance gap between visual and state-based approaches remains substantial. These findings motivate a compact, geometrically meaningful state representation that avoids the sample cost of pixel-based learning.

B. Representation Learning for Robot Control

One approach to bridging the visual-state efficiency gap is to learn compact representations from data. Contrastive methods such as CURL [9] train auxiliary objectives jointly with the RL policy, improving pixel-based sample efficiency on control benchmarks. Pre-trained visual encoders, such as R3M [4] and VIP [5] trained on large-scale human video, MVP [6] trained via masked autoencoding on ImageNet, provide frozen features that transfer to downstream manipulation, sometimes approaching state-based performance. World-model approaches such as DreamerV3 [10] and TD-MPC2 [11] learn latent dynamics models that enable planning in a compressed representation space. However, all of these methods require either large-scale pretraining datasets, joint training of the encoder alongside the policy, or learning an explicit dynamics model.

For articulated mechanisms specifically, learned estimators can recover joint-level state from visual input. Li et al. [12] propose a category-level approach that predicts part segmentation, normalized coordinates, and joint parameters from a single depth observation. Weng et al. [13] extend this paradigm to real-time pose tracking of articulated objects from point cloud sequences. Jiang et al. [14] reconstruct part-level geometry and articulation models from point cloud pairs using implicit neural representations. These methods provide structured state estimates suitable for control but depend on category-specific training data or learned reconstruction networks. By contrast, the proposed method constructs a joint-angle state estimator using purely geometric reasoning, requiring no training data beyond a single demonstration video of the target object.

C. Kinematic Structure Discovery

Sturm et al. [19] establish the foundational framework for articulation model estimation, using Bayesian model selection over kinematic graph hypotheses from noisy 6D pose observations. Subsequent work extends this paradigm through active exploration [20] and coupled recursive estimation during manipulation [21]. Katz et al. [22] recover 3D kinematic models of unknown objects via interactive segmentation and geometric motion analysis. These classical methods share the geometric reasoning philosophy of the proposed approach, but typically require a robot to physically interact with the object during the discovery phase.

Recent learning-based methods relax the interaction requirement. ScrewNet [23] estimates screw parameters from depth image sequences using a deep architecture, unifying revolute, prismatic, and helical joint types. Real2Code [24] formulates articulated reconstruction as code generation, using fine-tuned vision and language models to predict joint parameters. AutoURDF [25] constructs URDF files from point cloud time series via unsupervised cluster registration. Articulate-Anything [26] leverages vision-language foundation models to generate articulation models from diverse input modalities. Although these methods achieve impressive generalization, they rely on trained neural networks or foundation model inference. The proposed method instead operates entirely through geometric analysis—spectral clustering on distance invariance for segmentation, Kabsch-Umeyama alignment for joint estimation, and minimum spanning tree construction for topology—producing a compact joint-angle representation without any learned components.

III. Method

The proposed pipeline operates in two stages: an offline structure discovery stage (Sections III-B–III-E) that recovers the kinematic structure from an RGB-D demonstration using purely geometric reasoning, and an online state estimation stage (Section III-F) that maps each new frame to a compact joint-angle vector for policy learning.

A. Problem Formulation

Consider an articulated mechanism consisting of L rigid links connected by $L-1$ joints, whose kinematic structure is represented as a tree $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with link set $\mathcal{V} = \{1, \ldots, L\}$ and joint set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. Each joint $e \in \mathcal{E}$ is parameterized by its type $\tau_e \in \{\text{revolute}, \text{prismatic}\}$, axis direction $\mathbf{a}_e \in \mathbb{S}^2$, and axis position $\mathbf{o}_e \in \mathbb{R}^3$. The number of links L, the topology $\mathcal{G}$, and the joint parameters are all unknown a priori.

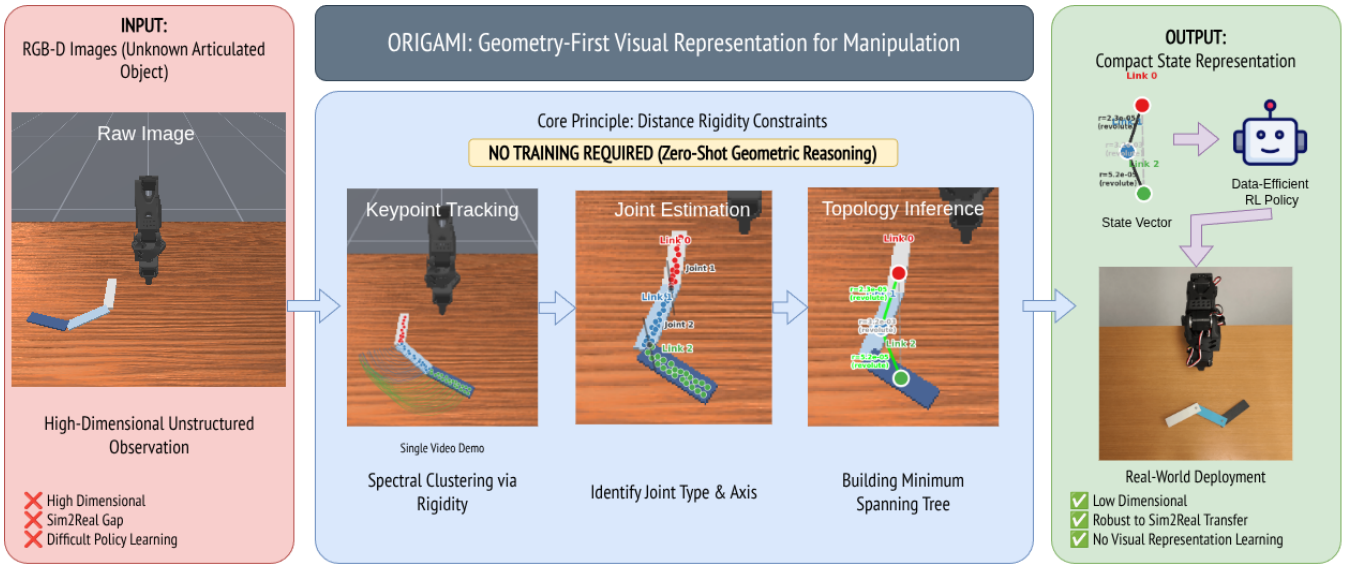

Fig. 1. Instead of collecting large-scale datasets, ORIGAMI uses a single RGB-D demonstration video for geometrical kinematic structure discovery, then feeds the discovered kinematic structure as the input to an RL policy. This state representation reduces reliance on visual generalization, improving robustness to visual domain gaps and sim-to-real transfer compared to purely vision-based RL

The goal is to enable robotic manipulation of passive articulated mechanisms – folding rulers, desk lamps, and pliers – whose kinematic structure is unknown at deployment time. Because such objects possess no actuators, a robot arm must reconfigure them through contact. A manipulation policy that operates directly on raw visual observations must implicitly disentangle the object's kinematic structure from the control strategy. The proposed method decouples these two problems: from RGB-D video of the object in motion, it constructs a kinematic state estimator ϕ that maps any single-frame observation to a compact joint-position vector $\mathbf{q} \in \mathbb{R}^{L-1}$ without any training, learned components, or category-level priors.

The discovered joint space provides a structured observation for goal-conditioned reinforcement learning. Given a target configuration specified as an RGB-D image $\mathbf{I}_{\text{goal}}$, the same estimator produces $\hat{\mathbf{q}}_{\text{goal}} = \phi(\mathbf{I}_{\text{goal}})$, and at each control step the policy observes the configuration displacement $\Delta\hat{\mathbf{q}} = \hat{\mathbf{q}}_{\text{goal}} - \hat{\mathbf{q}}_{\text{curr}}$. Although the estimated angles may deviate from the true configuration due to tracking noise and depth quantization, these errors are systematic properties of the estimator rather than of the particular configuration, and therefore cancel to first order in the difference $\Delta\hat{\mathbf{q}}$. Section III-F provides a formal analysis of this consistency property.

In the offline structure discovery stage, the object is observed through a demonstration sequence of T frames, from which N 3D keypoint trajectories $\mathbf{P} \in \mathbb{R}^{T\times N\times 3}$ with visibility indicators $\mathbf{V} \in \{0,1\}^{T\times N}$ are extracted. The pipeline imposes no constraint on the source of demonstration motion: the demonstration can be generated by a human manually articulating the mechanism, or by a simulator directly commanding the mechanism's joints; the only requirement is that each joint is exercised with sufficient range. From $(\mathbf{P}, \mathbf{V})$, the pipeline recovers (i) a partition $\mathcal{C} : \{1,\dots,N\} \to \{1,\dots,L\}$ assigning keypoints to rigid links (with L discovered automatically), (ii) the kinematic topology $\mathcal{G}$, and (iii) the joint parameters $\{(\tau_e, \mathbf{a}_e, \mathbf{o}_e)\}_{e\in\mathcal{E}}$, by exploiting the pairwise distance rigidity constraint:

$$\mathcal{C}(i) = \mathcal{C}(j) \iff \|\mathbf{p}_{i,t} - \mathbf{p}_{j,t}\| = \text{const}, \quad \forall\, t. \tag{1}$$

This rigidity constraint, together with the assumption that adjacent links are coupled by 1-DOF joints whose relative motion admits a screw parameterization, forms the geometric foundation of the method. At deployment time, the discovered structure maps each new RGB-D observation to a joint-angle vector $\hat{\mathbf{q}}$ via the same geometric procedure. The complete pipeline is illustrated in Fig. 1.

B. Keypoint Extraction and Tracking

Given an RGB-D video sequence of the mechanism in motion, an object mask on the initial frame is obtained using SAM v3 [15], and N keypoints are sampled on a uniform grid within the mask. These keypoints are tracked across all T frames by CoTracker3 [16], which provides 2D trajectories $\hat{\mathbf{P}} \in \mathbb{R}^{T\times N\times 2}$ along with per-point visibility scores $\mathbf{V} \in \{0,1\}^{T\times N}$. Each visible 2D keypoint is then lifted to 3D by querying the corresponding depth map and back-projecting through the known camera intrinsics, yielding the 3D trajectories $\mathbf{P} \in \mathbb{R}^{T\times N\times 3}$ in the camera frame.

Because point tracking is subject to high-frequency jitter, a Savitzky-Golay filter is applied independently to each keypoint trajectory, operating only over visible frames. This smoothing step preserves the underlying rigid body motion while suppressing per-frame noise that would otherwise degrade the downstream distance-based segmentation.

C. Rigid Body Segmentation

In practice, tracking noise and depth quantization make the distance invariance in Eq. (1) only approximate. For each pair (i, j) with $i < j$, attention is restricted to the set of mutually visible frames $\mathcal{T}_{ij} = \{t \mid V_{i,t} = 1 \wedge V_{j,t} = 1\}$, and pairs with $|\mathcal{T}_{ij}| < \lfloor T/4 \rfloor$ are discarded. The pairwise distance time series $d_{ij}(t) = \|\mathbf{p}_{i,t} - \mathbf{p}_{j,t}\|$ is computed for $t \in \mathcal{T}_{ij}$, and its variation is quantified via the Median Absolute Deviation (MAD),

$$m_{ij} = \mathrm{median}_{t\in\mathcal{T}_{ij}} \big| \, d_{ij}(t) - \mathrm{median}_{t\in\mathcal{T}_{ij}} \, d_{ij}(t) \, \big|, \quad (2)$$

which is preferred over the standard deviation for its robustness to outlier frames. For intra-link pairs, $m_{ij} \approx 0$; for cross-link pairs, the joint motion elevates m_{ij} well above the noise floor, provided the demonstration excites the corresponding joint sufficiently.

The MAD values are converted into a pairwise affinity matrix with entries $W_{ij} = \exp(-m_{ij}^2/\gamma^2)$ for valid pairs ($W_{ii} = 1$, and $W_{ij} = 0$ for insufficient co-visibility), where γ is set to the median of all nonzero off-diagonal MAD values. The number of links is determined via spectral analysis of the symmetric normalized graph Laplacian $\mathbf{L}_{\mathrm{sym}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$: each rigid body contributes an eigenvalue near zero, and L is identified by the largest spectral gap,

$$L = \underset{k\in\{1,\ldots,K_{\max}\}}{\arg\max} \; (\lambda_{k+1} - \lambda_k), \quad (3)$$

where $K_{\max}$ is a loose upper bound. The first L eigenvectors are row-normalized and clustered via k-means to yield the partition $\mathcal{C} : \{1, \ldots, N\} \to \{1, \ldots, L\}$. Spectral clustering may over-segment a single rigid body into multiple clusters, particularly when the eigenvalue gap is ambiguous or tracking noise elevates intra-link MAD values. To detect such spurious boundaries, the joint fitting procedure of Section III-D is applied to every candidate cluster pair, and the corresponding fitting residual (Eq. (7) or (8), depending on the classified joint type) is computed. A pair whose residual falls below a fraction α of the median residual across all pairs is merged, because a near-zero residual indicates that the two clusters undergo the same rigid body motion and therefore belong to a single link; in all experiments $\alpha = 0.1$. This merging step is repeated until no pair satisfies the criterion. Additionally, any cluster containing fewer than three keypoints is absorbed into its spatially nearest neighbor, as the Kabsch-Umeyama alignment in Section III-D requires at least three non-collinear points.

D. Joint Parameter Estimation

Given the partition $\mathcal{C}$, the kinematic parameters of each joint are estimated from the relative rigid body motion between link pairs. Let $\mathcal{I}_A = \mathcal{C}^{-1}(A)$ and $\mathcal{I}_B = \mathcal{C}^{-1}(B)$ denote the keypoint index sets of links A and B. A reference frame t_0 is selected, and for each frame t in which sufficient keypoints from both links are visible, the rigid transformation $(\mathbf{R}_t^B, \mathbf{t}_t^B) \in SE(3)$ aligning link B at frame t to its reference position is estimated as

$$\mathbf{R}_t^B = \mathbf{V}\,\mathrm{diag}(1, 1, \det(\mathbf{V}\mathbf{U}^\top))\,\mathbf{U}^\top, \quad \mathbf{t}_t^B = \bar{\mathbf{p}}_{t_0}^B - \mathbf{R}_t^B\,\bar{\mathbf{p}}_t^B, \quad (4)$$

where $\mathbf{H} = \sum_{i\in\mathcal{I}_B}(\mathbf{p}_{i,t_0} - \bar{\mathbf{p}}_{t_0}^B)(\mathbf{p}_{i,t} - \bar{\mathbf{p}}_t^B)^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is the cross-covariance matrix [17]. The same procedure yields $(\mathbf{R}_t^A, \mathbf{t}_t^A)$ for link A, and the relative transformation between the two links at that time is

$$\mathbf{R}_t^{AB} = (\mathbf{R}_t^A)^{-1}\mathbf{R}_t^B, \quad \mathbf{t}_t^{AB} = (\mathbf{R}_t^A)^{-1}(\mathbf{t}_t^B - \mathbf{t}_t^A). \quad (5)$$

Each per-frame Kabsch solve is wrapped in RANSAC to guard against mis-segmented keypoints near link boundaries.

From the sequence $\{(\mathbf{R}_t^{AB}, \mathbf{t}_t^{AB})\}_{t=1}^T$, the joint parameters are extracted. Each rotation is converted to axis-angle form $\theta_t\hat{\boldsymbol{\omega}}_t$: $\theta_t = \arccos\big((\mathrm{tr}(\mathbf{R}_t^{AB}) - 1)/2\big)$ and $[\hat{\boldsymbol{\omega}}_t]_\times = (\mathbf{R}_t^{AB} - \mathbf{R}_t^{AB\top})/(2\sin\theta_t)$, where $[\cdot]_\times$ denotes the skew-symmetric matrix. Frames with $\theta_t < \epsilon_\theta$ are discarded to avoid numerical instability near identity. The consensus axis direction $\mathbf{a}_e \in \mathbb{S}^2$ is obtained as the dominant left singular vector of the stacked unit vectors $\{\hat{\boldsymbol{\omega}}_t\}$.

a) Joint type classification.: The joint type is classified by comparing the accumulated rotation $\Theta = \sum_t |\theta_t - \theta_{t-1}|$ against the accumulated translation $\Delta = \sum_t \|\mathbf{t}_t^{AB} - \mathbf{t}_{t-1}^{AB}\|$: the joint is revolute if Θ/Δ exceeds a threshold ρ, and prismatic otherwise.

b) Axis position estimation.: For revolute joints, the axis position $\mathbf{o}_e \in \mathbb{R}^3$ is the least-squares solution to

$$(\mathbf{R}_t^{AB} - \mathbf{I})\,\mathbf{o}_e = -\mathbf{t}_t^{AB}, \quad t = 1, \ldots, T, \quad (6)$$

which identifies the fixed point of all observed relative rotations. For prismatic joints, the axis direction is taken as the dominant singular vector of $\{\mathbf{t}_t^{AB}\}$.

E. Topology Inference

The preceding steps recover joint parameters for every candidate link pair but do not specify which pairs are physically connected. The topology is determined by the fitting residual of each candidate pair to its estimated 1-DOF joint model. Because revolute and prismatic joints impose different geometric constraints, the residual is defined according to the joint type assigned in Section III-D.

For pairs classified as revolute, the residual measures the fixed-point constraint violation:

$$r_{AB}^{\mathrm{rev}} = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \big\|(\mathbf{R}_t^{AB} - \mathbf{I})\,\hat{\mathbf{o}}_e + \mathbf{t}_t^{AB}\big\|^2. \quad (7)$$

For pairs classified as prismatic, the residual measures the deviation from uniaxial translation:

$$r_{AB}^{\mathrm{pri}} = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \big[\big\|\mathbf{t}_t^{AB} - (\mathbf{t}_t^{AB}\cdot\hat{\mathbf{a}}_e)\,\hat{\mathbf{a}}_e\big\|^2 + \sigma^2\big\|\mathbf{R}_t^{AB} - \mathbf{I}\big\|_F^2\big], \quad (8)$$

where σ denotes the median pairwise keypoint distance and serves as a length scale that balances the translational and rotational terms. The first term penalizes off-axis translation: for an adjacent prismatic pair whose motion is purely along $\hat{\mathbf{a}}_e$, this term vanishes. The second term penalizes rotation

away from identity: for an adjacent prismatic pair with $\mathbf{R}_t^{AB} = \mathbf{I}$, this term likewise vanishes.

Adjacent links exhibit low residual because their relative motion conforms to a single screw, regardless of joint type. By contrast, non-adjacent pairs produce high residual because their motion compounds multiple joints and cannot be explained by any single 1-DOF model. This property holds even when a non-adjacent pair is assigned an incorrect joint type during classification: a revolute residual applied to compound motion fails to find a consistent fixed point, while a prismatic residual applied to compound motion incurs a large rotation penalty from the intervening revolute joints. The minimum spanning tree of the complete graph with edge weights r_{AB} therefore yields the correct kinematic tree $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The root of the tree is assigned to the link whose centroid exhibits the least total displacement across the demonstration, $\ell_{\text{base}} = \arg\min_\ell \sum_t \|\bar{\mathbf{p}}_{\ell,t} - \bar{\mathbf{p}}_{\ell,1}\|$, reflecting the physical prior that the base link remains approximately stationary while distal links undergo larger excursions during manipulation.

F. Online State Estimation

At deployment time, the discovered kinematic model (partition $\mathcal{C}$, topology $\mathcal{G}$, joint parameters) serves as a real-time state estimator ϕ. Given a new RGB-D frame, the same keypoint tracking and depth lifting procedure from Section III-B produces current 3D keypoint positions, which are assigned to links via $\mathcal{C}$. One link is designated as the base, and the relative transformation $\mathbf{R}_t^{AB}$ for each joint $e = (A, B) \in \mathcal{E}$ is computed via Kabsch-Umeyama as in the offline stage. The joint angle is extracted by projecting this rotation onto the consensus axis:

$$q_e(t) = \operatorname{atan2}\big((\mathbf{a}_e \times \mathbf{v}) \cdot \mathbf{v}',\ \ \mathbf{v} \cdot \mathbf{v}'\big), \tag{9}$$

where $\mathbf{v}$ is a unit vector perpendicular to $\mathbf{a}_e$ (fixed at discovery time) and $\mathbf{v}' = \mathbf{R}_t^{AB}\mathbf{v}$. This formulation avoids the numerically unstable per-frame axis extraction. For prismatic joints, $q_e(t) = \mathbf{t}_t^{AB} \cdot \mathbf{a}_e$. The resulting state vector

$$\hat{\mathbf{q}}(t) = \phi(\mathbf{I}_t) = [q_{e_1}(t), \ldots, q_{e_{L-1}}(t)]^\top \in \mathbb{R}^{L-1} \tag{10}$$

replaces the high-dimensional visual observation for policy learning.

a) Consistency analysis.: The estimator ϕ may produce joint angles offset from the true configuration $\mathbf{q}^*$ due to imperfections in tracking, depth sensing, and segmentation. However, these errors are predominantly systematic: they arise from fixed properties of the estimator—the discovered axis direction and position, segmentation boundaries, and tracker bias—rather than from the particular configuration being observed. Decomposing the estimate as $\hat{\mathbf{q}} = \mathbf{q}^* + \boldsymbol{\epsilon}(\mathbf{q}^*)$ and noting that $\boldsymbol{\epsilon}$ varies slowly with $\mathbf{q}^*$, the displacement observed by the goal-conditioned policy satisfies

$$\Delta\hat{\mathbf{q}} = \Delta\mathbf{q}^* + \boldsymbol{\epsilon}(\mathbf{q}^*_{\text{goal}}) - \boldsymbol{\epsilon}(\mathbf{q}^*_{\text{curr}}) \approx \Delta\mathbf{q}^*, \tag{11}$$

so the policy observes approximately correct relative displacements even when the absolute estimates carry non-trivial bias. More formally, the residual error is $\mathbf{J}_\epsilon\,\Delta\mathbf{q}^* + O(\|\Delta\mathbf{q}^*\|^2)$, where $\mathbf{J}_\epsilon = \partial\boldsymbol{\epsilon}/\partial\mathbf{q}^*$. The dominant error sources (axis direction bias, axis position offset) are configuration-independent to first order, yielding $\|\mathbf{J}_\epsilon\| \approx 0$. This cancellation property holds for both revolute and prismatic joints: the revolute angle computed via Eq. (9) depends on the fixed axis $\mathbf{a}_e$ and reference vector $\mathbf{v}$, while the prismatic displacement $\mathbf{t}_t^{AB} \cdot \mathbf{a}_e$ is a linear projection onto the fixed axis direction; in both cases, the bias arises from the axis estimate rather than from the configuration, and therefore cancels to first order in the difference $\Delta\hat{\mathbf{q}}$. This cancellation property eliminates the need for any calibration or learned correction: the raw geometric estimates are directly usable as policy observations. Furthermore, the low dimensionality of the representation ($L-1$ scalars compared to raw images or $N \times 3$ keypoint coordinates) reduces the exploration burden, enabling sample-efficient learning. The resulting displacement vector $\Delta\hat{\mathbf{q}} = \hat{\mathbf{q}}_{\text{goal}} - \hat{\mathbf{q}}_{\text{curr}}$ serves as the observation for a reinforcement learning policy; the complete training protocol is described in Section IV-A.

IV. Experiments

These experiments test whether the joint-angle representation discovered by ORIGAMI closes the sample efficiency gap between pixel-based and privileged-state reinforcement learning on goal-conditioned articulated object manipulation. Section IV-A describes the shared experimental protocol. Sections IV-C and IV-D present results on revolute and prismatic joints, respectively.

A. Setup

a) Mechanisms and environments.: Two mechanism categories are considered: a 3-link folding ruler with 2 revolute joints in a chain topology, and three PartNet-Mobility [27] retractable utility knives each with a single prismatic joint. All simulation experiments use SAPIEN [27] via ManiSkill3 [8]. The robot is an SO-101 arm (5-DOF plus gripper) controlled in end-effector delta-position mode with fixed orientation, yielding a 3-dimensional action space $\mathbf{a} \in \mathbb{R}^3$.

b) Task and observations.: At each episode, initial and target mechanism configurations are sampled uniformly within joint limits, subject to a minimum displacement threshold. Every baseline receives the robot's proprioceptive state (joint positions and velocities); the methods differ in how the current and target object configurations are represented. The GT State oracle reads both configurations directly from the simulator as ground-truth joint-position displacements $\Delta\mathbf{q}^* \in \mathbb{R}^{n_j}$; in real-world experiments, these values are obtained from AprilTag markers tracked by a calibrated camera. ORIGAMI and AutoURDF each process the current and target 256×256 RGB-D images (converted to point clouds for AutoURDF) through their respective estimators to obtain $\Delta\hat{\mathbf{q}} \in \mathbb{R}^{n_j}$. DrQ-v2 and CNN+SAC receive the current and target RGB-D images directly and encode them with a CNN trained end-to-end with the policy. An episode is deemed successful when all link positions fall within $1\,\text{cm}$ of the target configuration.

c) Reward.: The reward functions are task-specific. For the folding ruler, the per-step reward combines a weighted root-link pose error and joint-angle errors: $r_t = -10\,(\|\mathbf{p}_{\text{root}} - \mathbf{p}^*_{\text{root}}\|_2 + d_q(\mathbf{q}_{\text{root}}, \mathbf{q}^*_{\text{root}})) - \sum_{i=1}^{2} |\theta_i - \theta^*_i| + 10 \cdot \mathbf{1}[\text{success}]$, where d_q denotes quaternion distance and θ_i is the i-th joint angle. The higher weight on the root-link term reflects the empirical finding that correcting the root pose dominates task success. For the utility knife, the reward sums the position errors of the blade and handle links: $r_t = -(\|\mathbf{p}_{\text{blade}} - \mathbf{p}^*_{\text{blade}}\|_2 + \|\mathbf{p}_{\text{handle}} - \mathbf{p}^*_{\text{handle}}\|_2) + 10 \cdot \mathbf{1}[\text{success}]$. Both tasks use a success threshold of 1 cm and no time penalty.

d) Demonstration for structure discovery.: A single demonstration of T=250 frames is collected per mechanism by commanding random waypoint trajectories in which all joints move simultaneously, with each joint traversing at least 30° of its range. This motion pattern avoids the artificial separability of sequential single-joint activation.

e) Baselines: Six observation modalities are compared under SAC [28] with an identical MLP policy (3 hidden layers of 256 units, selected following the DrQ-v2 [2] default):

1) **GT State** (oracle): ground-truth joint-position displacement $\Delta\mathbf{q}^* \in \mathbb{R}^{n_j}$ concatenated with end-effector pose.
2) **ORIGAMI**: discovered joint-position displacement $\Delta\hat{\mathbf{q}} \in \mathbb{R}^{n_j}$ from the geometric state estimator, using the same observation dimensionality as the oracle.
3) **AutoURDF** [25]: joint-position displacement extracted from the URDF produced by AutoURDF's cluster-based registration pipeline. RGB-D observations are converted to point clouds to match AutoURDF's expected input format.
4) **R3M+SAC** [4]: a frozen R3M ResNet-18 encoder maps current and goal RGB images to 512-dimensional feature vectors, concatenated with end-effector pose and fed to the same MLP policy. The encoder is not finetuned during RL training.
5) **DrQ-v2** [2]: CNN encoder trained end-to-end with the actor and critic on current and goal RGB-D images, with random shift augmentation.
6) **CNN+SAC**: the same CNN encoder architecture without data augmentation.

B. Structure Discovery Accuracy

Before evaluating downstream policy performance, this subsection directly measures the accuracy of ORIGAMI's geometric pipeline against simulator ground truth. Table I reports results on the folding ruler (two revolute joints) and on Knife 101057 (one prismatic joint). Six metrics are evaluated: Adjusted Rand Index (ARI) for rigid-body segmentation quality, joint-axis direction error (degrees), joint-axis position error (mm, revolute only), joint-type classification accuracy, and two joint-position estimation metrics: range error (the absolute error in the estimated total range of motion) and displacement RMSE (the root-mean-square error of pairwise displacements $\Delta\hat{q} - \Delta q^*$ over 500 randomly sampled start/goal pairs).

TABLE I

STRUCTURE DISCOVERY ACCURACY AGAINST SIMULATOR GROUND TRUTH.

Metric	Ruler (Rev.)	Knife (Pris.)
Segmentation ARI	0.874	0.792
Axis Direction Error	1.71°	0.63°
Axis Position Error	0.62 mm	—
Joint Type Accuracy	100%	100%
Range Error	3.9°	4.2 mm
Displacement RMSE	4.20°	3.7 mm

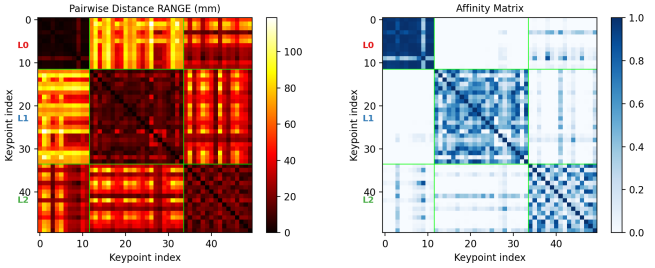

Fig. 2. Rigid body segmentation on the folding ruler. (Left) Pairwise distance range matrix D_{ij} (mm) between tracked keypoints, sorted by discovered link assignment. Dark entries indicate rigid pairs; bright entries span different links. (Right) The derived affinity matrix $W_{ij} = \exp(-D^2_{ij}/2\gamma^2)$ fed to spectral clustering. Three links are recovered with no mis-assigned keypoints.

ORIGAMI recovers the correct chain topology for the ruler (three links, two joints) and the correct two-link structure for the knife in all cases. The segmentation ARI of 0.874 on the ruler indicates that the spectral clustering step assigns the large majority of keypoints to their true links, with residual confusion concentrated near the joint boundaries. Axis direction and position errors are small in absolute terms (1.71° and 0.62 mm on the ruler), confirming that the Kabsch–Umeyama alignment followed by least-squares axis fitting produces geometrically faithful joint parameters from a single demonstration.

The range error and displacement RMSE quantify the accuracy of the online estimates observed by the policy. Although moderate in absolute magnitude, these errors are largely shared between the current and goal configurations and cancel in the displacement $\Delta\hat{\mathbf{q}}$, as analyzed in Section III-F; the displacement RMSE thus reflects the residual after cancellation. The near-oracle performance in Section IV-C confirms that a residual of this magnitude is compatible with effective goal-conditioned policy learning.

C. Revolute Joint: Folding Ruler

Figure 2 visualizes the intermediate segmentation step: pairwise distance invariance cleanly separates intra-link pairs (low variability) from inter-link pairs (high variability).

The learning curves in Figure 3 compare the effectiveness of the constructed state representation to that of the ground truth and alternative representations in RL. It reveals a clear hierarchy. ORIGAMI tracks the GT State oracle

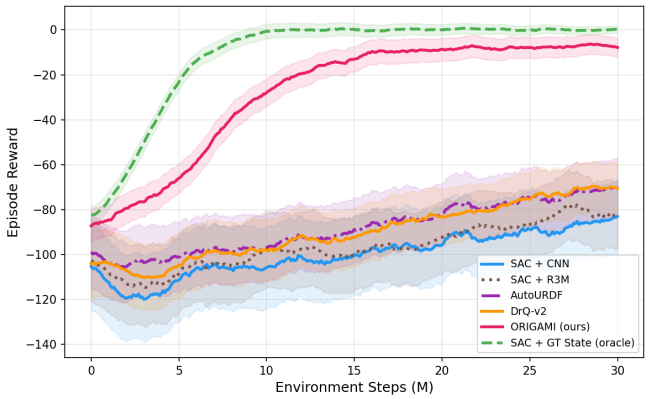

Fig. 3. Learning curves (mean ± 1 s.d. over 10 seeds) on the folding ruler task. ORIGAMI converges within ∼5M steps to near-oracle performance. AutoURDF plateaus at a level comparable to DrQ-v2, despite producing a low-dimensional state.

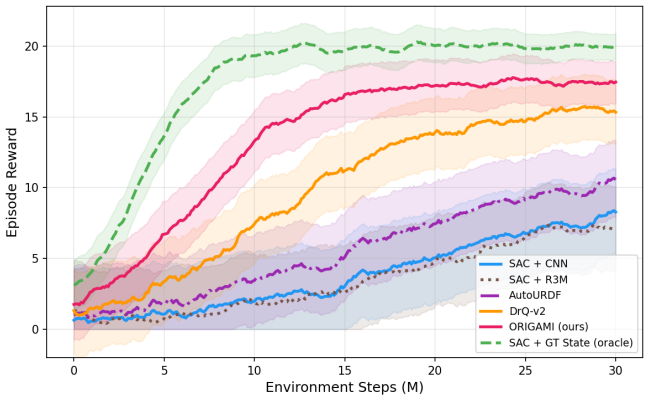

Fig. 4. Learning curves on Knife 101057 (prismatic joint, mean ± 1 s.d. over 10 seeds). ORIGAMI reaches near-oracle performance within 10M steps. AutoURDF and both visual baselines plateau at substantially lower success rates.

TABLE II

FOLDING RULER: SUCCESS RATE (%) AND RETURN (MEAN ± 1 S.D.) IN SIMULATION (30M STEPS, 100 EVALUATION EPISODES, 10 SEEDS) AND ON THE REAL ROBOT (50 EPISODES, 3 SEEDS).

Observation	Sim Success	Sim Return	Real Success
GT State	90.7±3.5	−13.2±2.1	53.0±5.0
ORIGAMI	64.2±7.8	−22.6±4.5	44.7±7.0
AutoURDF	25.5±8.1	−71.3±15.2	11.7±4.2
DrQ-v2	27.1±7.2	−79.5±18.0	9.3±3.5
CNN+SAC	13.0±6.5	−94.7±22.0	0.0±0.0

closely and converges within approximately 5M environment steps, confirming that the discovered joint-angle representation captures sufficient task-relevant information. DrQ-v2 improves over vanilla CNN+SAC through random shift augmentation but plateaus well below ORIGAMI, indicating that augmentation alone does not close the pixel-to-state gap. AutoURDF, despite also producing a low-dimensional joint-angle vector, performs comparably to DrQ-v2 rather than to the oracle. This result is attributable to AutoURDF's lack of temporal consistency: because its pipeline performs independent point cloud clustering and registration at each frame, the discovered link assignments and joint parameters can shift discontinuously between consecutive observations, injecting noise that undermines the sample efficiency advantage of a compact representation. ORIGAMI avoids this failure mode by maintaining persistent keypoint identities via CoTracker correspondences, yielding temporally smooth joint-angle estimates that allow the value function to propagate reward information reliably. The final success rates and returns in Table II confirm these trends quantitatively; the difference between ORIGAMI and all non-oracle baselines is statistically significant ($p < 0.01$, Welch's t-test across seeds).

a) Real robot.: The folding ruler task is additionally evaluated on a physical LeRobot SO-101 arm with a single RGB-D camera. The structure discovery demonstration is collected by manually articulating the ruler, and the geometric pipeline processes this recording without modification. AprilTag markers affixed to each link provide ground-truth joint angles for the oracle baseline and for evaluation. Policies trained in simulation are deployed zero-shot. As shown in the rightmost column of Table II, ORIGAMI retains the majority of its simulated performance and outperforms all visual baselines by a wide margin. CNN+SAC fails entirely under sim-to-real domain shift, and AutoURDF degrades to a level comparable to DrQ-v2. The robustness of ORIGAMI stems from the domain-invariant nature of its representation: keypoint tracks and their geometric relationships remain consistent across rendered and real images, unlike raw pixel features.

D. Prismatic Joint: Retractable Utility Knife

Three PartNet-Mobility retractable utility knives (IDs 101057, 101062, and 101085), each with a single prismatic joint, are used to test whether the pipeline generalizes beyond revolute joints. From a single demonstration per object, ORIGAMI correctly segments blade and handle and classifies the joint as prismatic.

The learning curves on Knife 101057 (Figure 4) and the success rates across all three knife instances (Table III) confirm that the trends observed on the folding ruler carry over to prismatic joints. ORIGAMI consistently tracks the oracle, while AutoURDF again falls short of its potential as a compact representation due to the same per-frame inconsistency discussed in Section IV-C. The differences between ORIGAMI and all non-oracle baselines are statistically significant across all three knife instances ($p < 0.01$).

a) Real robot.: Knife 101057 is further validated on the physical SO-101 arm, with AprilTag markers on blade and handle providing ground-truth prismatic displacement. As in the revolute case, ORIGAMI transfers robustly while all visual baselines degrade under the sim-to-real appearance gap (Table III, bottom row).

TABLE III

UTILITY KNIFE SUCCESS RATE (%, MEAN ± 1 S.D.) IN SIMULATION (30M STEPS, 100 EPISODES, 10 SEEDS) AND REAL WORLD (50 EPISODES, 3 SEEDS).

Observation	Simulation			Real
	101057	101062	101085	101057
GT State	75.3±4.2	73.8±5.8	69.1±4.4	55.7±6.3
ORIGAMI	52.1±8.3	49.6±7.1	50.3±7.3	37.3±9.9
AutoURDF	37.2±7.5	37.8±8.1	39.1±7.8	16.5±5.1
DrQ-v2	38.7±9.7	34.2±8.3	36.5±8.9	14.0±3.5
CNN+SAC	14.4±5.2	10.1±8.3	12.8±7.3	2.0±1.0

V. CONCLUSION

This paper presented ORIGAMI, a geometric pipeline that constructs compact representations for articulated object manipulation from a single RGB-D demonstration without task-specific training. Experiments on revolute and prismatic joints, in both simulation and on real hardware, demonstrated that the discovered representation closes much of the pixel-to-state efficiency gap and transfers zero-shot to a physical robot. The experimental evaluation is limited to planar manipulation; the underlying geometric machinery operates in full SE(3), but spatially articulated objects with branching topologies and self-occlusion have not yet been validated. In addition, all tested objects contain a single joint type, the evaluated tasks are restricted to goal-conditioned reconfiguration, and the pipeline assumes sufficient joint excitation during the demonstration. Extending the framework to 3D articulated manipulation, richer manipulation tasks, and active exploration of under-excited joints are promising directions for future work.

REFERENCES

[1] D. Yarats, A. Zhang, I. Kostrikov, B. Amos, J. Pineau, and R. Fergus, "Improving sample efficiency in model-free reinforcement learning from images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 10 674–10 681.

[2] D. Yarats, R. Fergus, A. Lazaric, and L. Pinto, "Mastering visual continuous control: Improved data-augmented reinforcement learning," in *International Conference on Learning Representations*, 2022.

[3] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas, "Reinforcement learning with augmented data," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 19 884–19 895.

[4] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3M: A universal visual representation for robot manipulation," in *Proceedings of The 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 205, 2022, pp. 892–909.

[5] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "VIP: Towards universal visual reward and representation via value-implicit pre-training," in *International Conference on Learning Representations*, 2023.

[6] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik, "Masked visual pre-training for motor control," *arXiv preprint arXiv:2203.06173*, 2022.

[7] T. Mu, Z. Li, S. Strzelecki, X. Yuan, Y. Yao, L. Liang, and H. Su, "When should we prefer state-to-visual DAgger over visual reinforcement learning?" in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[8] S. Tao, F. Xiang, A. Shukla, Y. Qin, X. Hinrichsen, X. Yuan, C. Bao, X. Lin, Y. Liu, T.-k. Chan, Y. Gao, X. Li, T. Mu, N. Xiao, A. Gurha, Z. Huang, R. Calandra, R. Chen, S. Luo, and H. Su, "ManiSkill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai," *arXiv preprint arXiv:2410.00425*, 2024.

[9] M. Laskin, A. Srinivas, and P. Abbeel, "CURL: Contrastive unsupervised representations for reinforcement learning," in *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 5639–5650.

[10] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," *arXiv preprint arXiv:2301.04104*, 2023.

[11] N. Hansen, H. Su, and X. Wang, "TD-MPC2: Scalable, robust world models for continuous control," in *International Conference on Learning Representations*, 2024.

[12] X. Li, H. Wang, L. Yi, L. J. Guibas, A. L. Abbott, and S. Song, "Category-level articulated object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3706–3715.

[13] Y. Weng, H. Wang, Q. Zhou, Y. Qin, Y. Duan, Q. Fan, B. Chen, H. Su, and L. J. Guibas, "CAPTRA: CAtegory-level pose tracking for rigid and articulated objects from point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 209–13 218.

[14] Z. Jiang, C.-C. Hsu, and Y. Zhu, "DITTO: Building digital twins of articulated objects from interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5616–5626.

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

[16] N. Karaev, Y. Makarov, J. Wang, N. Neverova, A. Vedaldi, and C. Rupprecht, "Cotracker3: Simpler and better point tracking by pseudo-labelling real videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 6013–6022.

[17] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, 1991.

[18] S. James, Z. Ma, D. Rovick Arrojo, and A. J. Davison, "RLBench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.

[19] J. Sturm, C. Stachniss, and W. Burgard, "A probabilistic framework for learning kinematic models of articulated objects," *Journal of Artificial Intelligence Research*, vol. 41, pp. 477–526, 2011.

[20] K. Hausman, S. Niekum, S. Osentoski, and G. S. Sukhatme, "Active articulation model estimation through interactive perception," in *IEEE International Conference on Robotics and Automation*, 2015, pp. 3305–3312.

[21] R. Martín-Martín and O. Brock, "Coupled recursive estimation for online interactive perception of articulated objects," *The International Journal of Robotics Research*, vol. 41, no. 8, pp. 741–777, 2022.

[22] D. Katz, M. Kazemi, J. A. Bagnell, and A. Stentz, "Interactive perception of articulated objects," in *IEEE International Conference on Robotics and Automation*, 2013, pp. 5003–5010.

[23] A. Jain, R. Lioutikov, C. Chuck, and S. Niekum, "ScrewNet: Category-independent articulation model estimation from depth images using screw theory," in *IEEE International Conference on Robotics and Automation*, 2021, pp. 13 670–13 677.

[24] Z. Mandi, Y. Weng, D. Bauer, and S. Song, "Real2Code: Reconstruct articulated objects via code generation," in *International Conference on Learning Representations*, 2025.

[25] J. Lin, L. Zhang, K. Lee, J. Ning, J. Goldfeder, and H. Lipson, "AutoURDF: Unsupervised robot modeling from point cloud frames using cluster registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[26] L. Le, J. Xie, W. Liang, H.-J. Wang, Y. Yang, Y. J. Ma, K. Vedder, A. Krishna, D. Jayaraman, and E. Eaton, "Articulate-anything: Automatic modeling of articulated objects via a vision-language foundation model," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 17 578–17 602.

[27] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang *et al.*, "Sapien: A simulated part-based interactive environment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 097–11 107.

[28] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. Pmlr, 2018, pp. 1861–1870.