Learning Dexterous Deformable Object Manipulation Through Cross-Embodiment Dynamics Learning

Author Names Omitted for Anonymous Review. Paper-ID [add your ID here]

***Abstract*—Dexterous manipulation of deformable objects remains a core challenge in robotics due to complex contact dynamics and high-dimensional control. While humans excel at such tasks, transferring these skills to robots is hindered by embodiment gaps. In this work, we propose using particle-based dynamics models as an embodiment-agnostic interface, enabling robots to learn directly from human-object interaction data. By representing both manipulators and objects as particles, we define a shared state and action space across embodiments. Using human demonstrations, we train a graph neural network dynamics model that leverages spatial locality and equivariance to generalize across differing embodiment shapes and structures. For control, we convert embodiment-specific joint actions into particle displacements via forward kinematics, enabling model-based planning in the shared representation space. We demonstrate that our approach transfers manipulation skills from humans to both low-DoF and high-DoF robot hands, achieving real-world clay reshaping without motion retargeting, expert demonstrations, or analytical simulation.**

I. Introduction

Dexterous manipulation of deformable objects is a fundamental yet unsolved problem in robotics. The challenge arises from the need to coordinate a high-dimensional action space, often involving many joints and coupled contacts, with complex object dynamics that are difficult to model or predict [5]. Humans perform such tasks effortlessly, adapting finger motions to stabilize, reorient, or reshape objects in-hand. Replicating this level of fine-grained control in robotic systems, however, remains elusive. This gap raises a natural question: Can we learn from human behavior to endow robots with similar dexterous capabilities, particularly for deformable object manipulation?

Leveraging human experience offers a promising path toward dexterous manipulation, but the embodiment gap remains a central challenge. Common strategies involve using motion retargeting to convert human motion into robots' action space, followed by imitation learning from expert demonstrations in the task domain [14, 3, 9], or reinforcement learning that treats natural human-object interaction as a prior or guidance signal [2, 13, 8, 7]. However, these approaches face key limitations: optimization-based retargeting is often imprecise, imitation learning depends on curated demonstrations, and reinforcement learning requires reward design and high-fidelity simulation, which are especially difficult in contact-rich deformable object manipulation tasks where fine motion control is essential.

In this work, we seek to enable robots to learn directly from human experience. The core idea is to define a ***universal state and action representation*** shared across human and robot embodiments. We propose to represent both human and robot hands as particles, and actions are defined as particle displacements. We train a graph-based dynamics model [10, 1, 11, 12, 16] to predict particle motion from human-object interaction data, leveraging *spatial locality* and *equivariance* to promote generalization across morphologies. For control, joint actions are sampled and mapped to particle motions via forward kinematics, enabling model-predictive planning through dynamics inference in the shared representation space. This structured approach allows robots to perform fine-grained manipulation of deformable objects directly from human-object interaction data, without requiring motion retargeting, expert demonstrations, or analytical simulation.

We evaluate our approach on two robotic hands with distinct kinematics, a 6-DoF PSYONIC Ability Hand and a 12-DoF Robot Era XHand, on real-world dough reshaping tasks. A single dynamics model trained solely on human-object interaction data enables both embodiments to perform fine-grained manipulation without additional robot data. These results provide preliminary findings that particle-based dynamics models can serve as a generalizable interface for cross-embodiment dexterous manipulation.

II. Method

A. Problem Formulation

Our goal is to learn dexterous deformable object manipulation skills that can generalize across robotic hands.

We formalize the general problem as follows. At each time step t, the end effector is in configuration $q_t \in \mathbb{R}^{n_e}$, where n_e is the number of degrees of freedom specific to embodiment e. The system transitions to a new state after executing control u_t, and the objective is to find a control sequence $\{u_t\}_{t=0}^{H-1}$ that minimizes a cost function over a finite horizon H. While conventional learning focuses on optimizing performance within a fixed embodiment, cross-embodiment skill learning introduces the additional challenge of bridging structural and dimensional mismatches in both configuration q_t and action u_t across different morphologies.

To enable generalization across embodiments, we define a ***shared action and state space*** in the form of particles. Let $\phi_e : \mathbb{R}^{n_e} \to \mathbb{R}^{3N}$ denote the embodiment-specific forward kinematics function that maps joint configuration q_t to the

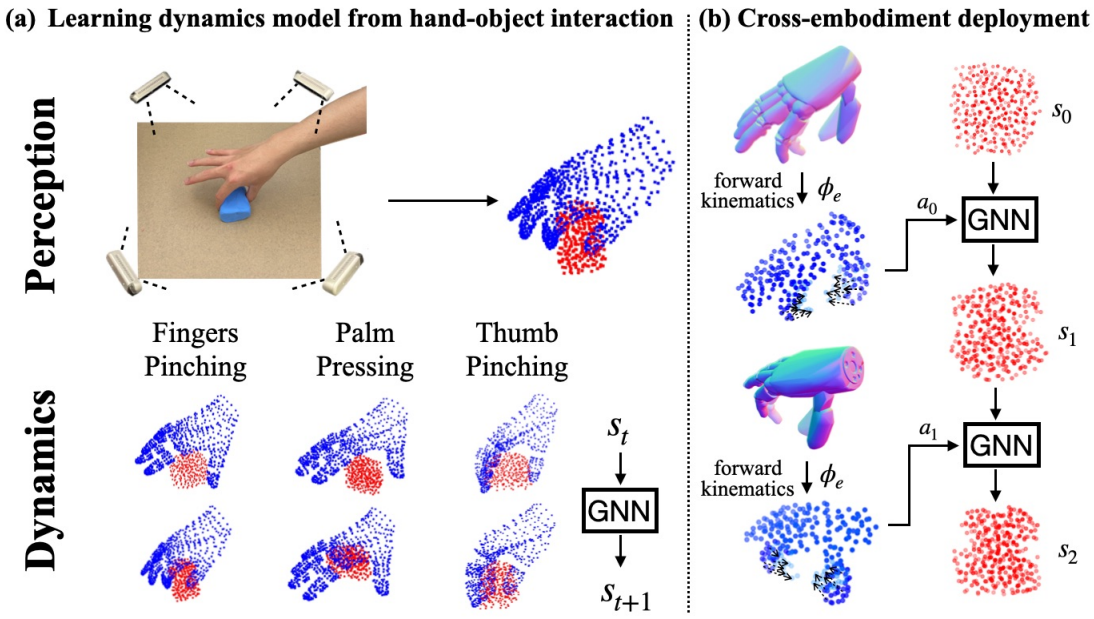

Fig. 1: Overall framework. The perception module captures a clear and consistent particle-based representation of the scene using four RGBD cameras. Leveraging this representation, our method learns a dynamics model from diverse hand-object interactions across multiple manipulation skills, including Fingers Pinching, Palm Pressing, and Thumb Pinching. During cross-embodiment planning, the learned dynamics model is employed to manipulate deformable objects toward desired target states.

3D positions of N particles representing the end effector. We define the action in the particle space as the displacement of these particles:

$$a_t = \phi_e(q_{t+1}) - \phi_e(q_t),$$

and the state in the particle space as the union of object and hand particles:

$$s_t = \left[s_t^{obj}; \phi_e(q_t)\right].$$

We then train a dynamics model f that operates in this shared space, predicting the next particle state given the current particle state and action:

$$\hat{s}_{t+1} = f(s_t, a_t).$$

The model can be trained with data from different embodiments, including robots and humans, due to the unified state and action representations. For control, we sample joint actions, convert them to particle motion using forward kinematics, and integrate the learned dynamics model with model-predictive control to minimize the task cost.

B. Perception

The perception module in our approach aims to uniformly sample particles from both the human hand and the object. To achieve this, we utilize 4 well-calibrated cameras for comprehensive RGBD images as well as point-cloud information.

For hand perception, we extract hand mesh in 3D space utilizing a multi-view hand mesh reconstruction model [15]. We then apply farthest point sampling (FPS) on this mesh to uniformly select 200 representative points, which serve as the final hand particle representation. For object perception, we follow the particle sampling procedure in [11].

To mitigate visual occlusion between hand and object during data collection, we adopt key-frame perception strategy: hand mesh is reconstructed before contact and at the deepest deformation point while object is captured before and after the interaction.

C. Dynamics Model

To enable generalization across embodiments, we model the object's dynamics as a graph-based neural network that predicts particle motions. We use DPI-Net [6], which leverages message-passing over the particle graph to compute the object's forward dynamics.

The GNN implements a message-passing update that aggregates local features over the graph edges to predict each particle's motion. Specifically, the graph state at each time step is represented as $s_t = \langle O_t, E_t \rangle$ with O_t as vertices and E_t as edges. For each particle in the graph, $o_{t,i} = \langle x_{i,t}, c^o_{i,t} \rangle$, where $x_{i,t}$ is the particle position i at time t, and $c^o_{i,t}$ is the particle's attributes at time t, including the group information (belongs to hand or object). In addition, edges between particles are denoted as $e_k = \langle u_k, v_k \rangle$, where $1 \leq u_k, v_k \leq |O_t|$ are the receiver and sender particle indices respectively. Given the graph, where any particles are connected within a certain radius, we can first use node encoder f_O^{enc} and edge encoder f_E^{enc} to extract node and edge features:

$$c^o_{i,t} = f_O^{enc}(o_{i,t}), c^e_{k,t} = f_E^{enc}(o_{u_k,t}, o_{v_k,t}, d^r_k)$$

where d^r_k denotes edge's attributes (e.g. length). Then, the features are propagated through edges in multiple steps. Denote $\epsilon^l_{k,t}$ and $h^l_{i,t}$ are propagating influence from edge k and node i

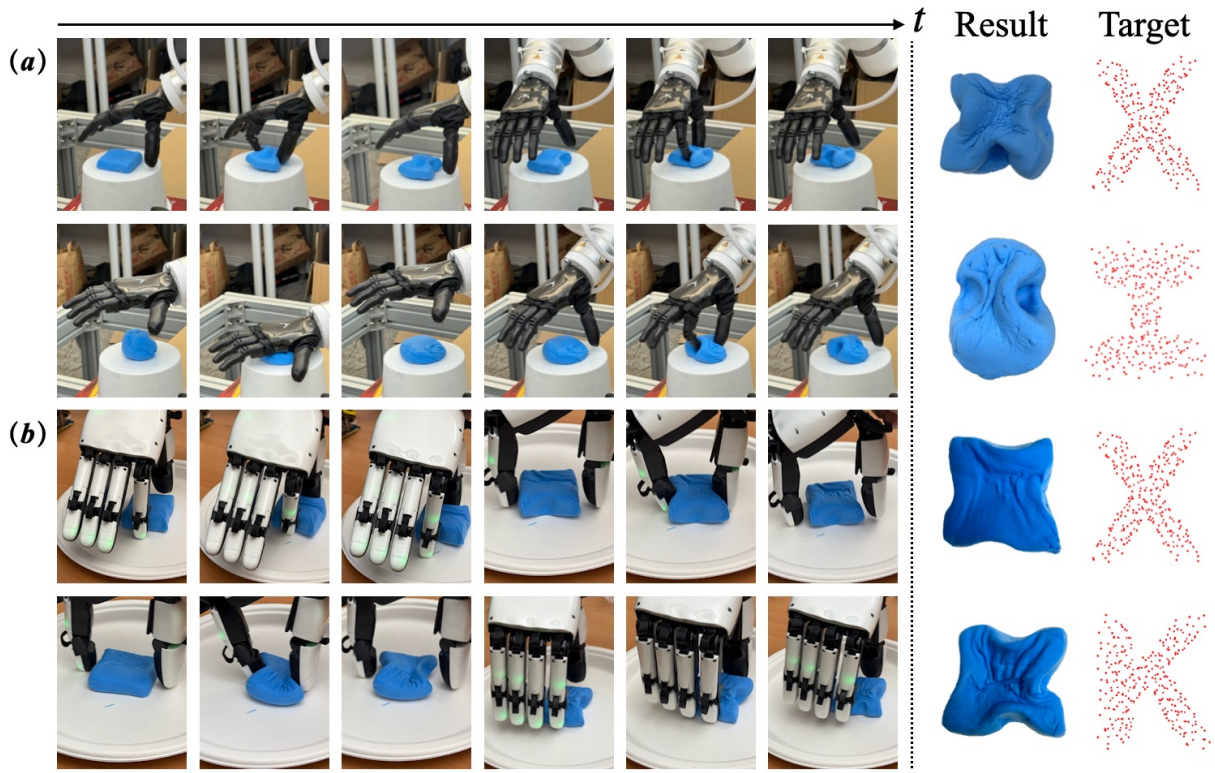

Fig. 2: Qualitative results of cross-embodiment deployment. (a) Ability Hand (6-DoF) and (b) XHand (12-DoF) utilize the same particle-space dynamics model learned from human demonstration. For each trial, the hand successfully reshapes the deformable clay toward the target shape using a combination of ***Fingers Pinching***, ***Palm Pressing***, and ***Thumb Pinching*** skills.

at step l, respectively. At step 0, initialize $h^0_{i,t} = 0, i = 1...|O|$. For each step $1 \leq l \leq L$:

$$\epsilon^l_{k,t} = f_E(c^e_{k,t}, h^{l-1}_{u_k,t}, h^{l-1}_{v_k,t}), k = 1...|E|$$
$$h^l_{i,t} = f_O(c^o_{i,t}, \sum_{k \in \mathcal{N}_i} \epsilon^l_{k,t}, h^{l-1}_{i,t}), i = 1...|O|$$

where $\mathcal{N}_i$ is the neighbor index set of particle i, f_O denotes the node propagator, and f_E denotes the edge propagator. Then the future state at time $t+1$ is predicted as

$$\hat{o}_{i,t+1} = f^{dec}_O(h^L_{i,t}), i = 1...|O|$$

The particle-based graph network incorporates strong inductive biases. First, spatial locality is enforced by restricting message passing to local neighborhoods. Second, equivariance is achieved through the use of relative coordinates and shared update functions, making predictions invariant to global translations and rotations. These properties support generalization across embodiments when configurations are projected into the shared particle space.

D. Model-Predictive Control

Human hand motions often lie in low-dimensional manifolds in the entire configuration space [4]. To leverage this structure, we define low-dimensional action parameterizations for efficient planning: (i) **Fingers Pinching**: involving rotational motion around the z-axis and relative movement between the index finger and thumb; (ii) **Palm Pressing**: characterized by rotational motion around the z-axis and translational motion along the z-axis; (iii) **Thumb Pinching**: composed of rotational motion around the z-axis and actuation of thumb-specific degrees of freedom.

To perform model-based control, we sample a set of control sequence $\{u_t\}^{H-1}_{t=0}$ from the robot hand's action space of each skill, which are then converted to particles in shared state space through forward kinematics . We evaluate the sampled action sequences using the dynamics model f learned from human-object interaction data and execute the lowest-cost actions. The cost function is defined as the Earth Mover's Distance (EMD) between the predicted final object state $\hat{s}^{obj}$ and the target object state s^{obj}_{goal}.

III. Experiments

In this section, we seek to answer the following key questions:

i. Does the learned dynamics model generalize across different hand morphologies?
ii. Does the learned dynamics model enable effective planning and control for dexterous manipulation?

A. Physical Setup

1) Platform.: The experimental platform consists of a XArm7 robotic arm with an Ability Hand and a XHand. Four Intel RealSense cameras are utilized for perception.All devices are connected to a workstation with a NVIDIA RTX 4090 GPU for both data collection and evaluation.

2) Protocols: For dynamics model training, we take 30 minutes to collect human hand demonstrations for ***Fingers Pinching*** skill, and 15 minutes for ***Palm Pressing*** and ***Thumb Pinching*** skills. These demonstrations are used to jointly train a single dynamics model across all skills. During evaluation, we select 3 alphabet letters "X", "K", and "I" as target shapes, and we run 5 trials per hand (Ability Hand and XHand) for each target shapes.

B. Real-World Manipulation of Deformable Objects

Methods	CD↓	EMD↓
Human Subjects	0.0104 ± 0.0013	0.0077 ± 0.0010
Ability Hand	0.0109 ± 0.0008	0.0080 ± 0.0006
XHand	**0.0100 ± 0.0005**	**0.0076 ± 0.0004**

TABLE I: Quantitative results using different embodiments. Numbers are averaged over all tested shapes.

During real-world cross-embodiment deployment, the robot hand aims to reshape the deformable clay toward the target shape using a combination of skills. As shown in Fig. 2, both *(a)* Ability Hand and *(b)* XHand sequences can reshape the clay well in the use of three learned skills-***Fingers Pinching***, ***Palm Pressing***, and ***Thumb Pinching***-to successively carve, spread, and compress clay. Despite the hands' kinematic disparity, the same particle-space dynamics model transfers with no retraining. This supports the claim that the graph-based dynamics model enables cross-embodiment planning via model-predictive control.

In addition, the graph-based dynamics model enables robot hands a near-human accuracy. We report the evaluation result in Table I. Our proposed embodiment-agnostic dynamics model enables robot hand to achieve near-human accuracy and high repeatability. XHand attains the lowest Chamfer Distance (0.0100 ± 0.0005) and Earth Mover Distance (0.0076 ± 0.0004), edging slightly ahead of human subjects, while the lower-DoF Ability Hand follows closedly at 0.00109 ± 0.0008 (CD) and 0.0080 ± 0.0006 (EMD).The small performance gap demonstrates that the learned particle-space dynamics model transfers effectively across embodiments with very different kinematics. Notably, the standard deviations shrink as we move from humans to Ability Hand, and XHand, revealing that the model-predictive controller generates more consistent motions than human manipulators.

IV. Conclusion

This work demonstrates that a particle-space dynamics model learned solely from human–object interactions can serve as an embodiment-agnostic interface for dexterous manipulation of deformable objects. By projecting both human and robot hands into a shared particle representation and combining this with model-predictive control, we successfully acquire three manipulation skills on two robot hands with distinct kinematics. These results highlight the potential of general dynamics models as a unified interface for cross-embodiment manipulation.

References

[1] Bo Ai et al. "RoboPack: Learning Tactile-Informed Dynamics Models for Dense Packing". In: *Robotics: Science and Systems*. 2024.

[2] Zerui Chen et al. "Vividex: Learning vision-based dexterous manipulation from human videos". In: *arXiv preprint arXiv:2404.15709* (2024).

[3] Xuxin Cheng et al. "Open-TeleVision: Teleoperation with Immersive Active Visual Feedback". In: *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*. Ed. by Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard. Vol. 270. Proceedings of Machine Learning Research. PMLR, 2024, pp. 2729–2749. URL: https://proceedings.mlr.press/v270/cheng25b.html.

[4] Thomas Feix et al. "The GRASP Taxonomy of Human Grasp Types". In: *IEEE Trans. Hum. Mach. Syst.* 46.1 (2016), pp. 66–77.

[5] Sizhe Li et al. "DexDeform: Dexterous Deformable Object Manipulation with Human Demonstrations and Differentiable Physics". In: *ICLR*. OpenReview.net, 2023.

[6] Yunzhu Li et al. "Learning Particle Dynamics for Manipulating Rigid Bodies, Deformable Objects, and Fluids". In: *ICLR (Poster)*. OpenReview.net, 2019.

[7] Priyanka Mandikal and Kristen Grauman. "DexVIP: Learning Dexterous Grasping with Human Hand Pose Priors from Video". In: *Conference on Robot Learning, 8-11 November 2021, London, UK*. Ed. by Aleksandra Faust, David Hsu, and Gerhard Neumann. Vol. 164. Proceedings of Machine Learning Research. PMLR, 2021, pp. 651–661. URL: https://proceedings.mlr.press/v164/mandikal22a.html.

[8] Yuzhe Qin et al. "DexMV: Imitation Learning for Dexterous Manipulation from Human Videos". In: *ECCV (39)*. Vol. 13699. Lecture Notes in Computer Science. Springer, 2022, pp. 570–587.

[9] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. "VideoDex: Learning Dexterity from Internet Videos". In: *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*. Ed. by Karen Liu, Dana Kulic, and Jeffrey Ichnowski. Vol. 205. Proceedings of Machine Learning Research. PMLR, 2022, pp. 654–665. URL: https://proceedings.mlr.press/v205/shaw23a.html.

[10] Haochen Shi et al. "RoboCook: Long-Horizon Elasto-Plastic Object Manipulation with Diverse Tools". In: *CoRL*. Vol. 229. Proceedings of Machine Learning Research. PMLR, 2023, pp. 642–660.

[11] Haochen Shi et al. "RoboCraft: Learning to see, simulate, and shape elasto-plastic objects in 3D with graph networks". In: *Int. J. Robotics Res.* 43.4 (2024), pp. 533–549.

[12] Haochen Shi et al. "RoboCraft: Learning to See, Simulate, and Shape Elasto-Plastic Objects with Graph Networks". In: *Robotics: Science and Systems*. 2022.

[13] Liangzhi Shi et al. *Learning Adaptive Dexterous Grasping from Single Demonstrations*. 2025. arXiv: 2503.20208 [cs.RO]. URL: https://arxiv.org/abs/2503.20208.

[14] Chen Wang et al. "DexCap: Scalable and Portable Mocap Data Collection System for Dexterous Manipulation". In: *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*. Ed. by Dana Kulic et al. 2024. DOI: 10.15607/RSS.2024.XX.043. URL: https://doi.org/10.15607/RSS.2024.XX.043.

[15] Lixin Yang et al. "Multi-view Hand Reconstruction with a Point-Embedded Transformer". In: *CoRR* abs/2408.10581 (2024).

[16] Kaifeng Zhang et al. "AdaptiGraph: Material-Adaptive Graph-Based Neural Dynamics for Robotic Manipulation". In: *Robotics: Science and Systems*. 2024.