Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

Emerging Extrinsic Dexterity in Cluttered Scenes via Dynamics-aware Policy Learning

Yixin Zheng[*1,2,3], Jiangran Lyu[*3,4,†], Yifan Zhang[3], Jiayi Chen[3,4], Mi Yan[3,4], Yuntian Deng[5]
Xuesong Shi[3], Xiaoguang Zhao[1], Yizhou Wang[4], Zhizheng Zhang[2,3,✉], He Wang[2,3,4,✉]

[1]Institute of Automation, Chinese Academy of Sciences [2]Beijing Academy of Artificial Intelligence
[3]Galbot [4]Peking University [5]Shanghai Jiao Tong University

[*] Equal contribution [†] Project Lead ✉ Corresponding authors

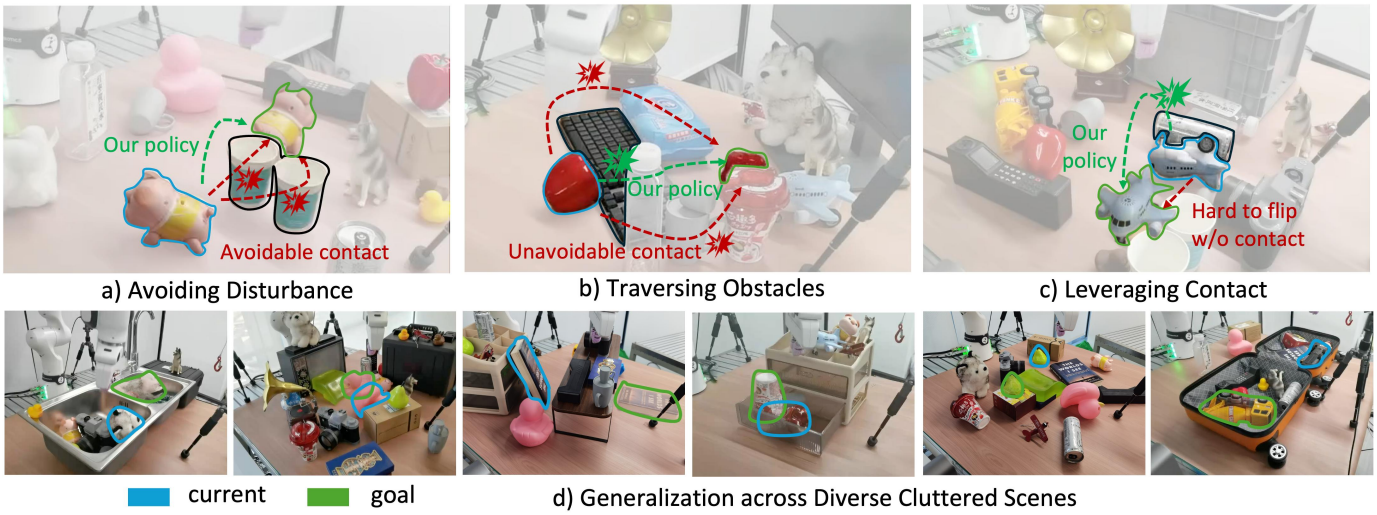

Fig. 1: Given the current object state (blue), the policy aims to rearrange objects to the goal state (green). To achieve this in cluttered scenes, the robot policy leverages extrinsic dexterity selectively. (a) The policy avoids unnecessary contacts to minimize disturbance when free-space motion suffices. (b) When contact is unavoidable, it robustly traverses obstacles under contact-rich dynamics. (c) It actively leverages environmental contacts to accomplish otherwise difficult manipulations (e.g., flipping objects). (d) The approach generalizes across diverse cluttered environments.

***Abstract*—Extrinsic dexterity leverages environmental contact to overcome the limitations of prehensile manipulation. However, achieving such dexterity in cluttered scenes remains challenging and underexplored, as it requires selectively exploiting contact among multiple interacting objects with inherently coupled dynamics. Existing approaches lack explicit modeling of such complex dynamics and therefore fall short in non-prehensile manipulation in cluttered environments, which in turn limits their practical applicability in real-world environments. In this paper, we introduce a Dynamics-Aware Policy Learning (DAPL) framework that can facilitate policy learning with a learned representation of contact-induced object dynamics in cluttered environments. This representation is learned through explicit world modeling and used to condition reinforcement learning, enabling extrinsic dexterity to emerge without hand-crafted heuristics or complex reward engineering. We evaluate our approach in both simulation and real-world. Our method outperforms prehensile manipulation, human teleoperation, and prior representation-based policies by over 25% in success rate on unseen simulated cluttered scenes with varying densities. The real-world success rate reaches around 50% across 10 cluttered scenes, while a practical grocery deployment further demonstrates robust sim-to-real transfer and applicability.**

Project Page: https://pku-epic.github.io/DAPL

I. Introduction

Robotic manipulation in cluttered scenes demands strategies beyond prehensile actions. Objects are often tightly packed and occluded in cluttered environments, where incidental and coupled contacts are pervasive, making reliable grasps difficult and collision-free execution highly constrained. In such cases, effective manipulation often relies on extrinsic dexterity—the ability to selectively leverage or avoid contacts with the surrounding environment through non-prehensile actions such as pushing, sliding, or toppling—to achieve task goals that grasping alone cannot.

Existing non-prehensile approaches either rely on highly specific designs or simplify away the complexity of contact-rich interactions. Model-based planning and hand-designed primitives [7, 30, 37] do not scale beyond narrowly defined settings, while reinforcement learning methods [23, 40] are typically confined to simplified contact scenarios. Even recent geometry-centric representation learning approaches, such as CORN[8] and UniCORN[9], remain brittle when faced with dense clutter and multiple interacting contacts. All these works lack explicit modeling of complex dynamics and therefore fall

short in manipulation in cluttered environments.

As discussed above, achieving extrinsic dexterity in cluttered scenes remains underexplored and challenging, as it requires selectively exploiting beneficial contacts while avoiding disruptive ones. For example, a policy may need to bypass light objects (Fig.1(a)), navigate through unavoidable obstacles (Fig.1(b)), or leverage stable neighbors to flip a target object (Fig.1(c))—tasks that cannot be solved by grasping alone. In such cluttered scenarios, success is not determined by geometry alone, but by how objects respond once contact occurs—whether they slide, topple, or transfer motion to surrounding clutter. These complex interactions depend on underlying object dynamics and cannot be reliably inferred from static observations or sparse rewards. This motivates us to decouple dynamics representation learning from task-specific control and to equip the policy with an explicit, learned representation that models contact-induced object motion.

In this paper, we introduce a **Dynamics-Aware Policy Learning (DAPL)** framework for more effective non-prehensile manipulation in cluttered scenes. It facilitates the emergence of extrinsic dexterity through contact-induced scene dynamics representation. Specifically, our approach learns this representation via a physical world model trained to predict object motion under interaction, capturing how contacts among multiple objects influence future dynamics. The learned dynamics representation is then used to condition reinforcement learning, enabling policies to reason about the consequences of contact and to selectively exploit environmental interactions. We further introduce curriculum learning into the DAPL framework, which uses policy rollout trajectories to iteratively refine the dynamics representation. As a result, extrinsic dexterity emerges naturally rather than being hand-designed or shaped by complex rewards.

We evaluate our approach extensively in both simulation and the real-world. To enable systematic training and benchmarking, we introduce **Clutter6D**, a simulation environment for 6D object rearrangement in cluttered scenes with varying densities. In simulation, DAPL significantly outperforms prehensile manipulation baselines, human teleoperation, and prior representation-based policies, achieving over 25% improvement in success rate on unseen cluttered scenes. After training in simulation, our policy can be zero-shot deployed in the real world. Across 10 diverse cluttered scenes, it achieves approximately 50% success—comparable to human teleoperation—while demonstrating higher efficiency (mean execution time 42.6s vs. 55.9s). These results highlight the robustness and generalization of our method. Finally, we integrate the learned policy with a high-level planner and a grasping module to perform practical grocery retrieval tasks. This demonstrates the applicability and potential real-world impact of our approach.

In summary, our contributions are threefold:

- We study the problem of **non-prehensile object rearrangement in cluttered scenes**, where effective manipulation requires extrinsic dexterity to selectively leverage and avoid environmental contacts.
- We introduce **Dynamics-Aware Policy Learning (DAPL)**, a framework that equips policies with a learned representation of contact-induced scene dynamics via a physical world model and curriculum learning, enabling extrinsic dexterity to emerge without hand-designed primitives or complex reward engineering.
- We extensively evaluate our approach in both simulation and the real-world, including a new benchmark **Clutter6D**, zero-shot sim-to-real transfer across diverse cluttered scenes, and deployment in a practical grocery retrieval task.

II. Related Work

A. Extrinsic Dexterity

Extrinsic dexterity[11] refers to manipulation skills that augment a robot's intrinsic hand capabilities by exploiting external resources such as environmental contacts, gravity, and dynamic arm motions.[10, 19, 18] Prior work[3, 5, 6, 41, 21] has demonstrated that complex manipulation behaviors can be achieved with simple grippers, including in-hand reorientation, prehensile pushing, and shared grasping. However, these successes largely rely on hand-crafted trajectories, task-specific motion primitives, or explicit planning over contact modes, which require precise object pose estimation and extensive manual design. As a result, such methods do not scale well beyond narrowly defined tasks or object sets. Learning-based methods [37, 23, 40] improve generalization, but they remain constrained to simplified scenes and are still brittle in cluttered, contact-rich environments. Dengler et al. [14] study learning-based pushing in clutter; however, their method focuses on 2D collision avoidance and does not address the challenges of full 6D object rearrangement with complex, contact-rich dynamics.

B. Representation Learning for Robotic Manipulation

Prior work accelerates reinforcement learning via representation pre-training with high-dimensional observations, using pretext tasks such as geometric completion, pose prediction, or contrastive learning[8, 1, 2, 24, 27]. Recent self-supervised[42, 16, 17, 39] approaches based on patch-wise transformers learn strong geometric features for point clouds, but primarily model static shape and lack sensitivity to contact and interaction dynamics. To better support manipulation, CORN[8] introduces contact-centric object representations, later extended by UniCORN[9] to scene-level point clouds. However, these geometry-driven representations remain brittle in dense clutter, where effective manipulation requires selectively avoiding lightweight objects while exploiting heavier ones as external supports for extrinsic dexterity. Recent dynamics-aware pre-training methods, such as [26], learn temporal representations from videos in pixel space, but do not model object-level physical properties or interaction forces. As a result, they are insufficient for contact-rich robotic manipulation that hinges on physically grounded scene dynamics.

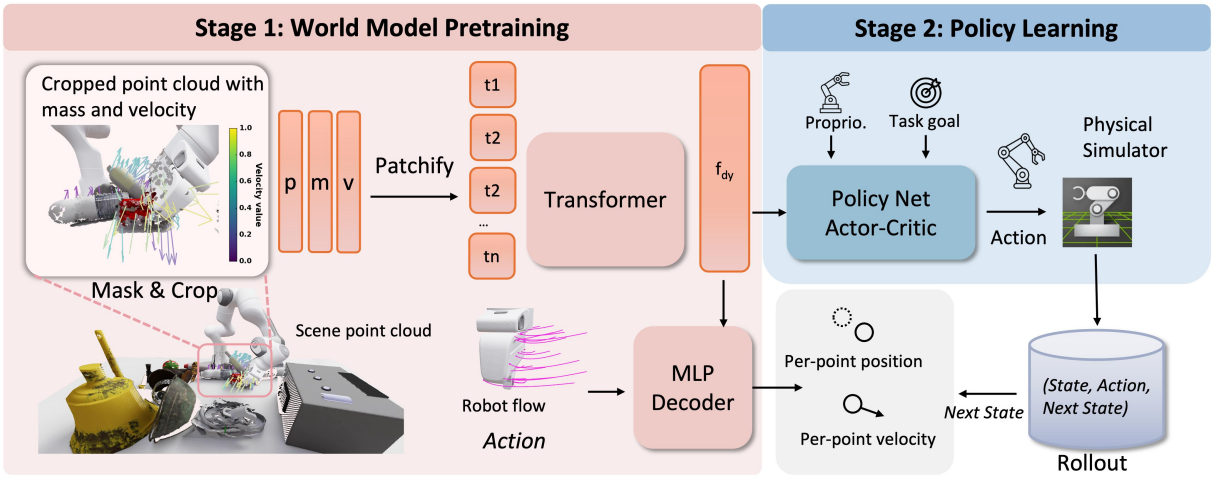

Fig. 2: Overview of the proposed two-stage learning framework. Stage 1 (World Model Pretraining): The model takes a cropped point cloud augmented with physical attributes (mass and velocity) as input. A Transformer-based architecture encodes these inputs into dynamics features (f_{dy}), which are used by an MLP decoder to predict future per-point positions and velocities conditioned on robot actions. Stage 2 (Policy Learning): The pre-trained dynamics representations (f_{dy}) are fed into an Actor-Critic policy network alongside proprioceptive data and task goals to facilitate efficient policy learning within a physical simulator.

III. Method

Our goal is to learn a non-prehensile manipulation policy that can selectively exploit or avoid environmental contacts to achieve object rearrangement in cluttered scenes. As illustrated in Fig. 2, our framework decouples the dynamics representation learning stage and RL policy learning stage. Specifically, our framework iteratively trains: (1) a physical world model using policy rollout interaction data to capture contact-induced object–scene dynamics; and (2) an RL policy that explores in a physical simulator while being conditioned on the learned dynamics representation.

A. Dynamics Representation Learning

To explicitly encode this physical prior, we learn a *physical world model* that predicts future object-centric dynamics conditioned on the current state and action. The world model serves as a physics-informed representation learner whose latent features are subsequently reused for downstream policy learning.

Physical Scene Representation. At each timestep t, the state consists of three components: (i) the target object point cloud $\mathcal{P}^{\text{obj}}$, (ii) the surrounding scene point cloud $\mathcal{P}^{\text{scene}}$, and (iii) the robot end-effector point cloud. The surrounding scene point cloud is spatially cropped within a local region centered at the target object, focusing the representation on contact-relevant geometry while reducing background redundancy. Each 3D point is augmented with physical attributes, resulting in a per-point feature:

$$\mathbf{x}_i = (\mathbf{p}_i,\ m_i,\ \mathbf{v}_i), \tag{1}$$

where $\mathbf{p}_i \in \mathbb{R}^3$ denotes the point position, m_i the point mass, and $\mathbf{v}_i$ the point velocity.

Architecture. As shown in Fig. 2, the world model consists of a dynamics encoder and decoder, built upon a patch-based transformer backbone. Following prior work on point cloud transformers [8, 9, 16], we first partition the point cloud into local patches. Patch centers are selected using Farthest Point Sampling (FPS), and each patch is formed by gathering the k-nearest neighbors (kNN) of the center point. Each patch is normalized by subtracting its center coordinate. A lightweight PointNet-style encoder embeds each normalized patch into a fixed-dimensional token. To restore global spatial information, we add sinusoidal positional embeddings based on the patch center coordinates. The resulting patch tokens are processed by a ViT, which models multi-object coupling effects. On top of the transformer features, a small MLP decoder predicts future point-wise dynamics, including positions and velocities.

World Model Training Objective. We train the world model with dense point-level supervision. Given predicted position $\hat{\mathbf{p}}_i^{t+1}$ and velocity $\hat{\mathbf{v}}_i^{t+1}$, we minimize

$$\mathcal{L}_{\text{dyn}} = \sum_i \left\|\hat{\mathbf{p}}_i^{t+1} - \mathbf{p}_i^{t+1}\right\|_2^2 + \lambda \left\|\hat{\mathbf{v}}_i^{t+1} - \mathbf{v}_i^{t+1}\right\|_2^2. \tag{2}$$

In cluttered scenes, most points exhibit near-zero velocities, leading to highly sparse supervision in the velocity field. As a result, optimizing only point-wise velocity losses can cause the model to collapse to trivial solutions that predict uniformly small velocities and fail to capture contact-induced motion. To address this, we treat the scene as a point distribution rather than independent points and introduce a variance-aware regularization over point-wise velocities. Specifically, we match the standard deviations of the predicted and ground-truth velocity fields:

$$\mathcal{L}_{\text{var}} = \left\|\text{Std}\big(\{\hat{\mathbf{v}}_i^{t+1}\}_i\big) - \text{Std}\big(\{\mathbf{v}_i^{t+1}\}_i\big)\right\|_2, \tag{3}$$

This regularization preserves the overall magnitude and spatial variability of motion in dynamic regions, preventing degenerate near-zero velocity collapse.

The overall world model training objective is defined as

$$\mathcal{L} = \mathcal{L}_{\text{dyn}} + \alpha \mathcal{L}_{\text{var}}, \tag{4}$$

where α controls the strength of variance-aware regularization.

B. Dexterous Policy Learning via RL

Observation and Action Spaces. At each timestep, the policy observes three components: a dynamics-aware scene representation, the robot proprioceptive state, and a task goal. The physical scene is processed by the dynamics encoder to produce a set of embeddings that capture contact-induced object dynamics. The robot proprioceptive observation consists of joint positions, joint velocities, and end-effector poses, while the task goal is represented as the relative pose between the target object and the desired goal configuration. The policy outputs continuous joint-space control commands, which are executed using an impedance controller.

Reward Design. Our reward design follows standard practice in non-prehensile manipulation (e.g., HAMNet [9]) and uses only lightweight shaping rather than heavy task-specific reward engineering. The reward consists of a sparse task success term and two lightweight shaping terms that encourage physical contact and goal-directed object motion. The contact term r_{contact} is defined based on the minimum distance d_{oe} between the target object and the end-effector, which encourages interaction:

$$r_{\text{contact}} = 1 - \tanh(d_{\text{oe}}). \tag{5}$$

The goal-reaching term r_{goal} measures the distance d_{og} between the object and the desired goal pose, and is activated once the end-effector distance d_{oe} falls below a threshold τ_d:

$$r_{\text{goal}} = \mathbb{I}(d_{\text{oe}} < \tau_d)\,(1 - \tanh(d_{\text{og}}))\,. \tag{6}$$

A sparse success reward is issued when the target object reaches the desired pose. To penalize disturbance to the scene, the reward is discounted by the displacement of non-target objects D_{disp} measured by Chamfer distance:

$$r_{\text{success}} = \mathbb{I}_{\text{success}}\,(1 - \beta D_{\text{disp}}) \tag{7}$$

All coefficients and thresholds use standard values and are provided in the Appendix.

C. Curriculum Learning with Policy Interaction

Rather than relying on a fixed offline dataset, we adopt an iterative curriculum that alternates between policy learning and world model refinement. We initialize the process by training an RL policy from scratch without a pretrained dynamics representation. In each curriculum iteration, we freeze the dynamics encoder and train the policy until both reward and success rate plateau. This typically takes about 2×10^4 RL steps per stage, with fewer steps required in later iterations. Once the policy reaches basic task coverage, we roll out the policy to collect approximately 60k interaction steps. These trajectories are intentionally imperfect and include substantial random collisions and suboptimal behaviors, which is beneficial for exposing diverse contact dynamics. The collected interaction data is then used to update the world model for 500k training steps until the validation loss stabilizes, improving its ability to capture contact-induced momentum transfer under realistic policy-induced distributions. The refined dynamics encoder is subsequently reused by the RL policy, yielding improved exploration efficiency and more stable policy learning. We repeat this alternating procedure until both the policy performance and the learned dynamics representation converge. This curriculum allows the world model and policy to co-evolve, progressively shifting from noisy exploratory interactions to task-relevant, physically consistent manipulation behaviors.

D. Clutter6D: Object Rearrangement Environment

To train and evaluate manipulation policy in clutter scenes, we introduce **Clutter6D**, a comprehensive 6D object rearrangement environment and benchmark. Clutter6D is designed to explicitly stress manipulation scenarios where extrinsic dexterity is necessary rather than optional. Increasing clutter density not only reduces free-space motion but also amplifies the coupling between object dynamics, making contact outcomes highly sensitive to physical properties. In such regimes, collision-free planning and grasp-centric strategies become unreliable, and successful manipulation requires selectively leveraging or avoiding environmental contacts. Compared to prior clutter benchmarks that emphasize collision avoidance or planar pushing, Clutter6D focuses on full 6D object rearrangement with multi-object contact and dynamic coupling. The use of task-oriented scene graphs enables systematic control over relational constraints while preserving scene diversity, making the benchmark suitable for evaluating contact-rich manipulation policies beyond grasping.

Environment Setup. Clutter6D is built on the IsaacLab simulator with PhysX as the physics backend. We curate high-quality object assets from Objaverse [12] and process them using mesh simplification and convex decomposition (PaMO[29] and CoACD[35]). Physical properties such as scale and mass, as well as semantic attributes including category, color, and material, are automatically generated[33] and verified through a human-in-the-loop process. From these, we select and normalize 10K assets to tabletop scale. Task scenarios are synthesized from object semantics and represented as task-oriented scene graphs (ToSG)[15], which specify spatial relations among objects. Scenes are instantiated by placing objects in topological order according to the scene graph, discarding placements that violate relational constraints.

Benchmark Setting. Clutter6D defines three difficulty tracks based on clutter density: *Sparse* (4 objects), *Moderate* (8 objects), and *Dense* (12 objects). The sparse track contains 1,024 scenes for training, while each track includes 128 held-out scenes for evaluation. Performance is evaluated using two metrics: (i) task success rate, where success is achieved if the target object reaches the desired pose within 0.05 m and 0.1 rad, and (ii) chamfer distance of non-target object offsets,

measuring unintended disturbance to the surrounding clutter. Episodes terminate upon task completion, object drop, or after 300 simulation steps.

IV. Experiments

In this section, we evaluate our method in both the proposed Clutter6D simulation benchmark and real-world scenarios.

A. Simulation Experiments

Setup and Baselines. We evaluate all methods on the Clutter6D benchmark across three difficulty levels: *Sparse* (4 objects), *Moderate* (8 objects), and *Dense* (12 objects). We compare our approach against three categories of baselines:

- **Prehensile Manipulation:** GraspGen + CuRobo [28, 34], a pipeline combining grasp synthesis with motion planning.
- **Human Teleoperation:** A baseline established by expert human operators using a Gello [38] interface.
- **Representation Learning+RL Policy:** We compare our method against two types of learning-based baselines: (i) general point cloud encoders, including Point2Vec [22] and Concerto [42], and (ii) encoders specifically tailored for non-prehensile manipulation: CORN [8] and UniCORN[9]. We also include a CORN-multi baseline which uses a CORN encoder to extract features from each individual object and fuses them into a global representation.

All learning-based methods share comparable policy networks to ensure a fair comparison. While the baselines use only standard geometric coordinates (x, y, z) as input, our approach leverages the full 7-dimensional physical features (including velocity and mass), highlighting the advantage of explicit physical reasoning and a dynamics-aware learning paradigm rather than architectural differences.

Main Results. Table I summarizes quantitative results across varying clutter densities, while Fig.9 shows representative qualitative rollouts. Our method consistently outperforms all baselines, with the performance gap becoming particularly pronounced in the *Dense* setting. While geometry-based methods such as CORN achieve reasonable success in sparse scenes (46.63%), their performance degrades sharply under heavy clutter (22.22%). In contrast, our approach maintains a robust success rate of **44.56%** in dense environments, effectively doubling the strongest baseline.

This advantage is further reflected in the Mean Offset (M.O., measured in cm) metric, which measures unintended disturbance to non-target objects. Our method achieves a favorable balance between task success and environment preservation. Specifically, while the CORN baseline manages a certain level of success, it incurs significantly higher disturbance (17.43 vs. our 12.65) due to its lack of dynamics reasoning. Conversely, lower offsets reported by other methods are largely artifacts of execution failure, whereas our policy maintains the highest success rate through precise, intentional interactions. Qualitatively, as shown in Fig.5, CORN relies on static geometric reasoning and often initiates excessive or poorly directed contacts, triggering unintended contact chains that cause the robot to become stuck in clutter. In contrast, our policy exhibits more selective and intentional interactions, exploiting environmental contacts when beneficial while actively avoiding unnecessary collisions. Together, these results support our core hypothesis that static geometric representations are insufficient in dense clutter, where momentum transfer and multi-object contact dynamics critically determine task success.

Training Efficiency and Convergence. Beyond final performance, Fig. 4 illustrates the learning dynamics of different methods. Our approach demonstrates significantly higher sample efficiency, reaching approximately 70% success within the first few thousand training iterations. In contrast, baselines relying on static geometric representations (e.g., CORN and UniCORN) converge substantially slower, while other representations fail to learn meaningful policies. These results indicate that the dynamics-aware representation provides a strong physical prior, allowing the policy to learn effective interaction strategies early without expending excessive samples on discovering basic contact physics.

Analysis of Dynamics Representation. We conduct ablation studies to assess the contribution of the proposed dynamics representation (Table II). Removing physical attributes such as velocity and mass leads to a substantial performance drop, highlighting the importance of reasoning about the *potential for motion* in cluttered manipulation. Additionally, replacing point-level dense prediction with object-level 6 DoF pose prediction leads to inferior performance, suggesting that pose supervision is too coarse to learn the intricate contact physics and local deformations essential for precise tracking. Moreover, replacing the dynamics prediction objective with simple autoencoder reconstruction significantly degrades performance, indicating that the effectiveness comes from dynamics learning other than static shape reconstruction. We further visualize world model predictions and observe close agreement with ground-truth object motion in Fig.6, confirming the quality of learned dynamics.

Effectiveness of Curriculum Learning. Fig. 7 shows the evolution of policy performance across curriculum iterations. Starting from 61.3%, the success rate steadily improves to **71.8%** after three iterations. Early policies generate diverse and imperfect interaction trajectories, which are used to refine the world model. The improved dynamics representation then provides more accurate conditioning for subsequent policy learning, leading to progressively more effective extrinsic dexterity.

Behavior Analysis for Extrinsic Dexterity. To examine whether the learned policy exhibits extrinsic dexterity beyond aggregate success metrics, we manually analyze 50 sampled simulation trajectories for each method. We measure unnecessary contacts, disturbance to non-target objects, and the occurrence of Avoid, Traverse, and Leverage behaviors. Table III summarizes the results. Compared with CORN, DAPL produces fewer unnecessary contacts and substantially fewer large translations and rotations of non-target objects. It

(a) Clutter6D Benchmark (b) Real-world Setup

Fig. 3: Overview of our proposed Clutter6D Benchmark and real-world setup. (a) Representative examples of the Clutter6D Benchmark with three levels of scene density: Sparse (4 objects), Moderate (8 objects), and Dense (12 objects). (b) The real-world experimental setup consisting of a Franka Research 3 robot and three Intel RealSense cameras for point cloud acquisition, along with the complete set of objects used in our real-world experiments.

Methods	**Action Type**	**Sparse**		**Moderate**		**Dense**	
		S.R. ↑	M.O. ↓	S.R. ↑	M.O. ↓	S.R. ↑	M.O. ↓
Teleoperation [38]	Mixed	50.0	3.13	40.0	7.49	20.0	21.34
GraspGen + CuRobo [28, 34]	Prehensile	26.6	–	15.6	–	3.13	–
Point2Vec [22]	Non-prehensile	6.89	5.09	1.95	3.36	0.78	5.35
Concerto [42]	Non-prehensile	3.13	1.65	1.56	2.90	0.39	7.56
CORN [8]	Non-prehensile	46.63	3.15	45.83	5.51	22.22	17.43
CORN-multi	Non-prehensile	35.93	2.73	15.38	3.92	11.83	12.06
UniCORN[9]	Non-prehensile	20.61	1.71	11.67	4.13	5.81	9.79
Ours	Non-prehensile	**71.88**	**2.59**	**51.04**	**2.7**	**44.56**	**12.65**

TABLE I: Quantitative results measured by success rate in the simulation benchmark. Note that Mean Offset (M.O.) for GraspGen + CuRobo is omitted because it relies on the CuRobo motion planner, which strictly prioritizes collision-free trajectories. Our method significantly outperforms all baselines across all clutter densities, maintaining high success rates even in dense environments where baselines experience a sharp performance decline.

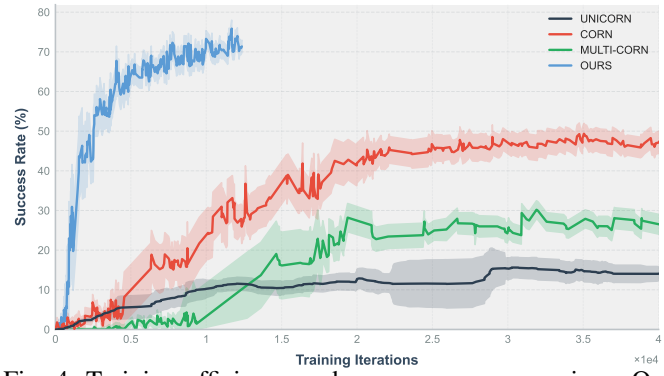

Fig. 4: Training efficiency and convergence comparison. Our method (blue) achieves significantly higher sample efficiency compared to geometry-based baselines, reaching about 70% success rate within the initial 10^4 iterations. This demonstrates that our dynamics-aware representation provides a robust physical prior, facilitating the rapid acquisition of complex interaction strategies without the need for exhaustive sampling of contact physics.

also exhibits Avoid, Traverse, and Leverage behaviors more frequently, suggesting that the learned policy develops purposeful contact strategies rather than relying on indiscriminate pushing.

Sensitivity to Noisy Inputs. We further evaluate policy sensitivity to noisy physical and goal inputs through controlled

Fig. 5: Compared to static geometric representation baseline CORN, which either triggers excessive contact chains leading to significant environment disturbance (middle left) or becomes ”stuck” due to a lack of dynamics reasoning (middle right), our policy (bottom row) exhibits selective and intentional interactions, successfully reorients the target while preserving scene stability.

Pretrain Task	Granularity	Velocity	Phys.	P.E. ↓	S.R. ↑	M.O. ↓
Recons.	Point-level	✗	✗	-	11.75	1.31
Recons.	Point-level	✓	✓	-	29.63	2.63
World Model	Object-level	✗	✗	3.1	14.13	3.27
World Model	Object-level	✓	✓	3.2	16.88	3.84
World Model	Point-level	✗	✗	4.1	42.00	4.91
World Model	Point-level	✓	✗	5.1	58.25	4.86
World Model	Point-level	✓	✓	**4.6**	**71.88**	**2.59**

TABLE II: Ablation Study on Pre-training Objectives and Input Modalities (Sparse Track). P.E.: Chamfer Distance prediction error (cm). Results highlight the indispensability of velocity and physical features, while further establishing that point-level dynamics modeling is a more effective pretext task than either object-level relative pose prediction or simple reconstruction.

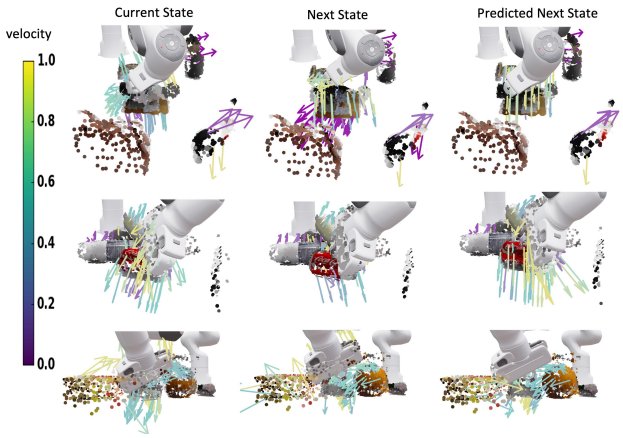

Fig. 6: Visualization of our world model predictions.

perturbations in simulation. As shown in Table IV, mass and velocity perturbations maintain similar success rates over a broad noise range, although larger noise gradually increases scene disturbance. In contrast, goal-pose noise causes a much sharper performance drop because it directly changes the task specification.

Analysis of Adaptive Behavior. Beyond quantitative performance gains, our policy exhibits adaptive interaction strategies that reflect a deep understanding of contact-induced dynamics. We demonstrate this through a controlled experiment where we maintain identical initial and goal configurations while modifying the objects' physical properties. As shown in Fig.8, when the pie is heavy and the Pringles can is light, the policy proactively exploits the pie as a stable anchor. It establishes deliberate contact with the pie to generate the necessary torque for object reorientation. Conversely, if the mass assignments are swapped, the policy autonomously adapts its trajectory to avoid the lightweight pie. It prevents unintended scene

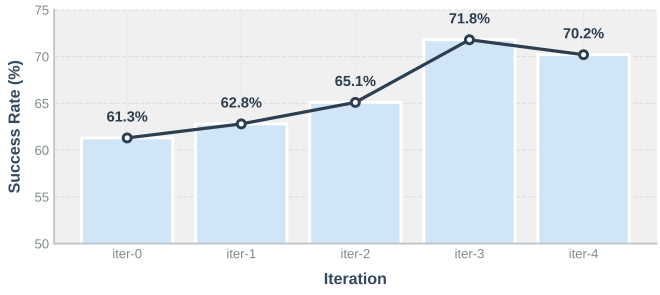

Fig. 7: Success rate over curriculum iterations. Performance steadily improves from 61.3% to 71.8%, demonstrating that iterative model refinement provides more accurate dynamics conditioning for learning effective manipulation strategies.

Method	Unnec. Contacts	Object Disturbance				Dexterous Behaviors (%)			
		Small Trans.	Small Rot.	Large Trans.	Large Rot.	Avoid	Traverse	Leverage	Simple
CORN	3.6	2.8	2.6	4.4	4.8	22	14	6	62
Ours	**2.4**	3.2	3.0	**1.8**	**1.2**	**52**	**38**	**42**	28

TABLE III: Behavior analysis over 50 sampled trajectories per method. Unnec. Contacts and Object Disturbance metrics are reported as average counts per trajectory. Small/Large Trans. count non-target objects with translations below/above 5 cm, while Small/Large Rot. count non-target objects with rotations below/above $45°$. Avoid, Traverse, and Leverage correspond to the dexterous behavior types illustrated in Fig. 1. Multiple behavior types may occur in one trajectory; Simple denotes trajectories in which none of the three is observed.

Noise Source	0.5%		5%		25%		50%	
	S.R.	M.O.	S.R.	M.O.	S.R.	M.O.	S.R.	M.O.
Mass	66.9	3.1	67.3	3.4	66.0	4.1	62.4	5.9
Velocity	64.0	3.6	64.0	3.2	61.1	3.9	58.5	4.4
Goal Pose	66.3	3.1	65.0	3.3	48.1	3.5	15.1	4.0
All Inputs	65.7	3.7	66.6	4.9	45.1	6.0	5.9	8.4

TABLE IV: Sensitivity to noisy physical and goal inputs under controlled simulation perturbations. Noise levels are reported as percentages of the corresponding input mean. S.R. denotes success rate and M.O. denotes mean offset (cm).

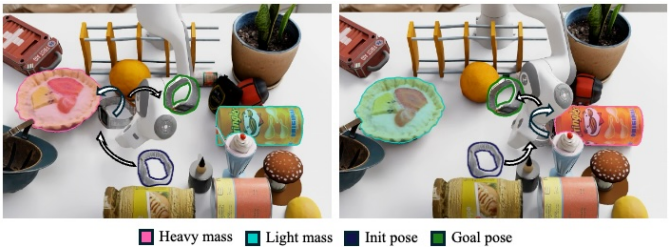

Fig. 8: Adaptive behavior guided by physical priors. In identical scenes, our policy autonomously adapts its interaction trajectory based on whether the pie or the Pringles can (pink/cyan mask) is assigned a heavy mass, transitioning between using objects as stable anchors versus avoiding them to prevent destabilization

Methods	S1		S2		S3		S4	
	S.R. ↑	M.T. ↓	S.R. ↑	M.T. ↓	S.R. ↑	M.T. ↓	S.R. ↑	M.T. ↓
Teleop. [25]	3/5	77	4/5	59	3/5	39	1/5	81
Ours	**2/5**	**51**	**1/5**	**26**	**3/5**	**33**	**1/5**	73
Methods	S5		S6		S7		S8	
	S.R. ↑	M.T. ↓	S.R. ↑	M.T. ↓	S.R. ↑	M.T. ↓	S.R. ↑	M.T. ↓
Teleop. [25]	3/5	28	2/5	39	4/5	20	0/5	90
Ours	4/5	37	3/5	59	2/5	24	2/5	28
Methods	S9		S10		Avg.			
	S.R. ↑	M.T. ↓	S.R. ↑	M.T. ↓	S.R. ↑	M.T. ↓		
Teleop. [25]	5/5	62	1/5	64	52%	55.9		
Ours	5/5	45	1/5	50	48%	42.6		

TABLE V: Quantitative results across 10 scenes. Mean (last column) denotes average performance across all scenes. SR ↑ indicates success rate and MT ↓ indicates mean execution time.

destabilization by utilizing the now-heavy Pringles can as the functional support instead. These results suggest that the policy reasons about how interactions unfold purely through a dynamics-aware representation rather than relying on geometric heuristics.

B. Real-world Experiments

System Setups. We deploy our policy on a Franka Research 3 robot equipped with three Intel RealSense cameras. Object segmentation is initialized using SAM2[32] on the first frame and tracked online with XMem[4]. Object poses are estimated using FoundationPose[36]. For sim-to-real transfer, we perform system identification of key physical parameters. Object mass is estimated using a vision–language model[33] that provides coarse physical priors, while velocities are obtained by temporal differentiation followed by EKF-based filtering [20]. To further bridge the sim-to-real gap, we employ a student-teacher distillation framework under perturbation. By training the student to recover the teacher's latent dynamics from observations injected with Gaussian noise, we force the policy to infer effective dynamics despite imperfect perception. Although these estimates are not precise, they provide sufficient priors for the policy to select appropriate interaction strategies. We compare our method against FACTR [25], a teleoperation baseline that enhances operator precision in contact-rich tasks through multimodal sensory integration. We evaluate performance on 10 diverse real-world cluttered scenes. A trial is considered successful if the target object reaches the desired pose with errors below 0.05 m and 0.1 rad within 90 seconds. We report success rate and mean execution time.

Results. As shown in Table V, our method achieves an average success rate of 48% across 10 scenes, approaching human teleoperation performance (52%). Notably, our approach achieves comparable success with more consistent execution times, demonstrating reliable real-world performance despite sensing noise and modeling inaccuracies. Although the dynamics representation relies on estimated physical attributes such as mass and velocity, precise estimation of these quantities is not required for effective deployment. In practice, coarse mass priors and noisy velocity estimates provide sufficient cues for the policy to differentiate stable supports from objects likely to be displaced. This observation is consistent with the sensitivity analysis in Table IV, where mass and velocity perturbations have limited impact on success rate compared with goal-pose noise.

C. Applications

To demonstrate practical utility in unstructured environments, we deploy our policy on a humanoid robot (Galbot G1) for a grocery retrieval task. The manipulation policy is trained entirely in a digital twin of the target shelf environment using Galbot, without access to real-world interaction data. We integrate a VLM-based planner (SoFar[31]) to translate natural language commands (e.g., "*Retrieve the cracker box*") into 6D target poses. As shown in Fig. 10, direct grasping is often infeasible in shelf scenarios due to limited reachability or objects exceeding the gripper's width. In this pipeline, our learned policy acts as a critical *pre-grasping* skill: it leverages extrinsic dexterity to slide and reorient the target object out of the clutter into a grasp-friendly configuration, enabling the downstream grasping module[13] to successfully complete the task.

V. Conclusion, Limitations and Future Direction

We studied non-prehensile object rearrangement in cluttered scenes, where effective manipulation requires extrinsic dexterity to selectively leverage and avoid environmental contacts. We proposed **Dynamics-Aware Policy Learning (DAPL)**, which equips reinforcement learning policies with a learned representation of contact-induced scene dynamics via a physical world model and curriculum learning. Furthermore, we introduced Clutter6D, a new benchmark tailored for evaluating 6-DoF non-prehensile manipulation across varying clutter densities, where DAPL achieves state-of-the-art performance. By reasoning about interaction outcomes beyond static geometry, our method enables robust extrinsic dexterity; importantly, in real-world experiments, DAPL achieves success rates comparable to human teleoperation while noticeably reducing average execution time.

Our approach has several limitations. The dynamics representation relies on approximate physical attributes (e.g., object mass and velocity), which may be noisy or unavailable in real-world settings. In addition, we focus on tabletop rearrangement with rigid objects and do not explicitly address articulated or deformable objects, nor long-horizon multi-stage tasks. Future work includes improving online dynamics estimation through richer sensing and system identification, extending dynamics-aware representations to articulated and deformable objects, and integrating DAPL with higher-level planning for long-horizon manipulation in complex environments. We believe dynamics-aware representations offer a promising direction

Fig. 9: Qualitative Results in the real world. The goal pose is shown transparently.

Fig. 10: Humanoid grocery retrieval using extrinsic dexterity. Our policy enables the Galbot G1 to manipulate objects in cluttered shelves by sliding and reorienting them into graspable configurations

for advancing dexterous manipulation beyond grasp-centric paradigms.

Acknowledgment

This work was partially supported by New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0122905).

References

[1] Chen Bao, Helin Xu, Yuzhe Qin, and Xiaolong Wang. Dexart: Benchmarking generalizable dexterous manipulation with articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21190–21200, 2023.

[2] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. in 2019 ieee. In *RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393, 2019.

[3] Nikhil Chavan-Dafle, Rachel Holladay, and Alberto Rodriguez. In-hand manipulation via motion cones. *arXiv preprint arXiv:1810.00219*, 2018.

[4] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European conference on computer vision*, pages 640–658. Springer, 2022.

[5] Xianyi Cheng, Eric Huang, Yifan Hou, and Matthew T Mason. Contact mode guided sampling-based planning for quasistatic dexterous manipulation in 2d. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6520–6526. IEEE, 2021.

[6] Xianyi Cheng, Eric Huang, Yifan Hou, and Matthew T Mason. Contact mode guided motion planning for quasidynamic dexterous manipulation in 3d. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2730–2736. IEEE, 2022.

[7] Xianyi Cheng, Sarvesh Patil, Zeynep Temel, Oliver Kroemer, and Matthew T Mason. Enhancing dexterity in robotic manipulation via hierarchical contact exploration. *IEEE Robotics and Automation Letters*, 9(1):390–397, 2023.

[8] Yoonyoung Cho, Junhyek Han, Yoontae Cho, and Beomjoon Kim. Corn: Contact-based object representation for nonprehensile manipulation of general unseen

objects. In *12th International Conference on Learning Representations, ICLR 2024*. International Conference on Learning Representations, ICLR, 2024.

[9] Yoonyoung Cho, Junhyek Han, Jisu Han, and Beomjoon Kim. Hierarchical and modular network on non-prehensile manipulation in general environments. *arXiv preprint arXiv:2502.20843*, 2025.

[10] Nikhil Chavan Dafle and Alberto Rodriguez. Sampling-based planning of in-hand manipulation with external pushes. In *ISRR*, pages 523–539, 2017.

[11] Nikhil Chavan Dafle, Alberto Rodriguez, Robert Paolini, Bowei Tang, Siddhartha S Srinivasa, Michael Erdmann, Matthew T Mason, Ivan Lundberg, Harald Staab, and Thomas Fuhlbrigge. Extrinsic dexterity: In-hand manipulation with external forces. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1578–1585. IEEE, 2014.

[12] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023.

[13] Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Wenhao Zhang, Heming Cui, Zhizheng Zhang, and He Wang. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. 2025. URL https://arxiv.org/abs/2505.03233.

[14] Nils Dengler, Juan Del Aguila Ferrandis, João Moura, Sethu Vijayakumar, and Maren Bennewitz. Learning goal-directed object pushing in cluttered scenes with location-based attention. *arXiv preprint arXiv:2403.17667*, 2024.

[15] Ning Gao, Yilun Chen, Shuai Yang, Xinyi Chen, Yang Tian, Hao Li, Haifeng Huang, Hanqing Wang, Tai Wang, and Jiangmiao Pang. Genmanip: Llm-driven simulation for generalizable instruction-following manipulation. In *CVPR*, 2025.

[16] Balázs Gyenes, Nikolai Franke, Philipp Becker, and Gerhard Neumann. Pointpatchrl–masked reconstruction improves reinforcement learning on point clouds. *arXiv preprint arXiv:2410.18800*, 2024.

[17] Chengkai Hou, Yanjie Ze, Yankai Fu, Zeyu Gao, Songbo Hu, Yue Yu, Shanghang Zhang, and Huazhe Xu. 4d visual pre-training for robot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8451–8461, 2025.

[18] Yifan Hou, Zhenzhong Jia, and Matthew T Mason. Fast planning for 3d any-pose-reorienting using pivoting. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1631–1638. IEEE, 2018.

[19] Yifan Hou, Zhenzhong Jia, and Matthew T Mason. Manipulation with shared grasping. *arXiv preprint arXiv:2006.02996*, 2020.

[20] Rudolf E Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82 (1):35–45, 1960. doi: 10.1115/1.3662552.

[21] Minchan Kim, Junhyek Han, Jaehyung Kim, and Beomjoon Kim. Pre-and post-contact policy decomposition for non-prehensile manipulation with zero-shot sim-to-real transfer. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10644–10651. IEEE, 2023.

[22] Karim Knaebel, Jonas Schult, Alexander Hermans, and Bastian Leibe. Point2vec for self-supervised representation learning on point clouds. 2023.

[23] Keita Kobashi and Masayoshi Tomizuka. Leveraging extrinsic dexterity for occluded grasping on grasp constraining walls. *arXiv preprint arXiv:2507.14721*, 2025.

[24] Hongzhuo Liang, Xiaojian Ma, Shuang Li, Michael Görner, Song Tang, Bin Fang, Fuchun Sun, and Jianwei Zhang. Pointnetgpd: Detecting grasp configurations from point sets. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3629–3635. IEEE, 2019.

[25] Jason Jingzhou Liu, Yulong Li, Kenneth Shaw, Tony Tao, Ruslan Salakhutdinov, and Deepak Pathak. Factr: Force-attending curriculum training for contact-rich policy learning. *arXiv preprint arXiv:2502.17432*, 2025.

[26] Hao Luo, Bohan Zhou, and Zongqing Lu. Pre-trained visual dynamics representations for efficient policy learning. In *European Conference on Computer Vision*, pages 249–267. Springer, 2024.

[27] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.

[28] Adithyavairavan Murali, Balakumar Sundaralingam, Yu-Wei Chao, Wentao Yuan, Jun Yamada, Mark Carlson, Fabio Ramos, Stan Birchfield, Dieter Fox, and Clemens Eppner. Graspgen: A diffusion-based framework for 6-dof grasping with on-generator training. *arXiv preprint arXiv:2507.13097*, 2025.

[29] Seonghun Oh, Xiaodi Yuan, Xinyue Wei, Ruoxi Shi, Fanbo Xiang, Minghua Liu, and Hao Su. Pamo: Parallel mesh optimization for intersection-free low-poly modeling on the gpu. In *Computer Graphics Forum*, page e70267. Wiley Online Library, 2025.

[30] Miquel Oller, Dmitry Berenson, and Nima Fazeli. Tactile-driven non-prehensile object manipulation via extrinsic contact mode control. *arXiv preprint arXiv:2405.18214*, 3, 2024.

[31] Zekun Qi, Wenyao Zhang, Yufei Ding, Runpei Dong, Xinqiang Yu, Jingwen Li, Lingyun Xu, Baoyu Li, Xialin He, Guofan Fan, et al. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. *arXiv preprint arXiv:2502.13143*, 2025.

[32] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint*

arXiv:2408.00714, 2024.

[33] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[34] Balakumar Sundaralingam, Siva Kumar Sastry Hari, Adam Fishman, Caelan Garrett, Karl Van Wyk, Valts Blukis, Alexander Millane, Helen Oleynikova, Ankur Handa, Fabio Ramos, et al. Curobo: Parallelized collision-free robot motion generation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8112–8119. IEEE, 2023.

[35] Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. Approximate convex decomposition for 3d meshes with collision-aware concavity and tree search. *ACM Transactions on Graphics (TOG)*, 41(4):1–18, 2022.

[36] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024.

[37] Albert Wu, Ruocheng Wang, Sirui Chen, Clemens Eppner, and C Karen Liu. One-shot transfer of long-horizon extrinsic manipulation through contact retargeting. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13891–13898. IEEE, 2024.

[38] Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12156–12163. IEEE, 2024.

[39] Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22193–22204, 2025.

[40] Shih-Min Yang, Martin Magnusson, Johannes A Stork, and Todor Stoyanov. Learning extrinsic dexterity with parameterized manipulation primitives. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5404–5410. IEEE, 2024.

[41] Weihao Yuan, Kaiyu Hang, Danica Kragic, Michael Y Wang, and Johannes A Stork. End-to-end nonprehensile rearrangement with deep reinforcement learning and simulation-to-reality transfer. *Robotics and Autonomous Systems*, 119:119–134, 2019.

[42] Yujia Zhang, Xiaoyang Wu, Yixing Lao, Chengyao Wang, Zhuotao Tian, Naiyan Wang, and Hengshuang Zhao. Concerto: Joint 2d-3d self-supervised learning emerges spatial representations. In *NeurIPS*, 2025.