ManiLong-Shot: Interaction-Aware One-Shot Imitation Learning for Long-Horizon Manipulation

Zixuan Chen[1*], **Chongkai Gao**[2], **Lin Shao**[2†], **Jieqi Shi**[1,3], **Jing Huo**[1†], **Yang Gao**[4,1,3]

[1]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2]School of Computing, National University of Singapore, Singapore
[3]School of Intelligence Science and Technology, Nanjing University, Suzhou, China
[4]School of Network Security and Information Technology, YiLi Normal University, Xinjiang, China
chenzx@nju.edu.cn, linshao@nus.edu.sg, huojing@nju.edu.cn

Abstract

One-shot imitation learning (OSIL) offers a promising way to teach robots new skills without large-scale data collection. However, current OSIL methods are primarily limited to short-horizon tasks, thus limiting their applicability to complex, long-horizon manipulations. To address this limitation, we propose ManiLong-Shot, a novel framework that enables effective OSIL for long-horizon prehensile manipulation tasks. ManiLong-Shot structures long-horizon tasks around physical interaction events, reframing the problem as sequencing interaction-aware primitives instead of directly imitating continuous trajectories. This primitive decomposition can be driven by high-level reasoning from a vision-language model (VLM) or by rule-based heuristics derived from robot state changes. For each primitive, ManiLong-Shot predicts invariant regions critical to the interaction, establishes correspondences between the demonstration and the current observation, and computes the target end-effector pose, enabling effective task execution. Extensive simulation experiments show that ManiLong-Shot, trained on only 10 short-horizon tasks, generalizes to 20 unseen long-horizon tasks across three difficulty levels via one-shot imitation, achieving a **22.8%** relative improvement over the SOTA. Additionally, real-robot experiments validate ManiLong-Shot's ability to robustly execute three long-horizon manipulation tasks via OSIL, confirming its practical applicability.

Website — https://sites.google.com/view/manilong-shot

1 Introduction

For robots to effectively integrate into daily life, they must rapidly learn and execute diverse long-horizon manipulation tasks. Daily chores, such as "setting the table" or "tidying the kitchen", exemplify these challenges, as they typically involve sequential interactions with multiple objects and comprise a series of sub-tasks. One-shot imitation learning (OSIL) (Duan et al. 2017; Finn et al. 2017; Valassakis et al. 2022) is a key technology for achieving this goal, allowing robots to acquire new skills from a single demonstration while avoiding costly retraining. However, despite its promise, existing OSIL methods often struggle to scale to long-horizon tasks. Many approaches are limited to short-horizon skills (Zhang and Boularias 2024; Vosylius and Johns 2025), require new tasks to be slight variations of training tasks (Xu et al. 2022), or rely on known 3D object models (Biza et al. 2023). These limitations significantly impede their application in complex, multi-stage, real-world manipulation scenarios. We draw inspiration from human learning: when faced with an unseen task such as *placing tableware*, a person naturally decomposes it into short-horizon primitives and infers the key interaction regions for each. For example: 1) pick up a plate (rim); 2) place the plate (target location on the table); 3) pick up a fork (handle). Each primitive is bounded by physical interactions, such as contact or release, and imitation is achieved by replicating actions on these regions. This raises the question: *Can a robot infer sub-task boundaries and critical interaction regions from an unannotated demonstration of an unseen long-horizon task, enabling efficient OSIL of its primitives?*

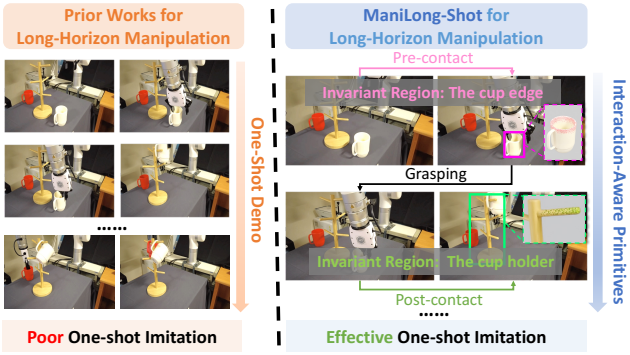

Figure 1: We introduce ManiLong-Shot, a novel framework for effective OSIL in long-horizon prehensile manipulation.

To this end, we propose a novel interaction-aware OSIL framework, **ManiLong-Shot**, inspired by human strate-

*This work was partially conducted during Zixuan Chen's visit to National University of Singapore.
†Corresponding author

gies and cognitive patterns when performing OSIL for long-horizon manipulation. ManiLong-Shot structures long-horizon manipulation tasks into a sequence of discrete primitives, each corresponding to a distinct phase of physical interaction between the robot's end-effector and the environment, namely: *pre-contact*, *grasping*, and *post-contact*, as illustrated in Fig. 1. The phase decomposition design builds upon prior work on sub-task decomposition for manipulation (James and Davison 2022; Chen et al. 2025b,e), and is founded on a key assumption: sub-task boundaries can be reliably identified through these interaction phases. Consequently, our framework primarily focuses on prehensile manipulation tasks, such as pick-and-place, and does not address purely non-prehensile behaviors.

ManiLong-Shot comprises three core modules: **First**, it enables flexible task decomposition of long-horizon tasks from a single demonstration, utilizing either a simple heuristic based on gripper status and joint velocity or the semantic reasoning capabilities of a Vision-Language Model (VLM). Both strategies are implemented and systematically evaluated within our framework. **Second**, after decomposing the one-shot demonstration into primitives, ManiLong-Shot focuses on robustly imitating each primitive. It employs an interaction-aware region prediction network that identifies crucial object surface regions for each interaction. For instance, in the cup-grasping task shown in Fig 1, the key action involves grasping the cup's edge—a region that consistently serves this function during the pre-contact phase, regardless of the cup's pose. Such regions are termed *invariant regions* (Zhang and Boularias 2024), defined as semantically and functionally stable interaction surfaces that generalize across diverse task scenes. **Finally**, ManiLong-Shot introduces an interaction-aware region matching network that aligns the predicted invariant regions from the demonstration with their counterparts in the current scene during execution. This matching yields the target pose for the robot's end-effector, enabling robust sequential execution of each primitive. Consequently, the robot can efficiently and reliably accomplish long-horizon tasks in a one-shot, fully end-to-end OSIL manner.

In summary, our contributions are as follows: (1) We propose **ManiLong-Shot**, a novel OSIL framework designed for long-horizon prehensile manipulation tasks, which structures tasks into sequential interaction primitives and enables effective end-to-end OSIL execution through an interaction-aware pipeline. (2) We introduce an interaction-aware task decomposition mechanism that utilizes VLMs or rule-based robot state transitions to derive interaction primitives from a single unannotated long-horizon demonstration. Each primitive undergoes a two-stage process to predict functionally invariant regions and perform region matching for accurate goal poses, ensuring robust task execution. (3) We construct the **RLBench-Oneshot** benchmark by selecting 30 tasks from the manipulation benchmark RLBench (James et al. 2020) for systematic evaluation of our framework. Furthermore, we demonstrate the effectiveness of ManiLong-Shot's OSIL capabilities on three real-world long-horizon manipulation tasks, validating its practical applicability.

2 Related Work

OSIL for Manipulation. One-shot imitation learning (OSIL) aims to learn from demonstrations of training tasks and generalize to novel tasks using only a single trajectory per new task, without additional training (Duan et al. 2017; Finn et al. 2017; Yu et al. 2018; Biza et al. 2023; Zhang and Boularias 2024; Bonardi, James, and Davison 2020). It is a key technique for enabling generalization in manipulation tasks. Early work employs meta-learning to transfer knowledge across diverse robotic tasks (Finn et al. 2017; Gao, Jiang, and Chen 2023; Yu et al. 2018), or trains transformer-based policies using expert trajectories (Mandi et al. 2022). Wen et al. (Wen et al. 2022b,a) adopt a "pre-training and adaptation" strategy to learn canonical representations, while others explore graph-based representations of task structure or scene geometry (Zhang and Boularias 2024; Kumar et al. 2023; Huang et al. 2019). However, most existing approaches are limited to simple, short-horizon tasks. Wu et al. (2024a) extend OSIL to long-horizon extrinsic manipulation by composing short-horizon, goal-conditioned primitives. In contrast, our work eliminates the reliance on predefined primitive libraries and infers the interaction process directly from a single demonstration, providing a more practical and scalable solution for long-horizon prehensile manipulation.

Task Decomposition for Long-horizon Tasks. A common approach to long-horizon manipulation is the "divide and conquer" strategy, which breaks complex tasks into shorter, relatively independent sub-tasks. These sub-tasks can be manually defined (Dalal, Pathak, and Salakhutdinov 2021; Gao et al. 2023), predefined by the environment (Chen et al. 2025d; Zhu et al. 2021), or learned automatically (Masson, Ranchod, and Konidaris 2016; Gao et al. 2025). However, ensuring their reusability remains a challenge—changes in task goals or environments can render predefined boundaries ineffective, hindering generalization and robustness (Wu et al. 2024a). Recent studies propose decomposition methods based on physical interactions between the robot gripper and the environment, demonstrating strong robustness and transferability (Chen et al. 2025b,e,a). With the rise of vision-language models (VLMs), research increasingly explores their reasoning and generative capabilities for sub-task planning (Huang et al. 2025; Ding et al. 2025). VLMs have been used to generate sub-goal sequences (Myers et al. 2024; Curtis et al. 2024; Zhu et al. 2025), synthesize executable manipulation code (Liang et al. 2023; Huang et al. 2023; Chen et al. 2025c), and identify consistent keypoints across instances (Fang et al. 2024), enhancing the flexibility and generalization of task decomposition. Our work builds on methods that leverage robot-object interactions (Chen et al. 2025b,e), proposing an interaction-aware framework for sub-task segmentation. This approach integrates heuristic changes in gripper state and VLMs to identify sub-task boundaries, providing a robust decomposition strategy for long-horizon prehensile manipulation tasks.

Invariant Correspondence for Manipulation. Establishing correspondences between seen and unseen scenarios (Zhang et al. 2024) and incorporating invariance (Brand-

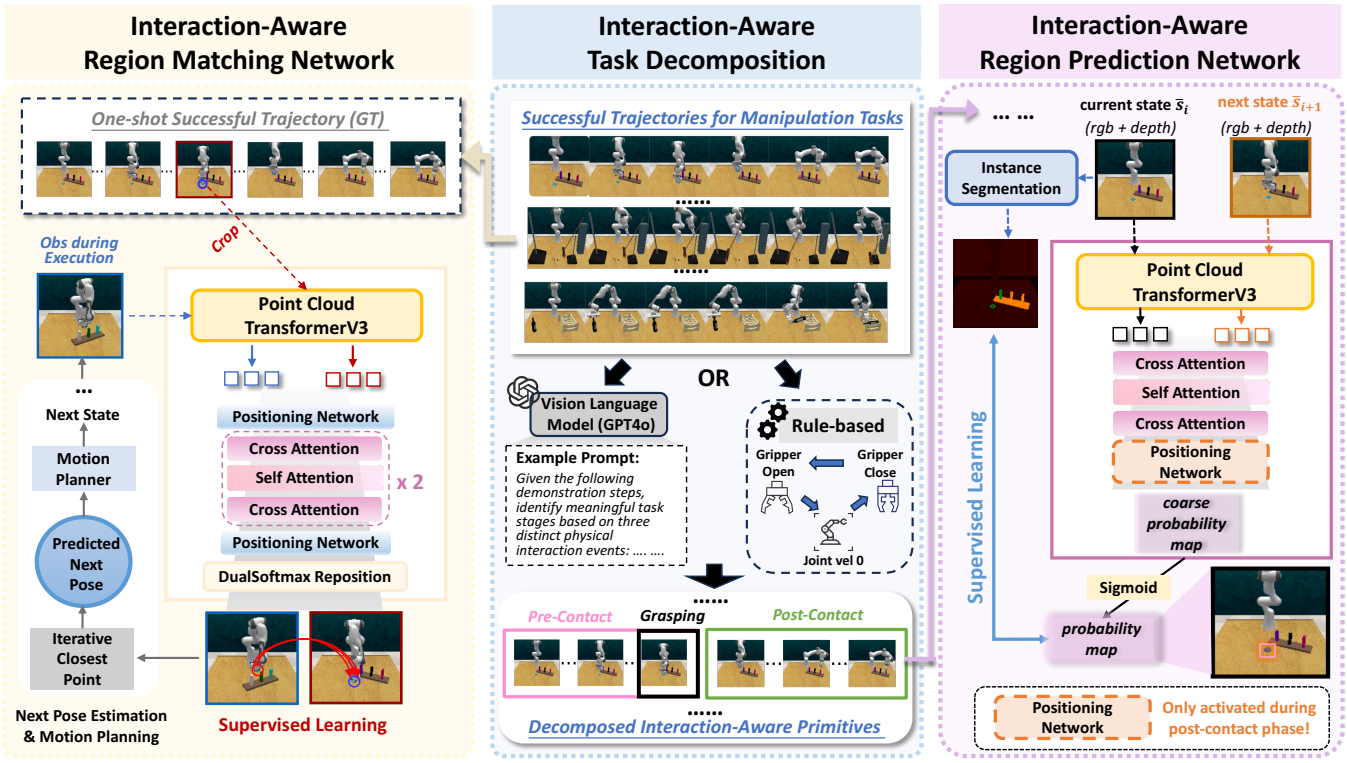

Figure 2: **The Overall Training Pipeline of ManiLong-Shot.** Best viewed when zoomed in.

stetter et al. 2021; Lyle et al. 2020; Gao et al. 2024) are crucial for generalization in robotic systems. In robotic manipulation, prior work has utilized self-supervised learning to build dense visual correspondences (Florence, Manuelli, and Tedrake 2018) or establish shape correspondences within object categories (Wen et al. 2022c), enabling grasp transfer across similar objects. Other approaches focus on learning viewpoint-invariant representations to facilitate action transfer from different perspectives (Graf et al. 2023). Recently, Zhang and Boularias(2024) proposed training neural networks to identify semantically invariant regions relevant to the robot's end-effector, enabling one-shot action transfer. Our work builds on this by predicting and matching invariant regions across multiple phases of interaction-aware primitives, supporting effective one-shot generalization even in complex long-horizon manipulation tasks.

3 Problem Formulation

Our work addresses the problem of enabling a robot to achieve one-shot generalization to novel long-horizon prehensile manipulation tasks, given only a single unannotated successful demonstration. We focus on a class of representative tasks in which the robot operates a parallel-jaw gripper to grasp objects and perform pick-and-place interactions with multiple objects in a tabletop environment. Under this setting, we formally define the short-horizon and long-horizon manipulation tasks considered in this work as follows: Based on the three physical interaction phases between the gripper and a single object—*pre-contact*, *grasping*, and *post-contact*—a task is defined as *short-horizon (SH)* if the gripper interacts with an object no more than once, i.e., the total number of such interactions is at most three. Otherwise, if the number of gripper-object interactions exceeds three, such as when manipulating multiple objects, the task is considered *long-horizon (LH)*. Formally, the manipulation tasks in this work are divided into two subsets: a SH set $\mathcal{T}^{\text{sh}}$, which provides abundant offline demonstrations for training, and a LH set $\mathcal{T}^{\text{lh}}$, which is unseen during training and requires one-shot generalization. During training, the robot learns from a dataset of demonstrations $\mathcal{D} = \{\tau_{\text{sh}}^{(k)} = \{(\bar{s}_t^{(k)}, \bar{a}_t^{(k)}, \bar{s}_{t+1}^{(k)})\}_{t=1}^{T_k}\}_{k=1}^{K}$, sampled from $\mathcal{T}^{\text{sh}}$, where each $\tau^{(k)}$ is a sequence of state-action pairs. At test time, the robot is provided with a single successful demonstration for each novel task in $\mathcal{T}^{\text{lh}}$, denoted as $\tau^{\text{lh}} = \{(\hat{s}_t, \hat{a}_t, \hat{s}_{t+1})\}_{t=0}^{H}$, where $\hat{a}_t$ transitions the system from $\hat{s}_t$ to $\hat{s}_{t+1}$. Given any state s encountered during execution of novel LH task, the robot aims to predict an 18-dimensional action $a = (\mathbf{T}, \lambda)$ based on τ^{lh}, where $\mathbf{T} \in \mathbb{R}^{4\times4}$ specifies the desired end-effector pose in $SE(3)$, and $\lambda \in \{0, 1\}^2$ encodes the gripper command and a collision flag. Low-level motion execution for $\mathbf{T}$ is handled by standard motion planners. The robot iteratively predicts actions and plans motions until the whole task is completed.

4 Method

Framework Overview. We propose **ManiLong-Shot**, an interaction-aware one-shot imitation learning (OSIL) frame-

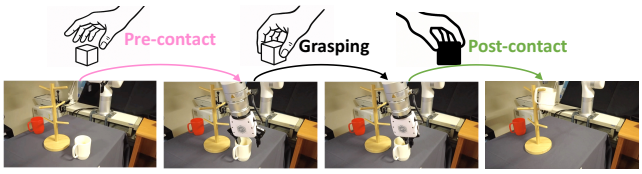

Figure 3: Visualization of the three physical interaction phases.

work for long-horizon prehensile manipulation. It consists of three modules: (1) **Interaction-aware Task Decomposition**, which decomposes a demonstration into primitives based on three physical interaction events; (2) **Interaction-aware Region Prediction Network**, which predicts functionally and semantically invariant regions for each subtask; and (3) **Interaction-aware Region Matching Network**, which aligns these regions with novel scene observations to ensure spatial correspondence. After matching, the system estimates and executes the next end-effector pose via motion planning, enabling sequential task execution. ManiLong-Shot achieves effective one-shot generalization to complex, long-horizon manipulation tasks. Fig. 2 illustrates the training pipeline of ManiLong-Shot.

Interaction-aware Task Decomposition. This process organizes successful demonstration trajectories into a sequence of interaction-aware primitives based on physical interaction phases. We focus on prehensile manipulation tasks with a gripper, drawing inspiration from human strategies in OSIL for long-horizon tasks. Following existing sub-task decomposition schemes (James and Davison 2022; Chen et al. 2025b,e), we categorize robot-object interactions into three phases: *pre-contact*, *grasping*, and *post-contact*. Fig. 3 illustrates these phases in the *Place 1 Cup on Cupholder* task with a real robot, highlighting their correspondence with human cognitive understanding of manipulation. Demonstration trajectories can be decomposed in two ways. One approach uses tailored prompts for vision-language models (VLMs) and structured representations of the trajectory to guide the automatic identification of interaction phases. In ManiLong-Shot, we employ GPT-4o (Hurst et al. 2024) as the underlying VLM. Alternatively, a rule-based method analyzes joint velocity and gripper state changes to infer phase boundaries throughout the trajectory. The *pre-contact* phase includes frames where the gripper remains open while approaching the object, ending when joint velocity drops to zero before closure, indicating precise alignment. The *grasping* phase captures the transition from an open to a closed gripper securing the object. The *post-contact* phase follows a successful grasp and involves transitioning from a closed to an open gripper while placing the object or interacting with others. Short-horizon tasks typically involve a single cycle of these phases, while long-horizon tasks repeat them until completion. Both decomposition methods are interchangeable in our framework: the rule-based method offers stability, while the VLM-based approach enables learning of semantically aware interaction patterns.

Interaction-aware Region Prediction Network. After extracting interaction-aware subtask primitives from demonstrations, ManiLong-Shot predicts functionally **interaction-invariant** regions across interaction stages. To achieve this, we train an Interaction-aware Region Prediction Network (Fig. 2(right)) using only successful short-horizon (SH) demonstrations, highlighting the system's data efficiency. The network builds on the *invariant region* concept from Zhang and Boularias (2024), where a 3D geometric subset is invariant if it preserves structure equivariant to $SE(3)$ transformations in states with the same optimal policy. For example, in the "insert onto square peg" task (Fig. 2), the "square ring" in the *pre-contact* phase consistently aligns with the grasp surface's principal axis, despite pose variations, demonstrating strong invariant properties.

Training the Interaction-aware Region Prediction Network utilizes demonstrations from the SH task set $\mathcal{T}^{\text{sh}}$. For each task $\mathcal{T}_i^{\text{sh}}$, one demonstration is divided into three segments: τ_{pre}^{sh}, τ_{grasp}^{sh}, and τ_{post}^{sh}. For each pair of consecutive states $\{\overline{s}_i, \overline{s}_{i+1}\}$, dense point clouds are generated from RGB-D observations and processed by the Point Cloud Transformer V3 (PTV3) (Wu et al. 2024b). This involves progressive downsampling with cross-attention across scenes ($\overline{s}_i$ and $\overline{s}_{i+1}$) and self-attention within each scene, enhancing inter-state correspondence and intra-state feature representation. The Positioning Network is activated only during the *post-contact* phase, where the robot aligns the grasped object with a target region using attention mechanisms. Since the invariant regions in the *pre-contact* and *grasping* phases are functionally similar, we train their prediction networks jointly. The network generates a coarse saliency map from learned spatial priors, refines it through point-wise sigmoid parameterization, and outputs an interaction probability distribution to activate the relevant region $\mathcal{I}(\overline{s}_i)$, yielding the predicted invariant region (square ring in Fig. 2). The pipeline is trained with supervised learning using ground-truth instance masks from simulation for accurate supervision.

Interaction-aware Region Matching Network. Another key module of ManiLong-Shot is the Interaction-aware Region Matching Network. While the Interaction-aware Region Prediction Network identifies invariant regions across interaction primitives in demonstrations, this module enables the robot to match these regions to its current execution state for action reproduction without additional training. The network architecture builds on prior work (Zhang and Boularias 2024), as illustrated in Fig. 2(left). It is trained on multiple SH task demonstrations without explicit task decomposition, with each state annotated with GT invariant regions derived from instance segmentation masks and RGB-D data. For a given SH task, consider a rollout trajectory τ_j^{sh} that differs from the demonstration τ_i^{sh}. For each state $\overline{s}_j$ in τ_j^{sh}, a state routing network (Zhang and Boularias 2024) selects the most similar state $\overline{s}_i$ in τ_i^{sh} based on feature similarity. The invariant region $I(\overline{s}_i)$ is cropped from its point cloud and fused with the full-scene point cloud of $\overline{s}_j$, jointly downsampled by the PTV3 model. The Positioning Network applies an initial attention step followed by a two-stage cross–self–cross attention module to enhance spatial and geometric feature alignment between the two clouds,

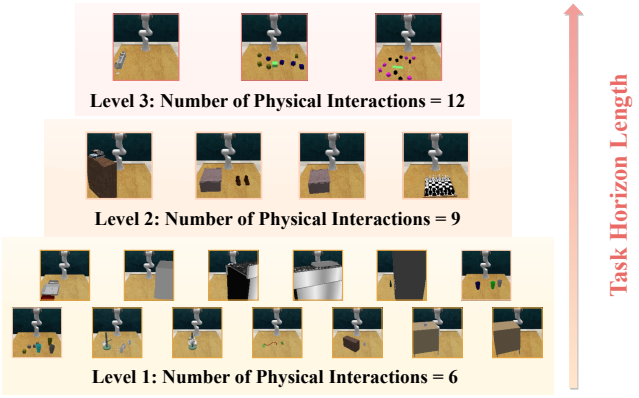

Figure 4: Visualization of 20 long-horizon manipulation tasks in **RLBench-Oneshot**, including 3 difficulty levels.

producing point-wise features over $\overline{s}_j$. Based on these features, a dual-softmax matching algorithm (Li and Harada 2022) computes a correspondence matrix $\mathbf{C}$ between the scene cloud of $\overline{s}_j$ and the invariant region cloud of $\overline{s}_i$. The network is trained using GT correspondence matrices and object position labels between τ_i^{sh} and τ_j^{sh} at each timestep.

After obtaining the correspondence matrix $\mathbf{C}$ between the invariant region $\mathcal{I}(\overline{s}_i)$ of the demonstrated state and the current execution state $\overline{s}_j$, we apply a correspondence-based pose regression algorithm (Zhang and Boularias 2024) to estimate the robot's action pose $\mathbf{T}_j$ at $\overline{s}_j$. The optimization objective is: $\mathbf{T}_j = \arg\min_{\mathbf{T}\in \mathrm{SE}(3)} \left\| \mathbf{T}\,\mathbf{T}_i^{-1}\, P_{\mathcal{I}(\overline{s}_i)}\, \mathbf{C} - P_{\overline{s}_j} \right\|$, where $\mathbf{T}_i$ denotes the demonstrated action pose at $\overline{s}_i$, $P_{\mathcal{I}(\overline{s}_i)}$ and $P_{\overline{s}_j}$ represent the point clouds of the invariant region in $\overline{s}_i$ and the full scene in $\overline{s}_j$, respectively. Once the next pose $\mathbf{T}_j$ is determined, we use the RRT-Connect algorithm (LaValle and Kuffner Jr 2001) for path planning, computing a collision-free trajectory from the current state to the target state and determining the next state observation after executing the predicted pose. This observation is then fed into the state routing network for comparison with frames from the demonstration trajectory, initiating the next round of invariant region matching. This iterative process continues until the entire task is completed.

Evaluation of ManiLong-Shot These three key modules together form the ManiLong-Shot framework. When evaluating unseen long-horizon (LH) tasks, ManiLong-Shot first uses the Interaction-Aware Task Decomposition Module to break down a successful demonstration into a sequence of interaction-aware primitives. For each interaction phase, the Interaction-Aware Region Prediction Network predicts the corresponding invariant region. During execution, the system captures real-time observations and employs the Interaction-Aware Region Matching Network to align the predicted regions with the current observations. Based on the correspondence matrix, it performs pose prediction and motion planning. This iterative process continuously refines poses and actions through region matching until the entire long-horizon task is completed.

5 Experiments

We conduct extensive experiments in both simulated environments and real-world settings to address the following questions: 1) Can ManiLong-Shot achieve high success rates on SH tasks encountered during training, under different initial conditions? 2) Given a single successful demonstration, can ManiLong-Shot effectively one-shot generalize to unseen LH manipulation tasks without any fine-tuning? 3) Can ManiLong-Shot perform OSIL effectively in real-world LH tasks through sim-to-real transfer? 4) How does the heuristic design of the key modules affect ManiLong-Shot's one-shot generalization performance on unseen LH tasks?

5.1 Experimental Setup

Simulation Benchmark. We select 30 tasks from the 100 tasks in the robotic manipulation simulation benchmark RLBench (James et al. 2020) to create the dedicated benchmark **RLBench-Oneshot** for evaluating the OSIL capabilities of robotic manipulation models. Based on the SH and LH task definitions in Sec. 3, the benchmark includes 10 SH tasks and 20 LH tasks. The SH tasks involve a single round of *pre-contact, grasping, and post-contact* interactions, with each task comprising 100 successful demonstration trajectories recorded from front, left/right shoulder, and wrist cameras using RGB-D observation. The LH tasks are categorized into three difficulty levels based on the task horizon length, as shown in Fig. 4: Level 1 includes 13 tasks with 6 physical interactions each; Level 2 includes 4 tasks with 9 interactions; Level 3 includes 3 tasks with 12 interactions. These levels comprehensively evaluate the model's OSIL ability on various LH tasks, with each LH task accompanied by one successful trajectory as a one-shot demonstration. More details on the RLBench-Oneshot benchmark tasks are provided in the Technical Appendix on our project website.

Real-World Setup. We validate ManiLong-Shot with a UFactory xArm7 robot, utilizing RGB-D observations from RealSense D435 and D415 cameras. For motion planning, we integrate MoveIt (Coleman et al. 2014) to guide the arm to the end-effector poses. We set up three LH manipulation tasks in the real world: *Stack Blocks*, *Stack Cups*, and *Place Cups*, each involving six physical interaction phases. For each task, we collect a successful trajectory through human teleoperation and conduct 5 one-shot generalization trials. In each trial, the object positions are slightly perturbed from those in the demonstration. This setup is used to evaluate the sim-to-real transfer of the ManiLong-Shot model pretrained in simulation, validating its effectiveness and robustness in real-world scenarios.

Baseline Models and Metrics. We select ARP (Zhang et al. 2025), 3DDA (Ke, Gkanatsios, and Fragkiadaki 2024), RVT2 (Goyal et al. 2024), and IMOP (Zhang and Boularias 2024) as baseline models for comparison. The first three are state-of-the-art (SOTA) IL-based multi-task models on the RLBench benchmark, used to evaluate ManiLong-Shot's performance on short-horizon (SH) tasks and its one-shot generalization on long-horizon (LH) tasks without fine-tuning. IMOP, the leading OSIL method for RLBench tasks,

Models	Avg. Success (%) ↑	Avg. Rank ↓	Turn Tap	Open Drawer	Put Groceries in Cupboard	Place Shape in Shape Sorter	Put Money in Safe	Close Jar	Place Wine	Light Bulb In	Insert Peg	Meat off Grill
RVT2	79.2	3.1	99.0 ± 1.5	74.0 ± 4.5	66.0 ± 5.6	35.0 ± 6.8	96.0 ± 2.0	**100.0 ± 0.0**	95.0 ± 2.4	88.0 ± 3.2	40.0 ± 6.2	99.0 ± 1.2
3DDA	85.0	2.9	**99.2** ± 1.6	89.6 ± 4.1	**85.6** ± 4.1	44.0 ± 4.4	97.6 ± 2.0	96.0 ± 3.6	93.6 ± 4.8	82.4 ± 2.0	65.6 ± 4.1	96.8 ± 1.6
ARP	86.6	2.7	100.0 ± 0.0	92.8 ± 2.2	69.6 ± 5.2	46.4 ± 6.3	94.4 ± 2.5	97.6 ± 1.5	92.0 ± 2.1	94.4 ± 1.8	**78.4** ± 3.9	96.0 ± 1.7
IMOP	65.24	3.9	51.2 ± 5.9	100.0 ± 0.0	46.4 ± 6.5	37.6 ± 7.0	96.0 ± 2.4	39.2 ± 6.7	96.0 ± 2.3	82.0 ± 4.1	12.0 ± 9.5	92.0 ± 2.7
ManiLong-Shot	**90.4** (**3.8%** ↑)	**1.9**	95.4 ± 2.1	**100.0** ± 0.0	82.0 ± 4.2	**48.0** ± 6.3	**100.0** ± 0.0	96.8 ± 1.9	**100.0** ± 0.0	**88.0** ± 2.8	44.0 ± 6.5	**100.0** ± 0.0

Table 1: **Performance on 10 Training SH Tasks.** We report the mean and standard deviation of success rates over 5 random seeds for each task, along with the average success rate and average rank across all tasks.

Models	Avg. Success (%) ↑	Avg. Rank ↓	Empty Container	Empty Dishwasher	Tray off Oven	Tray in Oven	Bottle in Fridge	Place 2 Cups	Remove 2 Cups	Straighten Rope	Slide & Place
RVT2+FT	4.1	3.7	1.6 ± 0.4	0.0 ± 0.0	0.0 ± 0.0	1.2 ± 0.3	4.0 ± 0.7	1.3 ± 0.5	2.7 ± 0.8	2.7 ± 0.8	2.7 ± 0.7
3DDA+FT	4.5	3.3	1.6 ± 0.5	1.3 ± 0.6	1.6 ± 0.7	1.2 ± 0.5	4.3 ± 0.9	1.3 ± 0.5	5.3 ± 1.0	5.3 ± 1.1	0.0 ± 0.0
ARP+FT	4.7	3.3	1.3 ± 0.6	2.7 ± 0.9	5.3 ± 1.1	0.0 ± 0.0	2.7 ± 0.7	0.0 ± 0.0	4.0 ± 0.9	4.0 ± 0.8	5.3 ± 1.2
IMOP	7.4	2.6	4.0 ± 0.8	1.3 ± 0.6	2.7 ± 0.9	5.3 ± 0.7	1.3 ± 0.5	2.7 ± 0.7	4.0 ± 0.8	4.0 ± 0.9	5.3 ± 1.0
ManiLong-Shot	**30.2** (**22.8%** ↑)	**1.0**	**28.0** ± 3.2	**42.7** ± 2.8	**37.3** ± 3.5	**48.0** ± 2.1	**18.7** ± 4.3	**14.7** ± 3.9	**24.0** ± 2.7	**30.7** ± 2.5	**26.7** ± 3.0

Models	Put Item	Take Item	Stack Cups	Stack Cups Blocks	Put 3 Books	Put Shoes	Take Shoes	Setup Chess	Set Table	Stack Blocks	Block Pyramid
RVT2+FT	28.0 ± 1.9	29.3 ± 2.1	2.7 ± 0.6	0.0 ± 0.0	2.7 ± 0.7	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	1.3 ± 0.5	0.0 ± 0.0	0.0 ± 0.0
3DDA+FT	29.3 ± 2.0	30.7 ± 1.9	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	1.3 ± 0.5	4.0 ± 0.8	1.3 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
ARP+FT	18.7 ± 1.5	37.3 ± 2.4	8.0 ± 1.2	0.0 ± 0.0	0.0 ± 0.0	2.7 ± 0.6	0.0 ± 0.0	1.3 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
IMOP	40.0 ± 2.0	38.7 ± 2.1	4.0 ± 1.0	1.3 ± 0.6	9.3 ± 1.4	1.3 ± 0.5	1.3 ± 0.5	1.3 ± 0.5	2.7 ± 0.6	1.3 ± 0.5	1.3 ± 0.5
ManiLong-Shot	**65.3** ± 2.4	**76.0** ± 1.8	**28.0** ± 3.1	**18.7** ± 4.0	**42.7** ± 2.7	**29.3** ± 2.9	**12.0** ± 4.6	**9.3** ± 4.8	**17.3** ± 3.5	**8.0** ± 4.9	**8.0** ± 4.9

Table 2: **OSIL Performance on 20 Unseen LH Tasks.** We report the mean and standard deviation of success rates over 5 random seeds for each task, along with the average success rate and average rank across all tasks. "FT" means "Fine-tuned with 5 demos".

serves as a key baseline for assessing ManiLong-Shot's OSIL performance on unseen LH tasks. Simulation evaluation metrics include average success rate and average rank. The success rate for each task is the mean and standard deviation from 25 trials with 5 random seeds, as RLBench-Oneshot generates different object layouts across trials. The average success rate is the mean across all tasks, while the average rank reflects overall performance by averaging rankings across tasks. For real-world robot experiments, we report the average success rate from 5 trials.

5.2 Main Results

Performance Comparison on SH Tasks To evaluate ManiLong-Shot on training SH tasks, we assess policies learned solely from offline demonstrations, without any access to one-shot examples. Table 1 reports the performance of ManiLong-Shot compared to several baselines on 10 SH tasks from the RLBench-Oneshot benchmark, which are also used during training. In our main experiments, the interaction-aware task decomposition module is implemented using a rule-based approach. During evaluation, the interaction-aware region prediction network identifies distinct interaction phases by monitoring changes in the robot's current state. Experimental results demonstrate that ManiLong-Shot achieves strong imitation learning performance, significantly outperforming three SOTA baselines. It reaches an average success rate of 90.4%, yielding a **3.8% improvement** over the best-performing baseline, 3DDA. ManiLong-Shot achieves the highest success rate on 6 out of 10 tasks, including precise manipulation (*Open Drawer*, 100% vs. ARP 92.8%) and multi-step coordination (*Put Money in Safe*, 100% vs. 3DDA 97.6%). These results highlight the effectiveness of ManiLong-Shot's interaction-aware mechanism in maintaining consistent imitation learning performance across different stages of interaction.

OSIL Performance on Novel LH Tasks Table 2 shows the OSIL performance of ManiLong-Shot and baselines on 20 novel LH tasks from **RLBench-Oneshot**. RVT2+FT, 3DDA+FT, and ARP+FT denote models fine-tuned on each task using five successful demonstrations. Despite this few-shot adaptation, these models struggle to generalize, particularly on tasks involving novel scenes and interactions. In contrast, for tasks that share scenes with training-time SH tasks (e.g., *Put Item in Drawer* and *Take Item out Drawer*, which share scenes with *Open Drawer*), the fine-tuned models achieve moderate generalization. Notably, ManiLong-Shot reaches an average success rate of 30.2% across all unseen LH tasks without task-specific fine-tuning, outperforming IMOP by **22.8%**. On the most challenging unseen LH tasks, ManiLong-Shot achieves significant results without fine-tuning, such as *Set Table* (17.4% vs. IMOP 0.0%), *Stack Blocks* (8.0% vs. IMOP 0.0%), and *Block Pyramid* (8.0% vs. IMOP 0.0%). Notably, *Stack Blocks* and *Block Pyramid* involve long interaction sequences and require precise manipulation. These results underscore the effectiveness of ManiLong-Shot's three core modules. For complex LH tasks, ManiLong-Shot decomposes tasks into interaction-aware primitives aligned with physical interac-

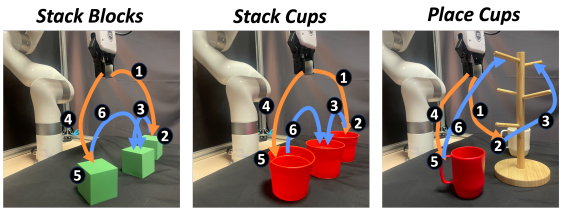

Figure 5: Three real-world LH tasks with their physical interaction visualizations.

Task	Success Rate	
	IMOP	**ManiLong-Shot**
Stack Blocks	60%	**80%**
Stack Cups	20%	**60%**
Place Cups	20%	**40%**
Average	33.3%	**60.0% (26.7% ↑)**

Table 3: **The OSIL Sim-to-Real Transfer Performance.**

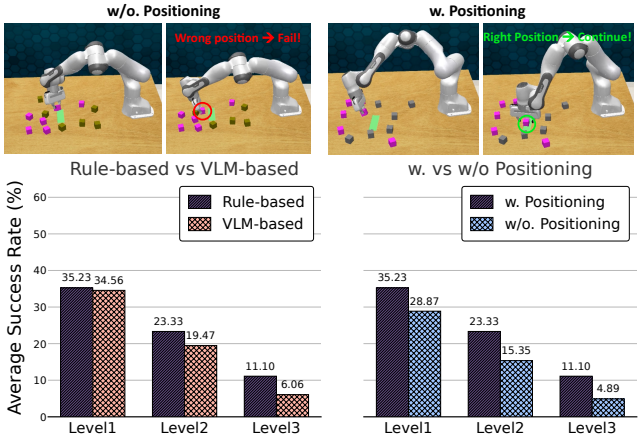

Figure 6: **Ablation Study of Key Modules in ManiLong-Shot.** Comparison of average success rates on the LH tasks across three difficulty levels in RLBench-Oneshot.

tion phases, enhancing robustness to variations in object poses and dynamics. By performing region prediction and matching at each phase, the system improves execution stability and overall success.

Real-world Experiments We evaluate ManiLong-Shot's OSIL performance in real-world LH manipulation tasks via sim-to-real transfer on three representative tasks: *Stack Blocks*, *Stack Cups*, and *Place Cups* (as illustrated in Fig. 5). For each task, a single successful demonstration trajectory is collected through human teleoperation of the robot arm, with the data collection process visualized in the figure. Table 3 reports the average success rates of ManiLong-Shot and IMOP over five OSIL trials on each task, with each evaluation performed under slightly varied object layouts. The results show that ManiLong-Shot achieves effective OSIL performance on real-world LH manipulation tasks, outperforming IMOP with an average success rate improvement of approximately 26.7%.

5.3 Ablation Studies

VLM-based vs. Rule-based. We investigate how different task decomposition strategies within ManiLong-Shot's interaction-aware task decomposition module affect OSIL performance on unseen LH tasks. As shown in Fig. 6(left), the VLM-based decomposition consistently underperforms the rule-based approach across all difficulty levels, primarily due to the instability of VLM reasoning. As task difficulty increases, leading to more primitives to identify and execute, the performance gap widens. Nevertheless, ManiLong-Shot with VLM-based decomposition still significantly outperforms all baseline models on unseen LH tasks, validating the framework's effectiveness in one-shot generalization. Detailed prompts for VLM are presented in Appendix.

W. Positioning vs. w/o. Positioning. We further investigate the effect of the positioning network in ManiLong-Shot's interaction-aware invariant region prediction module, particularly for invariant region prediction during the post-contact phase. As shown in Fig. 6(right), across all three difficulty levels, removing the positioning network in this phase may result in inaccurate placement of the grasped object. Such misalignment often disrupts the execution of subsequent sub-task primitives, which ultimately impedes the successful completion of the overall LH task and causes a notable decline in the system's OSIL performance on unseen LH tasks. Fig. 6(top) visualizes the differences in ManiLong-Shot's performance on the *Block Pyramid* task when placing the first block in the target region, comparing the cases with and without the positioning network.

6 Conclusion and Limitations

We present ManiLong-Shot, a novel OSIL framework tailored for long-horizon prehensile manipulation tasks. ManiLong-Shot leverages the physical interaction between the robot and the object to decompose tasks into *pre-contact*, *grasping*, and *post-contact* primitives, predicting interaction-aware invariant regions for each phase. Using these predictions, ManiLong-Shot performs region matching and pose prediction through correspondences during the one-shot learning phase. We evaluate the performance of ManiLong-Shot through extensive experiments on both simulated and real-world robots, demonstrating its effective and robust OSIL ability for long-horizon manipulation.

Limitations and Future Work. A primary limitation of ManiLong-Shot lies in its decomposition strategy, which defines manipulation in terms of three discrete phases grounded in physical contact. While effective for many long-horizon prehensile tasks, this limits applicability to tasks with extended post-contact tool use, such as wiping or pouring, where interactions are continuous and hard to decompose. Future work will focus on extending the framework to handle more general long-horizon tasks, cross-embodiment scenarios, and complex, temporally continuous behaviors.

Acknowledgements

This work is supported in part by National Natural Science Foundation of China (62192783, 62276128), Jiangsu Natural Science Foundation (BK20243051), Jiangsu Science and Technology Major Project (BG2024031), the Fundamental Research Funds for the Central Universities(14380128, KG202514), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Program of China Scholarship Council (Grant No. 202306190157).

References

Biza, O.; Thompson, S.; Pagidi, K. R.; Kumar, A.; van der Pol, E.; Walters, R.; Kipf, T.; van de Meent, J.-W.; Wong, L. L.; and Platt, R. 2023. One-shot imitation learning via interaction warping. *arXiv preprint arXiv:2306.12392*.

Bonardi, A.; James, S.; and Davison, A. J. 2020. Learning one-shot imitation from humans without humans. *IEEE Robotics and Automation Letters*, 5(2): 3533–3539.

Brandstetter, J.; Hesselink, R.; van der Pol, E.; Bekkers, E. J.; and Welling, M. 2021. Geometric and physical quantities improve e (3) equivariant message passing. *arXiv preprint arXiv:2110.02905*.

Chen, Y.; Chen, Z.; Chan, N. T.; Chen, J.; Yin, J.; Shi, J.; Gao, Y.; Li, Y.-L.; and Huo, J. 2025a. RoboHiMan: A Hierarchical Evaluation Paradigm for Compositional Generalization in Long-Horizon Manipulation. *arXiv preprint arXiv:2510.13149*.

Chen, Y.; Chen, Z.; Yin, J.; Huo, J.; Tian, P.; Shi, J.; and Gao, Y. 2025b. GravMAD: Grounded Spatial Value Maps Guided Action Diffusion for Generalized 3D Manipulation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Chen, Z.; Huo, J.; Chen, Y.; and Gao, Y. 2025c. RoboHorizon: An LLM-Assisted Multi-View World Model for Long-Horizon Robotic Manipulation. *arXiv preprint arXiv:2501.06605*.

Chen, Z.; Ji, Z.; Huo, J.; and Gao, Y. 2025d. SCaR: Refining skill chaining for long-horizon robotic manipulation via dual regularization. *Advances in Neural Information Processing Systems*, 37: 111679–111714.

Chen, Z.; Yin, J.; Chen, Y.; Huo, J.; Tian, P.; Shi, J.; Hou, Y.; Li, Y.; and Gao, Y. 2025e. DeCo: Task Decomposition and Skill Composition for Zero-Shot Generalization in Long-Horizon 3D Manipulation. *arXiv preprint arXiv:2505.00527*.

Coleman, D.; Sucan, I.; Chitta, S.; and Correll, N. 2014. Reducing the barrier to entry of complex robotic software: a moveit! case study. *arXiv preprint arXiv:1404.3785*.

Curtis, A.; Kumar, N.; Cao, J.; Lozano-Pérez, T.; and Kaelbling, L. P. 2024. Trust the PRoC3S: Solving Long-Horizon Robotics Problems with LLMs and Constraint Satisfaction. In *8th Annual Conference on Robot Learning*.

Dalal, M.; Pathak, D.; and Salakhutdinov, R. R. 2021. Accelerating robotic reinforcement learning via parameterized action primitives. *Advances in Neural Information Processing Systems*, 34: 21847–21859.

Ding, H.; Xu, Z.; Fang, Y.; Wu, Y.; Chen, Z.; Shi, J.; Huo, J.; Zhang, Y.; and Gao, Y. 2025. LaViRA: Language-Vision-Robot Actions Translation for Zero-Shot Vision Language Navigation in Continuous Environments. *arXiv preprint arXiv:2510.19655*.

Duan, Y.; Andrychowicz, M.; Stadie, B.; Jonathan Ho, O.; Schneider, J.; Sutskever, I.; Abbeel, P.; and Zaremba, W. 2017. One-shot imitation learning. *Advances in neural information processing systems*, 30.

Fang, X.; Huang, B.-R.; Mao, J.; Shone, J.; Tenenbaum, J. B.; Lozano-Pérez, T.; and Kaelbling, L. P. 2024. Keypoint Abstraction using Large Models for Object-Relative Imitation Learning. *arXiv preprint arXiv:2410.23254*.

Finn, C.; Yu, T.; Zhang, T.; Abbeel, P.; and Levine, S. 2017. One-shot visual imitation learning via meta-learning. In *Conference on robot learning*, 357–368. PMLR.

Florence, P. R.; Manuelli, L.; and Tedrake, R. 2018. Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation. In *Conference on Robot Learning*, 373–385. PMLR.

Gao, C.; Jiang, Y.; and Chen, F. 2023. Transferring hierarchical structures with dual meta imitation learning. In *Conference on Robot Learning*, 762–773. PMLR.

Gao, C.; Li, Z.; Gao, H.; and Chen, F. 2023. Iterative interactive modeling for knotting plastic bags. In *Conference on Robot Learning*, 571–582. PMLR.

Gao, C.; Xue, Z.; Deng, S.; Liang, T.; Yang, S.; Shao, L.; and Xu, H. 2024. RiEMann: Near real-time SE (3)-equivariant robot manipulation without point cloud segmentation. *arXiv preprint arXiv:2403.19460*.

Gao, C.; Zhang, H.; Xu, Z.; Zhehao, C.; and Shao, L. 2025. FLIP: Flow-Centric Generative Planning as General-Purpose Manipulation World Model. In *The Thirteenth International Conference on Learning Representations*.

Goyal, A.; Blukis, V.; Xu, J.; Guo, Y.; Chao, Y.-W.; and Fox, D. 2024. RVT2: Learning Precise Manipulation from Few Demonstrations. *RSS*.

Graf, C.; Adrian, D. B.; Weil, J.; Gabriel, M.; Schillinger, P.; Spies, M.; Neumann, H.; and Kupcsik, A. G. 2023. Learning dense visual descriptors using image augmentations for robot manipulation tasks. In *conference on Robot Learning*, 871–880. PMLR.

Huang, D.-A.; Nair, S.; Xu, D.; Zhu, Y.; Garg, A.; Fei-Fei, L.; Savarese, S.; and Niebles, J. C. 2019. Neural task graphs: Generalizing to unseen tasks from a single video demonstration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8565–8574.

Huang, J.; Xu, Y.; Wang, Q.; Wang, Q. C.; Liang, X.; Wang, F.; Zhang, Z.; Wei, W.; Zhang, B.; Huang, L.; et al. 2025. Foundation models and intelligent decision-making: Progress, challenges, and perspectives. *The Innovation*.

Huang, W.; Wang, C.; Zhang, R.; Li, Y.; Wu, J.; and Fei-Fei, L. 2023. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. In *Conference on Robot Learning*, 540–562. PMLR.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

James, S.; and Davison, A. J. 2022. Q-attention: Enabling efficient learning for vision-based robotic manipulation. *IEEE Robotics and Automation Letters*, 7(2): 1612–1619.

James, S.; Ma, Z.; Arrojo, D. R.; and Davison, A. J. 2020. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2): 3019–3026.

Ke, T.-W.; Gkanatsios, N.; and Fragkiadaki, K. 2024. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv preprint arXiv:2402.10885*.

Kumar, S.; Zamora, J.; Hansen, N.; Jangir, R.; and Wang, X. 2023. Graph inverse reinforcement learning from diverse videos. In *Conference on Robot Learning*, 55–66. PMLR.

LaValle, S. M.; and Kuffner Jr, J. J. 2001. Randomized kinodynamic planning. *The international journal of robotics research*, 20(5): 378–400.

Li, Y.; and Harada, T. 2022. Lepard: Learning partial point cloud matching in rigid and deformable scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5554–5564.

Liang, J.; Huang, W.; Xia, F.; Xu, P.; Hausman, K.; Ichter, B.; Florence, P.; and Zeng, A. 2023. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 9493–9500. IEEE.

Lyle, C.; van der Wilk, M.; Kwiatkowska, M.; Gal, Y.; and Bloem-Reddy, B. 2020. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*.

Mandi, Z.; Liu, F.; Lee, K.; and Abbeel, P. 2022. Towards more generalizable one-shot visual imitation learning. In *2022 International conference on robotics and automation (ICRA)*, 2434–2444. IEEE.

Masson, W.; Ranchod, P.; and Konidaris, G. 2016. Reinforcement learning with parameterized actions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Myers, V.; Zheng, C.; Mees, O.; Fang, K.; and Levine, S. 2024. Policy Adaptation via Language Optimization: Decomposing Tasks for Few-Shot Imitation. In *8th Annual Conference on Robot Learning*.

Valassakis, E.; Papagiannis, G.; Di Palo, N.; and Johns, E. 2022. Demonstrate once, imitate immediately (dome): Learning visual servoing for one-shot imitation learning. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 8614–8621. IEEE.

Vosylius, V.; and Johns, E. 2025. Instant Policy: In-Context Imitation Learning via Graph Diffusion. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wen, B.; Lian, W.; Bekris, K.; and Schaal, S. 2022a. Catgrasp: Learning category-level task-relevant grasping in clutter from simulation. In *2022 International Conference on Robotics and Automation (ICRA)*, 6401–6408. IEEE.

Wen, B.; Lian, W.; Bekris, K.; and Schaal, S. 2022b. You Only Demonstrate Once: Category-Level Manipulation from Single Visual Demonstration. *Robotics: Science and Systems 2022*.

Wen, H.; Yan, J.; Peng, W.; and Sun, Y. 2022c. Transgrasp: Grasp pose estimation of a category of objects by transferring grasps from only one labeled instance. In *European Conference on Computer Vision*, 445–461. Springer.

Wu, A.; Wang, R.; Chen, S.; Eppner, C.; and Liu, C. K. 2024a. One-shot transfer of long-horizon extrinsic manipulation through contact retargeting. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 13891–13898. IEEE.

Wu, X.; Jiang, L.; Wang, P.-S.; Liu, Z.; Liu, X.; Qiao, Y.; Ouyang, W.; He, T.; and Zhao, H. 2024b. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4840–4851.

Xu, M.; Shen, Y.; Zhang, S.; Lu, Y.; Zhao, D.; Tenenbaum, J.; and Gan, C. 2022. Prompting decision transformer for few-shot policy generalization. In *international conference on machine learning*, 24631–24645. PMLR.

Yu, T.; Finn, C.; Dasari, S.; Xie, A.; Zhang, T.; Abbeel, P.; and Levine, S. 2018. One-Shot Imitation from Observing Humans via Domain-Adaptive Meta-Learning. *Robotics: Science and Systems XIV*.

Zhang, J.; Herrmann, C.; Hur, J.; Polania Cabrera, L.; Jampani, V.; Sun, D.; and Yang, M.-H. 2024. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36.

Zhang, X.; and Boularias, A. 2024. One-Shot Imitation Learning with Invariance Matching for Robotic Manipulation. In *Proceedings of Robotics: Science and Systems*. Delft, Netherlands.

Zhang, X.; Liu, Y.; Chang, H.; Schramm, L.; and Boularias, A. 2025. Autoregressive action sequence learning for robotic manipulation. *IEEE Robotics and Automation Letters*.

Zhu, J.; Tie, C.; Cao, X.; Wang, Y.; Guo, J.; Chen, Z.; Chen, H.; Chen, J.; Xiao, Y.; Wu, R.; et al. 2025. AdaptPNP: Integrating Prehensile and Non-Prehensile Skills for Adaptive Robotic Manipulation. *arXiv preprint arXiv:2511.11052*.

Zhu, Y.; Tremblay, J.; Birchfield, S.; and Zhu, Y. 2021. Hierarchical planning for long-horizon manipulation with geometric and symbolic scene graphs. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 6541–6548. Ieee.