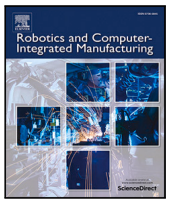
Robotics and Computer–Integrated Manufacturing 97 (2026) 103096

Contents lists available at ScienceDirect

Robotics and Computer-Integrated Manufacturing

journal homepage: www.elsevier.com/locate/rcim



LLM-driven symbolic planning and hierarchical imitation learning for long-horizon deformable object assembly☆

Jiaming Qi [a,1], Liang Lu [b,1], Fangyuan Wang [c], Hoi-Yin Lee [c], David Navarro-Alarcon [c], Zeqing Zhang [b,*], Peng Zhou [d,*]

[a] College of Mechanical and Electrical Engineering, Northeast Forestry University, Harbin, China
[b] Department of Computer Science, The University of Hong Kong, NT, Hong Kong
[c] Department of Mechanical Engineering, The Hong Kong Polytechnic University, KLN, Hong Kong
[d] School of Advanced Engineering, The Great Bay University, Dongguan, China

ARTICLE INFO

Keywords:
Large language models (LLMs)
Imitation learning
Symbolic planning
Deformable object manipulation
Robotic assembly

ABSTRACT

Long-horizon assembly tasks involving deformable objects pose substantial challenges for autonomous robots, stemming from infinite-dimensional state spaces, complex sequential dependencies, and high variability in real-world conditions. In this work, we propose a novel and robust framework that tightly integrates Large Language Model (LLM)-driven symbolic planning with hierarchical imitation learning to enable reliable and generalizable solutions for deformable object assembly. Our approach leverages the advanced reasoning capabilities of LLMs to translate natural language task instructions into structured symbolic task plans. This decomposition is initiated by a visual-language model (VLM) that generates descriptive subgoals from key visual frames of a human demonstration. Each subgoal is then automatically grounded in the robot's perception via a VLM query mechanism, ensuring precise and task-relevant state estimation. For execution, a 3D diffusion policy (DP3) conditioned on visual input and symbolic subgoals generates smooth, low-level action trajectories, bridging the gap between high-level symbolic reasoning and dexterous manipulation. We validate our hierarchical framework on a real-world round belt drive assembly benchmark, demonstrating significant improvements in success rates, error recovery, and generalization across diverse and perturbed initial conditions, compared to existing approaches. Our results highlight the potential of integrating LLM-based symbolic abstraction, targeted state querying, and diffusion-based visuomotor control for robust, autonomous assembly of deformable objects in unstructured environments.

1. Introduction

Deformable object manipulation [1–3] is a cornerstone challenge in robotics, crucial for diverse applications from surgical assistance [4] and textile handling [5] to domestic chores and industrial assembly [6,7]. Unlike rigid bodies, deformable objects possess an infinite-dimensional state space, making their perception, modeling, and control inherently complex [8]. This complexity is compounded in long-horizon assembly tasks, where robots must execute a sequence of interdependent manipulation steps, often requiring precise control and strategic deformation to achieve the desired final configuration. As shown in Fig. 1, we focus on the round belt assembly task, a representative manufacturing task that highlights the unique challenges of deformable object manipulation in real-world industrial contexts [9]. In this task, the robot must grasp, manipulate, and sequentially mount a flexible polyurethane belt onto a series of pulleys and idlers, requiring not only accurate perception and dexterous handling of the belt's shape, but also robust planning across multiple assembly steps. The round belt assembly exemplifies the demands of long-horizon, high-precision manipulation where deformation and environmental variation must be continuously managed for successful task completion. Current robotic systems [10] typically struggle with such tasks, limited by their inability to generalize to novel object geometries, adapt to unforeseen environmental variations, or effectively plan over extended operational sequences.

Recent advancements [11–14] in Large Language Models (LLMs) have unlocked unprecedented capabilities in high-level reasoning, symbolic planning, and human–robot interaction. LLMs can interpret natural language instructions, decompose complex goals into actionable

☆ This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62403211.
* Corresponding authors.
E-mail addresses: zzqing@connect.hku.hk (Z. Zhang), pzhou@gbu.edu.cn (P. Zhou).
[1] Equal contribution.

https://doi.org/10.1016/j.rcim.2025.103096
Received 20 June 2025; Received in revised form 15 July 2025; Accepted 16 July 2025
Available online 28 July 2025

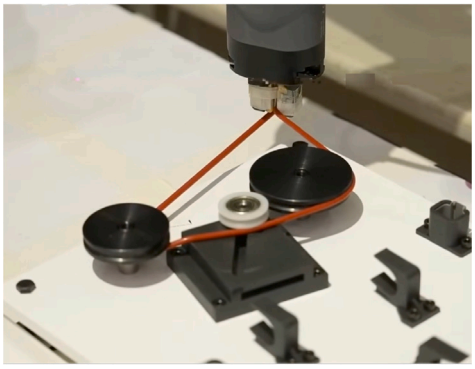


Fig. 1. The round belt assembly task. This task exemplifies the core challenges of deformable object manipulation, requiring precise control and strategic deformation to manage the belt's infinite-dimensional state space during the assembly sequence.

sub-tasks, and even generate code or logical sequences, offering a powerful avenue for abstract task planning in robotics [15–17]. Simultaneously, imitation learning has emerged as a promising paradigm for robotic control, enabling robots to acquire complex behaviors directly from human demonstrations without explicit analytical modeling. While powerful for learning specific skills, traditional imitation learning often struggles with long-horizon tasks due to compounding errors and a lack of hierarchical structure for handling sequential decision-making [18,19].

Manipulating deformable objects like belts presents a significant challenge for robotic systems. The complexity arises not only from their infinite-dimensional state space but also from the need to achieve specific, semantically meaningful configurations—for instance, ensuring a belt is 'properly seated' within a pulley's groove rather than just 'near' it. This requires a perception system that can move beyond simple geometric tracking to understand task-relevant states. Traditional methods, which often rely on complex physical models or hand-engineered visual features, struggle to robustly interpret these nuanced states, especially when faced with environmental variations like changes in lighting or object starting positions. Our framework directly addresses this perception gap by leveraging a Visual-Language Model (VLM). The VLM acts as a 'semantic sensor', capable of answering targeted natural language queries about the scene (e.g., 'Is the belt correctly seated on the first pulley?'). This ability to translate ambiguous visual information into clear, actionable state knowledge is how our system manages complex object deformations and adapts to real-time conditions, providing a level of robustness that is difficult to achieve with conventional approaches.

To address the challenges of long-horizon deformable object assembly, this paper introduces a novel integrated framework that unites the high-level cognitive abilities of large language models (LLMs) with the precise and adaptable control of a hierarchical imitation learning architecture. At the core of our approach, an LLM-driven symbolic planner functions as the high-level "brain", transforming abstract task instructions into a structured sequence of symbolic sub-goals. These sub-goals then direct a hierarchical imitation learning system, in which high-level policies choose appropriate manipulation strategies, while low-level policies execute fine-grained actions to achieve the desired deformations and placements. Crucially, our framework employs a 3D Diffusion Policy (DP3) for robust and generalizable low-level visuomotor control. DP3 leverages point cloud observations to generate smooth, adaptable action trajectories, enabling the robot to manipulate deformable objects with precision even in complex, variable environments. By tightly coupling LLM-driven symbolic planning with DP3-based continuous control, our system not only comprehends and plans for challenging deformable assembly tasks, but also executes them robustly across extended horizons—adapting to real-world variations and reliably recovering from minor errors and disturbances.

We validate our framework through extensive experiments on various challenging long-horizon deformable object assembly scenarios. Our results demonstrate that this integrated approach significantly improves success rates, reduces task completion times, and exhibits superior generalization capabilities compared to existing methods. By bridging the gap between high-level symbolic reasoning and low-level continuous control, our work represents a significant step towards enabling more intelligent, adaptable, and autonomous robotic systems for complex deformable object manipulation in real-world applications.

While our framework integrates several powerful pre-trained models, its core novelty lies not in the individual components, but in their synergistic integration to form a robust, closed-loop system for long-horizon deformable object manipulation. Our primary scientific contribution is the introduction of a goal-oriented state query mechanism where the LLM planner actively generates natural language questions to resolve its uncertainty, and the VLM provides grounded, semantic answers by interpreting the visual scene. This LLM–VLM interaction creates a "semantic sensor" that replaces brittle, hand-engineered state estimators and allows the system to precisely track its progress against a symbolic plan. This tight coupling between high-level symbolic reasoning and grounded visual perception is what enables the framework's superior error recovery and adaptability, marking a significant departure from open-loop planners or methods that rely on less expressive state representations.

The remainder of this paper is organized as follows: Section 2 provides a detailed overview of related work in deformable object manipulation, foundation models for task planning, and visual imitation learning. Section 3 presents the problem statement and system description. Section 4 introduces our integrated framework, describing the hierarchical combination of LLM-driven symbolic planning and imitation learning, including subgoal generation, the goal-oriented state query module, and low-level action generation via diffusion policy. Section 5 presents the experimental setup, evaluation protocol, and a comprehensive analysis of results. Finally, Section 6 concludes the paper and discusses future research directions.

2. Related work

2.1. Pre-trained foundation models for task planning

Early research efforts [20–22] have shown that large language models (LLMs) possess strong capabilities for tackling long-horizon planning problems, functioning effectively as high-level task planners. In contrast, traditional symbolic planning methods, such as those based on PDDL, rely on efficient search techniques to discover valid or optimal plans, but their flexibility is limited by the need for predefined, structured task specifications. More recent work [23,24] has sought to bridge this gap by combining LLMs with classical planners: LLMs are used to generate structured task descriptions, which are then input into PDDL-based planners to search for optimal solutions.

The integration of multimodal data has further expanded the scope of planning research. Unlike traditional methods that often rely on hand-engineered features or state estimators which can be brittle in the face of visual variations, several recent studies have utilized visual language models (VLMs) [25–28]. These models leverage strong reasoning and in-context learning abilities to extract complex spatial relationships and object affordances directly from raw sensory input for planning [29–32]. For instance, approaches like VoxPoser [12] and Instruct2Act [27] have shown that VLMs can interpret ambiguous natural language commands and ground them in complex 3D environments, a task that remains challenging for classical planners. Despite these advances, VLMs are still often limited in their ability to capture the fine-grained spatial details crucial for high-precision manipulation, such as the exact position and orientation of objects. Rather than relying solely on structured priors, our approach addresses this by using VLMs for high-level semantic understanding while employing an explicit grounding mechanism for task-relevant state verification—such as confirming whether a generated grasp pose is physically feasible for the robot to execute.

2.2. Deformable object manipulation

Robotic manipulation of deformable objects (DOs) is incredibly challenging due to their complex, infinite-dimensional states [33–36]. While crucial for many applications like assembly and surgery, traditional approaches often struggle. Early model-based methods use precise physical models (e.g., FEM) for planning [37,38]. However, these are often computationally expensive and difficult to accurately build for real-world scenarios. Learning-based approaches like Reinforcement Learning (RL) and Imitation Learning (IL) bypass explicit modeling [39,40]. RL can learn complex behaviors but is sample-inefficient and requires extensive reward engineering. IL, especially through behavior cloning, allows learning from demonstrations. While effective for short, specific tasks, scaling traditional IL to long-horizon deformable assembly with diverse goals is difficult due to compounding errors and a lack of high-level planning. Some hierarchical imitation learning attempts have been made, but they often lack robust abstract task representation. Despite advances in hybrid methods, robustly handling multi-step, long-horizon deformable object assembly—especially when strategic deformation is integral to the task—remains a significant open problem. Our work directly addresses this gap by integrating high-level intent with low-level execution.

2.3. Visual imitation learning

Imitation learning has become a prominent approach for enabling robots to acquire human-like abilities, leveraging large datasets of observation-action pairs collected from expert demonstrations. Due to the inherent difficulties in precisely estimating object states in real-world settings, visual modalities—particularly images—have become a reliable source of information. Although most existing methods have focused on 2D image-based policies [41–43], there is growing recognition of the value of 3D perception for robotic learning [44,45].

A new wave of 3D policy architectures, such as PerAct [46], RVT [47], ACT3D [44], and NeRFuser [48], have demonstrated strong performance in tasks characterized by low-dimensional control. Despite these advances, current 3D imitation learning frameworks face notable limitations. Firstly, many approaches reformulate the learning problem as one of prediction and planning, often extracting keyframe poses—a strategy that is less effective when extended to more complex, high-dimensional tasks. Secondly, the computational overhead of these models leads to slow inference speeds. For example, PerAct operates at 2.23 frames per second, which is insufficient for tasks that require rapid or continuous control in dynamic environments. Similarly, 3D Diffuser Actor [49] achieves only 1.67 FPS, primarily due to the overhead from attention mechanisms and differences in experimental setup.

In light of these challenges, our work aims to develop a versatile and efficient 3D imitation learning policy. The goal is to support a wide range of robotic tasks, seamlessly scaling from high-dimensional to low-dimensional control scenarios, while maintaining real-time performance.

3. Problem statement

This work addresses the challenge of enabling a single-arm robotic manipulator to autonomously complete long-horizon deformable object assembly tasks for a round belt drive system, guided by human language instructions. Specifically, given a human demonstration trajectory D and a visual observation I (such as an image or point cloud), the objective is to generate a sequence of action primitives $a_1, a_2, \ldots, a_n$ based on the current visual input I_t such that the robot autonomously completes the specified manipulation task. At each time step, the system infers the current state s^t from the available observations and, leveraging both learned policies and symbolic planning, decomposes the long-horizon task into a sequence of interpretable subgoals. Each action primitive a_i in the sequence is selected and executed to incrementally satisfy these subgoals g, ultimately achieving the overall assembly objective.

We formalize this as a symbolic planning problem, defined by the tuple $(\mathcal{S}, s^t, g, \mathcal{A}, \gamma)$:

- $\mathcal{S}$: The symbolic state space, representing a discrete set of world states. Each state $s \in \mathcal{S}$ is described by a set of predicates that capture the properties and relationships of objects in the environment.
- $\mathcal{A}$: The set of symbolic actions, where each action can take object instances as symbolic parameters and modify their properties accordingly.
- γ: The symbolic state transition function, which maps the current state and an action to the next state.
- $s^t \in \mathcal{S}$: The current symbolic state of the environment, inferred from perception and demonstration.
- $g_n \subset \mathcal{S}$: The set of subgoal symbolic states, representing the desired outcome of the assembly process.

The planning task is described in PDDL (Planning Domain Definition Language), where the PDDL problem specifies the initial and goal states, and the PDDL domain defines the object types, predicates, symbolic actions, and transition dynamics. Each symbolic action in PDDL includes its parameters, preconditions, and effects.

In our framework, both the current state s^t and the goal conditions g are inferred from the human tele-operated demonstration D and visual perception I. While the PDDL domain description is assumed to be predefined by a human expert, the reasoning over D and I enables automatic extraction of task-relevant state representations and goals. Once the symbolic initial state, goal state, and transition function γ are specified, a symbolic planner is employed to compute the action sequence $(a_1, a_2, \ldots, a_n)$ required to complete the round belt assembly task.

4. Methodology

Our proposed framework addresses long-horizon deformable object assembly through a hierarchical integration of symbolic planning and imitation learning. As shown in Fig. 2 at the high level, we leverage human demonstrations, visual-language models (VLMs), and large language models (LLMs) to decompose complex assembly tasks into a sequence of interpretable symbolic subgoals. These subgoals are then mapped to executable robot actions via a low-level policy. For robust action generation, we employ a 3D Diffusion Policy (DP3) that conditions on both visual observations and symbolic subgoals to produce smooth and generalizable trajectories in complex manipulation scenarios. This tight coupling between high-level reasoning and low-level control enables scalable, interpretable, and data-efficient solutions for long-horizon, deformable object assembly tasks.

4.1. Subgoal generation

To facilitate efficient long-horizon manipulation, we decompose the assembly task into a sequence of subgoals (see Table 1) via a combination of human demonstration, vision-language reasoning, and language model abstraction. First, a human tele-operated demonstration is performed for the round belt drive system assembly. Leveraging human knowledge and prior experience, we predefine a sequence of key image frames $\{I_1^*, I_2^*, \ldots, I_N^*\}$, each capturing a crucial step or transition in the manipulation process:

$$I_1^*, I_2^*, \ldots, I_N^* = f_{\text{prior}}(D) \tag{1}$$

where D denotes the demonstration trajectory and f_{prior} represents the extraction of subgoal-related frames based on human priors.

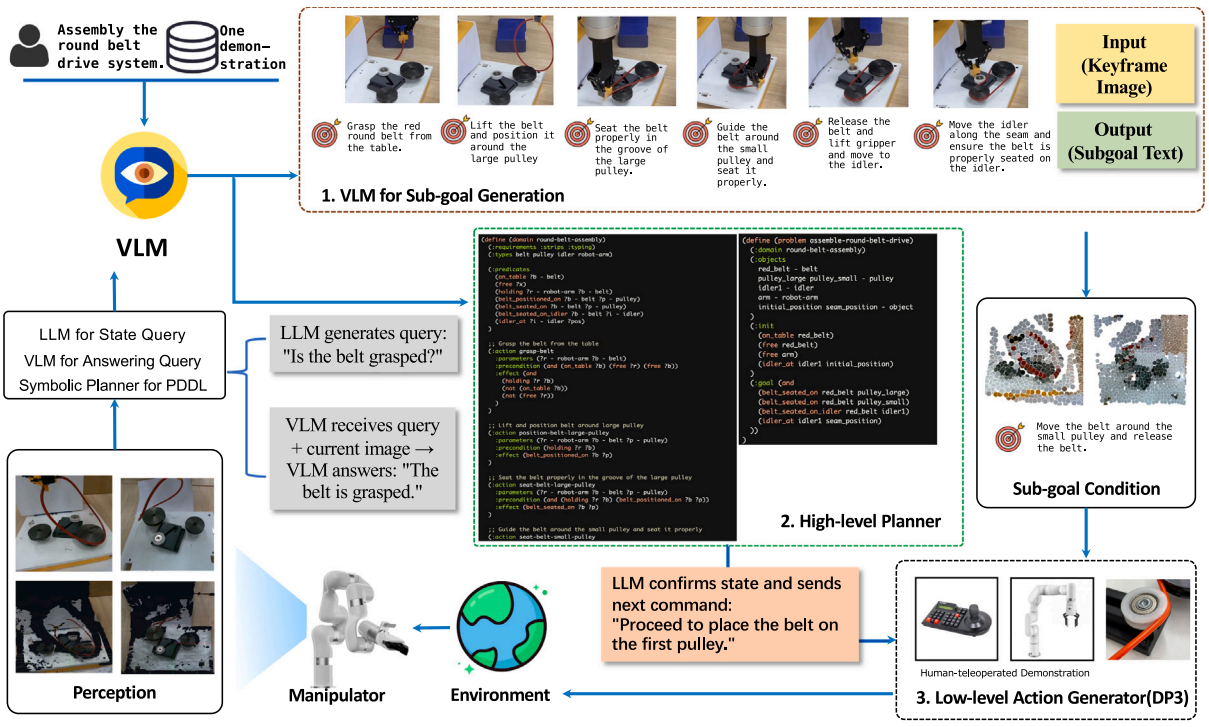


Fig. 2. Our Hierarchical Framework for Deformable Object Assembly, guided by LLMs and Imitation Learning. A human demonstration provides initial input. The framework then employs VLMs for sub-goal generation and LLM–VLM interaction for state querying and answering. A symbolic planner, utilizing PDDL, creates high-level task plans. Finally, a low-level action generator (DP3) executes the detailed manipulation actions, allowing the robot to assemble the round belt drive system.

Next, we employ a Visual-Language Model (VLM) to infer descriptive subgoal texts $\{\tilde{g}_1, \tilde{g}_2, \ldots, \tilde{g}_N\}$ from these key frames:

$$\tilde{g}_1, \tilde{g}_2, \ldots, \tilde{g}_N = f_{\text{VLM}}(I_1^*, I_2^*, \ldots, I_N^*) \tag{2}$$

The VLM bridges visual perception and natural language, providing concise subgoal descriptions that correspond to each key step in the demonstration. Subsequently, we utilize a Large Language Model (LLM) to convert each subgoal description into a symbolic subgoal state $\{g_1, g_2, \ldots, g_N\}$:

$$g_1, g_2, \ldots, g_N = f_{\text{LLM}}(\tilde{g}_1, \tilde{g}_2, \ldots, \tilde{g}_N) \tag{3}$$

Here, the LLM leverages *in-context learning* (ICL)—a powerful capability that enables the model to perform new tasks by conditioning on a prompt containing a handful of instruction-subgoal decomposition examples, without requiring any additional training or parameter updates. By providing a carefully constructed prompt with several annotated demonstrations, the LLM dynamically learns to generalize the subgoal abstraction process to novel instructions and scenarios.

By this hierarchical process, the high-level assembly task is modularized into a sequence of symbolic subgoal states, each representing a key transitional condition in the environment. This decomposition simplifies planning and enables incremental, interpretable robot control for long-horizon assembly.

4.2. Goal-oriented state query module

To acquire the symbolic state s_i^{tl} relevant to each subgoal, we first input the decomposed subgoal into a large language model (LLM). The LLM analyzes the subgoal to determine the essential pieces of information required for its completion. For example, given the subgoal "Lift the belt into the air and position it around the larger pulley", the LLM identifies key checkpoints such as whether the belt is currently grasped and if it is in proper contact with the large pulley. In-context learning (ICL) is applied to guide the LLM in extracting these specific state queries based on provided examples.

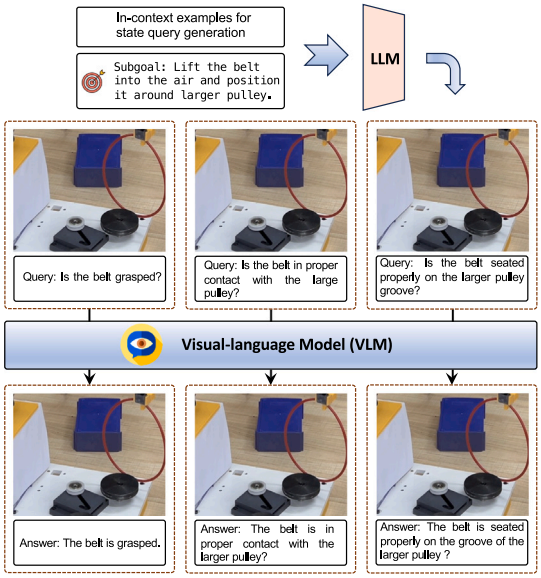


Fig. 3. Illustration of the LLM-driven state query and VLM-based answering mechanism within our framework. For a given sub-goal (e.g., "Lift the belt into the air and position it around larger pulley"), the LLM generates specific queries about the current state (e.g., "Is the belt grasped?", "Is the belt in proper contact with the large pulley?"). A Visual-Language Model (VLM) then processes the visual input to provide accurate answers to these queries (e.g., "The belt is grasped", "The belt is in proper contact with the larger pulley?"), enabling robust state perception for planning and execution.

Once the relevant state queries are formulated by the LLM, a visual-language model (VLM) is employed to ground these queries in the robot's perception. The VLM receives both the linguistic state queries and visual data from the robot's sensors, enabling it to directly assess the status of task-relevant objects and spatial arrangements within the environment (see Fig. 3).

Table 1
Symbolic states for each assembly key step (Subgoal).

Subgoal state	Symbolic state representation

Initial state	`(on_table red_belt)`
	`(free red_belt)`
	`(free arm)`
	`(idler_at idler1 initial_position)`

Grasp the red round belt from the table.	`(holding arm red_belt)`
	`(not (on_table red_belt))`
	`(free red_belt)`

Lift the belt and position it around the large pulley.	`(holding arm red_belt)` `(belt_positioned_on red_belt pulley_large)`
Seat the belt properly in the groove of the large pulley.	`(holding arm red_belt)` `(belt_seated_on red_belt pulley_large)`

Guide the belt around the small pulley and seat it properly.	`(holding arm red_belt)`
	`(belt_seated_on red_belt pulley_large)`
	`(belt_seated_on red_belt pulley_small)`

Release the belt and lift gripper and move to the idler.	`(free arm)`
	`(not (holding arm red_belt))`
	`(belt_seated_on red_belt pulley_large)`
	`(belt_seated_on red_belt pulley_small)`

Move the idler along the seam and ensure the belt is properly seated on the idler.	`(free arm)`
	`(belt_seated_on red_belt pulley_large)`
	`(belt_seated_on red_belt pulley_small)`
	`(belt_seated_on_idler red_belt idler1)`
	`(idler_at idler1 seam_position)`

Rather than requesting broad scene or image descriptions from the VLM, we leverage targeted state queries, as demonstrated in the accompanying figure. For instance, to verify progress on the belt assembly task, the system poses queries such as "Is the belt grasped?" or "Is the belt in proper contact with the large pulley?" This focused querying allows the VLM to deliver concise, task-specific answers, increasing the precision and reliability of the state assessment compared to open-ended descriptions.

To adapt the VLM for this specialized querying, we fine-tune it—using Sphinx as the backbone model—on a dataset generated in the Isaac Gym simulator. This dataset includes diverse scenarios and annotated state queries covering object accessibility, gripper status, and the condition of articulated objects. This enables the model to robustly answer whether objects are reachable, whether grippers are holding objects, and whether components such as pulleys or drawers are in the correct state, across a variety of assembly and manipulation contexts.

4.3. Low-level action generation via diffusion policy 3D

With a sequence of symbolic subgoals $\{g_1, g_2, \ldots, g_N\}$ specified by the high-level planner, we must ground each subgoal in continuous, executable robot actions. To achieve robust and generalizable low-level control, we employ a 3D Diffusion Policy (DP3) as our visuomotor policy, leveraging a small set of expert demonstrations that capture complex skill trajectories in deformable object assembly (see Fig. 4).

DP3 consists of two key modules: **perception** and **decision**. In the perception module, the environment is observed through point cloud data acquired from different view depth cameras, which is processed into compact 3D visual features. Point clouds are generated by converting depth images into 3D coordinates using camera intrinsics and extrinsics. To focus on task-relevant information and improve efficiency, points outside a bounding box around the workspace are cropped, and further downsampled via farthest point sampling (FPS). The resulting subset (typically 1024 points) is encoded into a 64-dimensional vector o using a lightweight MLP-based encoder with max-pooling and LayerNorm for stability.

The decision module of DP3 is a conditional denoising diffusion model that generates action sequences conditioned on a symbolic subgoal g_n, current 3D visual feature o_t, the current robot pose q_t. The process starts from a random Gaussian noise vector a^K and iteratively denoises it over K steps to produce a noise-free action a^0, following:

$$a^{k-1} = \alpha_k \left(a^k - \gamma_k \, \epsilon_\theta(a^k, k, g_n, o_t, q_t)\right) + \sigma_k \mathcal{N}(0, \mathbf{I}), \tag{4}$$

where ϵ_θ is the denoising network, and $\alpha_k, \gamma_k, \sigma_k$ are schedule-dependent coefficients. The DP3 policy is trained by minimizing the mean squared error between the predicted and true noise, using a diffusion process applied to expert action data:

$$\mathcal{L} = \mathrm{MSE}\left(\epsilon^k, \epsilon_\theta\left(\overline{\alpha_k} a^0 + \overline{\beta_k}\epsilon^k, k, g_n, o_t, q_t\right)\right), \tag{5}$$

where $\overline{\alpha_k}$ and $\overline{\beta_k}$ define the noise schedule for each diffusion step.

For each symbolic subgoal g_n, the DP3 policy π_{DP3} generates a sequence of low-level actions $\{a_t, a_{t+1}, \ldots, a_{t+K}\}$ conditioned on the current observation o_t and subgoal:

$$a_t, a_{t+1}, \ldots, a_{t+K} = \pi_{\mathrm{DP3}}(o_t, g_n) \tag{6}$$

DP3 demonstrates remarkable generalization from limited expert data, especially in 3D tasks that require precise manipulation in the presence of deformable objects. Its use of point clouds as the primary scene representation allows the policy to generalize beyond the specific training configurations, as evidenced in prior benchmarks such as MetaWorld Reach.

Integration with Symbolic Planning. In our hierarchical framework, the high-level PDDL planner (driven by LLM-generated symbolic subgoals) specifies the sequence of skill primitives and their associated symbolic states. For each subgoal, the DP3 policy is invoked to realize the required manipulation at the action level, ensuring smooth, temporally coherent, and robust execution of complex assembly steps. This tight coupling between symbolic reasoning and diffusion-based continuous control enables our system to tackle long-horizon deformable object assembly with both high-level interpretability and low-level dexterity.

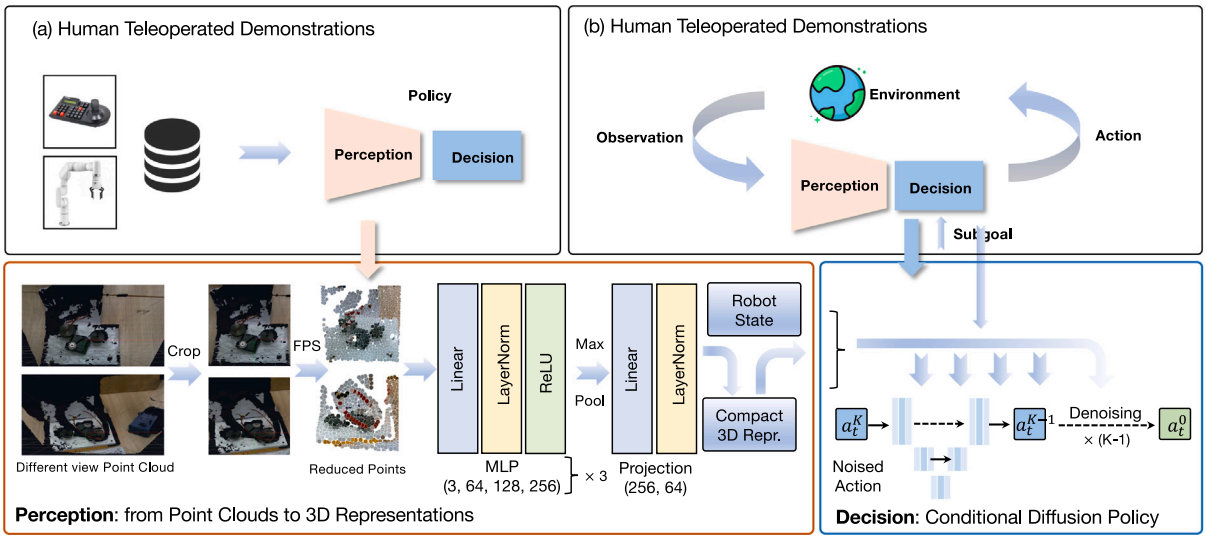


Fig. 4. The learning pipeline for low-level action generation. Human teleoperated demonstrations (a) provide the foundational data. The system's perception module (b) processes multi-view point cloud data, reducing it to a compact 3D representation via layers including MLP, ReLU, and Max Pool. This compact representation, combined with robot state and environment information, forms the input for the decision module. The decision module employs a Conditional Diffusion Policy (b)[cite: 2], which iteratively refines a noised action (a_t^K) through a denoising process to arrive at the final robot action (a_t^0)[cite: 2], effectively translating perceived states and sub-goals into executable movements.

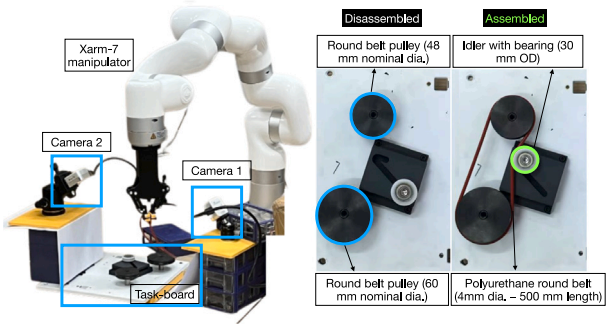


Fig. 5. Experimental setup for deformable object assembly tasks. The setup includes a **Xarm-7 manipulator**, a **Task-board** workspace, and three **Cameras (Front view, side view1 and side view2)** for visual feedback. The task involves assembling a **polyurethane round belt** onto **round belt pulleys** and an **idler with bearing**, shown in both disassembled and assembled configurations.

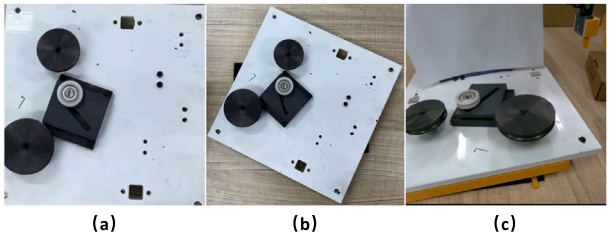


Fig. 6. Three different initial task board configurations for the experiments: (a) the task board is in the standard orientation; (b) the task board is rotated 30 degrees clockwise; (c) one side of the task board is raised by 1 cm.

5. Experiment

5.1. Experimental setup

We evaluate our proposed framework on a physical robotic platform tailored for deformable object assembly, as illustrated in Fig. 5. The system comprises an Xarm-7 manipulator equipped with a two-finger gripper, enabling precise and dexterous manipulation. The workspace is defined by a custom Task-board, which incorporates all necessary assembly components. To facilitate robust perception, three cameras (Front Camera, Side Camera 1, and Side Camera 2) are strategically arranged to provide comprehensive visual coverage for both the perception modules and experiment monitoring. The primary assembly task involves mounting a polyurethane round belt (4 mm diameter, 500 mm length) onto a system consisting of two round belt pulleys (48 mm and 60 mm nominal diameters) and an idler with bearing (30 mm OD). Fig. 5 depicts both the initial (disassembled) and final (assembled) states, reflecting the long-horizon and high-precision requirements of deformable object assembly.

5.2. Evaluation protocol

To systematically assess the robustness and generalizability of our framework, we consider three distinct initial task board configurations, as shown in Fig. 6: (a) standard orientation (STD), (b) rotated 30 degrees clockwise (ROT30), and (c) one side elevated by 1 cm (TILT1). Both assembly and disassembly tasks are evaluated under each condition. For each trial, the robot must complete the full assembly (or disassembly) sequence without human intervention. Additionally, in challenging cases (case 4 and 5), we introduce random human perturbations by removing the red belt from the gripper during execution to evaluate the system's error recovery ability.

We adopt the following quantitative metrics for evaluation:

- **Success Rate**: The proportion of trials in which the task is completed successfully.
- **Completion Time**: The average time taken to complete the assembly or disassembly task.
- **Error Recovery Rate**: For cases involving external disturbances, the percentage of trials in which the system successfully recovers and completes the task after the belt is removed from the gripper.

To demonstrate the effectiveness of our method, we compare the following approaches:

- **Ours (LLM-driven symbolic planning + DP3 diffusion policy)**: Our full hierarchical framework combining LLM-based high-level planning with diffusion-policy-based low-level control.
- **LLM symbolic planning + DP3**: Hierarchical planning using LLM with DP3, but without the state query mechanism.
- **Flat Diffusion Policy (No Hierarchy)**: End-to-end DP3 policy without subgoal decomposition or symbolic reasoning.

Table 2
Performance of different approaches on different initial settings for assembly tasks of round belt drive system.

Method	Success rate (%)			Completion time (s)		
	STD	ROT30	TILT1	STD	ROT30	TILT1
IBC-3D	13.3 ± 3.3	0.0 ± 0.0	2.2 ± 1.9	110.7 ± 7.3	/	103.8 ± 8.2
BCRNN-3D	16.7 ± 3.3	4.4 ± 1.9	11.1 ± 3.9	115.2 ± 6.5	123.9 ± 7.8	108.4 ± 7.1
Flat-DP3	24.4 ± 1.9	8.9 ± 1.9	10.0 ± 5.8	128.8 ± 4.7	135.6 ± 5.4	120.3 ± 5.9
LLM+DP3	63.3 ± 12.0	52.2 ± 17.0	51.1 ± 11.7	145.4 ± 3.8	160.6 ± 4.4	130.7 ± 4.3
LLM+DP3+Query (Ours)	**73.3** ± 3.3	**67.8** ± 1.9	**63.3** ± 3.3	**160.9** ± 1.6	**180.2** ± 1.3	**145.8** ± 1.9

Table 3
Performance of different approaches on different initial settings for disassembly tasks of round belt drive system.

Method	Success rate (%)			Completion time (s)		
	STD	ROT30	TILT1	STD	ROT30	TILT1
IBC-3D	15.6 ± 1.9	1.1 ± 1.8	3.3 ± 3.3	62.4 ± 2.0	65.4 ± 2.1	59.3 ± 2.1
BCRNN-3D	24.4 ± 1.9	2.2 ± 3.9	14.4 ± 3.9	67.5 ± 2.5	71.1 ± 2.6	63.6 ± 2.2
Flat-DP3	31.1 ± 6.9	10.0 ± 3.3	24.4 ± 1.9	75.2 ± 1.7	80.1 ± 2.0	70.5 ± 1.5
LLM+DP3	62.2 ± 10.2	50.0 ± 15.3	61.1 ± 8.4	84.7 ± 1.3	89.2 ± 1.4	81.6 ± 1.2
LLM+DP3+Query (Ours)	**84.4** ± 5.1	**72.2** ± 5.1	**68.9** ± 1.9	**90.2** ± 0.8	**94.3** ± 0.9	**87.8** ± 0.7

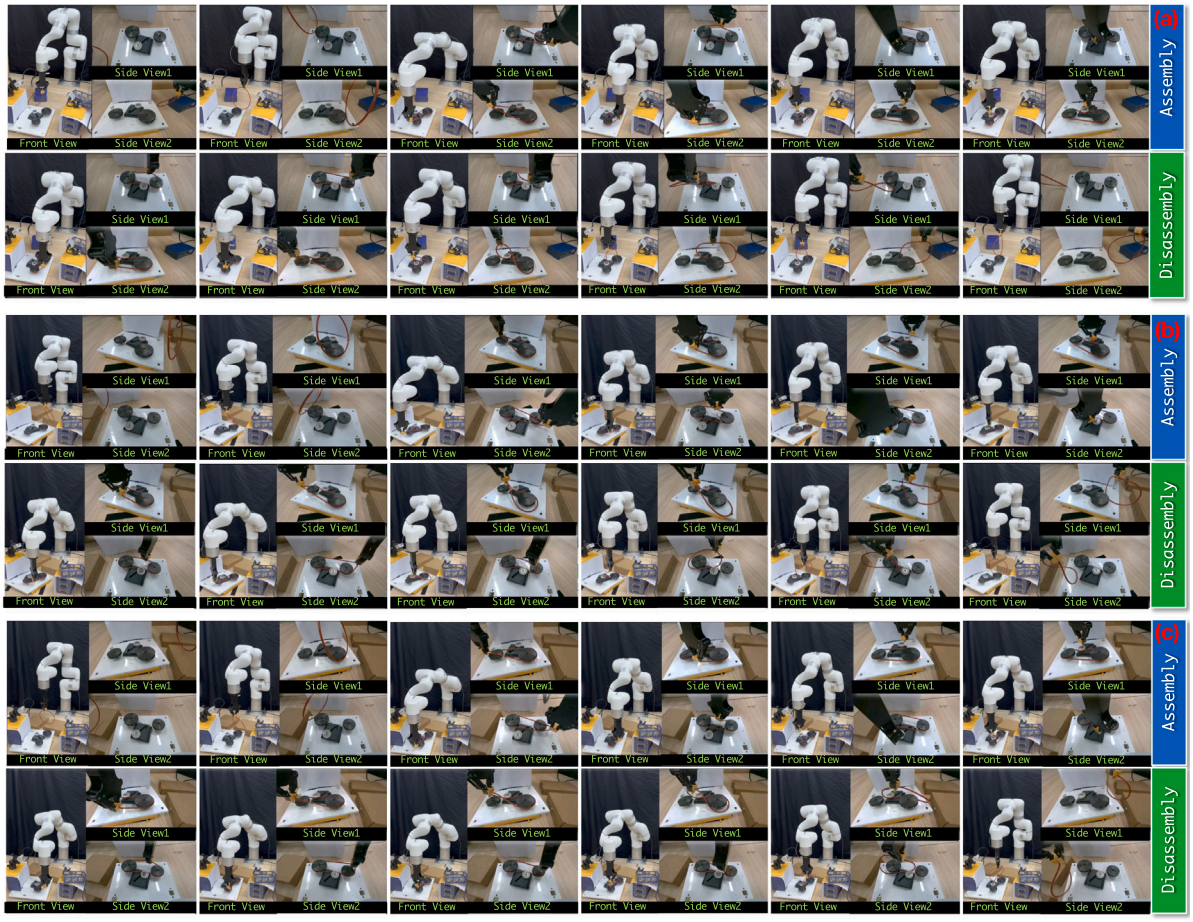


Fig. 7. Qualitative results for assembly and disassembly tasks of the round belt drive system using our **LLM+DP3+Query** approach under different initial experimental settings: (a) standard orientation (STD), (b) rotated 30° clockwise (ROT30), and (c) one side elevated by 1 cm (TILT1).

- **IBC** [41] + **3D**: Baseline based on Implicit Behavior Cloning in a 3D visuomotor policy setting.
- **BCRNN** [50] + **3D**: Baseline using a Bidirectional Convolutional Recurrent Neural Network for low-level control.

5.3. Results and analysis

Assembly tasks. Table 2 reports the performance of all methods under three initial settings (a) standard orientation (STD), (b) rotated 30° clockwise (ROT30), and (c) one side elevated by 1 cm (TILT1) for the assembly tasks, and Fig. 7 shows qualitative results for assembly and disassembly tasks of the round belt drive system using our **LLM+DP3+Query** approach under different initial experimental settings: (STD, ROT30, TILT1). Our method consistently achieves the highest success rates across all settings, reaching 73.3 ± 3.3% (STD), 67.8 ± 1.9% (ROT30), and 63.3 ± 3.3% (TILT1). The strongest baseline, LLM+DP3, achieves lower rates in all cases (e.g., 63.3 ± 12.0% for STD, 52.2 ± 17.0% for ROT30). Notably, classical learning-based methods

Fig. 8. Experimental results of error recovery in assembly tasks for the round belt drive system using our **LLM+DP3+Query** approach. The figure illustrates cases where, following external disturbances that remove the belt from the gripper, the system successfully recovers and completes the assembly task.

Table 4
Error recovery rate (%) of Different Approaches on Standard (STD) Initial Setting for assembly and disassembly tasks.

Method	Error recovery rate (%)	
	Assembly	Disassembly
IBC-3D	5.0	7.0
BCRNN-3D	36.0	38.0
Flat-DP3	52.0	56.0
LLM+DP3	73.0	78.0
LLM+DP3+Query (Ours)	**89.0**	**92.0**

(IBC-3D and BCRNN-3D) perform much worse, especially under non-standard settings, with IBC-3D even failing completely in the ROT30 scenario.

In terms of completion time, our method takes moderately longer to finish each task (e.g., 160.9 ± 1.6 s for STD), which is expected given the additional reasoning and querying steps involved. However, this moderate increase is justified by the substantial boost in success rates and robustness. Across all methods, completion time follows the trend: TILT1 < STD < ROT30, indicating that rotational disturbances (ROT30) are more challenging than minor tilts (TILT1). Flat-DP3 provides moderate success rates, but consistently trails behind the LLM-based approaches, highlighting the importance of high-level language reasoning and hierarchical planning.

Disassembly tasks. Table 3 shows that our method again outperforms all baselines for disassembly, achieving 84.4 ± 5.1% (STD), 72.2 ± 5.1% (ROT30), and 68.9 ± 1.9% (TILT1) success rates. LLM+DP3 remains the strongest baseline, but the gap between our method and all baselines widens further in more challenging (ROT30, TILT1) scenarios.

Our approach also exhibits slightly higher completion times (e.g., 90.2 ± 0.8 s for STD), again reflecting the overhead of more robust reasoning and error checking. Similar to assembly, all methods complete TILT1 scenarios faster than STD and ROT30, confirming that moderate tilts are less disruptive than large rotations. Classical baselines (IBC-3D, BCRNN-3D, Flat-DP3) show consistently lower success rates and are less robust to initial configuration disturbances.

Error recovery. Table 4 summarizes the error recovery rates under the standard initial setting for both assembly and disassembly. Our method achieves the highest recovery rates in both tasks (89% for assembly and 92% for disassembly), demonstrating strong robustness to execution errors. Fig. 8 shows corresponding experimental results of error recovery in assembly tasks using our **LLM+DP3+Query** approach. The figure illustrates cases where, following external disturbances that remove the belt from the gripper, the system successfully recovers and completes the assembly task. LLM+DP3 also performs well but is consistently surpassed by our approach. In contrast, prior learning-based methods (IBC-3D, BCRNN-3D, Flat-DP3) exhibit significantly lower recovery rates, with IBC-3D almost always failing to recover.

Overall, these results clearly demonstrate the effectiveness and robustness of our proposed method across both assembly and disassembly tasks, especially under challenging initial conditions. The integration of LLM-based reasoning and dynamic querying substantially improves both task success rates and error recovery, at the cost of only a moderate increase in execution time. This trade-off is highly favorable for real-world applications where reliability and adaptability are essential.

6. Conclusion

This paper introduced a novel framework that effectively tackles long-horizon deformable object assembly by integrating LLM-driven symbolic planning with hierarchical imitation learning. We successfully leveraged LLMs for high-level task understanding and symbolic decomposition, while hierarchical imitation learning provided robust, precise control for complex deformations. Our extensive experiments demonstrated the framework's superior performance in terms of success rates, task completion times, and generalization capabilities across various challenging scenarios. This work represents a significant advance toward more autonomous and intelligent robotic systems for manipulating deformable materials in unstructured environments. Future work will focus on several key directions to build upon this foundation. First, we plan to further optimize the Visual-Language Model (VLM) by fine-tuning it on larger and more varied datasets of manipulation scenarios. This will enhance its ability to answer state queries with higher precision, which is crucial for even more complex belt manipulation tasks. Second, we will work to extend the framework's applicability beyond linear objects. A critical next step is to test its generalization on different categories of deformable objects, such as planar items (e.g., textiles) and volumetric objects (e.g., bags), which present unique manipulation challenges. Finally, to operate robustly in dynamic, real-world settings, we will focus on integrating deeper, real-time visual feedback loops that can trigger dynamic re-planning. This would allow the system to intelligently adjust its symbolic plan and actions in response to unforeseen changes or perturbations, aiming for truly versatile and adaptive robotic manipulation.

CRediT authorship contribution statement

Jiaming Qi: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Liang Lu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Conceptualization. **Fangyuan Wang:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Hoi-Yin Lee:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **David Navarro-Alarcon:** Writing – review & editing, Supervision, Methodology. **Zeqing Zhang:** Writing – review & editing, Supervision, Investigation. **Peng Zhou:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.rcim.2025.103096.

Data availability

No data was used for the research described in the article.

References

[1] P. Jiménez, Survey on model-based manipulation planning of deformable objects, Robot. Comput.-Integr. Manuf. 28 (2) (2012) 154–163.

[2] A. Monguzzi, T. Dotti, L. Fattorelli, A.M. Zanchettin, P. Rocco, Optimal model-based path planning for the robotic manipulation of deformable linear objects, Robot. Comput.-Integr. Manuf. 92 (2025) 102891.

[3] H. Yin, A. Varava, D. Kragic, Modeling, learning, perception, and control methods for deformable object manipulation, Sci. Robot. 6 (54) (2021) eabd8803.

[4] D. Navarro-Alarcon, Y.-h. Liu, J.G. Romero, P. Li, On the visual deformation servoing of compliant objects: Uncalibrated control methods and experiments, Int. J. Robot. Res. 33 (11) (2014) 1462–1480.

[5] Z. Weng, P. Zhou, H. Yin, A. Kravberg, A. Varava, D. Navarro-Alarcon, D. Kragic, Interactive perception for deformable object manipulation, IEEE Robot. Autom. Lett. (2024).

[6] X. Xu, D. Zhu, J. Wang, S. Yan, H. Ding, Calibration and accuracy analysis of robotic belt grinding system using the ruby probe and criteria sphere, Robot. Comput.-Integr. Manuf. 51 (2018) 189–201.

[7] S. Li, P. Zheng, S. Liu, Z. Wang, X.V. Wang, L. Zheng, L. Wang, Proactive human–robot collaboration: Mutual-cognitive, predictable, and self-organising perspectives, Robot. Comput.-Integr. Manuf. 81 (2023) 102510.

[8] J. Borras, G. Alenya, C. Torras, A grasping-centered analysis for cloth manipulation, IEEE Trans. Robot. 36 (3) (2020) 924–936.

[9] S. Liu, L. Wang, X.V. Wang, Sensorless force estimation for industrial robots using disturbance observer and neural learning of friction approximation, Robot. Comput.-Integr. Manuf. 71 (2021) 102168.

[10] Y. Zhang, K. Ding, J. Hui, S. Liu, W. Guo, L. Wang, Skeleton-rgb integrated highly similar human action prediction in human–robot collaborative assembly, Robot. Comput.-Integr. Manuf. 86 (2024) 102659.

[11] Y. Jin, D. Li, J. Shi, P. Hao, F. Sun, J. Zhang, B. Fang, et al., Robotgpt: Robot manipulation learning from chatgpt, IEEE Robot. Autom. Lett. 9 (3) (2024) 2543–2550.

[12] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, L. Fei-Fei, Voxposer: Composable 3d value maps for robotic manipulation with language models, 2023, arXiv preprint arXiv:2307.05973.

[13] B. Gao, J. Fan, P. Zheng, Empower dexterous robotic hand for human-centric smart manufacturing: A perception and skill learning perspective, Robot. Comput.-Integr. Manuf. 93 (2025) 102909.

[14] S. Liu, Z. Liu, L. Wang, X.V. Wang, Vision-language-conditioned learning policy for robotic manipulation.

[15] H.-Y. Lee, P. Zhou, A. Duan, C. Yang, D. Navarro-Alarcon, Iterative shaping of multi-particle aggregates based on action trees and vlm, IEEE Robot. Autom. Lett. 10 (7) (2025) 7102–7109, http://dx.doi.org/10.1109/LRA.2025.3572426.

[16] F. Wang, A. Duan, P. Zhou, S. Huo, G. Guo, C. Yang, D. Navarro-Alarcon, Explicit-implicit subgoal planning for long-horizon tasks with sparse rewards, IEEE Trans. Autom. Sci. Eng. (2025).

[17] F. Wang, S. Lyu, P. Zhou, A. Duan, G. Guo, D. Navarro-Alarcon, Instruction-augmented long-horizon planning: Embedding grounding mechanisms in embodied mobile manipulation, Proc. AAAI Conf. Artif. Intell. 39 (14) (2025) 14690–14698, http://dx.doi.org/10.1609/aaai.v39i14.33610, URL https://ojs.aaai.org/index.php/AAAI/article/view/33610.

[18] A. Hussein, M.M. Gaber, E. Elyan, C. Jayne, Imitation learning: A survey of learning methods, ACM Comput. Surv. 50 (2) (2017) 1–35.

[19] C. Li, P. Zheng, P. Zhou, Y. Yin, C.K. Lee, L. Wang, Unleashing mixed-reality capability in deep reinforcement learning-based robot motion generation towards safe human–robot collaboration, J. Manuf. Syst. 74 (2024) 411–421.

[20] S. Kambhampati, K. Valmeekam, M. Marquez, L. Guan, On the role of large language models in planning, in: Tutorial Presented at the International Conference on Automated Planning and Scheduling, ICAPS, Prague, 2023.

[21] L. Guan, K. Valmeekam, S. Sreedharan, S. Kambhampati, Leveraging pre-trained large language models to construct and utilize world models for model-based task planning, Adv. Neural Inf. Process. Syst. 36 (2023) 79081–79094.

[22] J. Ao, F. Wu, Y. Wu, A. Swikir, S. Haddadin, Llm as bt-planner: Leveraging llms for behavior tree generation in robot task planning, 2024, arXiv preprint arXiv:2409.10444.

[23] B. Liu, Y. Jiang, X. Zhang, Q. Liu, S. Zhang, J. Biswas, P. Stone, Llm+ p: Empowering large language models with optimal planning proficiency, 2023, arXiv preprint arXiv:2304.11477.

[24] T. Silver, V. Hariprasad, R.S. Shuttleworth, N. Kumar, T. Lozano-Pérez, L.P. Kaelbling, Pddl planning with pretrained large language models, in: NeurIPS 2022 Foundation Models for Decision Making Workshop, 2022.

[25] H. Geng, S. Wei, C. Deng, B. Shen, H. Wang, L. Guibas, Sage: Bridging semantic and actionable parts for generalizable manipulation of articulated objects, 2023, arXiv preprint arXiv:2312.01307.

[26] X. Zhang, Z. Altaweel, Y. Hayamizu, Y. Ding, S. Amiri, H. Yang, A. Kaminski, C. Esselink, S. Zhang, Dkprompt: Domain knowledge prompting vision-language models for open-world planning, 2024, arXiv preprint arXiv:2406.17659.

[27] S. Huang, Z. Jiang, H. Dong, Y. Qiao, P. Gao, H. Li, Instruct2act: Mapping multi-modality instructions to robotic actions with large language model, 2023, arXiv preprint arXiv:2305.11176.

[28] D. Song, J. Liang, A. Payandeh, A.H. Raj, X. Xiao, D. Manocha, Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models, IEEE Robot. Autom. Lett. (2024).

[29] S. Akiyama, R.F.J. Dossa, K. Arulkumaran, S. Sujit, E. Johns, Open-loop vlm robot planning: An investigation of fine-tuning and prompt engineering strategies, in: First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024, 2024.

[30] B. Wang, J. Zhang, S. Dong, I. Fang, C. Feng, Vlm see, robot do: Human demo video to robot action plan via vision language model, 2024, arXiv preprint arXiv:2410.08792.

[31] A. Mei, G.-N. Zhu, H. Zhang, Z. Gan, Replanvlm: Replanning robotic tasks with visual language models, IEEE Robot. Autom. Lett. (2024).

[32] J. Gao, B. Sarkar, F. Xia, T. Xiao, J. Wu, B. Ichter, A. Majumdar, D. Sadigh, Physically grounded vision-language models for robotic manipulation, in: 2024 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2024, pp. 12462–12469.

[33] J. Zhu, A. Cherubini, C. Dune, D. Navarro-Alarcon, F. Alambeigi, D. Berenson, F. Ficuciello, K. Harada, J. Kober, X. Li, et al., Challenges and outlook in robotic manipulation of deformable objects, IEEE Robot. Autom. Mag. 29 (3) (2022) 67–77.

[34] Z. Zhang, G. Chen, W. Chen, R. Jia, G. Chen, L. Zhang, J. Pan, P. Zhou, A joint learning of force feedback of robotic manipulation and textual cues for granular materials classification, IEEE Robot. Autom. Lett. (2025).

[35] P. Zhou, P. Zheng, J. Qi, C. Li, H.-Y. Lee, Y. Pan, C. Yang, D. Navarro-Alarcon, J. Pan, Bimanual deformable bag manipulation using a structure-of-interest based neural dynamics model, IEEE/ASME Trans. Mechatronics (2024).

[36] P. Zhou, P. Zheng, J. Qi, C. Li, H.-Y. Lee, A. Duan, L. Lu, Z. Li, L. Hu, D. Navarro-Alarcon, Reactive human–robot collaborative manipulation of deformable linear objects using a new topological latent control model, Robot. Comput.-Integr. Manuf. 88 (2024) 102727.

[37] J. Sanchez, K. Mohy El Dine, J.A. Corrales, B.-C. Bouzgarrou, Y. Mezouar, Blind manipulation of deformable objects based on force sensing and finite element modeling, Front. Robot. AI 7 (2020) 73.

[38] P. Zhou, J. Qi, A. Duan, S. Huo, Z. Wu, D. Navarro-Alarcon, Imitating tool-based garment folding from a single visual observation using hand-object graph dynamics, IEEE Trans. Ind. Inform. 20 (4) (2024) 6245–6256.

[39] J. Matas, S. James, A.J. Davison, Sim-to-real reinforcement learning for deformable object manipulation, in: Conference on Robot Learning, PMLR, 2018, pp. 734–743.

[40] X. Lin, Y. Wang, J. Olkin, D. Held, Softgym: Benchmarking deep reinforcement learning for deformable object manipulation, in: Conference on Robot Learning, PMLR, 2021, pp. 432–448.

[41] P. Florence, C. Lynch, A. Zeng, O.A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, J. Tompson, Implicit behavioral cloning, in: Conference on Robot Learning, PMLR, 2022, pp. 158–168.

[42] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, S. Song, Diffusion policy: Visuomotor policy learning via action diffusion, Int. J. Robot. Res. (2023) 02783649241273668.

[43] S. Haldar, J. Pari, A. Rai, L. Pinto, Teach a robot to fish: Versatile imitation from one minute of demonstrations, in: Robotics: Science and Systems, 2023.

[44] T. Gervet, Z. Xian, N. Gkanatsios, K. Fragkiadaki, Act3d: Infinite resolution action detection transformer for robotic manipulation, 2023, arXiv preprint arXiv:2306.17817. 1 (3).

[45] Y. Ze, G. Yan, Y.-H. Wu, A. Macaluso, Y. Ge, J. Ye, N. Hansen, L.E. Li, X. Wang, Gnfactor: Multi-task real robot learning with generalizable neural feature fields, in: Conference on Robot Learning, PMLR, 2023, pp. 284–301.

[46] M. Shridhar, L. Manuelli, D. Fox, Perceiver-actor: A multi-task transformer for robotic manipulation, in: Conference on Robot Learning, PMLR, 2023, pp. 785–799.

[47] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, D. Fox, Rvt: Robotic view transformer for 3d object manipulation, in: Conference on Robot Learning, PMLR, 2023, pp. 694–710.

[48] G. Yan, Y.-H. Wu, X. Wang, Nerfuser: Diffusion guided multi-task 3d policy learning.
[49] T.-W. Ke, N. Gkanatsios, K. Fragkiadaki, 3D diffuser actor: Policy diffusion with 3d scene representations, 2024, arXiv preprint arXiv:2402.10885.
[50] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, R. Martín-Martín, What matters in learning from offline human demonstrations for robot manipulation, 2021, arXiv preprint arXiv:2108.03298.