LABDEX: A HIERARCHICAL BENCHMARK FOR DEXTEROUS MANIPULATION IN LABORATORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous laboratories hold great promise for accelerating scientific discovery. To achieve this vision, robots are supposed to dexterously manipulate diverse labware and instruments and execute long-horizon, state-dependent experimental procedures. Yet existing benchmarks do not jointly capture dexterous hand use, real-world laboratory interactions, and multi-stage experimental procedures, limiting systematic training and evaluation. To bridge this gap, we introduce **LabDex**, a real-world dataset and benchmark for dexterous manipulation in chemistry laboratories, organized around a hierarchical task taxonomy spanning atomic skills, compositional tasks, and long-horizon experiments. First, LabDex is cross-platform and, for the first time, unifies real-world and simulation platforms under a common framework, providing standardized task definitions, demonstrations, and evaluation protocols. Second, LabDex systematically organizes chemistry laboratory operations into three interconnected levels: *Atomic Skills*, which characterize fundamental dexterous manipulation capabilities; *Compositional Skills*; and *Long-Horizon Laboratory Workflows*. This hierarchical design not only supports the evaluation of end-task performance, but also enables the analysis of how fundamental dexterous skills compose and influence more complex laboratory operations. We conduct cross-level evaluations of representative robot learning methods in both real-world and simulation environments. The experimental results validate the effectiveness of the LabDex task design and demonstration data, and show that the benchmark supports the training and systematic evaluation of existing robotic policies across laboratory dexterous manipulation tasks at different levels, providing a foundation for further research and development of autonomous laboratory robots.

1 INTRODUCTION

Scientific laboratory automation Jiang et al. (2022); Abolhasani & Kumacheva (2023); Wang et al. (2022) requires robots to reliably interact with diverse instruments, containers, and materials. However, the execution of complex experimental procedures remains constrained by the robot's underlying manipulation capabilities. Compared with parallel-jaw grippers, multi-finger dexterous hands provide richer contact configurations and greater manipulation flexibility, making them a promising embodiment for laboratory operations. Despite recent progress in dexterous manipulation, chemical laboratory scenarios remain insufficiently explored in existing datasets and benchmarks.

Chemical laboratory operations naturally exhibit a clear hierarchical structure. A complete experimental procedure can typically be decomposed into fundamental atomic operations with relatively independent objectives, which can be reused across different experimental tasks. Laboratory dexterous manipulation therefore involves not only the execution of individual fundamental operations, but also the ability to compose multiple operations and complete complex experimental procedures. Moreover, evaluating only the final outcome of a complete experimental procedure makes it difficult to distinguish a robot's capabilities at different levels. Hierarchical modeling and evaluation of chemical laboratory operations can therefore provide a more fine-grained characterization of robotic dexterous manipulation capabilities across different levels of task complexity.

Recent efforts have developed robotic manipulation environments, task suites, and demonstration datasets for laboratories, while general-purpose dexterous manipulation benchmarks have expanded

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison with existing datasets and benchmarks.

Dataset / Benchmark	Laboratory Domain	Dexterous Hand	Simulation	Real Robot	Task Diversity		
					Atomic	Compositional	Long-horizon
Bi-DexHandsChen et al. (2022)	✗	✓	✓	✗	✓	✓	✗
DexArtBao et al. (2023)	✗	✓	✓	✗	✓	✗	✗
DexCapWang et al. (2024)	✗	✓	✗	✓	✓	✓	✗
DexVerseYao et al. (2026)	✗	✓	✓	✗	✓	✓	✓
Chemistry3DLi et al. (2024)	✓	✗	✓	✗	✓	✓	✗
AutoBioLan et al. (2025)	✓	✗	✓	✗	✓	✓	✗
LabUtopiaLi et al. (2026)	✓	✗	✓	✗	✓	✓	✓
BioProVLA-AgentDu et al. (2026)	✓	✗	✗	✓	✓	✓	✗
LabDex (Ours)	✓	✓	✓	✓	✓	✓	✓

toward increasingly diverse and complex skills. However, the intersection of laboratory manipulation and multi-finger dexterity remains insufficiently explored. As shown in Table 1, existing benchmarks Li et al. (2024); Lan et al. (2025); Li et al. (2026); Du et al. (2026) primarily employ parallel-jaw grippers or other non-dexterous robotic platforms, providing limited support for evaluating multi-finger manipulation capabilities. Conversely, existing dexterous manipulation benchmarks Chen et al. (2022); Bao et al. (2023); Wang et al. (2024); Yao et al. (2026) mainly focus on generic object interactions rather than laboratory-specific operations and workflows. Moreover, current benchmarks provide limited evaluation across atomic skills, compositional operations, and long-horizon laboratory procedures, making it difficult to characterize dexterous manipulation capabilities at different levels of task complexity.

In this work, we introduce **LabDex**, a hierarchical benchmark for dexterous manipulation in chemistry laboratories. Rather than treating laboratory tasks simply as independent evaluation instances, LabDex organizes laboratory manipulation hierarchically from the perspective of capability composition. Specifically, we divide the benchmark into three interconnected levels: *Atomic Skills*, which characterize reusable, fine-grained dexterous manipulation primitives, such as precise grasping, insertion, and liquid dispensing; *Compositional Skills*, which combine multiple atomic skills into reusable laboratory operations; and *Long-Horizon Laboratory Workflows*, which require robots to sequentially execute multiple compositional skills to complete full laboratory tasks. Built upon unified simulation and real-world robotic platforms, LabDex provides standardized laboratory assets, task definitions, demonstration data, and evaluation protocols across all three levels.

Beyond constructing the dataset and benchmark, we also investigate how robotic dexterous manipulation capabilities scale from fundamental skills to long-horizon laboratory operations. Based on the proposed hierarchical evaluation framework, we systematically evaluate representative machine learning methods across all task levels. More importantly, we analyze the relationships among atomic skills, compositional skills, and complete laboratory workflows, revealing how failures propagate across different task levels and identifying the key capability bottlenecks that limit long-horizon laboratory manipulation. We envision LabDex not only as a standardized benchmark for laboratory dexterous manipulation, but also as a capability-centered evaluation framework that facilitates future research on autonomous laboratory robots.

Our contributions are summarized as follows:

- We introduce **LabDex**, a dataset and benchmark for laboratory dexterous manipulation that organizes operations into three hierarchical task levels: *Atomic Skills*, *Compositional Skills*, and *Long-Horizon Laboratory Workflows*.
- We develop a comprehensive and unified real-world and simulation evaluation platform comprising standardized laboratory assets, hierarchical task definitions, demonstration datasets, and evaluation protocols for laboratory dexterous manipulation.
- We conduct comprehensive evaluations of representative robot learning methods and, for the first time, systematically analyze how atomic dexterous capabilities compose into long-horizon laboratory manipulation, providing new insights into the capability bottlenecks of autonomous laboratory robots.

2 Related Work

2.1 Robotic Manipulation for Laboratories

Automated laboratories integrate machine learning, robotics, and modular platforms to improve experimental efficiency, and accelerate scientific discovery Jiang et al. (2022); Abolhasani & Kumacheva (2023). Recent work increasingly integrates cognitive reasoning with embodied AI to enhance system autonomy and adaptivity Zhang et al. (2025); Boiko et al. (2023). Representative systems include the Chemputer, which executes complex organic syntheses on modular hardware using a chemical programming language Steiner et al. (2019); ChemCrow, which equips LLMs with expert-designed chemistry tools for planning and executing multi-step synthesis tasks M. Bran et al. (2024); and autonomous mobile robots that conduct exploratory synthetic chemistry using standard laboratory instruments Dai et al. (2024). However, existing systems often depend on predefined protocols, bespoke hardware, fixed interfaces, and task-specific workstations, limiting their flexibility, generalizability, and scalability across laboratory operations.

2.2 Laboratory Manipulation Benchmarks

Laboratory benchmarks require standardized tasks, reproducible settings, and comparable evaluation protocols. Chemistry3D provides a simulation toolkit with real-time visualization of chemical reactions Li et al. (2024). LabUtopia combines multi-physics simulation, procedural laboratory generation, and hierarchical tasks ranging from atomic actions to long-horizon mobile manipulation Li et al. (2026). AutoBio digitizes biological instruments with dedicated physics and rendering for dynamic interfaces and transparent materials, evaluating VLA policies across multiple difficulty levels Lan et al. (2025). However, existing laboratory manipulation benchmarks remain limited in terms of manipulation embodiment, experimental platforms, and task hierarchy.

3 The LabDex Dataset and Benchmark

Fig. 1 provides an overview of LabDex. We next introduce the hierarchical task structure, unified real-world and simulation platforms, dataset statistics, and evaluation protocols.

3.1 Hierarchical Task Structure

LabDex tasks are derived from representative chemical laboratory procedures and are organized hierarchically according to their operational objectives and task complexity.

Atomic Skills. Atomic Skills are fundamental operations with explicit objectives that can be independently evaluated and reused across tasks, including grasping, placing, insertion, pouring, stirring, and pressing. Each task has clearly defined initial states, goal states, and success criteria, and can be instantiated with different laboratory objects.

Compositional Skills . Compositional Skills consist of multiple Atomic Skills executed sequentially according to the operational logic of laboratory procedures, with the goal of completing an operation with a specific experimental function. For example, the robot may grasp a container, transfer material, and return the container. This level primarily evaluates the robot's ability to sequentially execute and compose fundamental skills.

Long-Horizon Laboratory Workflows . Long-Horizon Laboratory Workflows consist of multiple Compositional Skills connected according to experimental protocols and represent complete or relatively complete experimental procedures. These tasks typically involve more laboratory objects and longer action sequences, and evaluate the robot's ability to complete complex experimental procedures.

3.2 Unified Real-World and Simulation Platform

Real-World Hardware System. The real-world platform consists of a robotic execution system, visual sensing devices, and teleoperation equipment. The main hardware components are as follows:

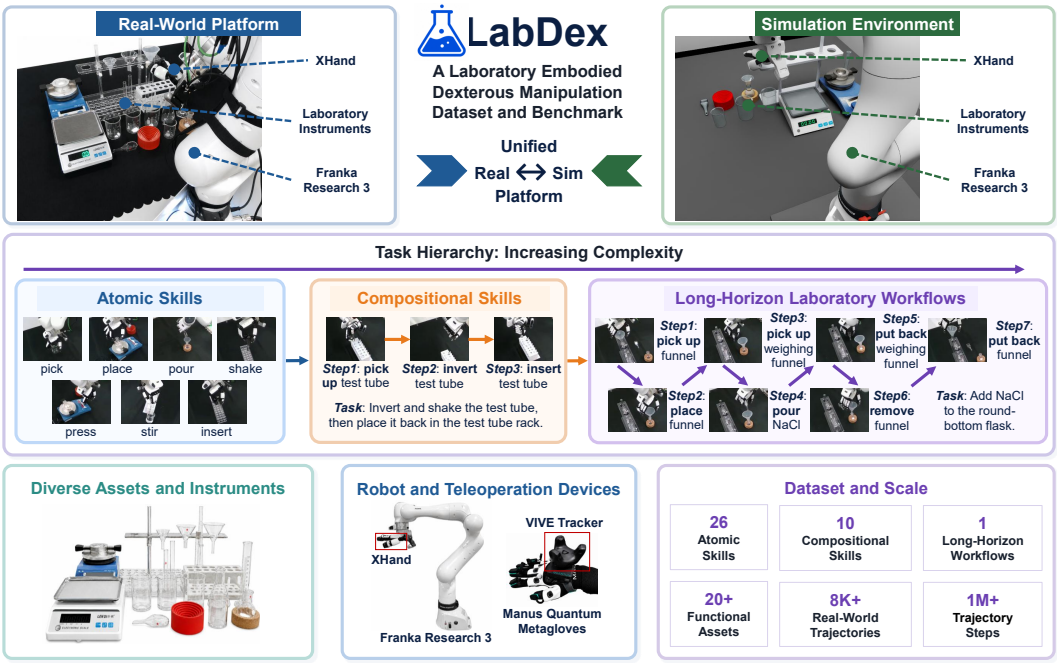

Figure 1: Overview of LabDex. LabDex is a hierarchical benchmark built on unified real-world and simulation platforms. It contains three task levels: *Atomic Skills*, *Compositional Skills*, and *Long-Horizon Laboratory Workflows*. The dataset is collected using a Franka Research 3 robot arm equipped with an XHand dexterous hand and covers a diverse set of laboratory objects.

- **Franka Research 3 robot arm and XHand dexterous hand:** These serve as the primary robotic platform for executing laboratory manipulation tasks.
- **Intel RealSense D435i cameras:** The platform is equipped with three cameras providing wrist-view, left-front-view, and head-view observations.
- **Manus Quantum data glove:** The glove is used to teleoperate the XHand dexterous hand.
- **VIVE Tracker and Lighthouse:** These devices are used to teleoperate the Franka Research 3 robot arm.

Teleoperation System. The LabDex teleoperation system is developed based on LeFranXWeng et al. (2025). The robot arm is controlled using the VIVE Tracker. Specifically, the system computes the change in the tracker pose relative to its initial pose and maps this relative transformation to the target pose of the robot end effector. Inverse kinematics is then used to obtain the corresponding target joint positions of the robot arm. The dexterous hand is controlled using the Manus Quantum data glove. Human hand joint motion captured by the glove is retargeted to the target joint positions of the XHand, enabling coordinated teleoperation of the robot arm and dexterous hand.

Real-World Data Collection Pipeline. The data collection pipeline of LabDex is illustrated in Fig. 2. The human operator controls the robot arm and dexterous hand using the VIVE Tracker and Manus Quantum data glove, respectively. During demonstration collection, the operator segments and annotates the recorded task online according to predefined atomic-task boundaries. As a result, each complete demonstration preserves the full long-horizon laboratory workflow while also providing the corresponding compositional-skill and atomic-skill segments, resulting in a demonstration dataset with an inherent hierarchical structure.

Simulation Data Collection Pipeline. We construct the simulation environment using Isaac Sim 6.0.1. As illustrated in Fig. 2, we first extract a set of key end-effector poses and dexterous-hand joint angles from real-world demonstrations. These poses and joint configurations are then manually selected and validated so that the robot can reproduce the corresponding tasks in simulation by

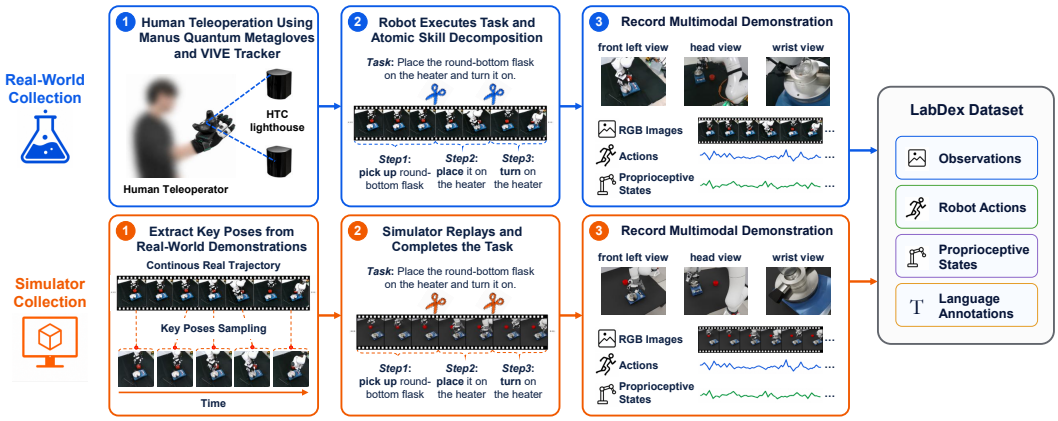

Figure 2: Real-world and simulation data collection pipeline.

sequentially executing them as waypoints. To support randomization of object positions and orientations, we further compute the relative transformation between each key pose and the manipulated object, and represent the key poses in the object coordinate frame. When the initial pose of the manipulated object changes, the corresponding key poses can be remapped to the world coordinate frame according to the new object pose, enabling the robot to complete the task under varying object configurations. Meanwhile, the key-pose sequence is aligned with the atomic-task boundaries, allowing the generated simulation trajectories to be naturally segmented into atomic skills and thereby forming a hierarchical dataset consistent with the real-world data.

Demonstration Dataset. We use the real-world and simulation platform described above to collect expert demonstrations. Each trajectory contains the following data modalities:

- **Robot proprioceptive states:** These include the joint positions, joint velocities, and end-effector pose of the robot arm, as well as the joint positions and joint torques of the dexterous hand.
- **Robot actions:** The robot-arm action can be represented either as 7-dimensional target joint positions or as a 9-dimensional end-effector pose. The latter consists of a 3-dimensional end-effector position and a 6-dimensional rotation representation formed by the first two rows of the rotation matrix. The dexterous-hand action includes 12-dimensional target joint positions and 75-dimensional human hand joint-position data. Depending on the policy formulation, either joint-space or Cartesian-space representations can be selected as the action space.
- **RGB images:** These consist of multi-view videos captured by three Intel RealSense D435i cameras, corresponding to the wrist, left-front, and head views. All images have a resolution of 640×480.

All modalities are synchronously recorded at 20Hz.

3.3 Dataset Tasks and Statistics

LabDex consists of 7 atomic-skill categories, instantiated into 26 atomic skills across over 20 laboratory objects. These atomic tasks are further organized into 10 compositional skills and 1 long-horizon laboratory workflows. Overall, the real-world dataset contains more than 8K trajectories and over 1M trajectory steps, with at least 200 trajectories collected for each task. Visualizations of selected tasks are shown in Fig. 3. More details and visualizations can be found in Appendix C.

3.4 Evaluation Protocol

The positions of the manipulated objects are randomly initialized within a region determined by the spatial range covered during training data collection. A task instance is considered successful if, after the final task stage is completed, the current state remains within the tolerance threshold of the target state for two consecutive seconds which is consistent with prior work Li et al. (2026); Gong

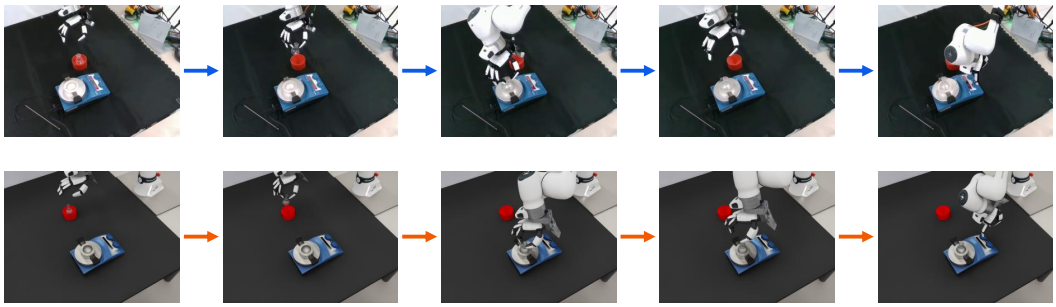

Figure 3: Visualization of a selected *compositional skill* in real-world and simulation environments.

Table 2: Average performance of atomic skill categories. We evaluate a total of 26 atomic skill instances, with each task evaluated over 50 trials for every model. Complete results for all atomic skills are provided in Appendix B.

Model	Pick	Place	Insert	Pour	Press	Shake	Stir	Overall
DP	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.02
ACT	0.25	0.64	0.25	0.23	0.00	**0.90**	0.00	0.34
$\pi_{0.5}$	**0.46**	**0.92**	**0.30**	**0.66**	**0.12**	0.88	**0.46**	**0.57**

et al. (2023); He et al. (2024); Hu et al. (2024); Liu et al. (2024). In addition, the execution must not cause unintended disturbances or unsafe interactions with the environment, such as knocking over unrelated objects or spilling liquid onto the workspace.

For *Atomic Skills*, LabDex uses task success rate as the evaluation metric. For *Compositional Skills* and *Long-Horizon Laboratory Workflows*, in addition to reporting the overall success rate of the complete task, we measure the success rate of each atomic skill and compute the average number of atomic skills completed per evaluation rollout.

4 Experiments

4.1 Experimental Setup

Models. To benchmark the performance of existing algorithms in LabDex, we select three representative models: ACT Zhao et al. (2023), Diffusion Policy Chi et al. (2025) and $\pi_{0.5}$ Physical Intelligence et al. (2025).

Training Details. For DP and ACT, we directly adopt the implementations provided in the LeRobotCadene et al. (2026) repository, while $\pi_{0.5}$ is implemented using the official OpenPI codebase. All model training and deployment are conducted on NVIDIA RTX 3090 GPUs. Unless otherwise specified, all models are trained using 200 real-world trajectories per task. Specifically, DP and ACT are trained on a single GPU with a batch size of 16. The number of training steps is determined according to the task level: 100,000 steps for *Atomic Skills*, 200,000 steps for *Compositional Skills*, and 400,000 steps for *Long-Horizon Laboratory Workflows*. For $\pi_{0.5}$, we use two GPUs with a total batch size of 16. The model is trained for 30,000 steps on *Atomic Skills*, 60,000 steps on *Compositional Skills*, and 100,000 steps on *Long-Horizon Laboratory Workflows*. In addition, ACT and $\pi_{0.5}$ condition their predictions on a single observation frame, whereas DP uses two consecutive observation frames as input. The action chunk sizes of DP, ACT, and $\pi_{0.5}$ are set to 48, 48, and 50, respectively. All models use joint angles as actions.

Evaluation Details. Each task is evaluated over 50 trials for every model. The positions of the manipulated objects and the robot arm poses are randomized within the ranges covered by the training data. All models are evaluated under the same initialization and task settings.

More details can be found in Appendix A.

Table 3: Performance of compositional skills. Seq. SR denotes the success rate of each atomic skill during sequential execution, while Ind. SR denotes the success rate when evaluated independently.

Task	**Sub-task**	**DP**		**ACT**		$\pi_{0.5}$	
		Seq. SR	**Ind. SR**	**Seq. SR**	**Ind. SR**	**Seq. SR**	**Ind. SR**
Transfer Beaker Water	Pick Right Beaker	0.00	0.38	0.70	0.38	0.62	0.86
	Pour Beaker Water	0.00	0.00	0.48	0.70	0.32	0.94
	Place Beaker	0.00	0.00	0.48	0.96	0.00	0.96
	Avg. Len.	0.00		**1.66**		0.94	
Weigh 100ml Beaker	Pick 100ml Beaker	0.00	0.00	0.12	0.02	0.72	0.82
	Place 100ml Beaker on Balance	0.00	0.00	0.12	0.56	0.70	0.94
	Avg. Len.	0.00		0.24		**1.42**	
Weigh 200ml Beaker	Pick 200ml Beaker	0.00	0.06	0.62	0.36	0.90	1.00
	Place 200ml Beaker on Balance	0.00	0.00	0.54	0.74	0.88	0.90
	Avg. Len.	0.00		1.16		**1.78**	
Shake and Return Test Tube	Pick Test Tube	0.00	0.00	0.12	0.08	0.14	0.06
	Shake Test Tube	0.00	0.00	0.02	0.90	0.14	0.88
	Place Test Tube	0.00	0.00	0.00	0.08	0.02	0.26
	Avg. Len.	0.00		0.14		**0.30**	
Stir Beaker Solution	Take Glass Rod	0.00	0.02	0.26	0.24	0.14	0.22
	Stir with Glass Rod	0.00	0.00	0.08	0.00	0.08	0.46
	Place Glass Rod	0.00	0.00	0.00	0.60	0.02	0.52
	Avg. Len.	0.00		**0.34**		0.24	
Transfer Cylinder Water	Pick Graduated Cylinder	0.00	0.00	0.00	0.20	0.00	0.04
	Pour Cylinder Water	0.00	0.00	0.00	0.00	0.00	0.78
	Place Graduated Cylinder	0.00	0.00	0.00	0.68	0.00	0.92
	Avg. Len.	0.00		0.00		0.00	
Set Funnel on Round Flask	Pick Transfer Funnel	0.00	0.00	0.60	0.84	0.94	0.88
	Place Funnel on Round Flask	0.00	0.00	0.00	0.02	0.10	0.20
	Avg. Len.	0.00		0.60		**1.04**	
Transfer NaCl to Flask	Pick Weighing Funnel	0.00	0.00	0.12	0.00	0.14	0.04
	Pour NaCl to Round Flask	0.00	0.00	0.00	0.00	0.02	0.26
	Place Weighing Funnel	0.00	0.00	0.00	0.88	0.02	1.00
	Avg. Len.	0.00		0.12		**0.18**	
Set Funnel on Stand	Remove Funnel from Round Flask	0.00	0.00	0.28	0.40	0.46	0.52
	Place Funnel on Stand	0.00	0.00	0.10	0.28	0.24	0.22
	Avg. Len.	0.00		0.38		**0.70**	
Overall Avg. Len.		0.00		0.52		**0.73**	

4.2 Experimental Results

Results of Atomic Skills. The experimental results for *Atomic Skills* are presented in Tab. 2. Complete results for all atomic skills are provided in Appendix B. Overall, $\pi_{0.5}$ achieves the highest average success rate of 0.57, substantially outperforming ACT at 0.34 and DP at 0.02. $\pi_{0.5}$ performs best on most tasks and operation categories, particularly on placement tasks, where it achieves an average success rate of 0.92. In contrast, ACT is more sensitive to the initial object positions during evaluation: it may achieve high success rates in certain regions of the workspace, while its performance degrades considerably in others. This localized performance variation can also allow ACT to outperform $\pi_{0.5}$ on a small number of challenging tasks.

In terms of operation type, placement tasks are generally easier. These tasks typically require the robot only to move an already grasped object to a target region, whereas grasping and insertion tasks impose greater demands on finger configurations and object-contact relationships, resulting in lower success rates. Pressing tasks achieve the lowest success rate, which may be attributed to the lack of tactile sensing.

Table 4: Performance of long-horizon laboratory workflows. Seq. SR and Ind. SR follow the same definitions as in the compositional-skill evaluation.

Task	Sub-task	DP		ACT		$\pi_{0.5}$	
		Seq. SR	**Ind. SR**	**Seq. SR**	**Ind. SR**	**Seq. SR**	**Ind. SR**
Funnel-Assisted NaCl Addition Experiment	Pick Transfer Funnel	0.00	0.00	0.40	0.84	0.54	0.88
	Place Funnel on Round Flask	0.00	0.00	0.00	0.02	0.02	0.20
	Pick Weighing Funnel	0.00	0.00	0.00	0.00	0.02	0.04
	Pour NaCl to Round Flask	0.00	0.00	0.00	0.00	0.02	0.26
	Place Weighing Funnel	0.00	0.00	0.00	0.88	0.02	1.00
	Remove Funnel from Round Flask	0.00	0.00	0.00	0.40	0.00	0.52
	Place Funnel on Stand	0.00	0.00	0.00	0.28	0.00	0.22
	Avg. Len.	0.00		0.40		**0.62**	

In terms of manipulated objects, tasks involving glass rods, test tubes, and graduated cylinders, which have small grasping regions or elongated geometries, are generally more challenging. These objects are more sensitive to grasp locations, finger-closing patterns, and end-effector pose errors, and therefore requiring greater manipulation accuracy.

Results of Compositional Skills. The experimental results for *Compositional Skills* are presented in Tab. 3. Overall, $\pi_{0.5}$ achieves the highest average number of completed atomic skills, reaching 0.73, compared with 0.52 for ACT, while DP fails to complete any atomic skill across all compositional tasks. The results reveal two main bottlenecks that hinder the completion of compositional skills. First, some compositional tasks are limited by the initial *Pick* skill, particularly for difficult-to-grasp objects such as round-bottom flasks, graduated cylinders, test tubes, and glass rods. *Insert* requires precise spatial alignment and is therefore sensitive to pose and object-state errors. Moreover, when transitioning from *Pour* to *Place*, $\pi_{0.5}$ may repeatedly move back and forth after pouring, suggesting insufficient memory of whether the preceding operation has been completed. As indicated by the Ind. SR values in the Tab. 3, even when subsequent atomic skills achieve relatively high success rates under independent evaluation, their success rates during sequential execution remain substantially lower. These results highlight precision and cross-skill memory as two key bottlenecks in compositional execution.

Results of Long-Horizon Laboratory Workflows. The results for *Long-Horizon Laboratory Workflows* are presented in Tab. 4. Overall, $\pi_{0.5}$ achieves the highest average number of completed atomic skills, reaching 0.62, compared with 0.40 for ACT, while DP fails to complete any atomic skill. However, none of the three models is able to complete the full workflow, indicating that existing methods still face substantial challenges in long-horizon laboratory manipulation. Consistent with the observations from the compositional-skill experiments, the main bottleneck remains the *Insert* skill following the initial funnel grasp, which requires high manipulation precision. Consequently, even though subsequent atomic skills achieve relatively strong performance when evaluated independently, the overall workflow success rate remains zero.

Results in Simulation. In the simulation experiments, we collect 50 demonstrations per task for model training. Partial results for Compositional Skills are presented in Tab. 5, showing an overall trend consistent with that observed in the real-world experiments, with $\pi_{0.5}$ achieving the best performance. For tasks involving grasping and pouring with a graduated cylinder, failures in the initial grasp constitute the primary bottleneck, preventing the execution of subsequent operations. This observation is consistent with the findings from the real-world experiments.

More experimental results can be found in Appendix B.

4.3 Ablation Study

Training Data Scale. The experiment under different amounts of training data is conducted on the atomic task: "Pick 200ml Beaker", with the results presented in Tab. 6. As the number of demonstrations increases from 50 to 200, the performance of both ACT and $\pi_{0.5}$ improves. In

Table 5: Performance of compositional skills in simulation. Each task is evaluated over 50 trials for every model.

Task	Sub-task	DP	ACT	$\pi_{0.5}$
Weigh 200ml Beaker	Pick 200ml Beaker	0.00	0.26	0.58
	Place 200ml Beaker on Balance	0.00	0.20	0.54
	Avg. Len.	0.00	0.46	**1.12**
Transfer Cylinder Water	Pick Graduated Cylinder	0.00	0.00	0.00
	Pour Cylinder Water	0.00	0.00	0.00
	Place Graduated Cylinder	0.00	0.00	0.00
	Avg. Len.	0.00	0.00	0.00

Table 6: Performance under different numbers of demonstrations.

Episodes	DP	ACT	$\pi_{0.5}$
50	0.00	0.00	0.30
100	0.00	0.14	0.34
200	0.00	0.36	1.00

Table 7: Performance under different object distractions.

Distractor Objects	Succ. Rate
None	1.00
100ml beaker	0.00
Graduated cylinder	0.42
Graduated cylinder and conical flask	0.16

Table 8: Performance on novel objects.

Object	Novel Object	Succ. Rate
200ml beaker	✗	0.88
100ml beaker	✓	0.96
Reagent bottle	✓	0.20
Weighing funnel	✓	0.22
Three-neck flask	✓	0.00

particular, $\pi_{0.5}$ achieves a success rate of 1.00 with 200 demonstrations, while ACT improves from 0 to 0.36. These results indicate that increasing the amount of demonstration data can effectively improve task learning.

Object Distractors. The object-distractor experiment is conducted on the atomic task: "Pick 200ml Beaker" using $\pi_{0.5}$, with the results shown in Tab. 7. Without distractor objects, the model achieves a success rate of 1.00. When a graduated cylinder is added to the scene, the success rate drops to 0.42, and further decreases to 0.16 when an erlenmeyer flask is also included. In particular, adding only a 100 ml beaker, which is highly similar to the target object, reduces the success rate to 0. This suggests that visually and geometrically similar objects cause stronger interference with target recognition and selection.

Novel-Object Generalization. We jointly train $\pi_{0.5}$ on six grasping tasks and evaluate its generalization to novel objects. As shown in Tab. 8, it achieves a success rate of 0.96 on an unseen 100 ml beaker, demonstrating effective generalization to objects with similar shapes and manipulation requirements. However, performance drops substantially as geometric differences from the training objects increase, indicating that its generalization remains largely limited to similar objects.

5 Conclusion

We introduce **LabDex**, a hierarchical benchmark for dexterous manipulation in chemistry laboratories. LabDex is built on unified real-world and simulation platforms and organizes laboratory manipulation tasks into three interconnected levels: *Atomic Skills*, *Compositional Skills*, and *Long-Horizon Laboratory Workflows*. We evaluate representative robot learning methods on LabDex. The experimental results demonstrate that reliable fundamental grasping, stable skill composition, and stronger generalization capabilities remain key challenges in realizing autonomous laboratory robots. We hope that LabDex will provide a standardized foundation for training, evaluating, and comparing algorithms for laboratory dexterous manipulation, while facilitating the development of more reliable and generalizable autonomous laboratory robotic systems.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, designing or providing feedback on research methodology or experiments, cleaning and reformating dataset, supporting qualitative and thematic data analysis or interpreting results. Formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses and implementing methods are not applicable to this work. Additionally, we used generative AI tools for editing a research paper to improve readability. We have reviewed all AI-assisted work.

REFERENCES

Milad Abolhasani and Eugenia Kumacheva. The rise of self-driving labs in chemical and materials sciences. *Nature Synthesis*, 2(6):483–492, 2023.

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

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Remi Cadene, Simon Aliberts, Francesco Capuano, Michel Aractingi, Adil Zouitine, Pepijn Kooijmans, Jade Choghari, Martino Russi, Caroline Pascal, Steven Palma, et al. Lerobot: An open-source library for end-to-end robot learning. *arXiv preprint arXiv:2602.22818*, 2026.

Yuanpei Chen, Tianhao Wu, Shengjie Wang, Xidong Feng, Jiechuan Jiang, Zongqing Lu, Stephen Marcus McAleer, Hao Dong, Song-Chun Zhu, and Yaodong Yang. Towards human-level bimanual dexterous manipulation with reinforcement learning. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Tianwei Dai, Sriram Vijayakrishnan, Filip T Szczypiński, Jean-François Ayme, Ehsan Simaei, Thomas Fellowes, Rob Clowes, Lyubomir Kotopanov, Caitlin E Shields, Zhengxue Zhou, et al. Autonomous mobile robots for exploratory synthetic chemistry. *Nature*, 635(8040):890–897, 2024.

Zhaohui Du, Zhe Wang, Hongmei Fei, Xiwen Cao, Ting Xiao, Qi Wang, Huanbo Jin, Jiaming Gu, Quan Lu, and Zhe Liu. Bioprovla-agent: An affordable, protocol-driven, vision-enhanced vla-enabled embodied multi-agent system with closed-loop-capable reasoning for biological laboratory manipulation. *arXiv preprint arXiv:2605.07306*, 2026.

Ran Gong, Jiangyong Huang, Yizhou Zhao, Haoran Geng, Xiaofeng Gao, Qingyang Wu, Wensi Ai, Ziheng Zhou, Demetri Terzopoulos, Song-Chun Zhu, et al. Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20483–20495, 2023.

Haoran He, Chenjia Bai, Ling Pan, Weinan Zhang, Bin Zhao, and Xuelong Li. Learning an actionable discrete diffusion policy via large-scale actionless video pre-training. *Advances in Neural Information Processing Systems*, 37:31124–31153, 2024.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

Yibin Jiang, Daniel Salley, Abhishek Sharma, Graham Keenan, Margaret Mullin, and Leroy Cronin. An artificial intelligence enabled chemical synthesis robot for exploration and optimization of nanomaterials. *Science advances*, 8(40):eabo2626, 2022.

Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*, 2025.

Zhiqian Lan, Yuxuan Jiang, Ruiqi Wang, Xuanbing Xie, Rongkui Zhang, Yicheng Zhu, Peihang Li, Tianshuo Yang, Tianxing Chen, Haoyu Gao, et al. Autobio: A simulation and benchmark for robotic automation in digital biology laboratory. *arXiv preprint arXiv:2505.14030*, 2025.

Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025.

Rui Li, Zixuan Hu, Wenxi Qu, Jinouwen Zhang, Zhenfei Yin, Sha Zhang, Xuantuo Huang, Hanqing Wang, Tai Wang, Jiangmiao Pang, et al. Labutopia: High-fidelity simulation and hierarchical benchmark for scientific embodied agents. *Advances in Neural Information Processing Systems*, 38, 2026.

Shoujie Li, Yan Huang, Changqing Guo, Tong Wu, Jiawei Zhang, Linrui Zhang, and Wenbo Ding. Chemistry3d: Robotic interaction benchmark for chemistry experiments. *arXiv preprint arXiv:2406.08160*, 2024.

Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Pengju An, Xiaoqi Li, Kaichen Zhou, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation. *Advances in Neural Information Processing Systems*, 37:40085–40110, 2024.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature machine intelligence*, 6(5):525–535, 2024.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Sebastian Steiner, Jakob Wolf, Stefan Glatzel, Anna Andreou, Jarosław M Granda, Graham Keenan, Trevor Hinkley, Gerardo Aragon-Camarasa, Philip J Kitson, Davide Angelone, et al. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science*, 363 (6423):eaav2211, 2019.

Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, 2022.

Zhengyang Kris Weng, Matthew L Elwin, and Han Liu. Levr: A modular vr teleoperation framework for imitation learning in dexterous manipulation. *arXiv preprint arXiv:2509.14349*, 2025.

Yunchao Yao, Zhuxiu Xu, Tianqi Zhang, Zixian Liu, Sikai Li, Zhenyu Wei, Feng Chen, Dihong Huang, Kechang Wan, Chenyang Ma, et al. Dexverse: A modular benchmark for multi-task, multi-embodiment dexterous manipulation. *arXiv preprint arXiv:2607.08751*, 2026.

Sha Zhang, Suorong Yang, Tong Xie, Xiangyuan Xue, Zixuan Hu, Rui Li, Wenxi Qu, Zhenfei Yin, Tianfan Fu, Di Hu, et al. Position: Intelligent science laboratory requires the integration of cognitive and embodied ai. *arXiv preprint arXiv:2506.19613*, 2025.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems*, 2023.

A Implementation Details

Additional Details about Hardware Systems. The LabDex real-world platform consists of a robotic execution system, visual sensing devices, teleoperation equipment and two workstations. The main hardware components are as follows:

- **Franka Research 3 robot arm and XHand dexterous hand:** These serve as the primary robotic platform for executing laboratory manipulation tasks.
- **Intel RealSense D435i cameras:** The platform is equipped with three cameras providing wrist-view, left-front-view, and head-view observations.
- **Manus Quantum data glove:** The glove is used to teleoperate the XHand dexterous hand.
- **VIVE Tracker and Lighthouse:** These devices are used to teleoperate the Franka Research 3 robot arm.
- **Ubuntu workstation:** This workstation serves as the main control and data collection host, and is responsible for recording experimental data, controlling the robot arm and dexterous hand, and deploying robotic policies.
- **Windows workstation:** This workstation connects to the Manus Quantum data glove, receives the captured human hand motion data, and transmits the data to the Ubuntu workstation.

Simulation Details. Our simulation environment is built on Isaac Sim 6.0.1. We reconstruct the real-world setup in simulation and reproduce the positions of the three cameras through camera calibration. We also create corresponding 3D assets for the laboratory objects involved in the manipulation tasks. During simulation data collection, we extract the dexterous-hand joint angles at contact from the real-world demonstrations and apply them in simulation, which helps achieve stable object contact. After minor manual adjustments to the key poses and joint configurations, the resulting waypoint-based policy can be used to collect simulation trajectories in batch with a success rate exceeding 99%.

Additional Details about Baseline Models. We select three representative models: ACT, Diffusion Policy and $\pi_{0.5}$.

- **ACT.** ACT integrates multi-view visual observations with the robot's current joint states through a transformer architecture and outputs a fixed-length chunk of future target joint positions. During execution, predictions from overlapping chunks are aggregated through temporal ensembling, enabling smoother control while alleviating error accumulation over long task horizons.
- **Diffusion Policy.** Diffusion Policy models robot control as a conditional denoising diffusion process. Given recent observations, it iteratively refines a noisy action sequence into a future control trajectory.
- $\pi_{0.5}$**.** $\pi_{0.5}$ is a vision-language-action model designed for open-world robotic manipulation. Given visual observations, language instructions, and robot states, it first predicts a high-level subtask and then generates a low-level action chunk using a flow-matching action expert.

In addition, we evaluate two additional dexterous-hand VLA models Kawaharazuka et al. (2025); Li et al. (2025) on several Atomic Skills. We fine-tune both models on LabDex using the official training scripts provided in their open-source repositories. However, neither model achieves meaningful task success under our evaluation setting. We think that this may result from the substantial discrepancy between the LabDex environment and the embodiments and data distributions used during their pretraining, making adaptation to our platform particularly challenging. We therefore do not include these models in the full benchmark evaluation.

Training Details. The number of training steps is determined according to the task level: 100,000 steps for *Atomic Skills*, 200,000 steps for *Compositional Skills*, and 400,000 steps for *Long-Horizon Laboratory Workflows*. For $\pi_{0.5}$, we use two GPUs with a total batch size of 16. The model is trained for 30,000 steps on *Atomic Skills*, 60,000 steps on *Compositional Skills*, and 100,000 steps on *Long-Horizon Laboratory Workflows*. These training-step settings are chosen to keep the effective number of training epochs approximately consistent across the three task levels for each model, accounting for differences in dataset size.

Table 9: Performance of atomic skills. Each task is evaluated over 50 trials for every model.

Type	Task	DP	ACT	$\pi_{0.5}$
Pick	Pick Round Flask	0.00	0.02	0.12
	Pick Right Beaker	0.38	0.38	0.86
	Pick 100ml Beaker	0.00	0.02	0.82
	Pick 200ml Beaker	0.06	0.36	1.00
	Pick Test Tube	0.00	0.08	0.06
	Take Glass Rod	0.02	0.24	0.22
	Pick Transfer Funnel	0.00	0.84	0.88
	Pick Graduated Cylinder	0.00	0.20	0.04
	Pick Weighing Funnel	0.00	0.00	0.04
	Remove Funnel from Round Flask	0.00	0.40	0.52
	Average	0.05	0.25	**0.46**
Place	Place Round Flask	0.08	0.00	0.78
	Place Beaker	0.00	0.96	0.96
	Place 100ml Beaker on Balance	0.00	0.56	0.94
	Place 200ml Beaker on Balance	0.00	0.74	0.90
	Place Graduated Cylinder	0.00	0.68	0.92
	Place Weighing Funnel	0.00	0.88	1.00
	Average	0.01	0.64	**0.92**
Insert	Place Test Tube	0.00	0.08	0.26
	Place Glass Rod	0.00	0.60	0.52
	Place Funnel on Round Flask	0.00	0.02	0.20
	Place Funnel on Stand	0.00	0.28	0.22
	Average	0.00	0.25	**0.30**
Pour	Pour Beaker Water	0.00	0.70	0.94
	Pour Cylinder Water	0.00	0.00	0.78
	Pour NaCl to Round Flask	0.00	0.00	0.26
	Average	0.00	0.23	**0.66**
Press	Turn On Heater	0.00	0.00	0.12
	Average	0.00	0.00	**0.12**
Shake	Shake Test Tube	0.00	0.90	0.88
	Average	0.00	**0.90**	0.88
Stir	Stir with Glass Rod	0.00	0.00	0.46
	Average	0.00	0.00	**0.46**
Overall Average		0.02	0.34	**0.57**

Evaluation Details. Each task is evaluated over 50 trials for every model. For Atomic Skills, we consider two initialization settings according to their positions in the complete task sequence:

- If the evaluated atomic skill is the first atomic operation in the task sequence, the robot arm and dexterous hand are reset to their default initial states, while only the positions of the manipulated objects involved in the skill are randomized.
- If the evaluated atomic skill is not the first atomic operation in the task sequence, all preceding atomic operations are first replayed to bring the environment to the required initial state. The end-effector pose is then randomized within the range covered by the training data.

Compositional Skills are evaluated using the same initialization protocol as Atomic Skills. For Long-Horizon Laboratory Workflows, only the positions of all manipulated objects involved in the task are randomized.

Table 10: Performance of compositional skills. Seq. SR denotes the success rate of each atomic skill during sequential execution, while Ind. SR denotes the success rate when evaluated independently.

Task	**Sub-task**	**DP**		**ACT**		$\pi_{0.5}$	
		Seq. SR	**Ind. SR**	**Seq. SR**	**Ind. SR**	**Seq. SR**	**Ind. SR**
Heat Round Flask	Pick Round Flask	0.00	0.00	0.00	0.02	0.00	0.12
	Place Round Flask	0.00	0.08	0.00	0.00	0.00	0.78
	Turn on Heater	0.00	0.00	0.00	0.00	0.00	0.12
	Avg. Len.	0.00		0.00		0.00	
Transfer Beaker Water	Pick Right Beaker	0.00	0.38	0.70	0.38	0.62	0.86
	Pour Beaker Water	0.00	0.00	0.48	0.70	0.32	0.94
	Place Beaker	0.00	0.00	0.48	0.96	0.00	0.96
	Avg. Len.	0.00		**1.66**		0.94	
Weigh 100ml Beaker	Pick 100ml Beaker	0.00	0.00	0.12	0.02	0.72	0.82
	Place 100ml Beaker on Balance	0.00	0.00	0.12	0.56	0.70	0.94
	Avg. Len.	0.00		0.24		**1.42**	
Weigh 200ml Beaker	Pick 200ml Beaker	0.00	0.06	0.62	0.36	0.90	1.00
	Place 200ml Beaker on Balance	0.00	0.00	0.54	0.74	0.88	0.90
	Avg. Len.	0.00		1.16		**1.78**	
Shake and Return Test Tube	Pick Test Tube	0.00	0.00	0.12	0.08	0.14	0.06
	Shake Test Tube	0.00	0.00	0.02	0.90	0.14	0.88
	Place Test Tube	0.00	0.00	0.00	0.08	0.02	0.26
	Avg. Len.	0.00		0.14		**0.30**	
Stir Beaker Solution	Take Glass Rod	0.00	0.02	0.26	0.24	0.14	0.22
	Stir with Glass Rod	0.00	0.00	0.08	0.00	0.08	0.46
	Place Glass Rod	0.00	0.00	0.00	0.60	0.02	0.52
	Avg. Len.	0.00		**0.34**		0.24	
Transfer Cylinder Water	Pick Graduated Cylinder	0.00	0.00	0.00	0.20	0.00	0.04
	Pour Cylinder Water	0.00	0.00	0.00	0.00	0.00	0.78
	Place Graduated Cylinder	0.00	0.00	0.00	0.68	0.00	0.92
	Avg. Len.	0.00		0.00		0.00	
Set Funnel on Round Flask	Pick Transfer Funnel	0.00	0.00	0.60	0.84	0.94	0.88
	Place Funnel on Round Flask	0.00	0.00	0.00	0.02	0.10	0.20
	Avg. Len.	0.00		0.60		**1.04**	
Transfer NaCl to Flask	Pick Weighing Funnel	0.00	0.00	0.12	0.00	0.14	0.04
	Pour NaCl to Round Flask	0.00	0.00	0.00	0.00	0.02	0.26
	Place Weighing Funnel	0.00	0.00	0.00	0.88	0.02	1.00
	Avg. Len.	0.00		0.12		**0.18**	
Set Funnel on Stand	Remove Funnel from Round Flask	0.00	0.00	0.28	0.40	0.46	0.52
	Place Funnel on Stand	0.00	0.00	0.10	0.28	0.24	0.22
	Avg. Len.	0.00		0.38		**0.70**	
Overall Avg. Len.		0.00		0.46		**0.66**	

Simulation Evaluation Protocol. For each task in simulation, we implement an automated success criterion consistent with that used in the real-world evaluation. In addition to satisfying the task-specific goal conditions, a rollout is considered successful only if the robot completes the task without causing unintended disturbances or unsafe interactions with the environment, such as knocking over unrelated objects or spilling liquid onto the workspace.

B Additional Experimental Results

Complete Real-World Atomic Skills Results. As shown in Tab. 9, we report the complete results for Atomic Skills in the real-world setting. Each task is evaluated over 50 trials for every model.

Table 11: Performance of atomic skills in simulation. Each task evaluated over 50 trials for every model.

Task	DP	ACT	$\pi_{0.5}$
Pick Round Flask	0.00	0.00	0.56
Pick 200ml Beaker	0.34	0.22	0.64
Place 200ml Beaker on Balance	0.08	0.32	0.02
Pick Graduated Cylinder	0.00	0.04	0.26
Pour Cylinder Water	0.52	0.74	0.72
Place Graduated Cylinder	0.00	0.08	0.92
Overall Average	0.19	0.26	**0.44**

Complete Real-World Compositional Skills Results. As shown in Tab. 10, we report the complete results for Compositional Skills in the real-world setting. Each task is evaluated over 50 trials for every model.

Simulation Atomic Skills Results. As shown in Tab. 11, we report the results for Atomic Skills in the simulation setting. Each task is evaluated over 50 trials for every model.

C Dataset Details

C.1 Laboratory Objects and Assets

LabDex incorporates a diverse collection of commonly used chemistry laboratory objects and supporting equipment. The manipulated objects include beakers of different sizes, round-bottom flasks, Erlenmeyer flasks, graduated cylinders, test tubes, glass rods, standard funnels, and a weighing funnel. The supporting assets include funnel stands, test-tube racks, a round-bottom-flask holder, cork rings, an electronic balance, and a laboratory heater. Several task scenes also contain multiple instances of similar objects, such as two beakers, to better reflect realistic laboratory settings and increase the difficulty of target-object selection.

These objects cover a wide range of geometries and manipulation requirements, including cylindrical and elongated objects, narrow-necked containers, transparent glassware, and objects with limited effective grasping regions. Together, they support diverse laboratory operations such as grasping, placing, insertion, pouring, stirring, weighing, and object transfer. Corresponding assets are constructed in the real-world and simulation platforms to maintain consistent task definitions and hierarchical task structures across both settings.

C.2 Atomic-Skill

Task Definitions. LabDex defines Atomic Skills as fundamental operations with explicit objectives that can be independently evaluated and reused across different laboratory tasks. The task name describes the operation to be performed and the corresponding object, such as picking up a 200 ml beaker, inserting a test tube into a test-tube rack, or pouring material from a graduated cylinder into a round-bottom flask. The initial state specifies the configurations of the robot and relevant laboratory objects at the beginning of the task, while the goal state describes the required object state or spatial relationship upon task completion. A task is considered successful only when the goal state satisfies the predefined tolerance and remains stable for the specified duration.

Atomic-Skill Categories. Atomic Skill Categories. We group all atomic skills into seven categories according to their operation types and provide a concise description of each task below.

Pick. Pick tasks require the robot to grasp a target object and remove it from its initial support or holder while maintaining a stable grasp.

- Pick Round Flask: Grasp the round-bottom flask and lift it from its holder.

- Pick Right Beaker: Grasp the beaker on the right and lift it from the table.
- Pick 100ml Beaker: Grasp the 100 ml beaker and lift it from the table.
- Pick 200ml Beaker: Grasp the target beaker and lift it from the table.
- Pick Test Tube: Grasp the test tube and remove it from the test-tube rack.
- Take Glass Rod: Grasp the glass rod and remove it from the white test-tube rack.
- Pick Transfer Funnel: Grasp the funnel and remove it from the transparent funnel stand.
- Pick Graduated Cylinder: Grasp the graduated cylinder and lift it from the table.
- Pick Weighing Funnel: Grasp the weighing funnel and lift it from the table.
- Remove Funnel from Round Flask: Grasp the funnel and remove it from the round-bottom flask.

Place. Place tasks require the robot to move a grasped object to a designated target region or laboratory device and release it in a stable state.

- Place Round Flask: Place the round-bottom flask on the heater and release it stably.
- Place Beaker: Return the grasped beaker to the table and release it stably.
- Place 100ml Beaker on Balance: Place the 100 ml beaker on the electronic balance.
- Place 200ml Beaker on Balance: Place the target beaker on the electronic balance.
- Place Graduated Cylinder: Return the graduated cylinder to the table and release it stably.
- Place Weighing Funnel the table: Return the weighing funnel to the table and release it stably.

Insert. Insert tasks require the robot to align a manipulated object with a target opening, rack, or holder and insert it into the designated position.

- Place Test Tube: Align the test tube with the target slot and insert it into the test-tube rack.
- Place Glass Rod: Return the glass rod to its designated position on the white test-tube rack.
- Place Funnel on Round Flask: Align the funnel with the opening of the round-bottom flask and insert it into the flask.
- Place Funnel on Stand: Align the funnel with the funnel stand and return it to the designated position.

Pour. Pour tasks require the robot to maintain a stable grasp on the source container while adjusting its orientation to transfer the contained material into a target container.

- Pour Beaker Water: Pour water from the grasped beaker into the beaker on the table.
- Pour Cylinder Water: Pour water from the graduated cylinder into the round-bottom flask.
- Pour NaCl to Round Flask: Transfer all NaCl from the weighing funnel into the round-bottom flask.

Stir. Stir tasks require the robot to manipulate a glass rod continuously inside a container while maintaining stable control of the tool.

- Stir with Glass Rod: Use the glass rod to stir the contents of the beaker located to the left of the white test-tube rack.

Press. Press tasks require the robot to accurately contact and activate a target control on a laboratory device.

- Turn on Heater: Press the control button to turn on the heater.

Shake. Shake tasks require the robot to maintain a stable grasp while repeatedly moving the manipulated object.

- Shake Test Tube: Maintain a stable grasp on the test tube and shake its contents for a short period.



(a) Pick 200ml Beaker (b) Place Round Flask (c) Pour Cylinder Water (d) Turn on Heater



(e) Place Test Tube (f) Shake Test Tube (g) Stir with Glass Rod

Figure 4: Examples of selected atomic skills.

C.3 Compositional-Skill

Compositional Skills consist of multiple Atomic Skills executed sequentially to complete a laboratory operation with a specific functional objective. The complete set of compositional tasks is described below. For each task, we list the constituent Atomic Skills in their execution order.

- Heat Round Flask: The robot picks up the round-bottom flask, places it on the heater, and turns on the heater.
- Transfer Beaker Water: The robot picks up the beaker on the right, pours the water into the beaker on the left, and returns the grasped beaker to the table.
- Weigh 100ml Beaker: The robot picks up the 100 ml beaker and places it on the electronic balance.
- Weigh 200ml Beaker: The robot picks up the 200 ml beaker and places it on the electronic balance.
- Shake and Return Test Tube: The robot picks up the test tube, shakes it, and returns it to the test-tube rack.
- Stir Beaker Solution: The robot removes the glass rod from the white test-tube rack, stirs the solution in the beaker, and returns the glass rod to the rack.
- Transfer Cylinder Water: The robot picks up the graduated cylinder, pours the water into the round-bottom flask, and returns the graduated cylinder to the table.
- Set Funnel on Round Flask: The robot removes the funnel from the transparent funnel stand and inserts it into the round-bottom flask.
- Transfer NaCl to Flask: The robot picks up the weighing funnel, pours all the NaCl into the round-bottom flask, and returns the weighing funnel to the table.
- Set Funnel on Stand: The robot removes the funnel from the round-bottom flask and returns it to the funnel stand.

C.4 Long-Horizon Laboratory Workflow

The Long-Horizon Laboratory Workflow consists of multiple Compositional Skills executed sequentially to complete a laboratory procedure. Compared with individual Compositional Skills, this workflow involves more objects, a longer operation sequence, and stronger dependencies between successive stages. The robot must maintain consistent object states throughout the workflow, as failures in an early stage may prevent subsequent operations from being executed.

Funnel-Assisted NaCl Addition Experiment. The robot first removes the funnel from the transparent funnel stand and places it on the round-bottom flask. It then picks up the weighing funnel, pours all the NaCl into the round-bottom flask through the funnel, and returns the weighing funnel

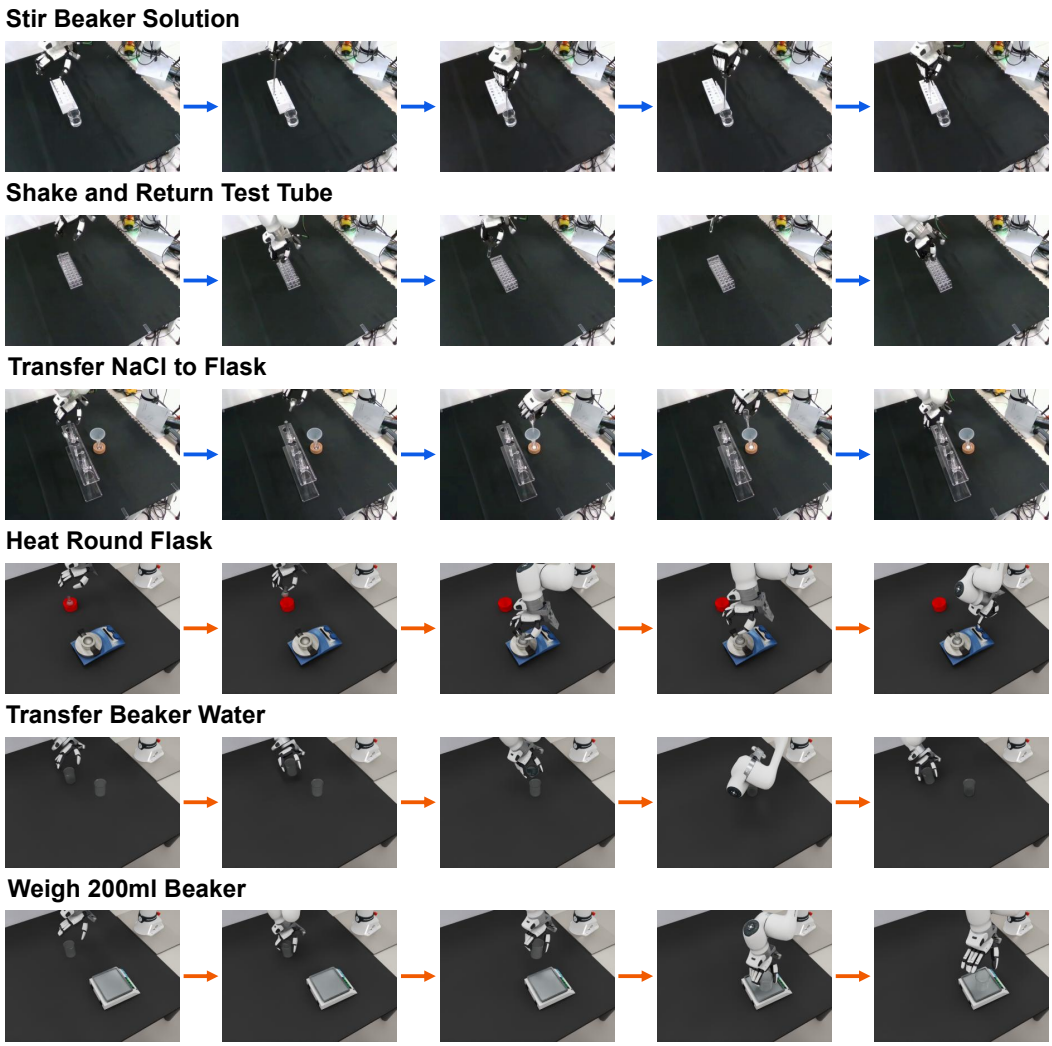

Figure 5: More visualizations of our tasks in real-world and simulation environments.

to the table. Finally, the robot removes the funnel from the round-bottom flask and returns it to the funnel stand.

The workflow consists of the following Atomic Skills: Pick Transfer Funnel; Place Funnel on Round Flask; Pick Weighing Funnel; Pour NaCl to Round Flask; Place Weighing Funnel; Remove Funnel from Round Flask; Place Funnel on Stand.

C.5 More Visualizations

More visualization of our tasks in real-world and simulation environments can be found in Fig. 4 and Fig. 5