ManiLadder: Benchmarking Robot Manipulation Through a Categorized and Multi-Level Task Ladder

Anonymous Author(s)
Affiliation
Address
email

Abstract: Recent advances in robotic manipulation have enabled increasingly complex and dexterous behaviors, yet existing benchmarks still lack a systematic framework for measuring progress in robot learning along the axis of task difficulty. We present ManiLadder, a large-scale and organized simulation benchmark for robotic manipulation. ManiLadder contains 112 carefully designed tasks stratified into 4 difficulty levels, with 8 categories in each level, spanning rigid-body and deformable-object manipulation, single-arm and dual-arm settings, and gripper and dexterous-hand end-effectors. To calibrate difficulty, we provide high-quality demonstrations for every task and train a unified flow-matching-based imitation learning policy under a standardized protocol, then use its evaluation success rate as an operational empirical difficulty signal. All tasks are iteratively refined until they fall into target difficulty ranges. We further evaluate advanced vision-language-action models, study transfer across difficulty levels and task categories, analyze data scaling behavior, and conduct real-world experiments showing that the designed tasks preserve their difficulty ordering beyond simulation. ManiLadder provides a calibrated ladder for measuring, comparing, and advancing robotic manipulation capabilities. Video results are on the website.

Keywords: Benchmark, Robot Manipulation, Simulation

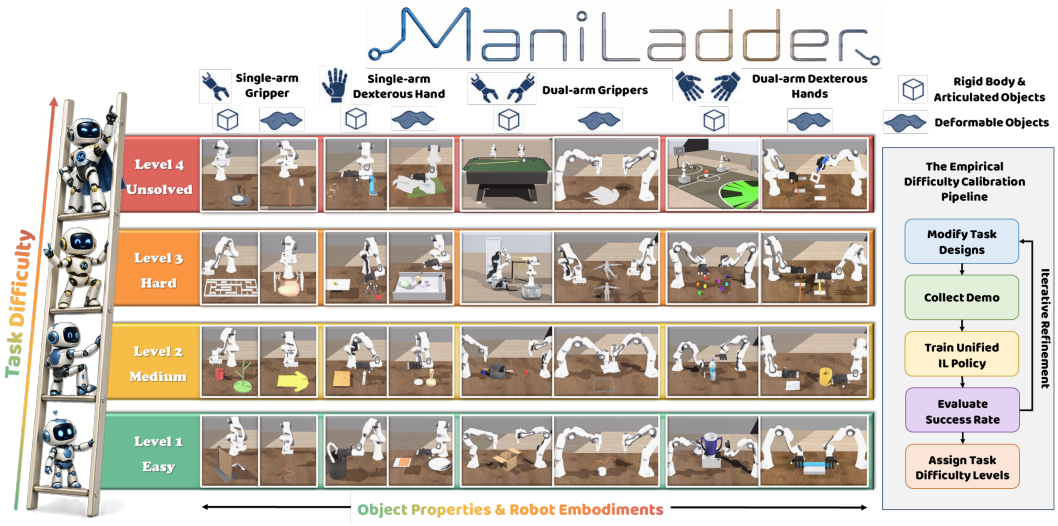

Figure 1: Overview of ManiLadder, a difficulty-calibrated simulation benchmark for robotic manipulation. ManiLadder organizes 112 tasks (only 32 are shown here) into a vertical ladder of four difficulty levels while maintaining broad horizontal diversity across robot embodiments, object properties, and skills. The difficulty of each task is calibrated and iteratively refined by the success rate of a standardized imitation-learning policy, enabling evaluation both of whether a model can solve a task and the level of manipulation difficulty it can reliably reach.

Submitted to the 10th Conference on Robot Learning (CoRL 2026). Do not distribute.

1 Introduction

Recent years have witnessed remarkable progress in the field of robotic manipulation, from more affordable and user-friendly hardware [1, 2], large-scale datasets [3, 4], to powerful foundation models [5, 6] and effective learning algorithms [7, 8]. These advances have enabled robots to perform highly complex and dexterous manipulation tasks in the real world, such as scrambling eggs [9], folding origami [10], and lab pipetting [11]. However, current benchmarks in robotic manipulation lag considerably behind this huge progress: most existing robotic simulation benchmarks [12, 13, 14, 15] primarily expand task diversity *horizontally*, by recombining relatively simple manipulation skills across different scenes and objects, such as pick-and-place, pushing, and door opening. While they are useful for evaluating generalization or semantic reasoning, they offer limited resolution for measuring the *vertical* growth of manipulation capability, i.e., they cannot reveal whether a model has truly expanded the frontier of solvable manipulation tasks. We therefore ask a capability-oriented question: **what level of manipulation difficulty can a model reliably solve?** Answering this question requires a benchmark that organizes manipulation tasks not only by combining simple skills, but also by calibrated levels of task difficulty, thereby exposing the boundaries of current methods and providing a concrete roadmap for future progress.

Constructing such a difficulty-aware benchmark raises two central challenges: (1) how to determine the difficulty level of a task, and (2) how to build diverse tasks whose difficulties are both high and well calibrated. For the first question, prior works define manipulation difficulty using subjective theoretical measures [16, 17], subskill compositions [18, 19], manually specified predicates [20, 21], or contact patterns [22, 23, 24], but these hand-designed abstractions are difficult to scale to diverse modern manipulation tasks. We instead adopt an objective and empirical view: under a standardized training protocol, what success rate can a robot learning algorithm achieve on the task? This view is increasingly practical as recent imitation learning methods [7, 1] become effective across a broad range of manipulation tasks. For the second challenge, although modern simulators [25, 26] and procedural task-generation tools [27] make it easier to create visually and physically diverse environments, they do not automatically calibrate task difficulty. Tasks can be made accidentally too easy or too hard due to artifacts in initial-state distributions, perception complexity, error tolerances, or object geometry. A meaningful difficulty ladder therefore requires careful and iterative, metric-driven task tuning, so that empirical difficulty matches the intended capability level.

In this work, we present ManiLadder, a large-scale difficulty-aware simulation benchmark that evaluates robotic manipulation along two complementary axes: calibrated task difficulty and manipulation diversity, as shown in Figure 1. ManiLadder contains 112 carefully designed tasks across 4 difficulty levels, organized by object properties, robot embodiments, and manipulation skills. The tasks cover both rigid-body and deformable-object manipulation, spanning four embodiment types: single-arm grippers, single-arm dexterous hands, dual-arm grippers, and dual-arm dexterous hands. Within each task suite, we further cover disparate manipulation skills or object types, resulting in 3 to 4 concrete tasks per suite. For example, in each deformable-object manipulation suite, we design tasks targeting 1D (e.g., rope), 2D (e.g., cloth), and 3D (e.g., clay) deformable object types, respectively. To stratify task difficulty, we provide high-quality demonstrations for every task and train a unified flow-matching-based [28] imitation learning policy RADIOFlow under a standardized protocol, using its evaluation success rate as an empirical difficulty signal. We then iteratively calibrate each task to the desired difficulty range by adjusting interpretable factors such as object geometry, initial randomization, perceptual difficulty, temporal coordination, and error tolerance. This process requires about four rounds of human refinement per task on average, yielding a benchmark that is not only broad in task coverage but also explicitly calibrated as a ladder of manipulation difficulty.

To highlight the flexibility and potential of ManiLadder for studying manipulation capability across different dimensions, we conduct a set of initial studies, including: 1) evaluating advanced VLA models [6] on ManiLadder; 2) analyzing transfer across difficulty levels, end-effectors, and object categories; 3) studying data scaling behaviors across different task families; 4) examining whether the difficulty ordering calibrated in simulation is preserved between simulation and real-world settings. Beyond these initial studies, ManiLadder provides a flexible foundation for studying a wider

Table 1: Comparison between ManiLadder and other simulation robot manipulation benchmarks focusing on task taxonomy and difficulty levels. Auto Demo means supporting automatic data generation. We use $\mathcal{G}$ for robots with grippers, $\mathcal{X}$ for robots with dexterous hands, $\mathcal{S}$ for single-arm robots, $\mathcal{D}$ for dual-arm robots, and $\mathcal{M}$ for mobile robots.

Benchmark	Simulator	No. of tasks	Auto Demo	Difficulty Taxonomy	Deformable objects	Robots
MetaWorld [29]	Mujoco [30]	50	✓	✗	✗	$\mathcal{S},\mathcal{G}$
FactoredWorld [31]	Mujoco [30]	19	✓	✗	✗	$\mathcal{S},\mathcal{G}$
COLOSSEUM [32]	V-Rep [33]	20	✓	✗	✗	$\mathcal{S},\mathcal{G}$
RoboCasa [14]	Mujoco [30]	100	✓	✗	✗	$\mathcal{S},\mathcal{X},\mathcal{G},\mathcal{D},\mathcal{M}$
Behavior 1K [15]	Isaac Sim [26]	1000	✗	✗	✓	$\mathcal{S},\mathcal{G},\mathcal{M}$
LIBERO [12]	Mujoco [30]	120	✗	✓	✗	$\mathcal{S},\mathcal{G}$
RoboLab [34]	Isaac Sim [26]	120	✓	✓	✗	$\mathcal{S},\mathcal{G}$
ManiLadder (ours)	Genesis [25]	112	✓	✓	✓	$\mathcal{S},\mathcal{X},\mathcal{G},\mathcal{D},\mathcal{M}$

range of questions in robot learning, including capability scaling, embodiment generalization, data efficiency, and continual learning. Overall, ManiLadder is not merely a large-scale collection of manipulation tasks, but a calibrated and organized difficulty ladder that measures how far robotic manipulation has climbed toward human-level dexterity, and how far it still has to go.

2 Related Works

2.1 Robot Manipulation Task Taxonomy

It is challenging to study taxonomy for robotic manipulation tasks and divide them into concise and well-defined categories, due to the extreme complexity arising from the tasks themselves, the diversity of the objects involved, and the variety of robot embodiment employed [24]. Some works use grasp types and contact modes [24, 35] or task primitives and skills [36, 37, 19] for task taxonomy. Others provide taxonomies for specific tasks, such as deformable object manipulation tasks [38] or bi-manual manipulation tasks [39]. However, the criteria underlying these task classifications are overly detailed and restrictive, making it difficult to accommodate complex, multi-stage tasks. In this work, we categorize robot manipulation tasks using two axes: the horizontal axis represents different tasks of the same difficulty level, distinguished by the type of manipulated object and the end effector employed; the vertical axis represents tasks of varying difficulty levels measured by success rates from a standard and unified policy and training protocol.

2.2 Robot Manipulation Benchmarks and Evaluations

There are numerous simulation robot manipulation benchmarks [12, 13, 32, 40, 41, 29, 42, 43, 14, 34] over the years. These studies target different aspects of manipulation research, such as task horizon [42, 14], generalization [32, 34], task transfer [29], and semantic reasoning [44]. However, most of them are recombining relatively simple manipulation skills across different scenes and objects, such as pick-and-place, pushing, and door-opening. There are also many real-world evaluation platforms [45, 46, 47, 48, 49, 50] or in-person challenges [51, 52] for robot manipulation tasks. These works try to ensure reproducibility with manuals for environment setups or enabling remote access to a centrally hosted evaluation platform. However, reproducibility and high participation barriers remain the greatest challenges faced by them. In this work, we choose to build ManiLadder in simulation, featuring the calibrated task difficulty levels and organized task categories.

3 The ManiLadder Benchmark

In this section, we describe how ManiLadder is constructed as a difficulty-calibrated simulation benchmark for robotic manipulation. We first introduce the shared basic scene setup and simulation interface, then describe task selection and organization principles, followed by the difficulty calibration process, and finally analyze the resulting benchmark properties and statistics.

3.1 ManiLadder Scene Setup

ManiLadder is built upon Genesis [25], a cross-platform simulator that supports rigid-body dynamics, deformable objects, and GPU-based parallel simulation. We deliberately focus on manipulation-centric challenges rather than scene-level diversity or navigation complexity. Therefore, most tasks

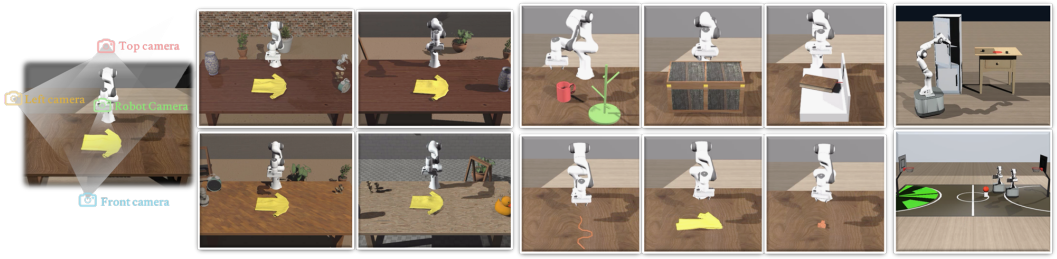

Figure 2: Left: the scene setup of task *FoldShirt*, serving as an example illustrating a typical task scenario in ManiLadder. We also show the domain randomization on the lighting, background, and distracting objects. Right: in each rigid-body object task suite, we design different tasks focusing on different manipulation skills, such as *HangMug* for pick-and-place, *OpenBoxRetrival* for articulated object manipulation, and *PivotBook* for non-prehensile manipulation. In each deformable object task suite, we design tasks including 1D, 2D, and 3D deformable objects, such as *StraightenRope* for rope manipulation, *UnfoldShirt* for cloth manipulation, and *ShapeTtoI* for clay manipulation. We also have a few mobile manipulation tasks in ManiLadder.

Figure 3: The objects and robots gallery in ManiLadder.

are instantiated in controlled table-top environments, with a small number of mobile manipulation scenarios included only when base motion is intrinsic to the task. We implement a unified simulation interface following the OpenAI Gym-style API [53] across all tasks, standardizing the policy-environment interaction protocol, including environment reset, observation retrieval, action execution, reward computation, success checking, and evaluation rollout. We set three third-view cameras and one or two first-view cameras on robots, and add domain randomization in lighting conditions, table and background textures, and distracting objects, as shown in Figure 2.

We curate object assets from multiple open-source repositories [46, 54] and custom designs to support both broad task coverage. These assets are rescaled, cleaned, and processed with convex decomposition to improve simulation stability. For deformable-object manipulation, we create assets in Blender and convert them into simulation-ready formats with Trimesh [55], and use different physical representations for different object categories: 1D ropes are represented as articulated chains of capsule-shaped rigid bodies, 2D cloths are simulated with a PBD solver [56], and 3D deformable materials are modeled with solvers such as MPM [57] for clay-like objects and SPH [58] for liquids. Together, these assets allow ManiLadder to cover a wide range of manipulation regimes, from rigid-body interaction to contact-rich deformable-object manipulation, as shown in Figure 3.

3.2 Task Selection and Taxonomy

ManiLadder is designed to measure the level of manipulation difficulty that a robot learning policy can reliably solve. To make such a comparison convincing, tasks in each level must be diverse enough as well as organized in a structured way, so that the performance differences mostly reflect changes associated with task difficulty. We design ManiLadder with a unified taxonomy: all levels contain unified task suites organized by object property, robot embodiment, and manipulation skill.

The first axis is **object physical property**. We divide tasks in each level into rigid-body manipulation and deformable-object manipulation. Rigid-body tasks include both objects with fixed geome-

Table 2: Manipulation skill coverage of ManiLadder. We group skills according to prior literature on manipulation-primitives and contact-rich manipulation taxonomy [59, 36, 60, 61].

Skill family	Primitive skills	# Tasks	Avg. Level
Prehensile manipulation	grasp, pick, hold, lift, carry, place, release	33	2.36
Non-prehensile manipulation	push, pull, slide, drag, shift, nudge, pivot, tilt, flip	9	2.22
Articulated-object manipulation	open, close, turn, twist, press, pull-handle, push-button	16	2.07
Contact-rich assembly	insert, extract, align, slide-along-contact, move-until-contact	17	2.88
Tool-use manipulation	scoop, sweep, wipe, hook, lever, strike, stir	15	3.20
Deformable-object manipulation	fold, unfold, hang, stretch, wrap, squeeze, flatten	48	2.50
Granular and fluid manipulation	pour, fill, transfer, scoop, spill-control, measure	16	2.50
Bimanual coordination	handover, stabilize-and-manipulate, dual-arm lift & pull & push	56	2.50
Dynamic manipulation	drop, toss, throw, catch, bounce, hit, swing	15	3.47

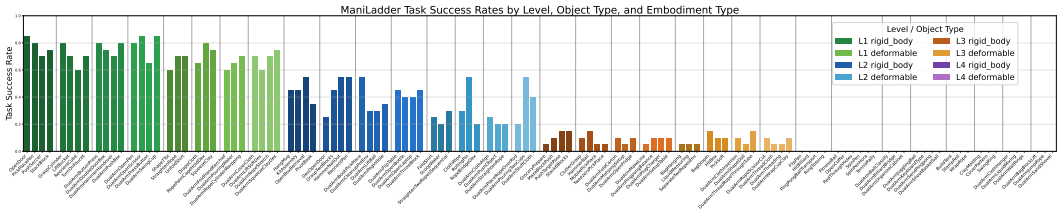

Figure 4: All task success rates in ManiLadder of the unified diffusion policy after calibration.

try, such as blocks and mugs, and articulated objects with movable parts, such as doors and drawers. Deformable-object tasks are grouped by geometric dimensionality: 1D deformables such as ropes and cables, 2D deformables such as cloths and towels, and 3D deformables such as clay, plasticine, and liquid. This distinction allows ManiLadder to cover qualitatively different interaction regimes. To the best of our knowledge, ManiLadder is the first simulation benchmark that contains more than 40 deformable object manipulation tasks within a well-organized structure.

The second axis is **robot embodiment**. We consider four end-effector configurations: single-arm gripper, single-arm dexterous hand, dual-arm gripper, and dual-arm dexterous hand. These embodiments introduce different manipulation capabilities and coordination requirements, ranging from prehensile grasping to dexterous in-hand and bimanual coordination. We use Franka Panda as the robot arm and LeapHand [2] as the dexterous hand. A small number of tasks further include mobile manipulation when base motion is intrinsic to the task, using TidyBot++ [62].

The third axis is **manipulation skill**. Within each object-property and embodiment category, we instantiate three to four tasks that cover distinct skills whenever possible, as shown in the middle of Figure 2. Our skill selection is a comprehensive consideration and summary of prior studies on manipulation primitives, contact-based manipulation behaviors, action representations, and grasp taxonomies [21, 20, 22, 23]. Table 2 summarizes the manipulation skills and their corresponding number of occurrences and levels in ManiLadder. This design avoids concentrating a task suite around a single operation and encourages each difficulty level to cover multiple sources of manipulation challenge. Overall, ManiLadder contains 112 manually curated tasks arranged in an aligned hierarchy of difficulty level, object property, embodiment, and skill.

3.3 Task Difficulty Calibration and Iterative Refinement

Assigning task difficulty by human intuition would make the benchmark subjective. We therefore acquire difficulty signals operationally using a standardized reference imitation learning policy: under the same number of demonstrations, model architecture, hyperparameters, training budget, and evaluation protocol, harder tasks are those on which the reference policy achieves lower success rates. We emphasize that ManiLadder does not aim to define the real and absolute, policy-independent task difficulty. Instead, for modern robot learning benchmarks, the practical question is often not whether a task is intrinsically difficult, but whether a policy class can reliably solve the task from a standardized amount of demonstration data and training budget. Under this view, a good and widely used reference policy could serve as a calibrated measuring instrument. For each task $\mathcal{T}$, we use the best evaluation success rate during training, denoted as $s(\mathcal{T})$, as the difficulty signal, and assign

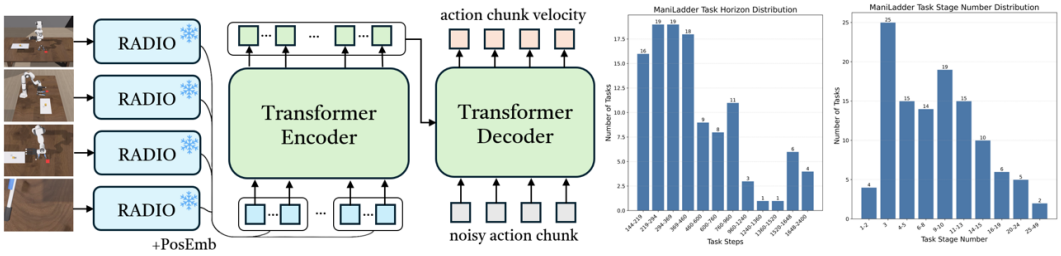

Figure 6: Left: the RADIOFlow architecture. Note, we do not use the state as input. This is intended to ensure fair comparison of tasks across different levels and categories, and to prevent information leakage resulting from specific task states. Right: the statistics of demonstrations in ManiLadder.

tasks to four levels according to:

$$\text{Level}(\mathcal{T}) = \begin{cases} 1, & s(\mathcal{T}) \in [60\%, 85\%], \\ 2, & s(\mathcal{T}) \in [20\%, 60\%), \\ 3, & s(\mathcal{T}) \in (0\%, 20\%), \\ 4, & s(\mathcal{T}) = 0\%. \end{cases}$$

Here, Level 1 contains tasks that are reliably learnable by the reference policy, while Level 4 contains tasks that remain unsolved under the same learning budget.

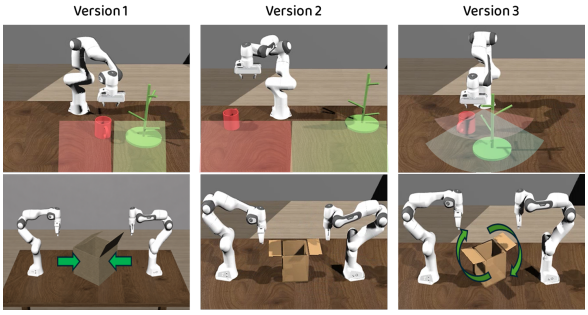

Figure 5: Examples of task calibration and refinement. Top: For the *HangMug* task, we progressively modify the initialization regions of the mug and rack to make the task more challenging. Bottom: For the *DualArm-CloseBox* task, we first reduce the box size, then introduce an initialization pose that requires coordinated dual-arm rotation before the box can be closed.

Demonstrations are collected using two complementary strategies. For contact-rich, dexterous, or hard-to-script tasks, we use human teleoperation with VR. For tasks with well-specified geometric or sequential structure, we use scripted policies to generate demonstrations in a scalable and reproducible manner. We show the statistics of all demonstrations in the right part of Figure 6. We then design a data-efficient imitation learning policy called RADIOFlow with RADIO [63] as the pretrained frozen visual encoder for multi-view RGB images input, and a transformer encoder-decoder architecture for flow-matching [28] action chunk denoising, as shown in Figure 6. Under this policy, we can only use 20 demonstrations for each task and get reasonable success rates as shown in Figure 4. Details are in Appendix A.

If a task does not fall into its intended difficulty interval, we iteratively refine it by modifying the task design, recollecting demonstrations, retraining the reference policy, and reevaluating its success rate, as illustrated in the right part of Figure 1. During this process, we adjust task difficulty through several human-interpretable factors, such as task horizon, spatial error tolerance, object complexity, and spatial-temporal coordination. An example is shown in Figure 5. On average, it requires about four rounds of such a refinement process to make a task calibrated, which shows a large gap between subjective human intuition and empirical difficulty signal under a standardized reference protocol.

3.4 Benchmark Analysis and Lessons

The construction of ManiLadder provides not only an evaluation platform, but also a large-scale empirical lens for studying what makes manipulation tasks difficult. We annotate each task along four task properties that frequently arise in task taxonomy:

1) Temporal score: the horizon of a task. To reduce the bias introduced by relying solely on demonstration horizon length, e.g., slower execution in some demonstrations, we define the temporal score T of a task as a weighted sum of demonstration steps N and task substages M: $T = 10 * \log N + M$;

2) Object complexity $C_{\text{geom}} \in \{1, \ldots, 5\}$, defined by statically inspecting the manipulated objects from tasks, assets, and object metadata. For a task $\mathcal{T}$ with manipulated objects $\mathcal{O}_{\mathcal{T}}$, we use the most complex task-critical object: $C_{\text{geom}}(\mathcal{T}) = \max_{o \in \mathcal{O}_{\mathcal{T}}} c_{\text{geom}}(o)$. The object score is rule-based:

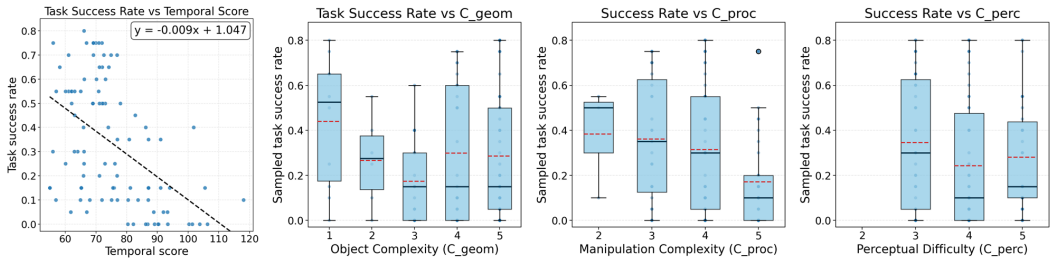

Figure 7: Task success rates with different task properties for all tasks in ManiLadder.

Table 3: Success rates (%) of $\pi_{0.5}$ [6] and increase/decrease values compared to the Diffusion Policy on each task suite in ManiLadder. We use $\mathcal{G}$ for robots with grippers, $\mathcal{X}$ for robots with dexterous hands, $\mathcal{S}$ for single-arm robots, and $\mathcal{D}$ for dual-arm robots.

	Rigid-Body Manipulation				Deformable-Object Manipulation			
	$\mathcal{SG}$	$\mathcal{SX}$	$\mathcal{DG}$	$\mathcal{DX}$	$\mathcal{SG}$	$\mathcal{SX}$	$\mathcal{DG}$	$\mathcal{DX}$
Level 1	41.25%(↓38.75%)	83.75% (↑13.75%)	65.00%(↓5.00%)	72.5%(↓6.25%)	71.67%(↑4.00%)	75.00%(↑1.70%)	40.00%(↓25.00%)	46.67%(↓20.00%)
Level 2	40.00%(↓15.00%)	32.50%(↓27.25%)	26.25%(↓11.25%)	18.33%(↓21.67%)	15.00%(↓10.00%)	26.67%(↓8.33%)	28.33%(↑6.67%)	13.33%(↓25.00%)
Level 3	5.75%(↓5.50%)	10.25%(↑1.50%)	0%(↓8.33%)	1.25%(↓7.5%)	13.33%(↑8.33%)	8.33%(↓3.34%)	5.00%(↓5.00%)	0%(↓7.5%)
Level 4	0%	0%	0%	0%	0%	0%	0%	0%

primitives $= 1$, simple rigid household objects $= 2$, thin/concave/asymmetric rigid objects $= 3$, articulated or multi-part objects $= 4$, and deformable or state-dependent objects $= 5$;

3) Manipulation complexity $C_{\text{proc}}(\mathcal{T}) = 1+B_{\text{stage}}+B_{\text{contact}}+B_{\text{coord}}+B_{\text{prec}}, C_{\text{proc}} \in \{1,\ldots,5\}$, where each binary term indicates whether the task requires multi-stage skill composition, sustained/constrained contact, multi-effector or tool-object coordination, and strict success tolerance, respectively, annotated from task and robot descriptions, success conditions, and coded thresholds;

4) Perception difficulty $C_{\text{perc}}(\mathcal{T}) = 1 + B_{\text{size}} + B_{\text{clutter}} + B_{\text{occlusion}} + B_{\text{hidden}}, \quad B_i \in \{0,1\}$. Here, B_{size} indicates small or thin task-critical objects/contact regions, B_{clutter} indicates distractors or crowded scenes, $B_{\text{occlusion}}$ indicates partially occluded targets or interaction regions, and B_{hidden} indicates visually hidden or ambiguous states such as deformable shape or internal container states. We assign these indicators by inspecting assets, scene layout, camera setup, and success conditions.

Figure 7 shows that task properties are generally predictive of reference-policy performance. The temporal score is negatively correlated with success rate, indicating that longer-horizon tasks are more difficult due to accumulated execution errors. Higher manipulation complexity and perception difficulty also tend to reduce success rates, suggesting that contact-rich procedures, coordination, strict tolerances, occlusion, and hidden states are key sources of difficulty. The variance further shows that task difficulty arises from multiple factors rather than any single property alone.

4 Experiments

In this section, we conduct experiments to demonstrate how ManiLadder can be used to evaluate and analyze robot learning methods from multiple perspectives: **Q1**: How well do state-of-the-art vision-language-action models perform across different levels of manipulation difficulty? **Q2**: How does knowledge transfer across difficulty levels, end-effector embodiments, and object categories? **Q3**: How does policy performance scale with the amount of demonstration data across different task families? and **Q4**: whether the difficulty order calibrated in simulation is preserved in the real world. The results of **Q4** are in the Appendix B.

For **Q1**, we report the success rates of $\pi_{0.5}$ [6] and the magnitude of change relative to RADIOFlow on each suite of ManiLadder, i.e., tasks with the same difficulty levels & object types & embodiments. We use LoRA [64] fine-tuning for $\pi_{0.5}$ [6] for $40K$ steps, with the same number of demonstrations as used to calibrate the task difficulty levels, and report evaluation results over 20 episodes.

Results are shown in Table 3. We observe that, in most cases, $\pi_0.5$ achieves lower success rates than RADIOFlow. This is mainly because $\pi_0.5$ uses only a single third-view camera (front), whereas RADIOFlow uses three views (front, left, and top), which can have a substantial impact on many tasks, especially those where the front camera alone cannot provide sufficient visual information. Notably, although the success rates of $\pi_0.5$ differ from those of RADIOFlow, they largely remain

Table 4: Transfer learning results under different settings. The first three rows are done on rigid-body suites. The values in parentheses on the right indicate the increase or decrease compared with isolated training.

L1 $\mathcal{SG}$ → L2 $\mathcal{SG}$	55.00%(↑15.00%)	L2 $\mathcal{SG}$ → L1 $\mathcal{SG}$	65.00%(↑23.75%)
$\mathcal{SG}+\mathcal{SX}$ → $\mathcal{DG}+\mathcal{DX}$	17.50%(↓4.79%)	$\mathcal{DG}+\mathcal{DX}$ → $\mathcal{SG}+\mathcal{SX}$	5.00%(↓31.25%)
$\mathcal{SG}+\mathcal{DG}$ → $\mathcal{SX}+\mathcal{DX}$	12.50%(↓13.00%)	$\mathcal{SX}+\mathcal{DX}$ → $\mathcal{SG}+\mathcal{DG}$	17.50%(↓15.625%)
Rigid $\mathcal{SG}$ → Deformable $\mathcal{SG}$	12.50%(↓2.5%)	Deformable $\mathcal{SG}$ → Rigid $\mathcal{SG}$	45.00%(↑5.00%)

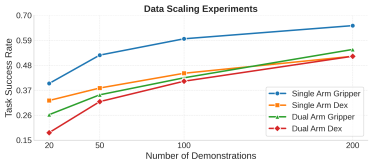

Figure 8: Data scaling experiments on Level 2 rigid body task suites.

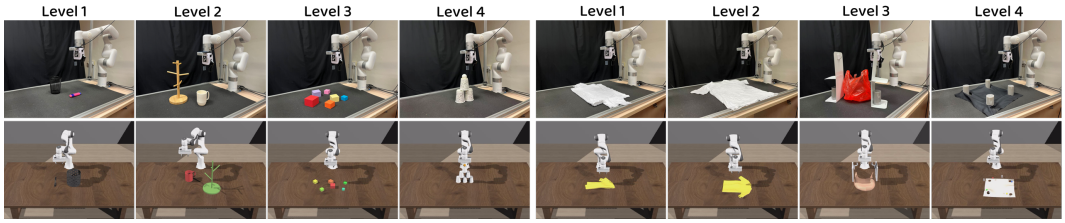

Figure 9: We choose 8 tasks from four different task difficulty levels for real-world experiments: *PickPlacePen*, *HangMug*, *Stack6Blocks*, *PingPongTransport*, *UnfoldShirt*, *FoldShirt*, *BagHanging*, and *BuildTent*, with corresponding simulation tasks at the bottom. Results are in Appendix B.

within the task-level intervals defined in Section 3. This suggests that the calibrated ladder captures difficulty factors that are not entirely specific to the reference policy, while we leave a more exhaustive cross-policy calibration study to future work.

For **Q2**, we report the transfer learning results of $\pi_{0.5}$ [6] on Level 1&2 tasks for experiments. For the transfer learning setting of $\mathcal{X} \rightarrow \mathcal{Y}$, we mean pretrain the model on suite $\mathcal{X}$ first and then finetune on $\mathcal{Y}$, and evaluate the success rates on $\mathcal{Y}$. Results are in Table 4. We can see that transfer learning across embodiments (Row 2 and 3) are not automatically beneficial: when the source and target suites differ substantially in embodiment, the source adaptation can bias the model toward incompatible action priors, leading to negative transfer. Transfer learning across different levels are beneficial (Row 1) bidirectionally, showing that within the same task category, tasks at different levels share reusable manipulation primitives and perceptual-action structures. Transfer across object properties (Row 4) exhibits mixed effects, indicating that rigid-body and deformable-object manipulation are neither fully separated nor universally transferable regimes.

For **Q3**, we report the results of data scaling experiments on four Level 2 task suites. Results are shown in Figure 8. As the number of demonstrations increases, the success rates generally improve, indicating that the tasks are learnable and not saturated under the evaluated data regimes

5 Conclusion and Limitations

We introduced ManiLadder, a difficulty-calibrated simulation benchmark for robotic manipulation that organizes 112 tasks across 4 difficulty levels and diverse task categories. By calibrating task difficulty with the success rate of a standardized imitation-learning policy and iteratively refining task designs, ManiLadder provides a policy-anchored ladder for measuring what level of manipulation difficulty the policy can reliably solve. We demonstrate that ManiLadder can support studies of capability scaling, transfer learning, data scaling, VLA evaluation, and sim-to-real difficulty ordering. We hope ManiLadder will serve as a structured testbed for understanding the boundaries of robot learning methods and guiding future progress toward more capable manipulation systems.

Our work has several limitations: 1) the current difficulty levels are calibrated using a single reference imitation-learning policy as RADIOFlow-anchored empirical learnability, not as absolute policy-independent task difficulty. Maintaining consistent task-success-rate intervals across multiple reference policies is challenging, and this has already been demonstrated by several prior works [34, 65, 44]. We chose the most widely used imitation learning policy method [7] to maximize the generality of the calibration results; 2) we did not involve reinforcement learning results for difficulty-level calibration. This is because reward engineering could be unfair between different tasks, thus the trained RL success rates cannot be fairly comparable across different tasks.

References

[1] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[2] K. Shaw, A. Agarwal, and D. Pathak. Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning. *arXiv preprint arXiv:2309.06440*, 2023.

[3] A. O'Neill, A. Rehman, et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *ICRA*, 2024.

[4] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Hu, X. Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[5] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[6] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, et al. Pi05: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[7] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *IJRR*, 2023.

[8] S. Ye, Y. Ge, K. Zheng, S. Gao, S. Yu, G. Kurian, S. Indupuru, Y. L. Tan, C. Zhu, J. Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

[9] S. AI. Learning by watching: A general-purpose vision-language-action world model for robotics. https://www.skild.ai/blogs/learning-by-watching, 2026.

[10] Sharpa. Sharpa origami demo. https://www.sharpa.com/pages/north, 2025.

[11] G. AI. Genesis ai official website. https://www.genesis.ai/, 2026.

[12] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *NeurIPS*, 2023.

[13] S. James, Z. Ma, D. Rovick Arrojo, and A. J. Davison. Rlbench: The robot learning benchmark & learning environment. *RAL*, 2020.

[14] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

[15] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pages 80–93. PMLR, 2023.

[16] B. R. Donald. On information invariants in robotics. *Artificial Intelligence*, 72(1-2):217–304, 1995.

[17] M. Ho, A. Farid, and A. Majumdar. Towards a framework for comparing the complexity of robotic tasks. In *WAFR*, 2022.

[18] C. Gao, Y. Jiang, and F. Chen. Transferring hierarchical structures with dual meta imitation learning. In *CoRL*. PMLR, 2023.

[19] S. Haresh, D. Dijkman, A. Bhattacharyya, and R. Memisevic. Clevrskills: Compositional language and visual reasoning in robotics. *NeurIPS*, 2024.

[20] C. R. Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling, and T. Lozano-Pérez. Integrated task and motion planning. *ANNU REV CONTR ROBOT*, 4(1):265–293, 2021.

[21] J. D. Morrow and P. K. Khosla. Manipulation task primitives for composing robot skills. In *Proceedings of International Conference on Robotics and Automation*, volume 4, pages 3354–3359. IEEE, 1997.

[22] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic. The grasp taxonomy of human grasp types. *IEEE Transactions on human-machine systems*, 46(1):66–77, 2015.

[23] D. Blanco-Mulero, Y. Dong, J. Borras, F. T. Pokorny, and C. Torras. T-dom: A taxonomy for robotic manipulation of deformable objects. *arXiv preprint arXiv:2412.20998*, 2024.

[24] I. M. Bullock and A. M. Dollar. Classifying human manipulation behavior. In *ICORR*. IEEE, 2011.

[25] G. Authors. Genesis: A universal and generative physics engine for robotics and beyond, December 2024.

[26] M. Mittal, P. Roth, J. Tigue, A. Richard, O. Zhang, P. Du, A. Serrano-Munoz, X. Yao, R. Zurbrügg, N. Rudin, et al. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025.

[27] C. Aeronautiques, A. Howe, C. Knoblock, I. D. McDermott, A. Ram, M. Veloso, D. Weld, D. W. Sri, A. Barrett, D. Christianson, et al. Pddl—the planning domain definition language. *Technical Report, Tech. Rep.*, 1998.

[28] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[29] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *CoRL*, 2020.

[30] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[31] O. Biza, T. Kipf, D. Klee, R. Platt, J.-W. van de Meent, and L. L. Wong. Factored world models for zero-shot generalization in robotic manipulation. *arXiv preprint arXiv:2202.05333*, 2022.

[32] W. Pumacay, I. Singh, J. Duan, R. Krishna, J. Thomason, and D. Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation. *arXiv preprint arXiv:2402.08191*, 2024.

[33] S. James, M. Freese, and A. J. Davison. Pyrep: Bringing v-rep to deep robot learning. *arXiv preprint arXiv:1906.11176*, 2019.

[34] X. Yang, R. Dagli, A. Zook, H. Hadfield, A. Goyal, S. Birchfield, F. Ramos, and J. Tremblay. Robolab: A high-fidelity simulation benchmark for analysis of task generalist policies. *arXiv preprint arXiv:2604.09860*, 2026.

[35] M. R. Cutkosky et al. On grasp choice, grasp models, and the design of hands for manufacturing tasks. *T-RO*, 1989.

[36] J. D. Morrow and P. K. Khosla. Manipulation task primitives for composing robot skills. In *Proceedings of International Conference on Robotics and Automation*, volume 4, pages 3354–3359. IEEE, 1997.

[37] E. Huang. *Robotic Manipulation Primitives*. PhD thesis, Carnegie Mellon University, 2021.

[38] D. Blanco-Mulero, Y. Dong, J. Borras, F. T. Pokorny, and C. Torras. T-dom: A taxonomy for robotic manipulation of deformable objects. *arXiv preprint arXiv:2412.20998*, 2024.

[39] F. Krebs and T. Asfour. A bimanual manipulation taxonomy. *IEEE Robotics and Automation Letters*, 7(4), 2022.

[40] C. Bao, H. Xu, Y. Qin, and X. Wang. Dexart: Benchmarking generalizable dexterous manipulation with articulated objects. In *ICCV*, 2023.

[41] Y. Zhu, J. Wong, A. Mandlekar, R. Martín-Martín, A. Joshi, S. Nasiriany, and Y. Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv:2009.12293*, 2020.

[42] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *RA-L*, 2022.

[43] R. Yang, H. Chen, J. Zhang, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025.

[44] S. Zhang, Z. Xu, P. Liu, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. *arXiv preprint arXiv:2412.18194*, 2024.

[45] M. Heo, Y. Lee, D. Lee, and J. J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. *IJRR*, page 02783649241304789, 2023.

[46] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *ICAR*. IEEE, 2015.

[47] Y. R. Wang, C. Ung, G. Tannert, J. Duan, J. Li, A. Le, R. Oswal, M. Grotz, W. Pumacay, Y. Deng, et al. Roboeval: Where robotic manipulation meets structured and scalable evaluation. *arXiv preprint arXiv:2507.00435*, 2025.

[48] P. Atreya, K. Pertsch, T. Lee, M. J. Kim, A. Jain, A. Kuramshin, C. Eppner, C. Neary, E. Hu, F. Ramos, et al. Roboarena: Distributed real-world evaluation of generalist robot policies. *arXiv*, 2025.

[49] Z. Zhou, P. Atreya, Y. L. Tan, K. Pertsch, and S. Levine. Autoeval: Autonomous evaluation of generalist robot manipulation policies in the real world. *arXiv preprint arXiv:2503.24278*, 2025.

[50] J. Luo, C. Xu, F. Liu, L. Tan, Z. Lin, J. Wu, P. Abbeel, and S. Levine. Fmb: a functional manipulation benchmark for generalizable robotic learning. *IJRR*, 2025.

[51] M. Buehler, K. Iagnemma, and S. Singh. *The DARPA urban challenge: autonomous vehicles in city traffic*, volume 56. Springer Science & Business Media, 2009.

[52] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman. Analysis and observations from the first amazon picking challenge. *TASE*, 2016.

[53] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[54] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *CVPR*, 2019.

[55] Dawson-Haggerty et al. trimesh. URL https://trimesh.org/.

[56] M. Müller, B. Heidelberger, M. Hennix, and J. Ratcliff. Position based dynamics. *Journal of Visual Communication and Image Representation*, 18(2):109–118, 2007. doi:10.1016/j.jvcir.2007.01.005.

[57] D. Sulsky, Z. Chen, and H. L. Schreyer. A particle method for history-dependent materials. *Computer Methods in Applied Mechanics and Engineering*, 118(1–2):179–196, 1994. doi:10.1016/0045-7825(94)90112-0.

[58] R. A. Gingold and J. J. Monaghan. Smoothed particle hydrodynamics: theory and application to non-spherical stars. *Monthly Notices of the Royal Astronomical Society*, 181(3):375–389, 1977. doi:10.1093/mnras/181.3.375.

[59] E. Huang, X. Cheng, Y. Mao, A. Gupta, and M. T. Mason. Autogenerated manipulation primitives. *The International Journal of Robotics Research*, 42(6):433–458, 2023. doi:10.1177/02783649231170897.

[60] P. Zech, E. Renaudo, S. Haller, X. Zhang, and J. Piater. Action representations in robotics: A taxonomy and systematic classification. *The International Journal of Robotics Research*, 38 (5):518–562, 2019. doi:10.1177/0278364919835020.

[61] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic. The GRASP taxonomy of human grasp types. *IEEE Transactions on Human-Machine Systems*, 46(1):66–77, 2016. doi:10.1109/THMS.2015.2470657.

[62] J. Wu, W. Chong, R. Holmberg, A. Prasad, Y. Gao, O. Khatib, S. Song, S. Rusinkiewicz, and J. Bohg. Tidybot++: An open-source holonomic mobile manipulator for robot learning. *arXiv*, 2024.

[63] M. Ranzinger, G. Heinrich, J. Kautz, and P. Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12490–12500, 2024.

[64] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[65] C. Gao, Z. Liu, Z. Chi, and L. Shao. Vla-os: Structuring and dissecting planning representations and paradigms in vision-language-action models. *arXiv preprint arXiv:2506.17561*, 2025.