RoboEval: Where Robotic Manipulation Meets Structured and Scalable Evaluation

Yi Ru Wang[1*] Carter Ung[2] Grant Tannert[1] Jiafei Duan[1]
Josephine Li[1] Amy Le[1] Rishabh Oswal[1] Markus Grotz[1]
Wilbert Pumacay[3] Yuquan Deng[1] Ranjay Krishna[1,3]
Dieter Fox[1†] Siddhartha Srinivasa[1,†]

[1]University of Washington [2]University of Houston [3]Allen Institute for AI
[†]Equal advising

https://robo-eval.github.io

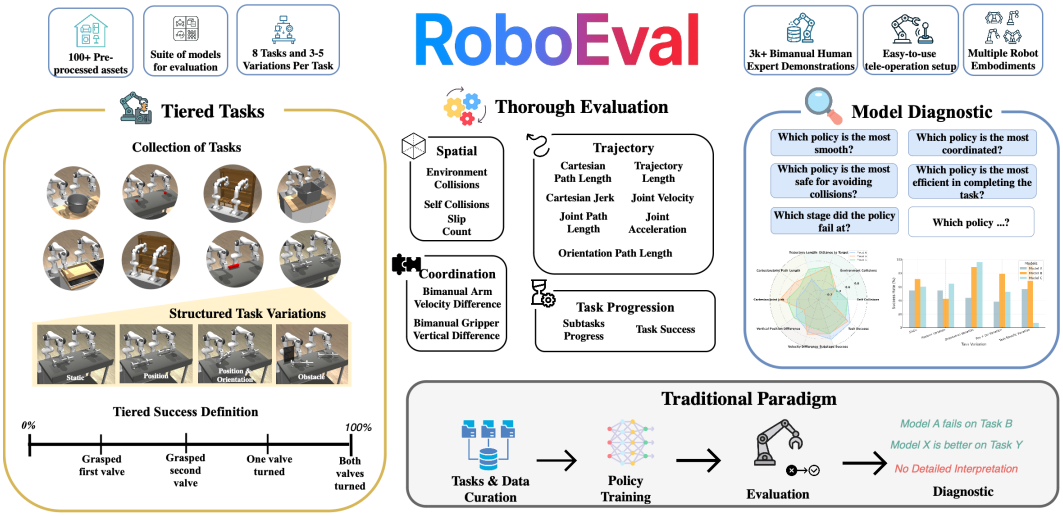


Figure 1: **Overview of RoboEval.** RoboEval is a structured and scalable simulation benchmark for bimanual manipulation, featuring 3,000+ human-collected demonstrations across 8 tasks, each with 3-5 variations. It includes a standardized asset library—collision meshes, annotated sites, and manipulable objects—for building and augmenting tasks with spatial perturbations and distractors. A VR-based teleoperation interface enables realistic data collection. For analysis, RoboEval provides rich evaluation tools that go beyond binary success, measuring task progression, coordination, trajectory efficiency, and spatial proximity.

Abstract: We present **RoboEval**, a simulation benchmark and structured evaluation framework designed to reveal the limitations of current bimanual manipulation policies. While prior benchmarks report only binary task success, we show that such metrics often conceal critical weaknesses in policy behavior—such as poor coordination, slipping during grasping, or asymmetric arm usage. RoboEval introduces a suite of tiered, semantically grounded tasks decomposed into skill-specific stages, with variations that systematically challenge spatial, physical, and coordination capabilities. Tasks are paired with fine-grained diagnostic metrics and 3000+ human demonstrations to support imitation learning. Our experiments reveal that policies with similar success rates diverge in how tasks are executed – some struggle with alignment, others with temporally consistent bimanual control. We find that behavioral metrics correlate with success in over

*yiruwang@cs.washington.edu

half of task-metric pairs, and remain informative even when binary success saturates. By pinpointing *when* and *how* policies fail, RoboEval enables a deeper, more actionable understanding of robotic manipulation – and highlights the need for evaluation tools that go beyond success alone.

Keywords: Benchmarking, Robot Learning, Bimanual Manipulation

1 Introduction

The advancement of general-purpose robotic agents hinges not only on better data and models, but also on systematic tools to evaluate and understand their behavior. Benchmarks in computer vision [1], natural language processing [2], and reinforcement learning [3, 4] have accelerated progress by standardizing task formulations and enabling reproducible comparisons. In robotics, benchmarks such as RLBench [5], Meta-World [6], and BiGym [7] test visuomotor policies across diverse settings. However, these efforts often reduce performance to binary success, offering limited insight into *how* policies behave, *why* they fail, or *what* capabilities they exhibit. Such coarse evaluation is particularly limiting in manipulation, where failure can stem from errors in perception, control, coordination, or temporal reasoning. As tasks grow more complex—spanning multiple stages, arms, and skills—understanding intermediate competencies becomes critical. Diagnostic tools are essential to identify bottlenecks, assess generalization, and inform principled algorithm design.

To address these gaps, we introduce **RoboEval**, a simulation benchmark and evaluation framework for fine-grained analysis of bimanual robotic manipulation. RoboEval features a suite of *tiered manipulation tasks*, each decomposed into semantically grounded stages targeting specific skills such as pushing, grasping, holding, rotating, lifting, etc. RoboEval is grounded in three core design principles. First, it emphasizes *structured complexity* through a hierarchical task organization with controlled variations in spatial layout, coordination, and object properties. Second, it enables *diagnostic interpretability* via outcome metrics that capture stagewise progress and task success, and behaviour metrics that capture spatial and temporal precision, trajectory properties, and bimanual coordination. Third, it supports *realistic supervision* through 3000+ human-collected demonstrations, enabling imitation and data-driven learning from expert behavior.

Through extensive experiments with state-of-the-art visuomotor policies, we show that RoboEval uncovers behavioral and structural differences that binary success alone fails to capture. Behavioral metrics correlate significantly with success in 59.4% of task-metric combinations, indicating that coordination quality, trajectory smoothness, and spatial precision are predictive of policy effectiveness in the majority of tasks. Even when policies achieve similar success rates, behavioral metrics reveal meaningful distinctions in how tasks are executed. Outcome metrics further expose structured failure modes: some policies consistently fail at specific substages—such as lifting or coordinated bimanual actions—and exhibit asymmetric failure patterns across arms. We also find that task difficulty modulates metric utility: binary success becomes uninformative for very easy or very hard tasks, whereas behavioral and stagewise metrics remain diagnostic, enabling a more nuanced evaluation of policy capabilities.

Our contributions are threefold. (1) We introduce **RoboEval**, a benchmark for dissecting manipulation capabilities via structured, skill-targeted tasks. (2) We propose **fine-grained evaluation metrics** for analyzing intermediate progress and coordination. (3) We release a **modular, extensible simulation framework** that supports reproducible research across imitation, reinforcement, and hybrid learning paradigms. Together, these components shift evaluation from binary outcomes to nuanced, skill-level understanding of robot behavior.

2 Related Works

Benchmarks for Robotic Manipulation. Significant advances have been made in benchmarking single-arm manipulation [8, 9, 5, 10, 11]. HumanoidBench [12] extends beyond bimanual manip-

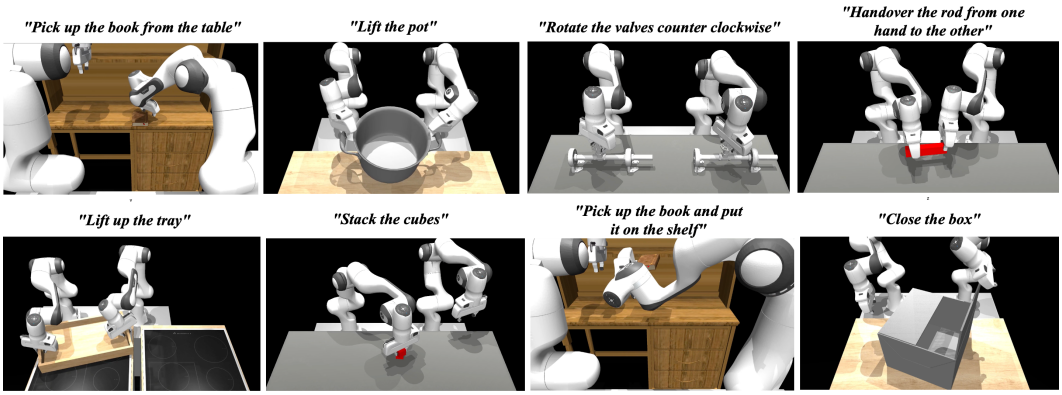


Figure 2: **Base tasks in RoboEval.** RoboEval introduces an initial suite of 8 bimanual manipulation tasks, each accompanied by 3–5 structured variations and over 500 human demonstrations. All tasks are instrumented with behavior metric logging and task-stage definitions to support fine-grained progress and outcome analysis. The benchmark is modular by design, allowing for seamless integration of new tasks to accommodate evolving research needs within the community.

ulation by benchmarking reinforcement learning algorithms for dexterous, whole-body humanoid manipulation. Other studies benchmark RL policies for dexterous hand use [13]. Meanwhile, some works focus on evaluation protocols in real-world scenarios, but these are often limited to single tasks, such as shoe lacing [14], and lack task diversity. Recent efforts such as Peract2 [15], BiGym [7], and RoboTwin [16] have moved toward scalable and data-rich bimanual manipulation. Peract2 uses scripted demonstrations, BiGym explores VR-based demonstrations, and RoboTwin introduces synthetic data generation via 3D generative models and LLMs. However, these approaches still fall short in systematically characterizing when, where, and why coordinated bimanual manipulation policies fail. Our work complements the existing efforts in benchmarking for bi-manual manipulation by providing a means for structured evaluation, and a path forward for unifying existing benchmarking tasks into a shared framework. We provide a comparison with existing benchmarks in Figure 1.

Evaluation Metrics for Manipulation. Robust evaluation metrics are fundamental for quantifying the capabilities of robotic manipulation policies, particularly as tasks increase in complexity, realism, and variability. A broad overview of evaluation methodologies for robotic grasping and manipulation is provided in [17], encompassing metrics such as binary success rate, task completion accuracy, spatial and temporal precision, and robustness to environmental perturbations. Evaluation frameworks for navigation agents, such as [18], further emphasize the need for domain-specific, fine-grained metrics. To assess generalization, RB2 [19] benchmarks performance across distinct physical environments, highlighting policy robustness under varying lab conditions. Similarly, the Colosseum Benchmark [10] evaluates manipulation methods under controlled perturbations to quantify generalization to unseen states and dynamics. However, despite progress in single-arm settings, the field lacks principled and fine-grained evaluation frameworks tailored to bimanual manipulation.

3 RoboEval Benchmark

RoboEval is a benchmark for evaluating bi-manual manipulation policies under diverse task settings. The first iteration consists of 8 base tasks and 3000+ human demonstrations. The tasks are derived from common tasks that humans perform in diverse settings, from service style tasks such as lifting a tray, to warehouse tasks like closing a box, to industrial tasks like rotating hand-wheels. Each task includes multiple variations—ranging from static setups to dynamic shifts in object pose and semantic context—designed to assess policy performance in a systematic manner. To facilitate research in imitation learning and demo-driven policy training, we provide a suite of raw expert human demonstrations, along with fine-grained evaluation metrics such as trajectory smoothness, environment collisions, etc. We provide an overview of RoboEval in Figure 1. In this section, we

Table 1: **Benchmark Comparison.** We compare six manipulation benchmarks across task design, evaluation, and data. **RoboEval** uniquely integrates tiered bimanual tasks, behavior metrics, task progression tracking, and human demonstrations.

Benchmark	**Task Features**				**Evaluation Features**			**Data Features**		
	Horizon	Tiered	Skills	Variations	Success	Behavior Metrics	Task Prog. Metrics	Human Demo	Demo-Driven	# Expert Human Demos
RLBench	Short–Med	✗	U, P, NP, QS	P, R	✓	✗	✗	✗	✓	0
Bigym	Short–Long	✗	B, P, QS	P, R	✓	✗	✗	✓	✓	2k
DexMimicGen	Short	✗	B, P, QS	P, R	✓	✗	✗	✓	✓	400
PerAct2	Short–Med	✗	B, P, NP, QS	P, R	✓	✗	✗	✗	✓	0
HumanoidBench	Short–Med	✗	U, B, P, NP, QS	–	✓	✗	✗	✗	✗	0
RoboTwin	Short–Med	✗	U, B, P, NP, QS	P, R, S	✓	✗	✗	✓	✓	300
Taskverse (Ours)	Short–Long	✓	U, B, P, NP, QS	P, R, Q, O	✓	✓	✓	✓	✓	3k+

Skills Legend: **U** = Unimanual, **B** = Bimanual, **P** = Prehensile, **NP** = Non-Prehensile, **QS** = Quasi-Static
Variation Legend: **P** = Position, **R** = Rotation, **S** = Size, **Q** = Quantity, **O** = Obstacles.

describe the design philosophy of the benchmark (Section 3.1), base tasks offered by the benchmark (Section 3.2), task and dataset statistics (Section 3.3), evaluation scoring (Section 3.4).

3.1 Design Philosophy

The goal of RoboEval is to serve as a comprehensive benchmark for evaluating learning-based bimanual manipulation. Its design is grounded in three core principles: **Diversity.** Real-world bimanual manipulation spans a broad spectrum of task styles, object geometries, and control challenges. RoboEval captures this diversity by including tasks with varying temporal complexity, coordination requirements, and semantic content—from non-prehensile pushing to tightly coupled lifting and handover behaviors. This ensures that policies are evaluated not only on isolated primitives but on their generality across manipulation. **Interpretability.** Traditional binary success metrics offer limited insight into policy behavior. RoboEval supports structured, fine-grained analysis through multi-dimensional evaluation metrics. These metrics enable deeper understanding of policy execution, identifying failure modes, behavior under variation, and qualitative differences across learning methods. **Extensibility.** The benchmark is designed for future-proof flexibility. Tasks, variation schemes, and evaluation protocols are modular and easily extensible. Researchers can modify existing tasks, create new ones, or integrate different embodiments and sensing modalities, making RoboEval adaptable to emerging research directions.

3.2 Tasks

Tasks in RoboEval are designed to span diverse settings and skill requirements, providing a systematic testbed for evaluating robotic manipulation capabilities. Each task is structured as a goal-conditioned episode with clearly defined success criteria and consists of object interaction in semantically grounded environments such as household, industrial, or tabletop settings. The task set includes both short-horizon objectives (e.g., object lifting, etc.) and long-horizon, multi-step tasks (e.g., clean up a desk by placing a book onto the bookshelf, etc.) with stage-wise progress checking.

Task Definition. Each task in RoboEval is defined by the tuple $\mathcal{T} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{G}, \rho_0, \mathcal{S}_{\text{success}})$. The state space $\mathcal{S}$ includes robot joint states, object poses, and environmental context; the action space $\mathcal{A}$ consists of continuous control inputs such as joint positions and end-effector delta displacements; and $\mathcal{P}$ denotes the transition dynamics governed by a physics simulator. The goal space $\mathcal{G}$ specifies the intended outcome of the task, while the success set $\mathcal{S}_{\text{success}} \subset \mathcal{S}$ defines binary completion based on thresholded geometric conditions (e.g., object pose alignment or contact). Tasks are initialized by sampling from an initial state distribution ρ_0. Agents in RoboEval learn from a dataset of expert demonstrations $\mathcal{D}_{\mathcal{T}} = \{(s_0, a_0, \ldots, s_T)\}$, collected via human teleoperation. To support fine-grained analysis, each task is instantiated with a parameterized family of variants $\mathcal{T}_\theta$ where $\theta \in \Theta$ modulates scene layout or semantic content.

Skill Diversity. Table 2 summarizes the taxonomy of skills required across tasks. Following the bimanual taxonomy in [20], we categorize tasks into coordination classes: *unimanual*, *bimanual uncoordinated*, *loosely coordinated*, *tightly coordinated symmetric*, and *tightly coordinated asymmetric*. Tasks span a wide range of motor and coordination demands, including single-axis control

Table 2: **Base Task Set in RoboEval.** We summarize the base tasks with their variation types, demonstration statistics, skill categories, and coordination structure. Variation types include static setups, spatial perturbations in position (Pos), rotation (Rot), combined (PR), and task-specific variants. Coordination structures span uncoordinated, loosely coordinated, and symmetric behaviors.

Task Name	Variations	# Demos	Traj Len	Skills	Coordination Type
Cube Handover	Static, Pos, Rot, PR, Vertical	511	93.631	grasp, hold	Loosely Coord.
Lift Pot	Static, Pos, Rot , PR	390	58.561	grasp, lift	Tight Sym.
Lift Tray	Static, Pos, Rot, PR, Drag	730	77.318	grasp, lift	Tight Sym.
Pack Box	Static, Pos, Rot, PR	312	123.016	push	Uncoord.
Pick Single Book From Table	Static, Pos, Rot, PR	359	103.364	grasp, lift	Loosely Coord.
Rotate Valve	Static, Pos, Rot, PR	456	112.484	grasp, rotate along axis	Uncoord.
Stack Single Book Shelf	Static, Pos, PR	199	187.280	push, grasp, lift, place	Loosely Coord.
Stack Two Block	Static, Pos, Rot, PR	400	108.368	grasp, hold, place	Loosely Coord.

(e.g., turning a valve), long-horizon motion (e.g., packing a box), high-precision alignment (e.g., inserting toast into a toaster), and synchronized dual-arm lifting (e.g., lifting and balancing a tray), etc. The benchmark is designed to probe capabilities in coordination, precision, and smooth execution trajectories across these axes.

Task Variations. The initial release of RoboEval includes a curated set of bimanual manipulation tasks with structured variations designed to probe robustness and generalization. Specifically, we introduce spatial perturbations—such as changes in object position and orientation—as well as physical obstacles that alter the geometry of the workspace. These variations challenge visuomotor policies to adapt their coordination strategies while preserving task semantics. Future extensions of RoboEval can build on this foundation to introduce additional variation modalities, including visual distractors, lighting changes, and perturbations of objects' physical properties.

Task Design. RoboEval features a modular task generation pipeline that facilitates efficient authoring and integration of new or external tasks with minimal code overhead. Its unified interface and built-in behavioral and outcome metrics enable principled evaluation and support the development of generalist manipulation policies. As summarized in Table 1, RoboEval distinguishes itself from existing benchmarks by offering tiered task variations, fine-grained behavioral and outcome metrics, and a large-scale repository of expert human demonstrations—providing a comprehensive platform for benchmarking bimanual manipulation.

3.3 Task and Dataset Statistics

In total, RoboEval introduces over 3,000 high-quality human expert demonstrations for bimanual manipulation, making it one of the largest collections of natural teleoperated bimanual demonstrations. These demonstrations were collected using a VR-based teleoperation system, enabling precise and dexterous control over dual-arm manipulators in diverse scenarios. Table 2 provides a breakdown of the task categories, associated variation schemes, and the number of demonstrations per task. The initial task suite in RoboEval spans core manipulation skills—including prehensile actions such as grasping and lifting, as well as non-prehensile strategies like pushing. These tasks are further characterized by varying spatial complexities, and task-specific variations, such as obstacles. Due to the natural variability inherent in human demonstrations, the dataset exhibits significant diversity in execution strategies, motion trajectories, and coordination styles. This variability is critical for robust learning and generalization. Importantly, RoboEval not only offers scale, but also supports fine-grained analysis by capturing rich multimodal signals—including proprioception, visual observations, and scene-annotated interaction states—enabling detailed diagnostics of policy behavior across spatial, temporal, and coordination axes.

3.4 Evaluation Scoring

We introduce four classes of metrics to systematically evaluate policy performance, encompassing both behavioral quality and task-level outcomes. Behavioral metrics are grouped into three axes: *trajectory*, *spatial*, and *coordination*. Outcome-driven metrics include *task progression* and *binary task success*.

Trajectory-Based Metrics. We compute *joint path length* and *Cartesian path length* as the cumulative displacement along the trajectory:

$$\mathcal{L}_{\text{joint}} = \sum_{t=1}^{T-1} \|q_{t+1} - q_t\|_2, \quad \mathcal{L}_{\text{cart}} = \sum_{t=1}^{T-1} \|x_{t+1} - x_t\|_2, \tag{1}$$

where q_t denotes the joint configuration and x_t the Cartesian end-effector position at timestep t. We also compute *joint jerk* and *Cartesian jerk*, defined as the average norm of the third-order finite difference of the trajectory, normalized by the control timestep Δt:

$$\text{Jerk}_{\text{joint}} = \frac{1}{T-3} \sum_{t=1}^{T-3} \left\| \frac{q_{t+3} - 3q_{t+2} + 3q_{t+1} - q_t}{(\Delta t)^3} \right\|_2, \quad \text{Jerk}_{\text{cart}} = \frac{1}{T-3} \sum_{t=1}^{T-3} \left\| \frac{x_{t+3} - 3x_{t+2} + 3x_{t+1} - x_t}{(\Delta t)^3} \right\|_2. \tag{2}$$

Spatial Metrics. To evaluate physical interaction quality and environmental safety, we monitor three key indicators: the number of *self-collisions* (contacts between the robot's own links), *environment collisions* (contacts with fixed scene elements such as tables or walls), and *object slips* (instances where a grasped object unintentionally changes state from contact to no contact relative to the gripper). These metrics reflect spatial precision, contact stability, and control reliability. High values may indicate poor trajectory execution or unstable grasping.

Coordination and Bimanual Metrics. Effective bimanual manipulation requires both spatial alignment and temporal synchronization between arms. Let $x_t^{(L)}, x_t^{(R)} \in \mathbb{R}^3$ denote the Cartesian positions of the left and right end-effectors at timestep t, and Δt the control interval.

(1) Height Discrepancy. We compute the mean absolute difference in the vertical (z-axis) positions:

$$\Delta z = \frac{1}{T} \sum_{t=1}^{T} \left| x_t^{(L)}[z] - x_t^{(R)}[z] \right|. \tag{3}$$

(2) Velocity Divergence. Let $v_t^{(L)} = \frac{x_{t+1}^{(L)} - x_t^{(L)}}{\Delta t}$ and $v_t^{(R)} = \frac{x_{t+1}^{(R)} - x_t^{(R)}}{\Delta t}$. We define:

$$\Delta v = \frac{1}{T-1} \sum_{t=1}^{T-1} \left\| v_t^{(L)} - v_t^{(R)} \right\|_2. \tag{4}$$

Lower values of Δz and Δv indicate better spatial and temporal coordination, respectively.

Task Progression and Outcome Metrics. We log *stage-wise success indicators* as binary flags corresponding to discrete phases of the task. Overall task success is measured as the proportion of episodes that achieve completion across evaluation rollouts.

4 Experiments

To validate our benchmark design and metric framework, we conduct experiments aimed at answering three core research questions. **RQ1** investigates how behavioral metrics complement the information provided by policy success rates. **RQ2** examines how outcome metrics reveal failure modes that reflect policy limitations or task bottlenecks. **RQ3** explores how task difficulty affects the informativeness of evaluation metrics. Our experiments evaluate both behavioral metrics—spatial, trajectory, and coordination—and outcome metrics such as task and substage success, across a range of manipulation tasks with varying difficulty. The analysis is structured to address each research question in turn, highlighting how multifaceted metrics and progressively challenging tasks together enable meaningful policy evaluation.

4.1 Experimental Setup

Models. We evaluate four models: ACT [21], Diffusion Policy [22], Behavior Cloning (BC), and OpenVLA [23]. ACT and Diffusion Policy follow their official implementations, with ResNet-18 visual encoders and autoregressive action prediction over a horizon of 16. BC uses a lightweight

Table 3: **Performance on Bimanual Tasks with Variations.** Success rates ($\mu \pm$ SE) for representative tasks under static, positional, orientational, compound, and task-specific perturbations. We compare OpenVLA [23], ACT [21], Diffusion Policy [22], and Behavior Cloning (BC).

Method	**Overall Metrics**			**Lift Tray**					**Stack Two Cubes**			
	Success	Rank	SPL	Static	Pos	Ori	P+O	T	Static	Pos	Ori	P+O
ACT	0.397 ± 0.010	1.15	0.290	1.00 ± 0.00	0.57 ± 0.06	0.76 ± 0.05	0.84 ± 0.04	0.32 ± 0.05	0.00 ± 0.00	0.08 ± 0.03	0.09 ± 0.03	0.27 ± 0.05
BC	0.090 ± 0.006	3.06	0.067	0.67 ± 0.05	0.16 ± 0.04	0.71 ± 0.05	0.16 ± 0.04	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
DIFFUSION	0.211 ± 0.008	2.24	0.161	0.67 ± 0.05	0.07 ± 0.03	0.63 ± 0.06	0.28 ± 0.05	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.07 ± 0.03
OPENVLA	0.116 ± 0.007	2.30	0.047	0.48 ± 0.07	0.22 ± 0.06	0.66 ± 0.07	0.38 ± 0.07	0.02 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.04 ± 0.03
Method	**Stack Single Book Shelf**			**Rod Handover**					**Lift Pot**			
	Static	Pos	P+O	Static	Pos	Ori	P+O	P+O+T	Static	Pos	Ori	P+O
ACT	0.21 ± 0.05	0.15 ± 0.04	0.04 ± 0.02	0.19 ± 0.05	0.85 ± 0.04	0.64 ± 0.06	0.27 ± 0.05	0.55 ± 0.06	1.00 ± 0.00	0.63 ± 0.06	0.60 ± 0.06	0.21 ± 0.05
BC	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.12 ± 0.04	0.00 ± 0.00	0.05 ± 0.03	0.00 ± 0.00
DIFFUSION	0.01 ± 0.01	0.00 ± 0.00	0.01 ± 0.01	0.05 ± 0.03	0.00 ± 0.00	0.07 ± 0.03	0.00 ± 0.00	0.00 ± 0.00	0.91 ± 0.03	0.09 ± 0.03	0.92 ± 0.03	0.16 ± 0.04
OPENVLA	0.00 ± 0.00	0.02 ± 0.02	0.00 ± 0.00	0.50 ± 0.07	0.10 ± 0.04	0.04 ± 0.03	0.06 ± 0.03	0.16 ± 0.05	0.14 ± 0.05	0.12 ± 0.05	0.06 ± 0.03	0.04 ± 0.03
Method	**Pack Box**				**Pick Book from Table**				**Rotate Valve**			
	Static	Pos	Ori	P+O	Static	Pos	Ori	P+O	Static	Pos	P+O	T
ACT	0.00 ± 0.00	0.80 ± 0.05	0.19 ± 0.05	0.25 ± 0.05	0.19 ± 0.05	0.27 ± 0.05	0.28 ± 0.05	0.37 ± 0.06	1.00 ± 0.00	0.00 ± 0.00	0.15 ± 0.04	0.07 ± 0.03
BC	0.00 ± 0.00	0.11 ± 0.04	0.00 ± 0.00	0.05 ± 0.03	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
DIFFUSION	0.00 ± 0.00	0.73 ± 0.05	0.16 ± 0.04	0.39 ± 0.06	0.35 ± 0.06	0.04 ± 0.02	0.17 ± 0.04	0.03 ± 0.02	0.97 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.04 ± 0.02
OPENVLA	0.24 ± 0.06	0.12 ± 0.05	0.02 ± 0.02	0.06 ± 0.03	0.02 ± 0.02	0.00 ± 0.00	0.02 ± 0.02	0.00 ± 0.00	0.58 ± 0.07	0.00 ± 0.00	0.00 ± 0.00	0.04 ± 0.03

convolutional encoder and MLP policy head. OpenVLA is fine-tuned from a pretrained `openvla-7b` checkpoint using LoRA on our task-specific data. All models are trained using the Adam optimizer with consistent hyperparameters unless otherwise noted. Full architectural and training details are provided in the Appendix.

Tasks and Variations. Our main experiments focus on 8 tasks (Table 3), each with 3–5 variations involving spatial or task-specific changes. *Static* denotes minimal scene changes, while *position* and *orientation* introduce spatial perturbations. Some tasks include unique variations—for example, `Lift Tray` features a drag-and-lift setup where the tray begins outside the bimanual workspace, requiring one arm to reposition it before lifting. `Rotate Valve` includes an obstacle blocking direct access to the valve, mimicking obstructed real-world settings.

4.2 Experimental Results

We organize our findings around three central questions: (1) *How do behavioral metrics complement the information provided by policy success rates?* (2) *How do outcome metrics reveal structured failure modes that reflect policy limitations or task bottlenecks?* (3) *How does task difficulty influence the informativeness of evaluation metrics?*

4.2.1 RQ1: How do behavioral metrics complement the information provided by policy success rates?

Behavioral metrics exhibit statistically significant correlations with success in 59.4% of task-metric pairs. Figure 3 presents a heatmap of point-biserial correlation coefficients between behavioral metrics and binary task success, where colored cells indicate statistically significant correlations ($p \leq 0.05$). We observe that 59.4% of task-metric combinations yield significant correlations, suggesting that behavioral metrics are predictive of success for a majority of tasks. In cases where no statistically significant correlation is observed, potential factors include limited samples of successful trajectories or the presence of multimodal solution strategies that result in high variability in behavioral metrics despite similar success outcomes.

Different tasks rely on distinct behavioral metrics to explain success, reflecting task-specific demands. Analyzing the correlation heatmap in Figure 3, we observe that while some metrics exhibit consistent correlation directions across multiple tasks, others vary substantially in both strength and sign depending on the task. This variability indicates that different tasks emphasize different behavioral capacities—such as coordination, efficiency, or stability—to achieve success. Diverging correlation patterns across tasks suggest that no single behavioral metric universally explains success; instead, task-specific demands shape which aspects of behavior are most predictive. This highlights the importance of multi-metric evaluation in understanding policy performance across diverse manipulation tasks.

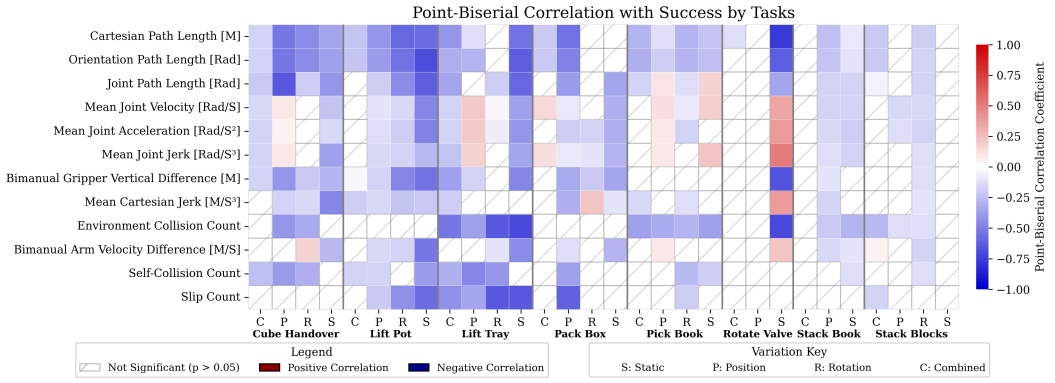
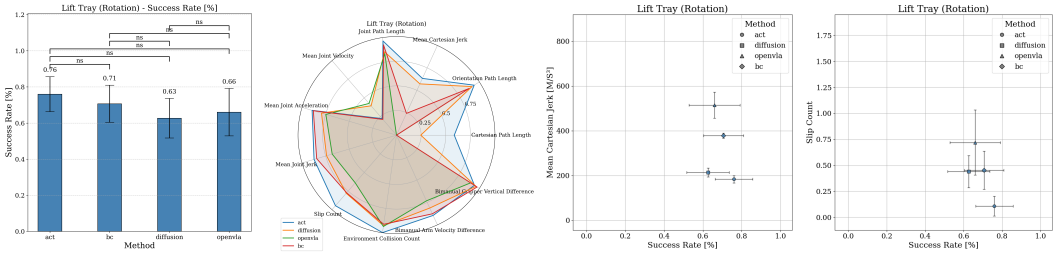


Figure 3: **Point-Biserial Correlation Between Behavioral Metrics and Trajectory Success.** We compute the point-biserial correlation between each behavioral metric and binary trajectory success across different task variations, highlighting only statistically significant correlations. Rows are sorted by the number of significant correlations per metric (descending), placing metrics most consistently associated with success at the top. Overall, 59.4% of metric-task pairs show statistically significant correlation, indicating that behavioral metrics are meaningfully related to success in the majority of tasks.



(a) (b) (c) (d)

Figure 4: **Behavioral metrics differentiate policies with similar success rates.** (a) Bar plot of success rates for the `Lift Tray (Rotation)` task, where no statistically significant differences are observed across policies. (b) Radial plot comparing policies along multiple behavioral metric dimensions, with values normalized and polarity-adjusted to fall within $[0, 1]$ such that higher values indicate better performance. (c) Scatter plot showing the mean Cartesian jerk per policy. (d) Scatter plot showing the mean slip count per policy. Error bars represent 95% confidence intervals; 'ns' denotes comparisons with $p > 0.05$.

Behavioral metrics differentiate policy performance even when success rates are similar. In some cases, point-biserial correlation between behavioral metrics and success is not statistically significant, particularly when policies achieve comparable success rates. For example, in the task `Lift Tray (Rotation)` (Figure 4a), all methods achieve similar success, leading to weak correlation signals. However, as shown in Figure 4b, behavioral metrics provide meaningful distinctions between policies. Using a normalized and polarity-adjusted radial plot, we observe that ACT spans the largest area, indicating overall stronger behavioral performance. Specifically, both ACT and Diffusion exhibit low mean Cartesian jerk, suggesting smoother motion, while ACT further achieves a lower slip count relative to other baselines. These differences, not captured by success rate alone, demonstrate the value of behavioral metrics for revealing nuanced policy capabilities.

4.2.2 RQ2: How do outcome metrics reveal structured failure modes that reflect policy limitations or task bottlenecks?

Outcome metrics decompose failure modes and reveal model-specific strengths and weaknesses. In Figure 5, we visualize the stagewise failure breakdown for six representative tasks using outcome metrics. These breakdowns reveal that ACT tends to succeed in early stages of manipula-

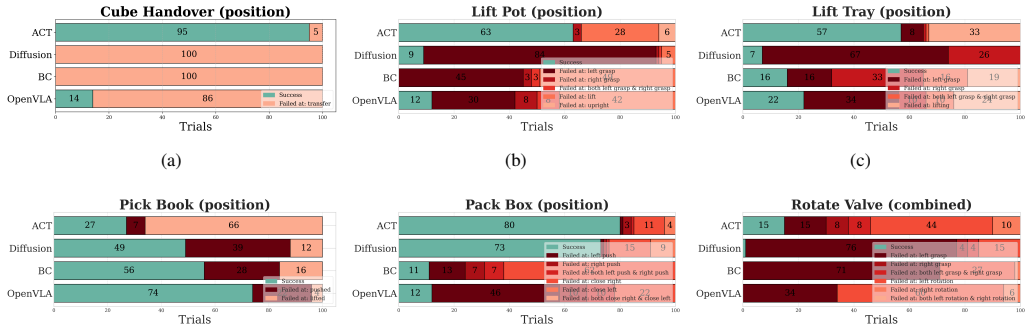


Figure 5: **Failure mode visualizations for six representative tasks.** (a) *Cube Handover*: failures concentrate in the transfer phase. (b) *Lift Pot*: most fail at the left-handle grasp. (c) *Stack Blocks*: errors arise during the second block grasp. (d) *Pick Book*: pushing fails for most, while ACT fails at the lift despite successful pushing. (e) *Pack Box*: BC/OpenVLA fail to contact the lid; ACT/Diffusion fail to close it. (f) *Rotate Valve*: failures occur at the left grasp and rotation.

tion—such as reaching and grasping—but struggles in later stages that require sustained control, including rotation, lifting, or coordinated bimanual motion. For other policies, we observe asymmetric failure patterns, particularly a higher incidence of failure in left-arm actions compared to right-arm ones. This trend is notable in tasks such as `Lift Pot (position)`, `Lift Tray (position)`, `Pack Box (position)`, and `Rotate Valve (combined)`, suggesting that some methods may suffer from imbalanced data quality or suboptimal policy representation for dual-arm coordination. These stagewise diagnostics provide interpretable insights into policy behavior beyond what aggregate success metrics can reveal.

Failures are not uniformly distributed across task stages; certain steps consistently dominate as failure points. In Figure 6, we observe that specific stages within a task account for a disproportionate number of failures across rollouts. This indicates that certain actions or transitions are more challenging for policies to execute reliably. The concentration of failures in these stages suggests they represent bottlenecks in task execution and may benefit from targeted data augmentation or curriculum learning focused on these subcomponents. Such stagewise insights can guide more efficient data collection and policy refinement strategies.

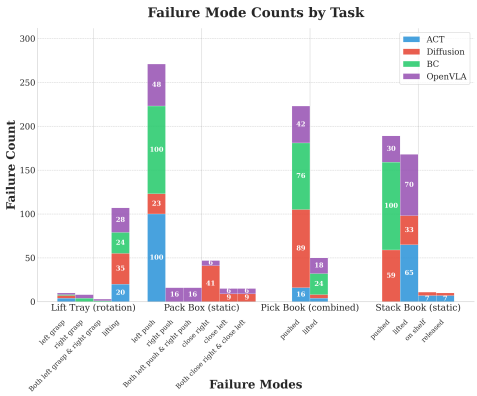
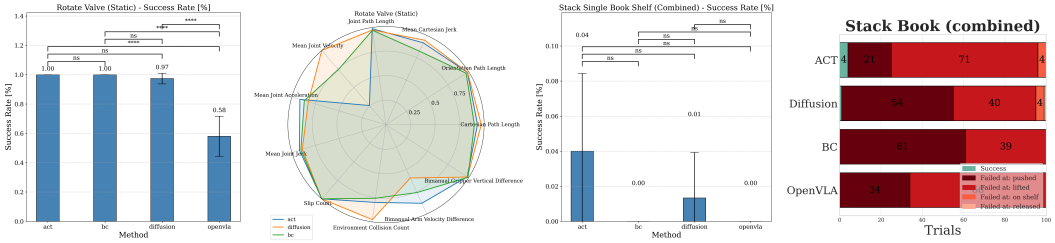


Figure 6: **Examples of tasks with dominant failure modes.** We visualize the total failure counts for each failure stage, aggregated across all baseline policy rollouts, for four representative tasks. Each task exhibits dominant failure modes, indicating that specific stages within the task are consistently more challenging. These concentrated failure patterns highlight bottlenecks in task execution that may benefit from focused analysis or targeted intervention during training.

4.2.3 RQ3: How does task difficulty influence the informativeness of evaluation metrics on policy performance?

Binary success is insufficient for evaluating policy performance on overly easy or difficult tasks. Table 3 presents average success rates across a set of base tasks for multiple policies. For tasks such as `Rotate Valve (static)` and `Lift Pot (static)`, nearly all policies achieve perfect success, offering limited insight into relative policy performance. Conversely, for more challenging tasks, all policies fail, again preventing meaningful comparison. In both cases, binary success becomes uninformative due to lack of variation. To address this, we introduce a tiered task variation framework that systematically adjusts structural complexity to control task difficulty. As complexity



(a) (b) (c) (d)

Figure 7: **Behavioral and outcome metrics provide complementary insights across task difficulties.** (a) Success rates for an easy task (`Rotate Valve (static)`) show ceiling effects, masking performance differences. (b) Behavioral metrics reveal Diffusion Policy's superior motion quality despite identical success. (c) In a hard task (`Stack Single Book Shelf (combined)`), uniformly low success rates offer little insight. (d) Stagewise failure analysis highlights push-to-edge and lifting stages as key bottlenecks, exposing specific policy shortcomings.

increases, we observe performance degradation across policies, revealing differences in robustness and generalization. Similarly, simplifying initially difficult tasks leads to non-zero success rates, enabling more nuanced analysis. This approach allows for more effective benchmarking by ensuring tasks fall within a difficulty range that separates policy capabilities.

Behavioral and outcome metrics provide complementary insights across the spectrum of task difficulty. For easy tasks where most policies achieve perfect or near-perfect success, binary success offers limited evaluative power. In these cases, behavioral metrics can reveal meaningful differences in policy quality. As illustrated in Figure 7, although three policies achieve similar success rates on an easy task, ACT demonstrates superior performance across multiple behavioral dimensions, indicating smoother and more stable execution. On the other end of the spectrum, in difficult tasks where all policies fail, success alone fails to convey progress. However, outcome metrics—such as stagewise task progression—highlight where failures most commonly occur. These insights can help pinpoint specific policy weaknesses and inform targeted improvements in policies.

5 Discussion

We introduced **RoboEval**, a diagnostic benchmark for bimanual manipulation that combines structured task variations, human-collected demonstrations, and fine-grained evaluation metrics. Our analysis reveals that task difficulty arises from long horizons, multimodal strategies, and coordination demands—factors that binary success rates fail to capture. By incorporating trajectory dynamics, spatial precision, and coordination metrics, RoboEval enables principled dissection of policy behavior across tasks and variation regimes. Future extensions will incorporate additional variation modalities, sim-to-real validation, and hosting a public benchmark suite with reproducible evaluation pipelines.

6 Limitations

Despite its diagnostic capabilities, RoboEval has several limitations. First, as a simulation-based benchmark, it is subject to physics artifacts such as unstable contacts or unmodeled dynamics unless parameters are carefully tuned—limiting direct transfer to hardware. Second, task scalability is constrained by the need for manually curated assets and environment setups; future integration of generative models for procedural asset generation may alleviate this bottleneck. Third, while the benchmark provides a large human demonstration dataset, data collection remains expensive. Scaling to broader task coverage may require leveraging pretrained visuomotor models or large-scale unlabeled interaction data to reduce reliance on expert demonstrations.

Acknowledgments

The authors would like to thank members of the Personal Robotics Lab (PRL) and Robotics and State Estimation Lab (RSELab) for fruitful discussions and insightful feedback on the manuscript. Yi Ru Wang is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). This work was (partially) funded by grants from the National Science Foundation NRI (#2132848), DARPA RACER (#HR0011-21-C-0171), and the Office of Naval Research (#N00014-24-S-B001 and #2022-016-01 UW). We gratefully acknowledge gifts from Amazon, Collaborative Robotics, Cruise, and others.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[2] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text, 2016. URL https://arxiv.org/abs/1606.05250.

[3] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym, 2016. URL https://arxiv.org/abs/1606.01540.

[4] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. de Las Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, T. Lillicrap, and M. Riedmiller. Deepmind control suite, 2018. URL https://arxiv.org/abs/1801.00690.

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

[6] 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 *Conference on robot learning*, pages 1094–1100. PMLR, 2020.

[7] N. Chernyadev, N. Backshall, X. Ma, Y. Lu, Y. Seo, and S. James. Bigym: A demo-driven mobile bi-manual manipulation benchmark, 2024. URL https://arxiv.org/abs/2407.07788.

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

[9] J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao, X. Yuan, P. Xie, Z. Huang, R. Chen, and H. Su. Maniskill2: A unified benchmark for generalizable manipulation skills. In *International Conference on Learning Representations*, 2023.

[10] W. Pumacay, I. Singh, J. Duan, R. Krishna, J. Thomason, and D. Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation, 2024. URL https://arxiv.org/abs/2402.08191.

[11] X. Li, K. Hsu, J. Gu, K. Pertsch, O. Mees, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024.

[12] C. Sferrazza, D.-M. Huang, X. Lin, Y. Lee, and P. Abbeel. Humanoidbench: Simulated humanoid benchmark for whole-body locomotion and manipulation, 2024.

[13] Y. Chen, Y. Geng, F. Zhong, J. Ji, J. Jiang, Z. Lu, H. Dong, and Y. Yang. Bi-dexhands: Towards human-level bimanual dexterous manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2804–2818, 2023.

[14] H. Luo and Y. Demiris. Benchmarking and simulating bimanual robot shoe lacing. *IEEE Robotics and Automation Letters*, 2024.

[15] M. Grotz, M. Shridhar, T. Asfour, and D. Fox. Peract2: Benchmarking and learning for robotic bimanual manipulation tasks, 2024. URL https://arxiv.org/abs/2407.00278.

[16] Y. Mu, T. Chen, Z. Chen, S. Peng, Z. Lan, Z. Gao, Z. Liang, Q. Yu, Y. Zou, M. Xu, L. Lin, Z. Xie, M. Ding, and P. Luo. Robotwin: Dual-arm robot benchmark with generative digital twins, 2025. URL https://arxiv.org/abs/2504.13059.

[17] R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic, et al. Deep learning approaches to grasp synthesis: A review. *IEEE Transactions on Robotics*, 39(5):3994–4015, 2023.

[18] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.

[19] S. Dasari, J. Wang, J. Hong, S. Bahl, Y. Lin, A. Wang, A. Thankaraj, K. Chahal, B. Calli, S. Gupta, D. Held, L. Pinto, D. Pathak, V. Kumar, and A. Gupta. Rb2: Robotic manipulation benchmarking with a twist, 2022. URL https://arxiv.org/abs/2203.08098.

[20] F. Krebs and T. Asfour. A bimanual manipulation taxonomy. *IEEE Robotics and Automation Letters*, 7(4):11031–11038, 2022. doi:10.1109/LRA.2022.3196158.

[21] 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.

[22] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[23] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.