

RoboEval: Where Robotic Manipulation Meets Structured and Scalable Evaluation

Yi Ru Wang^{1,2†}, Carter Ung¹, Christopher Tan¹, Grant Tannert¹, Jiafei Duan^{1,2}, Josephine Li¹, Anh Le¹
Rishabh Oswal¹, Markus Grotz¹, Wilbert Pumacay^{1,2}, Yuquan Deng¹, Ranjay Krishna^{1,2}
Dieter Fox^{1,2*}, Siddhartha Srinivasa^{1*}

Abstract—We introduce **ROBOEVAL**, a structured evaluation framework and benchmark for robotic manipulation that augments binary success with principled behavioral and outcome metrics. Existing evaluations often collapse performance into outcome counts, masking differences in execution quality and obscuring failure structure. **ROBOEVAL** provides eight bimanual tasks with systematically controlled variations, more than three thousand expert demonstrations, and a modular simulation platform for reproducible experimentation. All tasks are instrumented with standardized metrics that quantify efficiency, coordination, and safety/stability, as well as outcome measures that trace stagewise progress and localize failure modes. Through extensive experiments with state-of-the-art visuomotor policies, we validate these metrics by analyzing their stability under variation, discriminative power across policies with similar success rates, and correlation with task success. Project Page: <https://robo-eval.github.io>

I. INTRODUCTION

Robotic manipulation is most often evaluated by binary success [1], [2], [3], [4], [5]. Yet the true capabilities of a policy are reflected in its motions: the stability of a grasp, the smoothness of a trajectory, and the coordination between arms. Two policies may achieve the same success rate, one fluent and robust and the other brittle and erratic, but conventional evaluation renders them indistinguishable. This limitation becomes more pronounced as policies operate in increasingly complex environments. Without richer measures, we cannot explain how policies behave, why they fail, or what skills they possess. *In this work, we challenge the assumption that binary success adequately reflects policy performance, as policies may succeed while concealing critical failures beyond outcome-only evaluation.*

Behavioral metrics address this gap by quantifying execution quality rather than outcomes alone, including trajectory smoothness, contact stability, and coordination timing. Other domains have benefited from similar shifts toward richer evaluation signals, such as BLEU in language [6], FID in vision [7], and Elo ratings in games, which measure how well systems perform rather than only whether they succeed. The use of multi-dimensional metrics is well grounded in prior robotics literature. Classical standards and benchmarking efforts characterize performance along axes such as accuracy, efficiency, stability, and safety [8], [9], [10], [11]. Motivated

by this perspective, we organize behavioral metrics into three axes: *efficiency*, capturing time and resource usage; *safety/stability*, capturing robustness and avoidance of undesirable events such as collisions or slips; and *coordination*, capturing temporal and spatial coupling between actions. Complementing these, outcome-driven metrics measure *task progression* and *binary task success*, enabling analysis of both how a task is completed and whether it is completed.

A central question is whether behavioral metrics provide reliable and meaningful insight into policy performance. To be useful, they should be *skill-relevant*, reflecting capabilities that drive success; *discriminative*, distinguishing policies with similar success rates but different execution quality; and *consistent*, remaining informative under variation in task configuration and scene complexity.

We introduce **ROBOEVAL**, an evaluation framework for studying behavioral metrics in robotic manipulation. It includes eight bimanual tasks with controlled variations, over three thousand expert demonstrations collected via VR teleoperation, and a modular simulation platform for reproducible experimentation. Each task is instrumented with behavioral metrics across efficiency, safety or stability, and coordination, along with outcome metrics capturing task progression and binary success. Using this framework, we examine whether these metrics satisfy the proposed criteria by analyzing their stability under task variation, their ability to distinguish policies with similar success rates, and their relationship to task success. Beyond this study, **ROBOEVAL** serves as a platform for fine-grained evaluation and extension to new tasks and metrics. Through experiments with state-of-the-art visuomotor policies, we show that **ROBOEVAL** provides insights beyond success rates. Behavioral metrics reveal which aspects of execution contribute to robustness and distinguish policies with similar success but different execution quality. Outcome metrics expose structured failure modes, enabling more precise diagnosis of policy limitations. Task variations highlight when metrics remain informative and when success alone becomes insufficient.

Contributions. This work makes three contributions: (1) an evaluation framework that augments binary success with behavioral and outcome metrics for fine-grained analysis, (2) a validation study demonstrating that the metrics are stable, discriminative, and complementary to task success, and (3) a reusable benchmark with eight tasks, variations, and 3,000 demonstrations, supporting extensible research.

¹University of Washington, Seattle, WA, USA

²Allan Institute for Artificial Intelligence (AI2), Seattle, WA, USA

[†]Corresponding author: yiruwang@cs.washington.edu

*Dieter Fox and Siddhartha Srinivasa contributed equally as advisors.

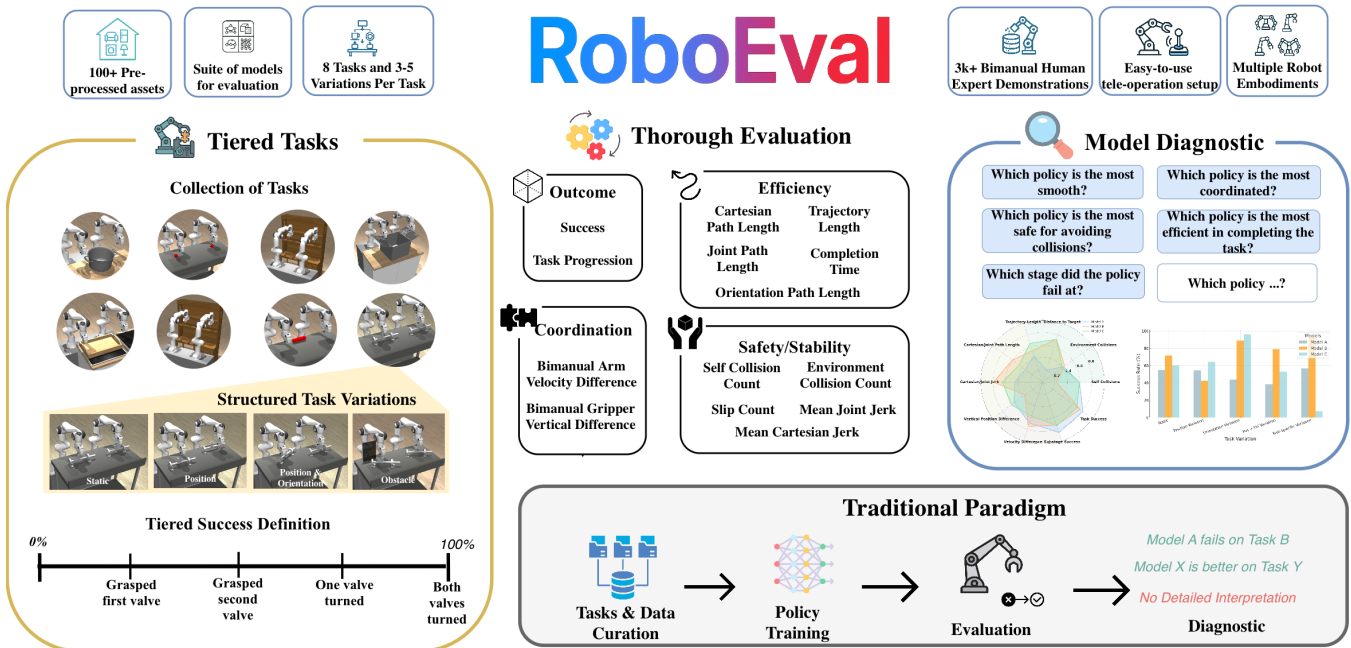


Fig. 1: **Overview of ROBOEval.** ROBOEval provides a principled evaluation framework that goes beyond binary success by integrating a rich set of *behavioral and outcome metrics*. These capture execution quality in terms of trajectory fluency, spatial precision, and bimanual coordination, as well as outcome measures of task progression and structured failures. The benchmark features 3,000 human-collected demonstrations across 8 tasks with 3–5 systematic variations each, a standardized asset library for constructing new tasks and perturbations, and a VR-based teleoperation interface for realistic data collection. Together, these components enable fine-grained policy diagnosis and reproducible evaluation.

II. RELATED WORKS

Benchmarks for Robotic Manipulation. Significant progress has been made in benchmarking single-arm manipulation [12], [13], [1], [14], [15]. Recent efforts such as Peract2 [4], BiGym [2], and RoboTwin [5] extend this to scalable and data-rich bimanual settings, using scripted, VR-collected, and synthetic data respectively. HumanoidBench [16] further broadens evaluation to whole-body humanoid manipulation, while other work studies dexterous hand use [17]. Despite these advances, existing benchmarks provide limited insight into when and why policies fail, particularly in coordinated bimanual settings. Real-world evaluation efforts are often restricted to single tasks [18], limiting diversity. Our work complements prior benchmarks by enabling structured, fine-grained evaluation and unifying diverse tasks within a common framework (Table I).

Evaluation Metrics for Manipulation. Evaluation metrics are central to assessing manipulation performance as tasks grow in complexity. Prior work surveys common metrics such as success rate, accuracy, and robustness [19], while related domains emphasize the importance of task-specific, fine-grained evaluation [20]. Benchmarks such as RB2 [21] and Colosseum [14] evaluate generalization under environmental variation. However, existing approaches remain fragmented, focusing on isolated measures of success or robustness without a unified framework that captures task progression, execution quality, and failure modes.

III. ROBOEval BENCHMARK

ROBOEval is designed to evaluate bimanual manipulation policies under controlled yet diverse conditions, enabling

principled analysis of both behavioral and outcome metrics. The first release includes 8 base tasks spanning service, warehouse, and industrial domains, such as lifting a tray, closing a box, and rotating a hand-wheel. Each task is paired with systematic variations in spatial configuration, from fixed layouts to randomized object poses, creating a controlled testbed for analyzing metric stability, discriminative power, and sensitivity to task complexity. To ground evaluation in real behavior, we provide 3000 human demonstrations across all tasks and variations. These support imitation learning and serve as baselines for evaluating policy fluency and precision. All environments are instrumented with fine-grained behavioral metrics across efficiency, safety/stability, and coordination, alongside outcome metrics that capture task progression and overall success. Together, these components form a modular and extensible simulation framework for reproducible, data-driven evaluation. An overview of ROBOEval is shown in Figure 1. The remainder of this section describes the design philosophy (Section III-A), base tasks (Section III-B), dataset statistics (Section III-C), and evaluation protocol (Section III-D).

III-A. Design Philosophy

ROBOEval is designed as a comprehensive benchmark for evaluating learning-based manipulation policies, grounded in three principles. *Diversity:* Real-world bimanual manipulation spans a wide range of task structures, object geometries, and control challenges. ROBOEval captures this through tasks with varying temporal complexity, coordination demands, and semantic content, from non-prehensile pushing to tightly coupled bimanual behaviors. This ensures

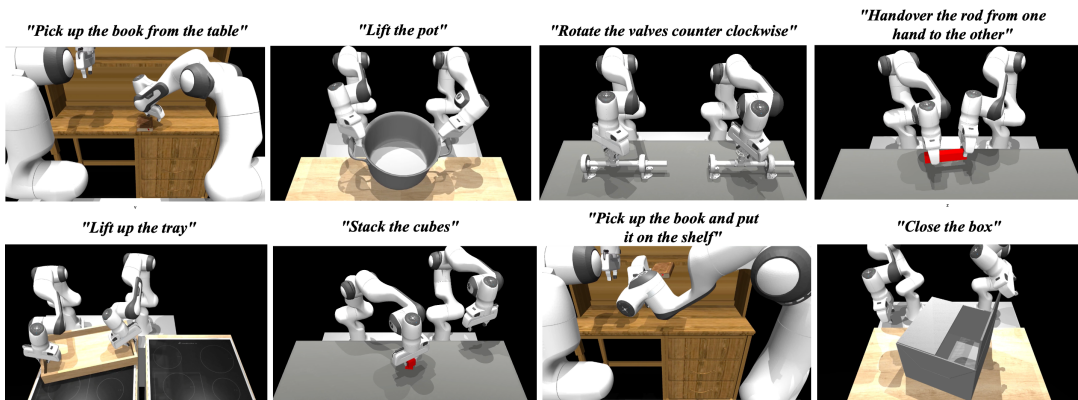


Fig. 2: **Base tasks in ROBOEVAL.** ROBOEVAL introduces an initial suite of 8 bimanual manipulation tasks, each accompanied by 3–4 structured variations. All tasks are instrumented with behavior metric logging and task-stage definitions to support fine-grained progress and outcome analysis. The benchmark is modular by design, allowing for seamless integration of new tasks to accommodate evolving research needs within the community.

TABLE I: **Benchmark Comparison.** We compare six manipulation benchmarks across task design, evaluation, and data. **ROBOEVAL** uniquely integrates tiered bimanual tasks, structured variations, behavior metrics, task progression tracking, and human demonstrations.

Benchmark	Task Features		Evaluation Features			Data Features	
	Tiered	Structured Variations	Success	Behavior Metrics	Task Prog. Metrics	Demo-Driven	# Expert Human Demos
RLBench	×	×	✓	×	×	✓	0
Bigym	×	×	✓	×	×	✓	2k
DexMimicGen	×	×	✓	×	×	✓	400
PerAct2	×	×	✓	×	×	✓	0
HumanoidBench	×	×	✓	×	×	×	300
RoboTwin	×	×	✓	×	×	✓	300+
ROBOEVAL (Ours)	✓	✓	✓	✓	✓	✓	3k+

TABLE II: **Base Task Set in ROBOEVAL.** We summarize the base tasks with their variation types and demonstration statistics. Variation types include static setups and spatial perturbations in position (Pos), rotation (Rot), and combined (PR).

Task Name	Variations	# Demos	Traj Len
Lift Tray	Static, Pos, Rot, PR	543	67.584
Stack Two Cubes	Static, Pos, Rot, PR	492	107.047
Stack Single Book Shelf	Static, Pos, Rot, PR	202	172.302
Rod Handover	Static, Pos, Rot, PR	408	93.529
Lift Pot	Static, Pos, Rot, PR	176	53.170
Pack Box	Static, Pos, Rot, PR	394	133.216
Pick Book From Table	Static, Pos, Rot, PR	366	105.984
Rotate Valve	Static, Pos, PR	349	119.610

evaluation reflects both individual skills and their generalization across tasks. *Interpretability:* Binary success provides limited insight into policy behavior. ROBOEVAL enables structured, fine-grained analysis through multi-dimensional behavioral and outcome metrics, allowing diagnosis of failure modes, sensitivity to variation, and qualitative differences between policies. *Extensibility:* ROBOEVAL separates task definitions, variation generation, and evaluation protocols through standardized interfaces. New tasks, variations, and metrics can be added modularly, enabling extension while preserving comparability across studies.

III-B. Tasks

Tasks in ROBOEVAL span diverse settings and skill requirements, forming a systematic testbed for evaluating robotic manipulation. Each task is a goal-conditioned episode with explicit success criteria reflecting common household, industrial, and tabletop interactions. The suite covers both short-horizon objectives (e.g., lifting an object) and long-horizon, multi-step activities (e.g., cleaning a desk by placing a book on a shelf), with stage-wise progress tracking to

capture partial success.

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 $\mathcal{D}_{\mathcal{T}} = \{(s_0, a_0, \dots, s_T)\}$, collected via human teleoperation. To support fine-grained analysis, each task is instantiated with a parameterized family of variants \mathcal{T}_{θ} where $\theta \in \Theta$ modulates scene layout or semantic content.

Skill Diversity. Table II summarizes the range of tasks included in the benchmark. Drawing on the bimanual taxonomy in [22], the suite is designed to span diverse motor and coordination demands, from single-arm actions such as turning a valve, to long-horizon interactions like packing a box, to tightly coordinated dual-arm motions such as lifting and balancing a tray. This diversity ensures that the benchmark probes capabilities in coordination, precision, and smooth execution across different task requirements.

Task Variations. The initial release of ROBOEVAL includes a curated set of bimanual manipulation tasks with structured spatial variations designed to probe robustness and generalization. These variations challenge visuomotor policies to adapt their coordination strategies while preserving task semantics. Future extensions of ROBOEVAL can build on this foundation by incorporating additional modalities, such as visual distractors, lighting changes, and variations in object physical properties.

Task Design. ROBOEVAL includes a modular task generation pipeline that allows new or external tasks to be authored and integrated with minimal code overhead. A unified interface and built-in behavioral and outcome metrics enable principled evaluation while supporting the development

of generalist manipulation policies. As shown in Table I, ROBOEVAL distinguishes itself from existing benchmarks through tiered task variations, fine-grained metrics, and a large-scale repository of expert human demonstrations, creating a comprehensive platform for benchmarking bimanual manipulation.

III-C. Task and Dataset Statistics

In total, ROBOEVAL introduces 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 II provides a breakdown of the task categories, associated variation schemes, and the number of demonstrations per task. The initial task suite in ROBOEVAL spans 8 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 amount of spatial complexities. 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. We capture proprioception, visual observations, and scene-annotated interaction states, enabling detailed diagnostics of policy behavior across the spatial, temporal, and coordination axes.

III-D. 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: *efficiency*, *safety/stability*, and *coordination*. Outcome-driven metrics include *task progression* and *binary task success*.

Efficiency-Based Metrics. We measure efficiency using both temporal and spatial properties of the trajectory. For temporal efficiency, we track the *trajectory length* (number of control steps) and *completion time* (total duration in seconds). For spatial efficiency, we measure using trajectory length in joint, Cartesian, and orientation spaces. Specifically, we compute *joint path length* and *Cartesian path length* as the cumulative displacement along the trajectory:

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

where q_t denotes the joint configuration and x_t the Cartesian end-effector position at timestep t .

Safety/Stability Metrics. We evaluate physical interaction quality and execution stability through both motion-based and contact-based indicators. For motion smoothness, we 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 :

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

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

We additionally track three contact-based indicators: the number of *self-collisions* (contacts between robot links), *environment collisions* (contacts with scene elements such as tables or walls), and *object slips* (instances where a grasped object unintentionally loses contact with the gripper). Together, these metrics capture motion smoothness, contact stability, and overall control reliability. High values typically indicate unstable execution or imprecise interactions.

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:

$$\Delta z = \frac{1}{T} \sum_{t=1}^T \left| x_t^{(L)}[z] - x_t^{(R)}[z] \right|. \quad (4)$$

(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:

$$\Delta v = \frac{1}{T-1} \sum_{t=1}^{T-1} \left\| v_t^{(L)} - v_t^{(R)} \right\|_2. \quad (5)$$

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.

IV. RESULTS

The goal of our experiments is to evaluate not only how policies perform on the benchmark, but also what can be learned from the proposed metrics themselves. To this end, we organize the study around three guiding research questions:

- **RQ1:** Do behavioral and outcome metrics provide insights beyond binary success rates (Section IV-B)?
- **RQ2:** Under what task or policy conditions are these metrics most informative (Section IV-C)?
- **RQ3:** How do their patterns and discriminative ability vary across domains and levels of complexity (Section IV-D)?

Addressing these questions requires evaluating representative policies across a diverse set of bimanual manipulation tasks and systematically varied configurations, allowing us to examine both overall performance and the diagnostic value of different metrics.

IV-A. Implementation Details

Models. We evaluate five policy architectures: ACT [23], Diffusion Policy [24], GR00T N1.6 [25], X-VLA [26], and $\pi_{0.5}$ [27]. ACT and Diffusion Policy use official LeRobot implementations with ResNet-18 encoders and a 16-step action horizon, trained as single-task specialists. GR00T N1.6 (3B) is a vision-language-action model, where we fine-tune the projector and action head from the pretrained checkpoint. X-VLA (0.9B) is a soft-prompted VLA with a Florence2 backbone and flow-matching policy head. $\pi_{0.5}$ combines SigLIP, Gemma-2B, and a Gemma-300M action head, fine-tuned from a pretrained checkpoint in both full and LoRA settings. All models use AdamW with consistent hyperparameters unless noted. We also include human teleoperation demonstrations as an upper-bound reference.

Tasks and Variations. Our main experiments focus on 8 tasks (Table II), each with 3-4 spatial variations. *Static* denotes minimal scene changes, while *position* and *orientation* introduce spatial perturbations.

IV-B. RQ1: To what extent do behavioral and outcome metrics provide insights beyond binary success rates?

Metrics offer complementary insights beyond binary success rates. Table III compares policies across success, task progression, and behavioral metrics on combined variations. Across eight tasks, the top-success model leads in only 34 of 104 other metric comparisons (33%), showing that success is a poor proxy for overall quality. This gap is especially clear in smoothness: ACT achieves the lowest Cartesian and joint jerk in all tasks, up to $3\times$ lower than the next-best policy in *Rotate Valve* (1.0 vs. 3.0 m/s^3), despite leading in success on only one task. Similarly, in *Pack Box*, Pi-0.5 and ACT have comparable success (0.54 vs. 0.48), yet ACT achieves $4\times$ lower jerk and $2.7\times$ shorter paths, revealing efficiency differences not captured by success. No single model dominates safety: collisions, self-collisions, and slips vary by task across ACT, Pi-0.5, and Diffusion. Task progression further distinguishes policies when success fails: in *Lift Tray*, three policies have 0% success but reach 35–48% progression, and in *Pick Single Book From Table*, 0%-success models range from 2% to 22%. Safety metrics also expose distinct failure modes: in *Stack Two Blocks*, GR00T and Diffusion have similar success (4%, 6%), but GR00T incurs far more self-collisions (23.6 vs. 2.0 per episode). Together, these results show that behavioral and outcome metrics provide a much richer view of policy competence than success alone.

Behavioral metrics cluster into interpretable families that capture distinct aspects of policy behavior. Figure 3(a) reports pairwise Spearman correlations across all task-variant bins, and Figure 3(b) presents a hierarchical clustering using correlation distance ($1 - |\rho|$). Within each evaluation-intent group, tightly correlated subsets emerge: path-length metrics (CPL, JPL, OPL; $\rho \geq 0.95$), jerk metrics (MCJ, MJJ; $\rho = 0.96$), and temporal metrics (CT, TL; $\rho = 0.94$) each merge below distance 0.07 in the dendrogram. However, the broader tree does not collapse

into a single branch. The temporal metrics cluster with the outcome measures (Success, TP) at distance 0.42, despite a *negative* correlation ($\rho \approx -0.6$), before joining the path-length group at 0.54, indicating that completion time relates more closely to task success than spatial path efficiency does. The two coordination metrics fall into separate branches: BAVD groups with the smoothness metrics at distance 0.51, while BGVD groups with path-length and collision metrics at 0.37, despite sharing $\rho = 0.55$, suggesting they capture complementary behavioral dimensions. Overall, the full metric suite spans multiple distinct branches, supporting the use of diverse metric families to evaluate policy quality beyond binary success.

Behavioral metrics benchmark policies against human demonstrations. The human row in Table III provides a reference for efficient and reliable execution from teleoperated demonstrations. Policies consistently underperform humans on efficiency: none match human completion time or trajectory lengths, with the fastest policies $3\text{--}10\times$ slower and trajectories $2\text{--}9\times$ longer. Humans also exhibit near-zero self-collisions and slips, while all policies incur higher counts, indicating lower reliability. In contrast, policies often appear smoother: ACT achieves lower Cartesian jerk than humans in all tasks, and Pi-0.5 in six of eight, with values as low as 1.0 m/s^3 versus 3.6 m/s^3 for humans. Overall, policies have lower Cartesian jerk in 26 of 48 comparisons and lower joint jerk in 21 of 48. This reflects smoothing from learned action generation rather than better control, as teleoperation introduces discrete corrections that increase jerk. These comparisons show that behavioral metrics reveal where policies approach or exceed human execution and where clear gaps remain, which success alone cannot capture.

IV-C. RQ2: Under what conditions are the metrics most informative?

Behavioral metrics differentiate policy performance even when success rates are similar. When policies achieve comparable success, binary outcomes alone cannot distinguish their execution quality, yet behavioral metrics reveal meaningful differences. In *Pack Box*, Pi-0.5 and ACT achieve overlapping success rates (0.54 ± 0.14 vs. 0.48 ± 0.14), but ACT produces $4\times$ lower cartesian jerk (6.8 vs. 27.6 m/s^3) and $2.7\times$ shorter cartesian path lengths (1.7 vs. 4.6 m), as shown in Figure 4(a) and (b). Similarly, in *Lift Tray*, Pi-0.5 and ACT have overlapping success (0.32 ± 0.14 vs. 0.26 ± 0.14), yet ACT achieves $5\times$ lower cartesian jerk (6.1 vs. 31.3 m/s^3) and $3.5\times$ shorter paths (1.4 vs. 4.8 m). More broadly, Table III shows that the highest-success model leads on only 33% of the remaining metrics across all eight tasks. Safety metrics such as slip count and environment collisions exhibit the most diversity, with three to four different models claiming the top rank across tasks.

Outcome metrics diagnose policy behavior and expose skill gaps when success rates are low. When overall success is limited, outcome metrics pinpoint where failures accumulate. In *Pack Box*, four of six policies achieve $\leq 12\%$ success, yet task progression separates them: Diffusion and

TABLE III: **Summary of performance metrics across all methods on the combined variation.** Bold values indicate the best result for each metric within the task. Values are reported as mean \pm 95% confidence interval. The best performing method across each metric is highlighted in green. The final row reports human demonstration statistics (computed from successful teleoperated demos only), which provide a reference band for efficient execution.

Method	Success \uparrow	TP \uparrow	BAVD \downarrow	BGVD \downarrow	CPL \downarrow	CT \downarrow	ECC \downarrow	JPL \downarrow	MCJ \downarrow	MJJ \downarrow	OPL \downarrow	SCC \downarrow	SC \downarrow	TL \downarrow
Cube Handover														
ACT	0.08 \pm 0.11	0.23 \pm 0.08	0.18 \pm 0.03	0.08 \pm 0.02	1.21 \pm 0.12	5.70 \pm 0.42	1.72 \pm 0.31	8.70 \pm 1.00	4.51 \pm 0.80	23.60 \pm 3.28	5.98 \pm 0.80	6.28 \pm 1.55	0.28 \pm 0.15	941.02 \pm 56.95
Diffusion	0.06 \pm 0.10	0.27 \pm 0.07	0.32 \pm 0.03	0.13 \pm 0.01	5.07 \pm 0.35	4.80 \pm 0.28	2.94 \pm 0.49	26.14 \pm 2.02	24.73 \pm 0.90	133.00 \pm 4.88	19.61 \pm 1.59	2.46 \pm 0.57	0.22 \pm 0.14	939.76 \pm 52.04
GR00T	0.00 \pm 0.07	0.01 \pm 0.01	0.50 \pm 0.01	0.10 \pm 0.01	4.69 \pm 0.26	9.52 \pm 0.26	2.90 \pm 0.49	33.15 \pm 1.11	42.62 \pm 1.49	235.44 \pm 4.45	23.34 \pm 0.97	25.88 \pm 1.91	0.00 \pm 0.07	976.10 \pm 23.48
Pi_0.5	0.40 \pm 0.14	0.65 \pm 0.10	0.34 \pm 0.03	0.05 \pm 0.01	2.22 \pm 0.44	3.66 \pm 0.68	1.32 \pm 0.47	11.22 \pm 2.31	18.14 \pm 1.78	88.53 \pm 7.98	8.00 \pm 1.68	2.30 \pm 0.99	0.26 \pm 0.13	620.58 \pm 120.85
Pi_0.5 (LoRA)	0.10 \pm 0.11	0.29 \pm 0.08	0.40 \pm 0.03	0.08 \pm 0.01	4.61 \pm 0.49	5.22 \pm 0.42	4.22 \pm 0.63	26.98 \pm 3.04	27.40 \pm 1.99	153.46 \pm 11.11	18.32 \pm 2.03	5.44 \pm 2.24	0.20 \pm 0.15	900.28 \pm 71.03
xVLA	0.00 \pm 0.07	0.04 \pm 0.03	0.56 \pm 0.04	0.15 \pm 0.02	5.51 \pm 0.68	8.50 \pm 3.26	3.40 \pm 0.59	31.99 \pm 3.01	31.00 \pm 3.20	146.84 \pm 13.51	23.25 \pm 2.47	13.68 \pm 2.53	0.04 \pm 0.09	962.24 \pm 42.87
Human	1.00 \pm 0.00	1.00 \pm 0.00	0.33 \pm 0.03	0.03 \pm 0.00	0.53 \pm 0.03	0.51 \pm 0.07	0.10 \pm 0.16	2.26 \pm 0.18	28.94 \pm 3.07	120.52 \pm 11.54	1.38 \pm 0.18	0.17 \pm 0.17	0.03 \pm 0.13	103.53 \pm 14.28
Lift Pot														
ACT	0.44 \pm 0.14	0.58 \pm 0.08	0.19 \pm 0.02	0.07 \pm 0.02	1.33 \pm 0.27	3.82 \pm 0.76	0.02 \pm 0.08	7.82 \pm 1.63	15.74 \pm 2.57	82.66 \pm 13.36	5.07 \pm 0.98	0.98 \pm 0.78	0.28 \pm 0.18	518.30 \pm 121.72
Diffusion	0.02 \pm 0.08	0.07 \pm 0.03	0.36 \pm 0.05	0.18 \pm 0.03	4.58 \pm 0.56	8.37 \pm 0.40	0.20 \pm 0.18	24.46 \pm 2.63	27.14 \pm 3.77	142.55 \pm 18.69	16.55 \pm 1.88	10.64 \pm 2.02	0.12 \pm 0.12	973.06 \pm 32.64
GR00T	0.06 \pm 0.10	0.30 \pm 0.05	0.51 \pm 0.02	0.09 \pm 0.02	2.13 \pm 0.30	7.43 \pm 1.22	4.54 \pm 0.76	21.44 \pm 2.66	28.17 \pm 1.91	231.37 \pm 6.96	13.14 \pm 1.64	5.88 \pm 2.26	1.34 \pm 0.53	628.14 \pm 92.83
Pi_0.5	0.52 \pm 0.13	0.41 \pm 0.08	0.34 \pm 0.03	0.07 \pm 0.02	2.73 \pm 0.62	3.86 \pm 0.78	0.06 \pm 0.10	14.23 \pm 3.13	27.71 \pm 2.54	139.02 \pm 12.27	9.53 \pm 2.06	0.80 \pm 0.67	0.28 \pm 0.15	550.78 \pm 117.31
Pi_0.5 (LoRA)	0.08 \pm 0.11	0.22 \pm 0.07	0.61 \pm 0.03	0.20 \pm 0.03	8.65 \pm 0.91	6.17 \pm 0.59	0.58 \pm 0.32	47.03 \pm 4.83	53.22 \pm 2.58	294.12 \pm 12.29	28.68 \pm 2.85	5.56 \pm 1.82	0.28 \pm 0.14	854.18 \pm 80.23
xVLA	0.00 \pm 0.07	0.21 \pm 0.05	0.67 \pm 0.06	0.22 \pm 0.04	5.31 \pm 0.84	8.29 \pm 3.48	2.26 \pm 0.69	30.12 \pm 3.74	36.35 \pm 3.80	173.17 \pm 16.93	21.79 \pm 2.89	4.96 \pm 1.66	0.32 \pm 0.20	747.02 \pm 90.35
Human	1.00 \pm 0.00	1.00 \pm 0.00	0.49 \pm 0.05	0.03 \pm 0.01	0.61 \pm 0.07	0.45 \pm 0.06	0.00 \pm 0.14	3.33 \pm 0.34	62.59 \pm 4.70	287.53 \pm 20.08	2.40 \pm 0.21	0.00 \pm 0.14	0.00 \pm 0.14	71.04 \pm 9.68
Lift Tray														
ACT	0.26 \pm 0.14	0.81 \pm 0.06	0.19 \pm 0.02	0.06 \pm 0.01	1.39 \pm 0.19	5.63 \pm 0.78	1.34 \pm 0.57	10.43 \pm 1.59	6.08 \pm 0.64	40.51 \pm 4.04	6.71 \pm 1.06	2.26 \pm 1.02	0.74 \pm 0.23	773.56 \pm 103.73
Diffusion	0.00 \pm 0.07	0.47 \pm 0.06	0.47 \pm 0.04	0.21 \pm 0.04	5.01 \pm 0.52	9.85 \pm 0.28	5.28 \pm 1.29	29.97 \pm 2.22	27.50 \pm 3.48	147.21 \pm 13.54	21.38 \pm 1.78	6.70 \pm 2.30	0.48 \pm 0.20	1000.00 \pm 0.00
GR00T	0.00 \pm 0.07	0.46 \pm 0.05	0.51 \pm 0.02	0.12 \pm 0.02	3.93 \pm 0.30	13.28 \pm 0.25	6.62 \pm 1.25	32.10 \pm 1.15	34.08 \pm 1.98	228.16 \pm 6.87	20.66 \pm 0.93	17.06 \pm 2.25	0.50 \pm 0.23	1000.00 \pm 0.00
Pi_0.5	0.32 \pm 0.14	0.75 \pm 0.07	0.52 \pm 0.05	0.14 \pm 0.03	4.83 \pm 0.84	5.87 \pm 0.94	2.62 \pm 0.99	29.67 \pm 5.06	31.27 \pm 2.31	181.95 \pm 11.41	11.42 \pm 3.78	5.22 \pm 2.07	0.92 \pm 0.22	719.16 \pm 115.13
Pi_0.5 (LoRA)	0.14 \pm 0.12	0.64 \pm 0.06	0.60 \pm 0.03	0.24 \pm 0.04	7.87 \pm 0.89	7.31 \pm 0.73	2.28 \pm 1.08	45.16 \pm 4.86	47.05 \pm 2.99	257.01 \pm 13.42	30.91 \pm 3.36	7.68 \pm 2.26	0.80 \pm 0.21	875.56 \pm 84.41
xVLA	0.00 \pm 0.07	0.35 \pm 0.04	0.63 \pm 0.06	0.24 \pm 0.04	7.01 \pm 0.92	10.49 \pm 0.43	4.88 \pm 0.94	37.46 \pm 3.72	32.82 \pm 4.39	151.68 \pm 16.81	28.46 \pm 3.09	10.46 \pm 2.68	0.16 \pm 0.12	1000.00 \pm 0.00
Human	1.00 \pm 0.00	1.00 \pm 0.00	0.39 \pm 0.04	0.03 \pm 0.00	0.52 \pm 0.03	0.56 \pm 0.03	0.00 \pm 0.06	2.87 \pm 0.13	26.77 \pm 2.93	153.00 \pm 16.82	2.04 \pm 0.08	0.00 \pm 0.06	0.00 \pm 0.06	83.62 \pm 3.96
Pack Box														
ACT	0.48 \pm 0.13	0.46 \pm 0.08	0.25 \pm 0.03	0.08 \pm 0.02	1.68 \pm 0.24	3.41 \pm 0.62	0.74 \pm 0.37	7.87 \pm 1.51	6.84 \pm 1.09	29.15 \pm 4.25	5.51 \pm 1.16	0.72 \pm 0.47	0.28 \pm 0.16	613.72 \pm 110.18
Diffusion	0.04 \pm 0.09	0.27 \pm 0.06	0.43 \pm 0.04	0.11 \pm 0.02	7.19 \pm 0.41	5.42 \pm 0.31	0.58 \pm 0.34	36.50 \pm 2.51	29.98 \pm 2.07	154.68 \pm 11.16	25.85 \pm 1.81	9.82 \pm 3.05	0.44 \pm 0.16	962.98 \pm 44.27
GR00T	0.00 \pm 0.07	0.07 \pm 0.03	0.50 \pm 0.02	0.24 \pm 0.04	4.77 \pm 0.36	10.75 \pm 0.20	1.04 \pm 0.47	29.29 \pm 1.37	40.49 \pm 3.08	207.72 \pm 9.51	19.50 \pm 1.10	13.30 \pm 2.92	0.30 \pm 0.19	1000.00 \pm 0.00
Pi_0.5	0.54 \pm 0.14	0.55 \pm 0.06	0.73 \pm 0.07	0.10 \pm 0.01	4.60 \pm 0.29	3.70 \pm 0.72	0.90 \pm 0.41	22.61 \pm 4.56	27.59 \pm 1.39	130.35 \pm 7.35	15.84 \pm 3.25	1.24 \pm 1.19	0.44 \pm 0.20	566.54 \pm 113.98
Pi_0.5 (LoRA)	0.12 \pm 0.12	0.35 \pm 0.06	0.68 \pm 0.04	0.17 \pm 0.03	7.92 \pm 0.65	6.01 \pm 0.44	1.18 \pm 0.47	45.24 \pm 3.63	42.69 \pm 2.35	225.24 \pm 11.40	31.91 \pm 2.61	7.54 \pm 2.64	0.62 \pm 0.23	923.82 \pm 62.68
xVLA	0.02 \pm 0.08	0.12 \pm 0.05	0.79 \pm 0.05	0.25 \pm 0.03	9.76 \pm 0.88	7.29 \pm 0.35	1.30 \pm 0.53	47.57 \pm 3.10	46.60 \pm 3.63	200.34 \pm 12.90	34.53 \pm 2.36	9.24 \pm 3.08	0.14 \pm 0.11	988.12 \pm 23.28
Human	1.00 \pm 0.00	1.00 \pm 0.00	0.51 \pm 0.04	0.05 \pm 0.00	1.30 \pm 0.06	0.68 \pm 0.04	0.84 \pm 0.37	5.38 \pm 0.34	28.24 \pm 1.50	134.77 \pm 6.73	3.35 \pm 0.21	0.18 \pm 0.16	0.03 \pm 0.07	129.66 \pm 8.00
Pick Single Book From Table														
ACT	0.34 \pm 0.14	0.55 \pm 0.13	0.51 \pm 0.09	0.25 \pm 0.02	1.13 \pm 0.24	5.79 \pm 1.41	1.58 \pm 0.46	6.70 \pm 1.43	8.59 \pm 1.32	39.56 \pm 5.31	3.33 \pm 0.70	0.18 \pm 0.14	0.10 \pm 0.11	504.46 \pm 121.23
Diffusion	0.02 \pm 0.08	0.15 \pm 0.09	0.59 \pm 0.04	0.24 \pm 0.03	4.42 \pm 0.49	11.20 \pm 0.83	5.52 \pm 1.27	28.17 \pm 2.59	21.24 \pm 1.98	136.43 \pm 10.11	17.38 \pm 1.69	4.26 \pm 1.50	0.04 \pm 0.09	977.16 \pm 61.77
GR00T	0.00 \pm 0.07	0.05 \pm 0.05	0.56 \pm 0.03	0.26 \pm 0.03	2.96 \pm 0.36	15.47 \pm 1.63	9.12 \pm 1.30	28.12 \pm 2.69	28.97 \pm 1.90	237.89 \pm 7.47	18.37 \pm 1.90	10.90 \pm 2.44	0.04 \pm 0.09	821.42 \pm 83.40
Pi_0.5	0.22 \pm 0.13	0.44 \pm 0.13	0.72 \pm 0.07	0.20 \pm 0.02	1.89 \pm 0.47	6.03 \pm 1.54	3.02 \pm 0.88	11.37 \pm 2.88	17.94 \pm 1.51	91.87 \pm 6.23	6.67 \pm 1.76	0.28 \pm 0.24	0.12 \pm 0.11	458.76 \pm 120.93
Pi_0.5 (LoRA)	0.00 \pm 0.07	0.22 \pm 0.09	0.67 \pm 0.08	0.20 \pm 0.02	3.65 \pm 0.64	9.11 \pm 1.46	5.80 \pm 1.44	24.05 \pm 4.15	26.09 \pm 1.80	156.80 \pm 9.46	14.76 \pm 2.55	2.36 \pm 1.38	0.10 \pm 0.11	705.64 \pm 114.01
xVLA	0.00 \pm 0.07	0.02 \pm 0.08	0.72 \pm 0.06	0.21 \pm 0.03	5.10 \pm 0.72	13.53 \pm 0.81	5.58 \pm 0.96	31.56 \pm 2.94	22.35 \pm 2.75	114.50 \pm 11.31	22.67 \pm 2.35	7.90 \pm 1.75	0.02 \pm 0.08	949.08 \pm 47.67
Human	1.00 \pm 0.00	1.00 \pm 0.00	0.87 \pm 0.00	0.18 \pm 0.00	0.76 \pm 0.00	1.20 \pm 0.00	0.00 \pm 0.00	4.09 \pm 0.00	30.32 \pm 0.00	138.21 \pm 0.00	1.90 \pm 0.00	0.0		

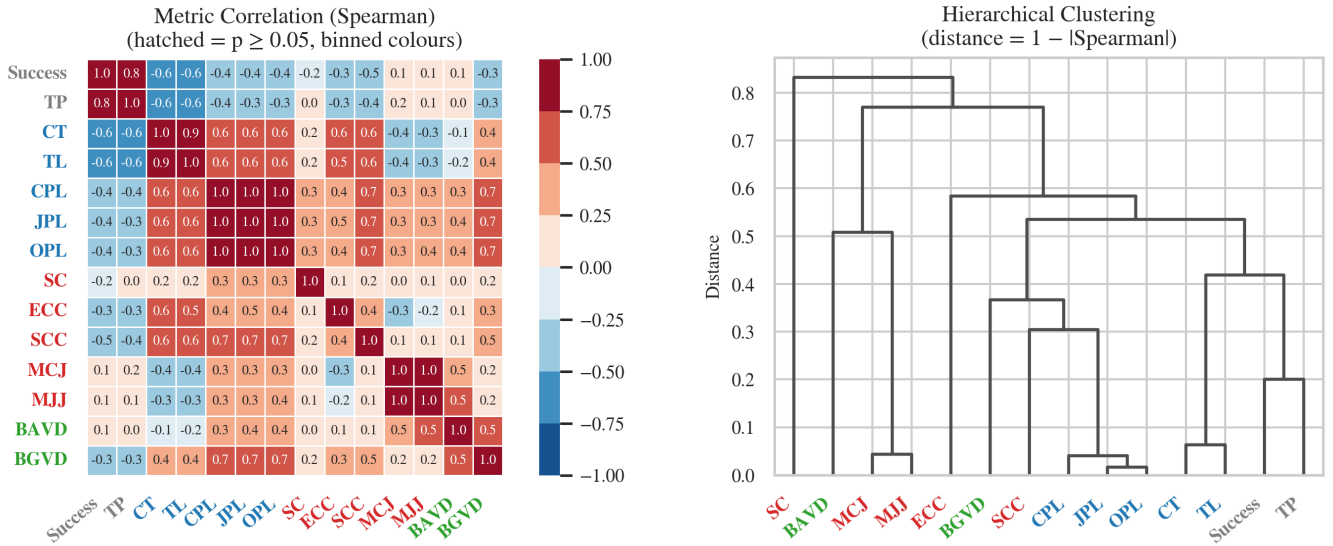


Fig. 3: **Correlation Structure of Behavioral Metrics.** From left to right, (a) Spearman correlation heatmap across all tasks and models. Hatched cells indicate non-significant correlations ($p \geq 0.05$). Color-coded labels denote metric groups: Outcome, Efficiency, Safety/Stability, and Coordination. (b) Average-linkage hierarchical clustering with distance $1 - |\rho_{\text{Spearman}}|$. The empirical clustering largely aligns with our predefined groupings, while confirming the metric suite is not collapsed into a single latent factor.

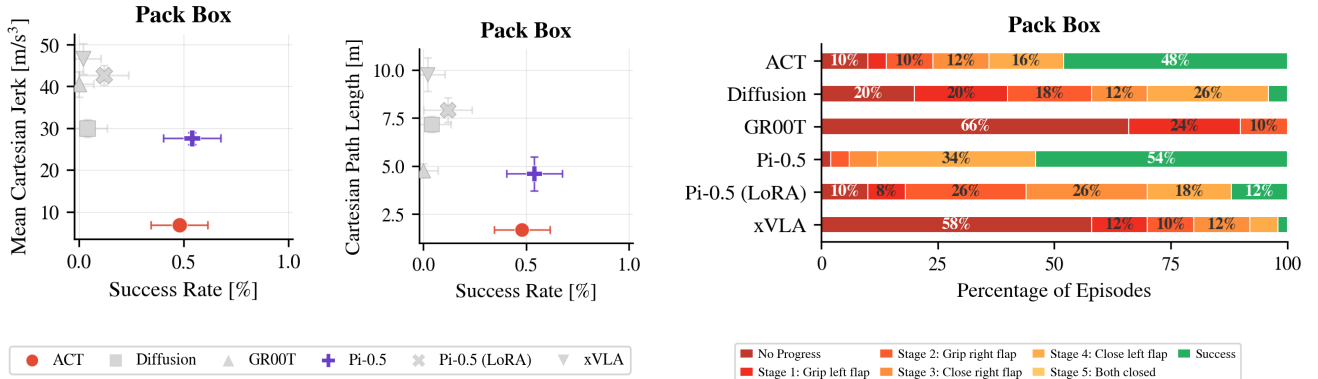


Fig. 4: **Behavioral metrics and failure modes for Pack Box.** From left to right, (a) Mean Cartesian jerk vs. success rate and (b) Cartesian path length vs. success rate, where each point represents a policy with 95% confidence intervals. **ACT** and **Pi-0.5** are highlighted; Pi-0.5 achieves the highest success rate, while ACT exhibits the lowest jerk and shortest path length, indicating smoother and more efficient motion despite slightly lower success. (c) Failure mode breakdown showing the farthest subtask stage reached by each policy’s episodes. Pi-0.5 succeeds in 54% of trials and ACT in 48%, while GR00T and xVLA fail predominantly at early grasping stages. Together, the scatter plots and failure analysis show that high success does not guarantee efficient motion: ACT’s lower jerk and shorter paths suggest qualitatively different execution strategies even among the top-performing policies.

turbations. This indicates that the metrics remain informative even under challenging conditions, rather than collapsing as all policies degrade uniformly.

Discriminative power is task-dependent. The task-level analysis (Figure 5, bottom) reveals that the discriminative strength of individual metrics varies substantially across tasks. For example, **SCC** is especially discriminating on **RotateValve** and **StackTwoBlocks**, where collision-avoidance behavior diverges sharply between policies, while end-effector collision count (**ECC**) spikes on **LiftPot**. Efficiency metrics such as **CPL** and **CT** show moderate, relatively uniform discrimination across tasks, whereas coordination metrics (**BAVD**, **BGVD**) are most discriminating on **RotateValve** and **StackTwoBlocks**. This task-dependence reinforces that a comprehensive metric suite is necessary: any single metric would under-represent policy

differences on certain tasks.

V. DISCUSSION

ROBOEVAL demonstrates that binary success alone is an incomplete measure of policy performance. By integrating behavioral metrics of fluency, precision, and coordination with outcome metrics that track stagewise progress and failure modes, the benchmark enables systematic analysis of how policies succeed, where they fail, and which execution factors underlie robustness. Our experiments show that these metrics are not only predictive of success but also expose distinctions hidden by success rates, validating the premise that richer evaluation reveals the true capabilities of a policy. Looking ahead, potential extensions include expanding variation modalities, and using the metrics as auxiliary objectives or reward-shaping signals. In this way, ROBOEVAL positions

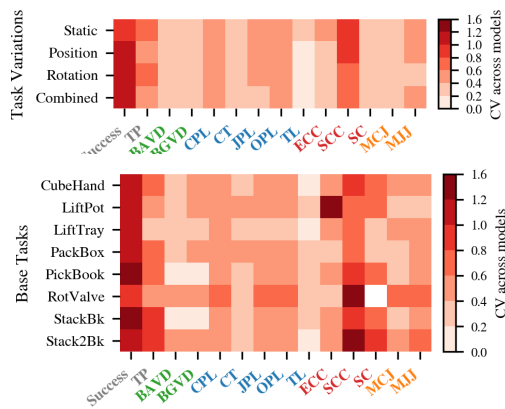


Fig. 5: Coefficient of variation (CV) of metrics across policies, grouped by task variation (top) and base task (bottom). Higher CV indicates greater discriminative power between policies. (Top) Metric discrimination is largely stable across variation complexity. (Bottom) Discrimination patterns are task-dependent: SCC is especially discriminating on RotateValve and StackTwoBlocks, while end-effector collision count (ECC) spikes on LiftPot. Success rate remains the most discriminating metric across all tasks.

metrics as both a diagnostic lens and a training signal, advancing the development of interpretable, robust, and data-efficient robot learning.

ACKNOWLEDGMENTS

Yi Ru Wang is supported by the Natural Sciences and Engineering Research Council of Canada Postgraduate Scholarships – Doctoral program (NSERC-PGSD). This work was partially supported by the National Science Foundation NRI program (#2132848), DARPA RACER (#HR0011-21-C-0171), the Office of Naval Research (#N00014-24-S-B001 and #2022-016-01 UW), and the DE-VCOM Army Research Laboratory (Award: W911NF-24-2-0191). We gratefully acknowledge support from Amazon and the Allen Institute for Artificial Intelligence (AI2), as well as gifts from Collaborative Robotics, Cruise, and other industry partners.

REFERENCES

- [1] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, “Rlbench: The robot learning benchmark & learning environment,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.
- [2] N. Chernyadev, N. Backshall, X. Ma, Y. Lu, Y. Seo, and S. James, “Bigym: A demo-driven mobile bi-manual manipulation benchmark,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.07788>
- [3] Z. Jiang, Y. Xie, K. Lin, Z. Xu, W. Wan, A. Mandlekar, L. Fan, and Y. Zhu, “Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning,” in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.
- [4] M. Grotz, M. Shridhar, T. Asfour, and D. Fox, “Peract2: Benchmarking and learning for robotic bimanual manipulation tasks,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.00278>
- [5] 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. [Online]. Available: <https://arxiv.org/abs/2504.13059>
- [6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [7] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in neural information processing systems*, vol. 30, 2017.
- [8] K. Kimble, J. Albrecht, M. Zimmerman, and J. Falco, “Performance measures to benchmark the grasping, manipulation, and assembly of deformable objects typical to manufacturing applications,” *Frontiers in Robotics and AI*, vol. 9, p. 999348, 2022.

- [9] J. Falco, K. Van Wyk, S. Liu, and S. Carpin, “Grasping the performance: Facilitating replicable performance measures via benchmarking and standardized methodologies,” *IEEE Robotics & Automation Magazine*, vol. 22, no. 4, pp. 125–136, 2015.
- [10] R. Madhavan, R. Lakaemper, and T. Kalmár-Nagy, “Benchmarking and standardization of intelligent robotic systems,” in *2009 International Conference on Advanced Robotics*. IEEE, 2009, pp. 1–7.
- [11] R. Bostelman, T. Hong, and J. Marvel, “Survey of research for performance measurement of mobile manipulators,” *Journal of Research of the National Institute of Standards and Technology*, vol. 121, p. 342, 2016.
- [12] 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.
- [13] 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.
- [14] W. Pumacay, I. Singh, J. Duan, R. Krishna, J. Thomason, and D. Fox, “The colosseum: A benchmark for evaluating generalization for robotic manipulation,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.08191>
- [15] 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.
- [16] C. Sferrazza, D.-M. Huang, X. Lin, Y. Lee, and P. Abbeel, “Humanoid-bench: Simulated humanoid benchmark for whole-body locomotion and manipulation,” 2024.
- [17] 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*, vol. 46, no. 5, pp. 2804–2818, 2023.
- [18] H. Luo and Y. Demiris, “Benchmarking and simulating bimanual robot shoe lacing,” *IEEE Robotics and Automation Letters*, 2024.
- [19] 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*, vol. 39, no. 5, pp. 3994–4015, 2023.
- [20] 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.
- [21] 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. [Online]. Available: <https://arxiv.org/abs/2203.08098>
- [22] F. Krebs and T. Asfour, “A bimanual manipulation taxonomy,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 031–11 038, 2022.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] J. Zheng, J. Li, Z. Wang, D. Liu, X. Kang, Y. Feng, Y. Zheng, J. Zou, Y. Chen, J. Zeng *et al.*, “X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model,” *arXiv preprint arXiv:2510.10274*, 2025.
- [27] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky, “ $\pi_{0.5}$: a vision-language-action model with open-world generalization,” 2025. [Online]. Available: <https://arxiv.org/abs/2504.16054>