Reliable and Scalable Evaluation for the Next Generation of Robotics

Apurva Badithela
Princeton University

I. Introduction

As robotic AI systems become more capable, the primary prerequisite for widespread adoption will be demonstrating safe and reliable operation [1]. Machine learning has significantly expanded and improved robot capabilities — from neural networks that form the backbone of perception to large foundation models that exhibit commonsense reasoning. The resulting models, however, are complex and opaque to analytic inspection, and deploying these systems in equally complex environments with a combinatorially large number of variations makes it challenging to predict when the system might fail, let alone provide any safety guarantees at design time. Thus, we need comprehensive and principled evaluation methodologies to validate system behavior and provide trustworthy assurances. **My goal is to advance a virtuous evaluation-design cycle where rigorous test and evaluation drives the development of reliable, robust, and safe robots, which are subsequently vetted with powerful evaluation frameworks as they take on new and richer capabilities.**

My work aims to advance general-purpose, rigorous robot evaluation using techniques from applied statistics, optimization, and formal methods, with applications spanning from robot manipulation to self-driving (see Figure 1). I have built reliable and scalable *end-to-end evaluation* frameworks for assessing robot foundation models [2, 3, 4, 5, 6, 7], including the first non-asymptotically valid evaluation framework for providing real-world assurances on robot behavior by combining large-scale *imperfect* simulation with small-scale real-world testing [2]. I have also developed fundamental task-relevant evaluation metrics to appropriately assess perception performance in the context of system-level safety (*modular evaluation*) [8, 9, 10, 11], and introduced a new paradigm of synthesizing closed-loop test plans for high-level task reasoning (*closed-loop evaluation*) [12, 13, 14, 15].

II. Current and Past Research

Reliable and Scalable Evaluation. Consider the problem of evaluating the mean performance of a robot policy on some target task and environment distribution (e.g., success rate, mean progress score). Mean estimation appears at each stage of the robot design lifecycle, from comparisons to baselines at the policy development stage to assessing real-world performance prior to deployment [16, 17]. Yet, rigorous evaluation of generalist robotic manipulation policies remains a challenge, especially as they are trained to work well in a wide-range of environments [18, 16, 19, 20, 21, 22]. For

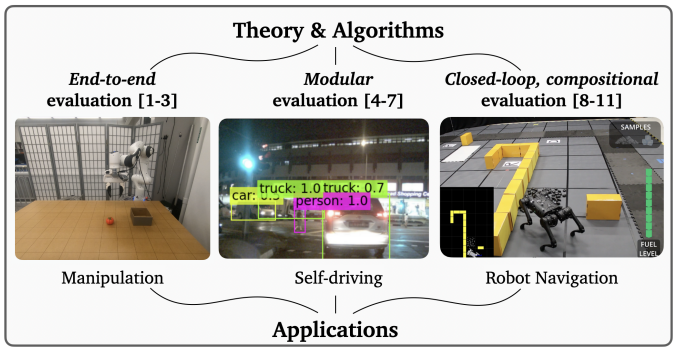

Fig. 1. I develop rigorous evaluation frameworks to systematically build and validate robot policies for safety, reliability, and robustness.

manipulation policies, real-world testing remains the gold-standard, but is expensive in time and labor. As a result, hardware testing is often limited to 20–30 trials, with performance typically summarized by empirical means without rigorous statistical analysis [16, 6]. In contrast, simulation offers a more scalable alternative [17, 23, 24, 25, 26]. However, simulations are imperfect due to poorer visual fidelity, model mismatch, and inaccuracies in contact modeling and object interaction [27, 28]. Policies are highly sensitive to these variations, making simulation a poor proxy of real behavior [29, 30, 31]. To address this evaluation challenge, I developed a statistically rigorous evaluation framework to *combine large-scale simulation with small-scale hardware tests for reliable inference of real-world performance* [2].

By establishing a real2sim formalism, I mathematically represent policy evaluation as a prediction powered inference problem [32, 33]. Each hardware rollout is paired with its simulated counterpart, and these paired evaluations are used to rectify evaluation bias in large-scale simulation, resulting in an unbiased estimates of real policy performance. Using this data, we can then construct non-asymptotically valid confidence intervals on the true mean using concentration inequalities [34]. Unlike empirical success rates, confidence intervals explicitly quantify uncertainty in real-world performance under finite evaluations. My method substantially reduces real-world evaluation effort required for state-of-the-art robot foundation models such as π_0 [19], *cutting the number of hardware trials required by 20-25%*, while returning tighter confidence intervals compared to hardware-only evaluation. Finally, the resulting confidence intervals are *statistically valid* in the finite amount of real-world and simulation trials. This research attracted the interest of Waymo and Amazon Robotics, leading

to invited talks at their research seminars. I have also explored other approaches to efficient policy evaluation such as statistically rigorous policy comparison frameworks that minimize the number of hardware trials required while controlling for false positives under binary success metrics [6] and partial progress scores [3], the use of latent policy embeddings in constructing offline proxies of real-world success rates to guide data collection for imitation learning [7], and the use of action-conditioned video models [4, 5].

Fundamental Evaluation Metrics for Perception. While the prior work focused on evaluation of end-to-end policies, *modular evaluation* becomes important as learning-enabled components are integrated as individual components in a larger system architecture. One such example is self-driving where deep neural networks are an integral backbone for perception tasks like object detection, segmentation, and tracking [35]. However, current approaches typically evaluate perception and downstream controllers in isolation, without accounting for the application-specific, system-level task (e.g., maintain safe distance from obstacles) [36, 37, 38, 39, 36]. Moreover, not all perception errors are equally safety-critical [40, 41]. To address these challenges, I introduce a confusion-based uncertainty metric to propagate semantic classification errors into probabilistic guarantees for system-level tasks [8]. I also propose task-relevant confusion matrices by incorporating logical formulas relevant to the downstream control logic and task specifications, thereby providing tighter system-level guarantees on specification satisfaction [9, 10]. Through rigorous quantitative analysis, I provide a compositional framework that enables designers to weigh trade-offs and select models best suited to their system architecture and task [11, 8].

Closed-loop, Compositional Test Plans for High-Level Task Reasoning. My work also considers evaluation frameworks that are compositional and closed-loop for testing rich semantic interactions (e.g., decision-making, high-level planning) that require evaluating agent behavior under feedback from its environment. For example, a mobile manipulator or a self-driving car should safely navigate environments with other agents while reasoning about how its actions will influence the behaviors of others. Such high-level reasoning often involves fundamentally discrete decisions — for example, selecting a route, deciding whether to yield or merge, or deciding to stop for a pedestrian that may or may not cross. In robot planning, high-level tasks and test objectives can be effectively stated in logical formalism (e.g., propositional, predicate, or temporal logic formalisms) [42]. My work uses this interpretable formalism and the mathematics of assume-guarantee reasoning [43] to design test campaigns that compose multiple test objectives into a single test instance [12, 14] or decompose complex testing requirements into simpler system- or component-level tests for a more nuanced failure analysis [14].

Evaluating these decision-making capabilities requires a closed-loop test environment that reacts dynamically to the robot's actions for a realistic assessment of its high-level reasoning capabilities in complex interactions. I developed a game-theoretic framework that programmatically synthesizes such test environments while treating the robot controller as a black box [15, 13]. These environments can include both obstacles and agents with rich physical dynamics navigating in the same space as the robot under test. The interaction is formalized as a general-sum game: the robot acts to accomplish its task and the test environment optimizes for an adversarial yet fair test. Using automata-theoretic representations of these objectives, I formulate the general-sum game as a network flow optimization with affine, yet coupled, constraints. A key insight is that the desired equilibrium condition can be expressed as an affine graph-cut constraint, reducing the overall formulation to a mixed-integer linear program (MILP), allowing us to leverage advanced MILP solvers [15]. From the optimization solution, we efficiently construct closed-loop test environments to evaluate high-level decision-making capabilities of the robot under test.

III. Future Research Agenda

Unreliable Simulators for Real-world Reliability. My prior work [2] used physics-based simulation, and is the beginning of a long-term research program in leveraging unreliable proxies for reliable inference. An important direction is active inference for targeted real-world evaluation: rather than uniformly sampling environments for hardware testing, we can prioritize scenarios where the simulator exhibits high uncertainty, and systematically characterize this uncertainty to design targeted statistical inference campaigns. Simultaneously, we need improved and scalable simulation. In contrast to physics-based simulation, action-conditioned video generation models are faster to setup, offer greater visual fidelity, and can be improved with more data [44, 45, 46]. Yet, they are prone to physically implausible rollouts that introduce evaluation bias. How do we leverage these powerful models for evaluation but correct for their imperfections?

Evaluation for Designing Better Policies. My earlier work explored using policy embeddings to inform data collection, but these representations were not consistently predictive of real-world performance across tasks [7]. A comprehensive evaluation campaign should optimize for efficiency and coverage to identify operational environments where the robot policy underperforms. Moreover, not all failure modes are equally consequential; assessing both the frequency and severity of failures is critical when prioritizing where to gather additional data for improving policy performance.

Re-thinking Metrics, Benchmarks, and Safety. To keep pace with rapidly advancing generalist robot policies, we must rethink evaluation metrics and benchmarks to ensure readiness for safe deployment. Instead of a single validation stage, we are more likely to require a curriculum evaluation strategy in which policies are tested on progressively more challenging environments, increasing in complexity only after demonstrating reliable performance in earlier stages. Identifying meaningful evaluation targets will require a careful examination of physical safety and commonsense reasoning, including in the presence of human interaction, and the development of principled methods to evaluate them.

References

[1] Jeannette M Wing. Trustworthy ai. *Communications of the ACM*, 64(10):64–71, 2021.

[2] Anonymous Authors. Anonymous title. *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

[3] Anonymous Authors. Anonymous title. *Submitted to Robotics: Science and Systems*, 2026.

[4] Anonymous Authors. Anonymous title. *Submitted to Robotics: Science and Systems*, 2026.

[5] Anonymous Authors. Anonymous title. *Submitted to ACM Computing Surveys*, 2026.

[6] Anonymous Authors. Anonymous title. *Proceedings of Robotics: Science and Systems*, 2025.

[7] Anonymous Authors. Anonymous title. *Submitted to IEEE Transactions on Robotics: Special Issue on Foundation Models for Robotics.*, 2025.

[8] Anonymous Authors. Anonymous title. In *2021 60th IEEE Conference on Decision and Control (CDC)*, 2021.

[9] Anonymous Authors. Anonymous title. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[10] Anonymous Authors. Anonymous title. *ACM Transactions on Cyber-Physical Systems*, October 2025.

[11] Anonymous Authors. Anonymous title. *ACM Transactions on Cyber-Physical Systems*, January 2025.

[12] Anonymous Authors. Anonymous title. In *NASA Formal Methods Symposium*, pages 133–155. Springer, 2022.

[13] Anonymous Authors. Anonymous title. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[14] Anonymous Authors. Anonymous title. In *NASA Formal Methods Symposium*, pages 278–294. Springer, 2023.

[15] Anonymous Authors. Anonymous title. *IEEE Open Journal of Control Systems (OJ-CSYS)*, October 2025.

[16] Hadas Kress-Gazit, Kunimatsu Hashimoto, Naveen Kuppuswamy, Paarth Shah, Phoebe Horgan, Gordon Richardson, Siyuan Feng, and Benjamin Burchfiel. Robot learning as an empirical science: Best practices for policy evaluation. *arXiv preprint arXiv:2409.09491*, 2024.

[17] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. In *Conference on Robot Learning*, pages 3705–3728. PMLR, 2025.

[18] Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, et al. A careful examination of large behavior models for multitask dexterous manipulation. *arXiv preprint arXiv:2507.05331*, 2025.

[19] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. In *Robotics: Science and Systems*, 2025.

[20] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[21] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[22] Pranav Atreya, Karl Pertsch, Tony Lee, Moo Jin Kim, Arhan Jain, Artur Kuramshin, Clemens Eppner, Cyrus Neary, Edward Hu, Fabio Ramos, et al. Roboarena: Distributed real-world evaluation of generalist robot policies. *arXiv preprint arXiv:2506.18123*, 2025.

[23] Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425*, 2024.

[24] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

[25] Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation. *arXiv preprint arXiv:2402.08191*, 2024.

[26] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

[27] Zhiyuan Zhou, Pranav Atreya, You Liang Tan, Karl Pertsch, and Sergey Levine. Autoeval: Autonomous evaluation of generalist robot manipulation policies in the real world. *arXiv preprint arXiv:2503.24278*, 2025.

[28] Elie Aljalbout, Jiaxu Xing, Angel Romero, Iretiayo Akinola, Caelan Reed Garrett, Eric Heiden, Abhishek Gupta, Tucker Hermans, Yashraj Narang, Dieter Fox, Davide Scaramuzza, and Fabio Ramos. The Reality Gap in Robotics: Challenges, Solutions, and Best Practices. December 2025.

[29] Jack Collins, David Howard, and Jurgen Leitner. Quantifying the reality gap in robotic manipulation tasks. In *2019 International Conference on Robotics and Automa-*

tion (ICRA), pages 6706–6712. IEEE, 2019.

[30] Peide Huang, Xilun Zhang, Ziang Cao, Shiqi Liu, Mengdi Xu, Wenhao Ding, Jonathan Francis, Bingqing Chen, and Ding Zhao. What went wrong? closing the sim-to-real gap via differentiable causal discovery. In *Conference on Robot Learning*, pages 734–760. PMLR, 2023.

[31] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.

[32] Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

[33] Tijana Zrnic and Emmanuel J Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62993–63010, 2024.

[34] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

[35] Peter Karkus, Boris Ivanovic, Shie Mannor, and Marco Pavone. Diffstack: A differentiable and modular control stack for autonomous vehicles. In *Conference on robot learning*, pages 2170–2180. PMLR, 2023.

[36] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[37] Anand Balakrishnan, Aniruddh G. Puranic, Xin Qin, Adel Dokhanchi, Jyotirmoy V. Deshmukh, Heni Ben Amor, and Georgios Fainekos. Specifying and evaluating quality metrics for vision-based perception systems. In *2019 Design, Automation and Test in Europe Conference & Exhibition (DATE)*, pages 1433–1438, 2019.

[38] Bowei Yan, Sanmi Koyejo, Kai Zhong, and Pradeep Ravikumar. Binary classification with karmic, threshold-quasi-concave metrics. In *International Conference on Machine Learning*, pages 5531–5540. PMLR, 2018.

[39] Oluwasanmi Koyejo, Nagarajan Natarajan, Pradeep Ravikumar, and Inderjit S Dhillon. Consistent multilabel classification. In *NeurIPS*, volume 29, pages 3321–3329, 2015.

[40] Sever Topan, Karen Leung, Yuxiao Chen, Pritish Tupekar, Edward Schmerling, Jonas Nilsson, Michael Cox, and Marco Pavone. Interaction-dynamics-aware perception zones for obstacle detection safety evaluation. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 1201–1210. IEEE, 2022.

[41] Kaustav Chakraborty, Zeyuan Feng, Sushant Veer, Apoorva Sharma, Boris Ivanovic, Marco Pavone, and Somil Bansal. System-level safety monitoring and recovery for perception failures in autonomous vehicles. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12885–12891. IEEE, 2025.

[42] Hadas Kress-Gazit, Georgios E Fainekos, and George J Pappas. Temporal-logic-based reactive mission and motion planning. *IEEE transactions on robotics*, 25(6):1370–1381, 2009.

[43] Thomas A Henzinger, Marius Minea, and Vinayak Prabhu. Assume-guarantee reasoning for hierarchical hybrid systems. In *International Workshop on Hybrid Systems: Computation and Control*, pages 275–290. Springer, 2001.

[44] Gemini Robotics Team, Coline Devin, Yilun Du, Debidatta Dwibedi, Ruiqi Gao, Abhishek Jindal, Thomas Kipf, Sean Kirmani, Fangchen Liu, Anirudha Majumdar, et al. Evaluating gemini robotics policies in a veo world simulator. *arXiv preprint arXiv:2512.10675*, 2025.

[45] 1X World Model Team. 1x world model: Evaluating bits, not atoms. Technical report, 1X, 2025.

[46] Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025.