Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

Beyond Binary Success: Sample-Efficient and Statistically Rigorous Robot Policy Comparison

David Snyder[1], Apurva Badithela[3], Nikolai Matni[1],
George Pappas[1], Anirudha Majumdar[3], Masha Itkina[2‡], and Haruki Nishimura[2‡]
[1]University of Pennsylvania, [2]Toyota Research Institute (TRI), [3]Princeton University.
[‡]Equal Advising. Corresponding Author: dsnyder5@seas.upenn.edu

***Abstract*—Generalist robot manipulation policies are becoming increasingly capable, but are limited in evaluation to a small number of hardware rollouts. This strong resource constraint in real-world testing necessitates both more informative performance measures and reliable and efficient evaluation procedures to properly assess model capabilities and benchmark progress in the field. This work presents a novel framework for robot policy comparison that is sample-efficient, statistically rigorous, and applicable to a broad set of evaluation metrics used in practice. Based on safe, anytime-valid inference (SAVI), our test procedure is *sequential*, allowing the evaluator to *stop early* when sufficient statistical evidence has accumulated to reach a decision at a pre-specified level of confidence. Unlike previous work developed for binary success, our unified approach addresses a wide range of informative metrics: from discrete partial credit task progress to continuous measures of episodic reward or trajectory smoothness, spanning both parametric and nonparametric comparison problems. Through extensive validation on synthetic and real-world evaluation data, we demonstrate up to 70% reduction in evaluation burden compared to standard batch methods and up to 50% reduction compared to state-of-the-art sequential procedures designed for binary outcomes, with no loss of statistical rigor. Notably, our empirical results show that competing policies can be separated more quickly when using fine-grained task progress than binary success metrics.**

I. Introduction

Recent advances in robot policy synthesis incorporate increasing complexity throughout the design process, training on large-scale datasets, using sophisticated, stochastic architectures, and requiring commensurate increases in training resources. These advances have led to significant improvements in solving dexterous and long-horizon tasks [9], inferring semantic information from context [34], and safely interacting with numerous other autonomous agents [43]. However, this complexity makes design decisions like the choice of dataset, the training or fine-tuning procedure, and the network architecture analytically opaque. The value of a novel intervention cannot be determined from first principles but instead requires rigorous analysis [2, 42] of its effect on empirical performance.

Unfortunately, rigorous analysis of empirical performance is a significant challenge in the robotics setting. Hardware evaluation is the gold-standard measure, but is expensive and slow to collect. Evaluations in simulation reduce the time burden, but continue to suffer from sim-to-real gaps, making them an imperfect proxy [3]. Further, evaluation metrics can be quite coarse. Binary measurement of task success or failure, the *de facto* standard for measuring the performance of robot manipulation policies (particularly on hardware), can obscure valuable information about the robot behavior [11, 42]. For instance, a policy that completes 90% of the task is clearly better than a policy that is frozen the whole time, yet their success rates would be identically 0%. Finally, rigorous evaluation must account for inherent uncertainty, including in environment configuration and policy action selection (which is often stochastic [20]). This uncertainty means that observed empirical performance is a noisy estimate of, and *not equivalent to*, the true expected performance of the policy.

Since the value of design interventions is implicitly counterfactual (see Figure 1), policy comparison is a particularly important form of evaluation to reliably determine the effects of design changes. For example: 'does a change in policy architecture [41, 60] or action tokenization [63] improve downstream performance?' Or: 'what set of expert data is the most valuable [85] to improve policy performance?' Importantly, comparison must account for uncertainty in evaluating both the new design and the baseline. Making rigorous, statistically assured decisions for these problems is necessary to ensure reliable progress within the field.

The costs of robot evaluation and the complexity of the robot policy design space motivate the development of a versatile framework for policy comparison. Ideally, such a framework would apply to very general performance measures, maintain statistical rigor, and ensure maximal sample efficiency via sequentialized evaluation. Current methods are deficient in at least one of these respects. Active learning [5] and asymptotic [16, 82] approaches are not statistically rigorous, particularly in small-data regimes. Batch and fully nonparametric methods [8, 27, 81, 82] are not maximally sample efficient: the former due to the batch evaluation, and the latter due to not tailoring to the comparison setting. Near-optimal approaches [49, 69, 74] are constrained to narrow metrics (e.g., binary success rate), and do not readily generalize.

This work presents **N**onparametric **S**equential **CO**mparison for **R**igorous **E**valuation (N-SCORE)[1] to compare robot policy performance. N-SCORE is designed to address the shortcomings of prior approaches by explicitly accounting for all three of the preceding desiderata – rigor, sample efficiency, and generality – within the decision rule synthesis. We summarize our **contributions** as follows:

[1]Project website and links to codebase and paper results can be found here.

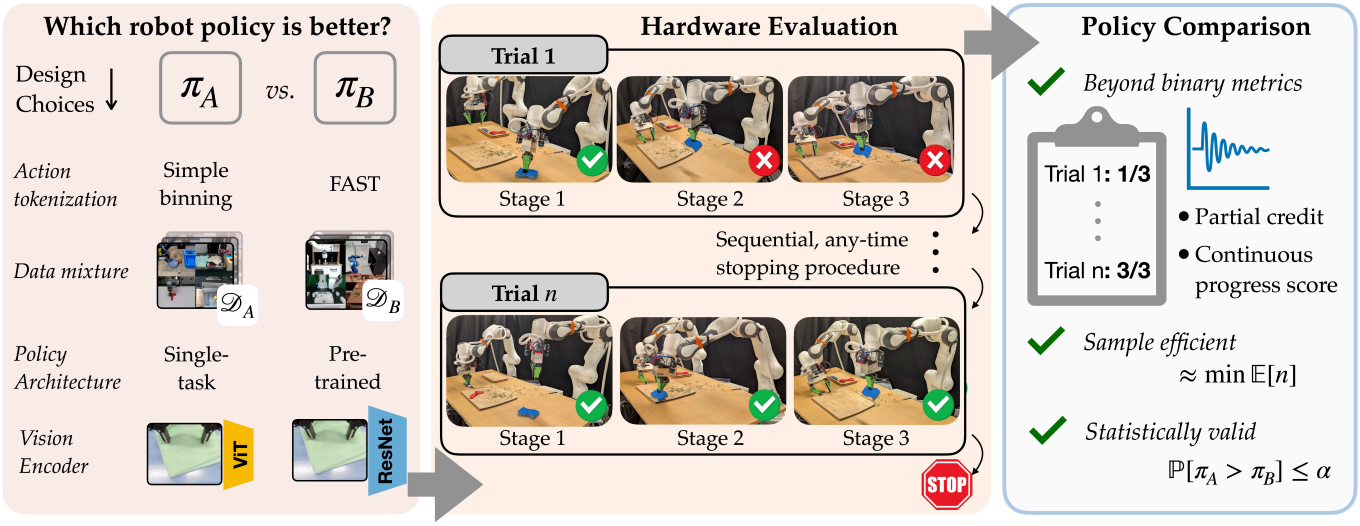

Fig. 1: The evaluation context of the N-SCORE procedure. (Left) We consider the general problem of policy comparison, which arises out of counterfactual design decisions in the policy synthesis process. (Middle) Evaluation on hardware or in high-fidelity simulation is the gold standard to assess the effect of such changes, but is costly to collect. (Right) N-SCORE is a sequential evaluation procedure that is statistically rigorous, sample efficient, and generalizes to rich, diverse measures of robot performance.

1) We introduce N-SCORE, a novel method for sequential policy comparison with general progress metrics.
2) We prove that N-SCORE is statistically rigorous in the sense of controlling Type-1 Error, and empirically demonstrate its improved sample efficiency compared to state-of-the-art (SOTA) baselines.
3) We comprehensively evaluate N-SCORE on large, high-impact evaluation datasets in robotics comprising over 4500 hardware evaluation rollouts and 2000 high-fidelity simulation rollouts [6, 9]. The results demonstrate significant savings in evaluation effort of up to 70% over batch methods and 50% over methods using binary success measures to rigorously certify policy improvements.

II. Related Work

Efficient evaluation is increasingly appreciated as a critical aspect of the robot policy design process. The fundamental concern of these approaches is the strict limit on available hardware data. A core focus lies in constructing proxy signals (e.g., using simulations or evaluator preferences) to circumvent this constraint. By contrast, statistical testing approaches seek to employ the available hardware data more efficiently. These approaches are complementary; improvements in one domain augment those in the other.

A. Robot Policy Evaluation

Hardware and Simulation Evaluation. In robot manipulation, evaluation constraints often limit evaluation to 10-60 trials; the statistical validity of such comparisons is often not reported [42, 69]. To address this constraint, recent works have introduced standardized benchmarks [23, 30, 39, 55, 84], cloud-based evaluation platforms [10, 54, 86, 87], and distributed evaluation infrastructure [6] to increase available data. Further, recent works have proposed active sampling of policy-task pairs to improve efficiency [5]. However, none of these approaches confer statistical rigor on downstream comparisons. Similarly, though policy evaluation via simulation [50, 53, 57, 64, 72] or via world models [29, 33, 51, 65, 70, 71, 73, 88] show promise, the sim-to-real gap [3] makes rigorous comparison difficult. Some recent works have proposed statistically rigorous methods to combine small-scale hardware testing with large-scale simulation evaluation [7, 56], but these do not address policy comparison.

Evaluation Metrics. Recent studies emphasize the importance of detailed rubrics and task progress scores [6, 9, 42] over traditional binary success metrics. In parallel, finetuning procedures [34] and policy ranking problems [6, 21] have introduced indirect signals of policy performance, including reinforcement learning (RL) rewards and human preferences. However, none of the approaches give rigorous methods for decision-making. Whereas the state-of-the-art procedure for policy comparison under binary success metrics [69] is rigorous and sample-efficient, current methods for more informative signals are not. This paper introduces a novel test procedure that scales rigorous evaluation to general metrics.

B. The Parametric Policy Comparison Problem

The *parametric* setting for policy comparison arises when the evaluator designs the performance measure to have definite distributional structure (e.g., binary success or discrete partial credit). This structure specifies a tradeoff between the generality, sample efficiency, and correctness of the comparison.

Classic tests [8, 13, 27] were designed to optimize power for Bernoulli (binary) outcomes. Some subsequent results use tools from sequential analysis [68] to improve the sample efficiency [48, 78, 79]. In this regime, the STEP proce-

dure [69] is state-of-the-art. Other developments extend the generality of these tests to broader classes of parametric distributions [49, 74]. However, simultaneously extending both parametric generality and sample efficiency is difficult, requiring an exact solution of the Wald-Bellman partial differential equation [14, 48, 75]. For more complex parametric families, finding this solution quickly becomes intractable [19, 25].

This motivates work that extends generality by making approximations in the large-data regime. A canonical example is Welch's t-Test [82], which is often used for batch comparison [9, 24, 83]. Further *asymptotic* results [12] can extend generality [16–18] *and* improve sample efficiency [44–47], at the cost of losing rigorous statistical assurances. This cost is significantly more pronounced in the low-data regime common to robotic evaluation [42]. Thus, in our setting, conclusions derived from asymptotic approaches are difficult to assess, as their validity is not guaranteed in the low-data regime.

C. Nonparametric Methods and Safe, Anytime-Valid Inference

Nonparametric methods seek to exchange some exploitable structure for generality in application. Within the batch regime, these are often termed 'distribution free,' as in the case of conformal prediction and related techniques [4, 67, 77]. However, they have two downsides: limited capacity to represent the underlying data distribution, and limited generalization to sequential settings. Kernel density estimation (KDE) addresses the first constraint by directly constructing a representation of the data-generating process [62]; see [15] for an overview. Critically, KDE is more efficient in low-dimensional settings [36] because it can quickly adapt to structure in the data. However, alone, it is not generally valid in finite samples [80].

To address the second constraint, work in the line of safe, anytime-valid inference (SAVI) considers *sequentialization* of estimation and decision problems, which act in an *online fashion* [66]. A core benefit is safety in the face of p-hacking [37], or 'data dredging' [28, 66]. Additionally, though not parametric, the framework does allow for methodological fine-tuning to the particular problem setting (as in [22, 52, 74]). Of these approaches, the WSR method [81] is closest to our own, considering the problem of estimating the mean of a general class of random variables (see Definition 1). Importantly, N-SCORE utilizes ideas from KDE to tailor a SAVI framework *specifically to the comparison problem*, achieving state-of-the-art generality and sample-efficiency.

III. Preliminaries

We model a robot that must complete a task while subject to stochastic uncertainty in its environment as a Partially Observable Markov Decision Process (POMDP) [38]. The uncertainty might arise in the initial environment configuration or in the robot's action selection. We assume the existence of a real-valued scalar performance measure R encoding the degree of success the robot demonstrates in completing the task. Conditioned on an arbitrary robot policy π mapping state observations to actions, the stochastic environment uncertainty is compressed to uncertainty over performance outcomes (i.e., by running the policy and measuring its level of success). Thus, all the potential complexities of the task, real-world system dynamics, and policy class are compressed to a (possibly complicated and *unknown*) distribution over the performance measure $\mathcal{D}_R$. We specify a standard regularity assumption necessary for all subsequent analysis:

Assumption 1 (Regularity (i.i.d))**.** In each evaluation trial, the environment conditions are drawn in an independent and identically distributed (i.i.d.) fashion from an underlying distribution of interest (which need not be explicitly represented by the evaluator). By implication, the resulting draws of evaluation scores $r \sim \mathcal{D}_R$ are i.i.d. as well.[2]

Our goal is to make rigorous comparisons for as general a class of $\mathcal{D}_R$ as possible. With this in mind, we make a single restriction: that R be a *progress metric*, defined as a performance measure that is bounded w.p. 1.

Definition 1 (Generalized Progress Metric)**.** A real-valued random variable $M(\omega)$ is termed a '*generalized progress metric*' if it is bounded; that is, $\mathbb{P}[M \in [0,1]] = 1$.[3]

Unless stated otherwise, we assume R is a progress metric as per Definition 1 but make no other structural assumptions.

We conclude this section with a formalization of the comparison problem in the robotics context. In the most general form, we are tasked with quickly and accurately comparing the means of two distributions over a progress metric R:

$$\text{Test: } \mathbb{E}_{r\sim\mathcal{D}_R^{[0]}}[r] < \mathbb{E}_{r\sim\mathcal{D}_R^{[1]}}[r].$$

This encompasses many practical problems, for example: comparing two policies with different architectures or trained on different data (π_0 and π_1), comparing distribution shift in task realizations for a single policy, or comparing the effects of different hyperparameters in training. Importantly, because the true means are unknown and not directly observable, a testing decision must be made on the basis of empirical performance in evaluation. Due to stochastic uncertainty in gathering this data, any decisions are inherently *probabilistic* over the collection of evaluation data. Thus, we seek statistical assurances that bound the probability of an incorrect inference:

$$\begin{aligned} &\text{If:} \quad \mathbb{E}_{\mathcal{D}_R^{[0]}}[R] \geq \mathbb{E}_{\mathcal{D}_R^{[1]}}[R] \\ &\text{Then: } \mathbb{P}\left[\text{Wrongly decide } \mathbb{E}_{\mathcal{D}_R^{[0]}}[R] < \mathbb{E}_{\mathcal{D}_R^{[1]}}[R]\right] \leq \alpha^*. \end{aligned} \tag{1}$$

The challenges to obtaining a result in the form of Equation (1) are twofold. First, the decision must be statistically rigorous—it must safeguard against inadvertently reporting statistical noise as genuine improvement. This can require a substantial number of evaluations when the difference is small or the desired confidence $1-\alpha^*$ is very high. However, the result also must be obtained as quickly as possible, so that the evaluator

[2] i.e., via a pushforward measure representing a rollout of π on the environment draw.

[3] This immediately extends to real-valued performance measures that are bounded in an interval $[A, B]$ via normalization [35, 61, 81].

can efficiently allocate hardware or computational resources to begin investigating other problems. Indeed, allocating limited evaluation resources is a major bottleneck in improving policy performance [69]. The sequential nature of N-SCORE balances statistical confidence and speed of decision-making, adaptively making decisions when precisely enough evidence has accumulated while minimizing unnecessary evaluation trials.

IV. Problem Formulation

Section III introduced the policy comparison problem, which seeks guarantees as in Equation (1). The challenge in designing a test is the complexity of $\mathcal{D}_R^{[\cdot]}$, as our procedure must be robust to *any such progress metrics arising in practice*.

A. The Formal Hypotheses

We formulate the problem in the context of frequentist statistical testing [12]. To do so, we construct a general null (skeptical) hypothesis and an alternative (desired) hypothesis which reflect *all possible* cases of Equation (1).

In the nonparametric case (see Section II-C), we consider the set $\mathcal{M}_{[0,1]}$ of all Lebesgue-measurable distributions on the real interval $[0,1]$ (i.e., all progress metrics per Definition 1). We partition the product measure $\mathcal{M}^2_{[0,1]}$ into two parts:[4]

$$S^- := \{(\mathcal{D}_R^{[0]}, \mathcal{D}_R^{[1]}) \in \mathcal{M}^2_{[0,1]} : \mathbb{E}_{\mathcal{D}_R^{[0]}}[R] \geq \mathbb{E}_{\mathcal{D}_R^{[1]}}[R]\} \text{ and}$$
$$S^+ := \{(\mathcal{D}_R^{[0]}, \mathcal{D}_R^{[1]}) \in \mathcal{M}^2_{[0,1]} : \mathbb{E}_{\mathcal{D}_R^{[0]}}[R] < \mathbb{E}_{\mathcal{D}_R^{[1]}}[R]\}.$$

Comparing the means amounts to formulating null and alternative hypotheses as $\mathcal{H}_0 : (\mathcal{D}_R^{[0]}, \mathcal{D}_R^{[1]}) \in S^-$ and $\mathcal{H}_1 : (\mathcal{D}_R^{[0]}, \mathcal{D}_R^{[1]}) \in S^+$, respectively.

These can be intuitively understood as "all possible true states of the world in which the novel innovation is not better" ($\mathcal{H}_0$, the 'skeptic'), and "all possible true states of the world in which the novel innovation is better" ($\mathcal{H}_1$, the 'desired result'). Equation (1) amounts to being $1-\alpha^*$ confident that the true state is *not* in S^-.

B. Measuring the Quality of an Evaluation Algorithm

Having formulated the test hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, we consider how to measure the quality of any potential evaluation scheme. These measures can be understood intuitively: the optimal evaluation protocol should *make correct decisions* and *make the decisions as quickly as possible*.

There are precise mathematical analogues of these desiderata. In particular, there are two types of errors which can be made. First, in the case that $\mathcal{H}_0$ is true, the evaluation algorithm may incorrectly decide that $\mathcal{H}_1$ is true. This is termed a false positive, or a 'Type-1 Error.' The rate at which such errors occur (under a particular decision-making protocol) is denoted $\alpha \in (0,1)$. Equation (1) precisely amounts to the statement "the Type-1 Error rate of our evaluation method must be less than α^*." In the case that $\mathcal{H}_1$ is true, the protocol may incorrectly decide that $\mathcal{H}_0$ is true. This is a false negative, or a 'Type-2 Error.' The rate at which this occurs is denoted $\beta \in (0,1)$. Semantically, a Type-1 Error corresponds to reporting a false discovery — that the new policy is better *when it actually is not*. A Type-2 Error corresponds to a missed discovery — failing to discern that the new policy is better *when it actually is*. These two error types combine to give a complete accounting of the algorithm's correctness. The last metric pertains to sample efficiency: how long the algorithm takes to make a decision, denoted $\mathbb{E}[N]$.[5] As presented, an optimal evaluation algorithm — one that is fast and correct — minimizes all three of the measures $\{\alpha, \beta, \mathbb{E}[N]\}$.

C. Efficient Tests as a Multi-Objective Optimization

Unfortunately, there are fundamental tradeoffs which prevent simultaneous minimization of all of these metrics. Therefore, we adopt the Neyman-Pearson testing approach [59], which normatively chooses the Type-1 Error rate as the quantity to be rigorously controlled. This amounts to requiring that $\alpha \leq \alpha^*$ be treated as a hard constraint. Having made this selection, the problem of designing an evaluation procedure reduces to finding (near-)optimal solutions to the following multi-objective optimization problem:

$$\gamma^* = \arg\inf_{\gamma \in \Gamma} \left[\mathbb{E}[N](\gamma) + \lambda \cdot \beta(\gamma)\right] \quad \text{s.t. } \alpha(\gamma) \leq \alpha^*. \tag{2}$$

Here, Γ denotes a space of decision-making rules (evaluation procedures) and $\lambda \in [0, \infty)$ is a nonnegative parameter trading off the expected time to decision and false negative rate.

Efficiently solving Equation (2) results in an evaluation procedure that is guaranteed to maintain statistical rigor and quickly and effectively detects changes in performance. This provides the roboticist with the capacity to quickly iterate on new innovations, while ensuring justified and well-calibrated confidence in reported performance improvements.

V. Methodology

Unfortunately, solving Equation (2) for γ^* directly is challenging, due to the high (infinite) dimension of the space of progress metrics. As such, we will rely on the SAVI framework (see Section II-C), which restricts to a policy class of near-optimal tests via tractable approximations. We motivate the N-SCORE procedure (as an instance of a SAVI method) via the construction of an intuitive process to distinguish S^- from S^+. Concretely, we will construct a scalar-valued dynamical system that behaves fundamentally differently when $\mathcal{H}_0$ is true than when $\mathcal{H}_1$ is true, based on the empirical evidence observed during evaluation. This difference is precisely in the sense of being stable in the former case and unstable in the latter. The testing problem then reduces to assessing the stability properties of the dynamical system. From this

[4] For a parametric family Θ (e.g., Bernoulli distributions), the same formalism holds with a restriction to $\Theta^2 \subseteq \mathcal{M}^2_{[0,1]}$

[5] For technical reasons, sample efficiency must be posed with respect to a measure over S^+. This amounts in practice to prior beliefs over features like the size of the gap in performance; the reader may safely assume, e.g., an 'uninformative' uniform measure.

development, the properties of N-SCORE can then be proven via a similar process as existing SAVI methods (Section V-B).

A. Constructing an 'Evidence Integrator'

A natural correlate to $\mathcal{H}_1$ is the difference in the evaluation reward; if we are on the n^{th} evaluation trial, we consider:

$$\Delta \text{ evidence}_n \approx \Big[\xi \cdot (r_{1,n} - r_{0,n})\Big],$$

where $\xi > 0$ is a positive scaling factor and $r_{0,n}$, $r_{1,n}$ are the observed progress values. Intuitively, the evidence for $\mathcal{H}_1$ is positive when $r_{1,n} > r_{0,n}$ and negative otherwise. Note that by assumed independence of the evaluation trials, this relationship does not depend on n. What remains is to design the appropriate rate of integration of this evidence. Results in the SAVI literature and across a variety of statistical estimation contexts suggest that the optimal aggregation rate [66] is *multiplicative*, resulting in a measure of aggregate evidence X_n (setting $X_0 = 1$ w.l.o.g.) that evolves according to:

$$X_{n+1} = (1 + \xi \cdot (r_{1,n} - r_{0,n}))X_n. \tag{3}$$

Therefore, when the evidence is positive for $\mathcal{H}_1$, the growth rate is greater than one, and the system is locally unstable. Conversely, when it is negative (i.e., in favor of $\mathcal{H}_0$), the growth rate is less than one and the system is stable. We can additionally select $\xi_n = g(\mathcal{F}_{n-1})$ online based on the evidence accumulated so far (the 'natural filtration' $\mathcal{F}_{n-1}$), as long as the choice of ξ_n is independent of $\{r_{0,n}, r_{1,n}\}$.

Intuitively, ξ_n modulates the degree of confidence in marginal changes to X_n. Large ξ_n grow the process more quickly when data is favorable, but are penalized more harshly when it is not. Identifying an effective strategy to modulate ξ_n online is central to improving sample efficiency across a broad array of evaluation problems. To do this, N-SCORE utilizes intuition from kernel density estimation and the structure of the discrete partial credit setting to optimize ξ_n; this is represented as a family N-SCORE$_K$, where $K \in \mathbb{N}$ is a real-valued parameter akin to the kernel bandwidth. By constructing explicit nonparametric representations of the independent $\mathcal{D}_R^{[i]}$, ξ_n may be selected to maximize the expected growth rate of the stochastic process X_n, thereby explicitly minimizing the time-to-decision subject to the Type-1 Error control of SAVI-structured algorithms (see Remark 1).

The last step is designating an evidence threshold, amounting to the idea that "X_n has become sufficiently unstable as to provide sufficient evidence that the true state of the world is not in $\mathcal{H}_0$." We designate the threshold to be $1/\alpha^*$, for a desired Type-1 Error rate α^* of the test procedure. That is,

$$\begin{aligned} &\text{If:} \quad X_n \geq \frac{1}{\alpha^*} \\ &\text{Then:} \quad \Big[\text{Stop at step } n, \text{ and conclude } \mathcal{H}_1 \text{ is true.}\Big]. \end{aligned} \tag{4}$$

This procedure is operationalized in Algorithm 1. As shown in Section V-B, this evaluation protocol efficiently balances time-to-decision and correctness, while rigorously controlling the Type-1 Error rate at tunable, pre-specified level α^*. These properties enable rigorous confidence in comparison problems (yielding statements of the form of Equation (1)) while minimizing the evaluation burden necessary to reliably form them.

Algorithm 1 N-SCORE Evaluation Protocol

Input:
Type-1 error limit $\alpha^* \in (0, 1)$, evaluation limit $N_{\max} > 0$.
Initialize:
$X_0 = 1$; $\bar{X} = 1$; $\xi_0 = 0$; $\mathcal{F}_0 = \{\emptyset\}$; $n = 1$.
while $\bar{X} < 1/\alpha^*$ **and** $n \leq N_{max}$ **do**
  Observe evaluation progress scores: $\{r_{0,n}, r_{1,n}\}$
  Compute increment: $a_{n-1} = 1 + \xi_{n-1}(r_{1,n} - r_{0,n})$
  Update martingale: $X_n \leftarrow a_{n-1} \cdot X_{n-1}$
  Update test statistic: $\bar{X} \leftarrow \max\{\bar{X}, X_n\}$
  Update filtration: $\mathcal{F}_n \leftarrow \mathcal{F}_{n-1} \cup \{r_{0,n}, r_{1,n}\}$
  Update $\xi_n \leftarrow \underset{[0,1]\subseteq\mathbb{R}}{\text{proj}}\big[g(\mathcal{F}_n)\big]$ // project $g(\mathcal{F}_n)$ to [0, 1]
  $n \leftarrow n + 1$
end while
if $\bar{X} < 1/\alpha^*$ **then**
  return Fail to Reject Null
else
  return Reject Null
end if

B. Theoretical Properties of Algorithm 1

We briefly present several key theoretical results of the evaluation protocol effectuated in Algorithm 1. The import of these results, along with proof sketches, are included *in situ*. Detailed proofs are deferred to Appendix A.

We begin with a critical lemma pertaining to a property of the evidence aggregation process in Equation (3). This property amounts to a statement that the stochastic process X_n is stable in expectation *for all* elements $h \in \mathcal{H}_0$.

Lemma 1 (Null Stability (NSM) Property)**.** *Consider the stochastic process $\{X_n\}$ defined in Equation (3), setting (w.l.o.g.) $X_0 = 1$. Then the expectation of $\{X_n\}$ is contracting in time with respect to the current value, for all $h \in S^-$ for any $\xi_n \in [0, 1]$. That is:*

$$\sup_{h\in S^-} \sup_{\xi_n\in[0,1]} \mathbb{E}_{r_0,r_1\sim h}\left[\frac{X_{n+1}}{X_n}\middle|\mathcal{F}_n\right] \leq 1. \tag{5}$$

Lemma 1 is necessary for ensuring stability of the process in Equation (3) because it rules out (with high probability) that evidence increments generated by any element of $\mathcal{H}_0$ will cause the process to grow by very much. This is a necessary step to ensure Type-1 Error control and enforce the hard constraint in Equation (2), which is formalized in Theorem 1.

Theorem 1 (Type-1 Error Control of Algorithm 1)**.** *Consider the evaluation procedure in Algorithm 1, utilizing the process defined in Equation (3). Then*

$$\mathbb{P}\Big[\mathcal{H}_0 \textit{ is true} \Big| \textit{ decide } \textbf{\textit{Reject Null}}\Big] \leq \alpha^* \tag{6}$$

Proof: The proof uses Lemma 1 in concordance with Ville's Inequality [76] to bound the probability of observing large values of X_n. After verifying several ancillary conditions and matching constants, the bound can be computed directly. More detail in included in Appendix A-B. ∎

Enforcing Type-1 Error control ensures high confidence (at level $1-\alpha^*$) that reported significant innovations and improvements to the state-of-the-art are strictly separated from the inherent noise in robotic evaluation.

Remark 1 (Efficient Optimization of ξ_n)**.** Consider the stochastic process family described in Equation (3), and let $\mathcal{F}_n$ be the natural filtration $\{(r_{0,i},\ r_{1,i})\}_{i=1}^{n-1}$ at step n. There exists an efficient algorithm to maximize (online) over $\{\xi_n\}_{n=1}^{N}$ the expected growth rate of $\{X_n\}_{n=1}^{N}$.

For brevity, further details about this optimization are deferred to Appendix A-C. However, we emphasize that this optimization has strong practical benefits, yielding a state-of-the-art near-optimal solution to Equation (2) for general progress metrics.

VI. Experimental Results

We evaluate N-SCORE across a broad suite of evaluation settings, including some of the largest available datasets for real-world robot comparison with task progress metrics, such as RoboArena [6] and the LBM 1.0 study [9]. To organize the results, we tie them to three core research questions (RQs):

1) (**RQ1: Sequential Comparison**) Do we observe sample efficiency benefits from sequential evaluation?
2) (**RQ2: Informative Metrics**) Are there sample efficiency benefits from using more informative evaluation metrics than coarse binary success?
3) (**RQ3: Technical Novelty**) What are the respective strengths of N-SCORE and baseline approaches?

The **Policy Comparison** component (RQ1 and RQ2) is intended to demonstrate *practical use cases* for the roboticist, translating the theoretical guarantees from Section V into practical savings. These necessarily focus on more realistic settings, prioritizing real-world hardware results and high-fidelity simulation evaluations. Such datasets are generally expensive to collect and are thus intended to be evaluated single-shot. This still allows robust, rigorous comparison of the underlying robot policies, but not of the baseline evaluation methods themselves. The **Baseline Comparison** component (RQ3), by contrast, illustrates the relative benefits of N-SCORE with respect to existing evaluation procedures. To answer this question, we must be able to run each of the evaluation procedures many times to get a reliable sense of their behavior, as the randomness in any individual dataset may be misleading. This necessitates *independent* redraws of the evaluation sequences and, therefore, a reliance on synthetic data (i.e., with known distribution parameters) and repeatable, low-cost simulation data (i.e., standard RL benchmark tasks).

For space constraints, additional properties of N-SCORE, including computational burden and scaling laws with the α parameter, are discussed in Appendix B. For all subsequent experiments, we use α in the standard range of $[0.01, 0.05]$. Further, any test on real datasets that fails to reach a decision is denoted by '–' in the associated tables of results. These conditions arise when empirical means are very close and the dataset is too small to separate the policies with sufficient confidence $1-\alpha^*$.

Method	Valid?	Bernoulli	Parametric	Continuous
STEP [69]	**Yes**	**SOTA**	**N/A**	**N/A**
θ-SAVI [74]	**Yes**	$\preceq$	**SOTA**	**N/A**
WSR [81]	**Yes**	$\preceq$	$\preceq$	$\preceq$
N-SCORE (ours)	**Yes**	$\preceq$	**SOTA**	**SOTA**

TABLE I: **Efficiency of Baseline Comparison methods for all evaluation contexts**. As we show in relation to RQ3, N-SCORE *complements* the STEP procedure, extending rigorous evaluation to performance measures beyond binary success and failure. Additionally, N-SCORE *Pareto-dominates* the θ-SAVI and WSR procedures (as indicated by the 'SOTA' labels), recovering the efficiency of the former procedure when parametric structure is present and the generality of the latter procedure when it is not.

A. Baseline Methods

We introduce three major baselines, proceeding in order of increasing generality. The **Baseline Comparison** (RQ3) context of each procedure is represented in Table I. The recently proposed STEP procedure [69] demonstrated state-of-the-art (SOTA) performance in the setting of binary success metrics. However, because STEP is tailored to this important but narrow setting, it does not generalize to more informative metrics. Another recent work proposed a safe, anytime-valid inference approach to parametric comparison problems, motivated by settings with discrete partial credit [74]. This method, termed θ-SAVI to denote its parametric nature, is more general than STEP, as it retains validity for any parametric comparison setting (defined in Section II-B). However, as with STEP, θ-SAVI cannot extend to nonparametric performance metrics such as continuous-valued progress scores. The most general evaluation procedure is the hypothesis-testing dual of the WSR estimation method [81], which is also most similar among the baselines to our approach. As WSR and N-SCORE are equally general, their comparison will center on sample efficiency.

B. ***Policy Comparison***: *Sequential Comparison (RQ1) and Informative Metrics (RQ2)*

Our experiments on real-world robotic datasets are designed to investigate the practical capacity of sequential methods and informative evaluation metrics, when paired with efficient algorithms like N-SCORE, to save the evaluator significant time and resources in their evaluation protocols. We decompose this problem into two parts: investigating the benefits of sequentiality under binary measures (in the vein of Snyder et al. [69]), and then investigating the relative benefit of informative metrics given the use of sequential procedures.

Method → / Task ↓	Progress Metrics N-SCORE$_K$	WSR	Binary Success/Failure Metrics STEP	N-SCORE$_2$	θ-SAVI	WSR
DumpVegetables	38	36	154	–	–	–
PutContainer	33	36	169	–	–	–
PutFruit	10	10	40	46	45	52
SeparateFruit	18	20	77	98	98	103
TurnUpsideDown	–	–	–	–	–	–
Total (Simulation)	598	604	1280	1488	1486	1510
BikeRotor	–	–	–	–	–	–
CutApple	29	29	–	–	–	–
CleanLitter	36	23	–	–	–	–
ClearCounter	16	15	23	25	30	46
SetUpBreakfast	12	13	15	19	19	17
Total (Hardware)	286	260	376	388	398	426

TABLE II: **Time-to-decision for all simulation (top) and hardware (bottom) policy comparisons for evaluation context in [9]**. All hardware (resp. simulation) tasks utilize $N = 50$ (resp. $N = 200$). Any comparison not reaching a decision is denoted "–" and counted at N for computing sample complexity. The total number of trials is twice the column sum (because there are two policies being compared). Thus, a comparison method may access up to the batch size of 500 (resp. 2000) evaluations on hardware (resp. in simulation). (Right) Sequential evaluation saves significant evaluation effort over batch methods, with STEP optimal. (Left) Using more informative partial credit metrics can provide substantial savings in evaluation effort *independent* of the sequential evaluation. For these evaluations, the savings w.r.t. the batch size constitute 50% on hardware and 70% in simulation.

1) Results on LBM 1.0 (Binary Evaluation) [9]: We begin by considering binary evaluation data for robot policies from the LBM 1.0 study results [9], presented on the right-hand side of Table II. For each of five hardware (resp. simulation) tasks, $N = 50$ (resp. $N = 200$) rollouts were collected for both a pretrained-and-finetuned policy and a single-task policy trained from scratch. Therefore, the batch hardware (resp. simulation) dataset contains a total of 500 (resp. 2000) rollouts. For each task, we test whether the pretrained policy outperforms the single-task policy. For each task and evaluation method, the time-to-decision (TTD) is recorded in terms of the number of evaluations *per policy* before a significant result at level $\alpha = 0.05$ was reached. On hardware (resp. simulation) tasks, we observe savings of 20-25% (resp. 25-35%) over the total batch evaluations, corresponding to savings of 100-125 hardware (resp. 500-700 simulation) rollouts.

2) Results on LBM 1.0 Partial Credit Evaluation [9]: We now consider informative, partial-credit metrics within the context of Barreiros et al. [9], moving to the left-hand side of Table II. Here, partial credit measures (e.g., completion rate of K subtasks) were pre-specified to provide fine-grained information on the policy performance and behavior. In every task, this varied between six and eight partial credit outcomes; every subtask was weighted equally in the scoring. All evaluation scores are linked to *the same rollouts* as in Section VI-B1.

As shown in the left-hand side of Table II, partial credit metrics yield significant practical savings that complement sequential evaluation. On hardware (resp. simulation), we observe 24-30% (resp. 50-60%) improvement in sample complexity over

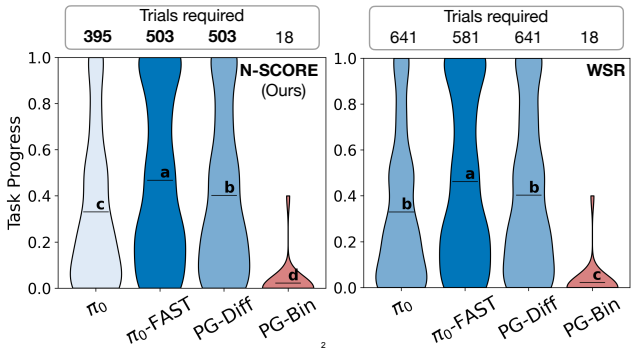

Fig. 2: **Policy performance comparisons on crowd-sourced real-world evaluations on the DROID [40] setup from RoboArena [6]**. The violin plots represent empirical distributions of observed results. Policies with different letters are statistically distinguishable by the method. Policies are compared at a global error bound of $\alpha = 0.05$ with a Bonferroni (union bound) correction [9].

binary measures (even when the latter uses SOTA methods like STEP). Total savings over batch procedures combine to 40-45% (resp. 70%), corresponding to a reduction of nearly 240 hardware (resp. 1,400 simulation) rollouts necessary to distinguish the policy with higher mean performance.

3) Multipolicy Ranking on the RoboArena [6] Dataset: We conclude the **Policy Comparison** discussion with results on multipolicy comparison (i.e., ranking). In this setting, $M \geq 2$ policies are to be ranked in terms of mean performance; this corresponds to requiring $\binom{M}{2}$ policy-pairwise comparisons. For each policy, N evaluations are gathered; using a Bonferroni correction (union bound) [9] on α^*, these datasets are *shared* in running each of the $\binom{M}{2}$ tests at level $2\alpha^*/(M(M-1))$. The sample complexity for a *policy* is the number of samples needed to decide all of the policy's $M-1$ pairwise tests.

We instantiate a policy ranking problem using four policies from the open-source RoboArena [6] benchmark (for more detail, see Appendix D-D). Because the evaluation scores are continuous, only N-SCORE and WSR can be applied. In Figure 2, we illustrate the ranking TTD, computed as the number of trials required for each policy to decide *all* of its pairwise comparisons. For this dataset, 641 rollouts are stored for each policy (corresponding to $\approx 2,560$ total hardware rollouts). N-SCORE is able to recover a full ranking of the policies at level $\alpha = 0.05$ using only 1,420 of the evaluations, a savings of nearly 45%, and equivalent to a significant investment (more than 1,000 rollouts) of hardware resources. It is particularly instructive to note that the PG-Bin policy was nearly immediately detected as being inferior to the other three, allowing the evaluator to stop further rollouts after fewer than 20 evaluations, more than an order of magnitude fewer than required for the three more closely-competing policies.

The preceding evaluations give strong empirical evidence for answering RQ1 and RQ2 – the **Policy Comparison** questions – in the affirmative: sequentialized evaluation reliably

	Bernoulli Data		Nonparametric Data	
	TTD (↓)	$1-\beta$ (↑)	TTD (↓)	$1-\beta$ (↑)
STEP	95.1	0.953	–	–
θ-SAVI	117.6	0.962	–	–
N-SCORE$_2$	117.9	0.965	–	–
N-SCORE$_{100}$	122.3	0.958	206.8	0.889
WSR	224.8	0.592	247.3	0.840

TABLE III: **Average time-to-decision (TTD) and empirical power** $(1-\beta)$ for each method on synthetic data. For Bernoulli data (left), TTD is averaged over all 35 alternative hypotheses, while power is averaged over the 9 hardest alternative hypotheses corresponding to a gap of 0.1 (see Figure A.3). For nonparametric data (right), TTD and power are averaged over 3000 independent redraws. Only N-SCORE$_{100}$ and WSR are valid in the nonparametric regime; the remaining methods cannot be adapted to this case. We observe that N-SCORE$_{100}$ outperforms WSR by approximately 15% on average in terms of time-to-decision, while improving the empirical power by approximately 5 percentage points.

improves sample efficiency on real-world data across a wide range of metrics and uncertainty distributions. Furthermore, implementing informative metrics, even those as simple as partial credit subtask completion, can have profound effects in accelerating evaluation. Using SOTA methods like STEP, θ-SAVI, and N-SCORE gives the potential for significant reductions (including more than 50% savings on the real-world data analyzed in Section VI-B2) in evaluation burden.

C. *Baseline Comparison: Optimizing Sample Complexity (RQ3) with Type-1 Error Control*

Having demonstrated the practical benefits of sequential evaluation and informative metrics, we investigate the **Baseline Comparison** problem: which sequential evaluation procedure should be preferred in a given comparison setting? Our central claim, illustrated in Table I, is that STEP and N-SCORE form a Pareto-optimal 'toolkit,' with N-SCORE combining the strongest respective features of the θ-SAVI approach (statistical power $(1-\beta)$, sample efficiency) and the WSR approach (nonparametric generality). We illustrate these properties on a broad range of progress metric distributions.

1) Synthetic Parametric Sequences: We begin by investigating Bernoulli sequences (corresponding in practice to binary success measures). As these correspond to a *parametric* setting, we seek to demonstrate that N-SCORE competes with θ-SAVI, while WSR lags in terms of sample efficiency.

The left-hand side of Table III shows the average time-to-decision for each evaluation method on artificially generated Bernoulli data comprising 35 different distributions with gaps in mean performance ranging from 10-50 percentage points. Each evaluation method is tested on each distribution 250 times (for a visualization of these hypotheses, see Figure A.3 in Appendix D-A). In each instance, the batch evaluation size was set to $N = 1,000$ samples per test to ensure a high probability of decision.

As shown in the table, STEP is near-optimal in this setting (as predicted in Table I), requiring 95 evaluation pairs on average to reach a decision. Notably, θ-SAVI and N-SCORE$_2$ are indistinguishable, requiring ≈ 118 evaluations apiece on average. We demonstrate that N-SCORE is robust to misspecification, in that applying N-SCORE$_{100}$ (a 50x overspecification in the number of potential different evaluation outcomes) to the data only adds 4 additional evaluations in expectation. By contrast, WSR requires nearly double the number of trials, suffering particularly on the hardest cases with performance gaps of 10 percentage points (see Figure A.3).

2) Synthetic Nonparametric Sequences: We proceed to the right-hand side of Table III, considering evaluation of nonparametric densities. Random polynomials of order 10 are generated, and then rectified into an appropriate density (i.e., are shifted and scaled to be everywhere nonnegative on $[0,1]$ and to integrate to 1). This process is analytically opaque, making it difficult to express the resulting family of distributions in any parametric form. Thus, we observe the reverse context to the Bernoulli case: the lack of generality of STEP and θ-SAVI means that only N-SCORE and WSR are applicable. We run $3,000$ comparison sequences of pairs of these distributions, limiting to cases where the gap in mean performance is at least 0.01. Again, $N = 1,000$ for each sequence. The right-hand side of Table III shows summary statistics for N-SCORE and WSR. There is a moderate gap in average time-to-decision, equal to savings of around 15% in evaluation burden, by using N-SCORE. These savings arise as a result of the ξ_n-optimization mechanism of Remark 1.

Because nonparametric tests are highly general-purpose, we can use N-SCORE to certify with high probability that N-SCORE's average time-to-decision is lower than that of WSR on the preceding datasets. Intuitively, each evaluation sequence is turned into an independent instance of the TTD metric for the evaluation procedure. Thus, resampling the synthetic data allows the construction of a dataset of times-to-decision for N-SCORE and WSR, which can itself be sequentially compared using the derived general progress metric $r_{i,n} = \frac{\text{ttd}_{i,n}}{N}$. For the synthetic nonparametric data, N-SCORE returns a decision at $\alpha = 0.05$ in approximately 525 evaluations, corresponding to the hypothesis: "the N-SCORE procedure has a lower time-to-decision than WSR on the class of nonparametric densities w.p. ≥ 0.95." This procedure can be repeated for the Bernoulli sequences of Section VI-C1, which also returns a significant result corresponding to N-SCORE having a significantly lower time-to-decision at $\alpha = 0.05$.

3) Real-valued Continuous Metrics: We conclude the **Baseline Comparison** setting by evaluating practical nonparametric data corresponding to policy rewards on the Mujoco [72] benchmark task suite. These continuous-valued episodic rewards have difficult-to-model underlying distributions, so only N-SCORE and WSR are applicable. Table IV shows the average time-to-decision to fully and partially rank policies synthesized using four popular RL algorithms (PPO, TD3, DDPG, and SAC) on each benchmark task. Average performance for each algorithm and task is included in Table A.II,

Mujoco Task	Full (Partial) Ranking (↓) N-SCORE$_{100}$	WSR	Significance ($\alpha = 0.01$) Full	Partial
Ant-v4	57 (39)	80 (55)	**Yes**	**Yes**
HalfCheetah-v4	66 (49)	91 (69)	**Yes**	**Yes**
Hopper-v4	46 (29)	61 (38)	**No**	**Yes**
Humanoid-v4	171 (58)	236 (75)	**Yes**	**Yes**
InvPendulum-v4	500 (500)	500 (500)	**No**	**No**
Pusher-v4	500 (257)	500 (347)	**No**	**Yes**
Walker2d-v4	50 (32)	65 (44)	**Yes**	**Yes**

TABLE IV: **Average TTD for full and partial RL policy ranking on Mujoco benchmarks.** (Left) Average ranking TTD for N-SCORE and WSR, for $N_{eval} = 500$ trials at $\alpha_{eval} = 0.05$. (Right) Significance of the **Baseline Comparison** of the average times-to-decision, with $N_{base} = 1,000$ and $\alpha_{base} = 0.01$. The presence of many instances of significant improvement ('Yes') suggests that N-SCORE significantly improves upon WSR in many practical nonparametric regimes.

averaged over 500k independent evaluation draws of each policy. These draws are partitioned into $1,000$ independent evaluation sequences of length $N = 500$ at $\alpha_{eval} = 0.05$. The left-hand side of Table IV shows the average time-to-decision for full (resp. partial) rankings of the policies for each task by N-SCORE and WSR. The sequence of times-to-decision for both partial and full rankings of the policies is then compared at $N_{base} = 1,000$ (corresponding to the $1,000$ independent partitioned sequences) at $\alpha_{base} = 0.01$. The right-hand side identifies all task / ranking pairs for which a significant difference between N-SCORE and WSR is found.

To contextualize the results, we note that in no instance does WSR empirically outperform N-SCORE, represented by the lower average times-to-decision across the entire left-hand side of Table IV. Conversely, on six of the seven benchmark tasks (corresponding to diverse evaluation metric distributions), N-SCORE significantly outperforms WSR in expected time-to-decision. We reiterate that the distribution over RL rewards is very complex (visualization of the reward distributions is given in Figure A.5 in Appendix D-B), and thus neither STEP nor θ-SAVI can be applied in this context.

These extensive experiments answer RQ3 by validating the desired **Baseline Comparison** properties of N-SCORE: we match the sample efficiency of θ-SAVI in parametric contexts (see Table II and Table III), while maintaining the generality of, and improving sample efficiency over, the WSR procedure in nonparametric contexts (e.g., Table IV and Figure 2).

We briefly summarize the central conclusions with respect to the research questions. As shown on real hardware and simulation data, as well as numerous synthetic experiments, sequential evaluation (RQ1) with informative progress metrics (RQ2) *accelerates* rigorous evaluation and saves the practitioner significant evaluation time, commonly on the order of nearly 30%, and in some cases exceeding 50%, of reasonable batch allotments of evaluation resources. To realize these benefits, (RQ3) shows that STEP and N-SCORE form a complementary, complete sample-efficient evaluation toolkit for binary success metrics and beyond.

VII. Limitations and Future Work

There are several current limitations of N-SCORE, which suggest the possibility for valuable future investigation. First, unlike STEP, any procedure using tools from safe, anytime-valid inference tends to achieve tighter Type-1 error control than specified, leaving some 'risk budget' unused. This both explains the gap to STEP in the regime of binary evaluation metrics and the capacity for robust generalization to complex and nonparametric measures. Developing a finite-N rectification to use the full available risk budget would be exceedingly valuable for a host of problems for which SAVI is currently applied. Similarly, the current method for optimizing ξ_n has connections to ideas in kernel density estimation, but at present the full insight of developments in the latter have not been applied to our approach. Using these tools and domain knowledge promises more efficiency and potential generalization to unbounded performance measures. Additionally, *verifying* the necessary assumption of i.i.d. data is important in robotic evaluation [42]. Some recent works in the statistical testing literature consider the related problem of testing exchangeability [26]; this could represent a valuable addition to a self-monitoring evaluation protocol.

We note that, while SAVI methods naturally hinder some avenues towards inadvertent data dredging, they rely crucially on i.i.d. evaluation data. Moreover, rigorous guarantees are only meaningful if the evaluation paradigm is similarly rigorous and reproducible. As such, these methods are inherently dependent on careful and measured evaluation procedures.

VIII. Conclusion

We have introduced and validated N-SCORE, a novel procedure for statistically rigorous sequential evaluation of robot policies that generalizes to metrics that go beyond binary success and failure rates. In so doing, we have highlighted the practical benefits of sequential methods and informative metrics to reduce practical evaluation burden, and situated our approach as a novel synthesis of two state-of-the-art sequential evaluation procedures. Each of these results is validated by substantial empirical evidence spanning synthetic and real-world evaluation data. The promise of such results is to both codify *and accelerate* progress within the field, by ensuring reliable performance improvements and minimizing the requisite evaluation burden necessary to confirm them.

Acknowledgments

D. Snyder acknowledges support from the Toyota Research Institute (TRI) and Penn AI Fellowship. A. Badithela is supported by the Presidential Postdoctoral Fellowship. The authors were also partially supported by the NSF Career Award #2044149, the NSF SLES Award #2331880, and the Sloan Fellowship. TRI provided funds to assist the authors with their research; this work solely reflects the opinions and conclusions of its authors, and not TRI nor any other Toyota entity.

References

[1] Joshua Achiam. Spinning Up in Deep Reinforcement Learning. 2018.

[2] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. volume 34, pages 29304–29320, 2021. doi: 10.48550/arXiv.2108.13264.

[3] 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. doi: 10.1146/annurev-control-031924-100130. URL https://www.annualreviews.org/content/journals/10.1146/annurev-control-031924-100130.

[4] Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, December 2022. URL http://arxiv.org/abs/2107.07511. arXiv:2107.07511 [cs].

[5] Abrar Anwar, Rohan Gupta, Zain Merchant, Sayan Ghosh, Willie Neiswanger, and Jesse Thomason. Efficient evaluation of multi-task robot policies with active experiment selection. *arXiv preprint arXiv:2502.09829*, 2025.

[6] 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.

[7] Apurva Badithela, David Snyder, Lihan Zha, Joseph Mikhail, Matthew O'Kelly, Anushri Dixit, and Anirudha Majumdar. Reliable and scalable robot policy evaluation with imperfect simulators. *arXiv preprint arXiv:2510.04354*, 2025.

[8] G. A. Barnard. Significance Tests for 2×2 Tables. *Biometrika*, 34(1-2):123–138, January 1947. ISSN 0006-3444. doi: 10.1093/biomet/34.1-2.123.

[9] 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.

[10] Stefan Bauer, Manuel Wüthrich, Felix Widmaier, Annika Buchholz, Sebastian Stark, Anirudh Goyal, Thomas Steinbrenner, Joel Akpo, Shruti Joshi, Vincent Berenz, et al. Real robot challenge: A robotics competition in the cloud. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 190–204. PMLR, 2022.

[11] Yoav Beck, Talia Herman, Marina Brozgol, Nir Giladi, Anat Mirelman, and Jeffrey M. Hausdorff. SPARC: a new approach to quantifying gait smoothness in patients with Parkinson's disease. *Journal of NeuroEngineering and Rehabilitation*, 15(1):49, June 2018. ISSN 1743-0003. doi: 10.1186/s12984-018-0398-3. URL https://doi.org/10.1186/s12984-018-0398-3.

[12] Peter J. Bickel and Kjell A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Volumes I-II Package*. Chapman and Hall/CRC, New York, December 2015. ISBN 978-1-315-36926-6. doi: 10.1201/9781315369266.

[13] R. D. Boschloo. Raised conditional level of significance for the 2 × 2-table when testing the equality of two probabilities. *Statistica Neerlandica*, 24(1):1–9, 1970. doi: 10.1111/j.1467-9574.1970.tb00104.x.

[14] Luis A. Caffarelli and S. Salsa. *A Geometric Approach to Free Boundary Problems*. American Mathematical Soc., 2005. ISBN 978-0-8218-3784-9. Google-Books-ID: YOzpBwAAQBAJ.

[15] Yen-Chi Chen. A Tutorial on Kernel Density Estimation and Recent Advances, September 2017. URL http://arxiv.org/abs/1704.03924. arXiv:1704.03924 [stat].

[16] Herman Chernoff. Sequential Tests for the Mean of a Normal Distribution. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4.1, pages 79–92. University of California Press, January 1961.

[17] Herman Chernoff. Sequential Test for the Mean of a Normal Distribution III (Small t). *The Annals of Mathematical Statistics*, 36(1):28–54, 1965. ISSN 0003-4851. Publisher: Institute of Mathematical Statistics.

[18] Herman Chernoff. Sequential Tests for the Mean of a Normal Distribution IV (Discrete Case). *The Annals of Mathematical Statistics*, 36(1):55–68, 1965. ISSN 0003-4851. Publisher: Institute of Mathematical Statistics.

[19] Herman Chernoff and A. John Petkau. Numerical Solutions for Bayes Sequential Decision Problems. *SIAM Journal on Scientific and Statistical Computing*, 7(1):46–59, 1986. doi: 10.1137/0907003.

[20] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, September 2025. ISSN 0278-3649. doi: 10.1177/02783649241273668. URL https://doi.org/10.1177/02783649241273668.

[21] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

[22] Brian Cho, Kyra Gan, and Nathan Kallus. Peeking with PEAK: Sequential, Nonparametric Composite Hypothesis Tests for Means of Multiple Data Streams, June 2024. URL http://arxiv.org/abs/2402.06122. arXiv:2402.06122 [stat].

[23] Jack Collins, Mark Robson, Jun Yamada, Mohan Sridharan, Karol Janik, and Ingmar Posner. Ramp: A benchmark for evaluating robotic assembly manipulation and planning. *IEEE Robotics and Automation Letters*, 9(1): 9–16, 2023.

[24] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking Deep Reinforcement Learning for Continuous Control, May 2016. URL http://arxiv.org/abs/1604.06778. arXiv:1604.06778 [cs].

[25] Michael Fauss, Abdelhak M. Zoubir, and H. Vincent Poor. Minimax Optimal Sequential Hypothesis Tests for Markov Processes. *The Annals of Statistics*, 48(5):2599–2621, 2020. ISSN 0090-5364. Publisher: Institute of Mathematical Statistics.

[26] Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk. Plug-in martingales for testing exchangeability on-line. 2012.

[27] R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922. ISSN 09528385. URL http://www.jstor.org/stable/2340521.

[28] Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, November 2024. ISSN 1369-7412. doi: 10.1093/jrsssb/qkae011. URL https://doi.org/10.1093/jrsssb/qkae011.

[29] 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.

[30] Minho Heo, Youngwoon Lee, Doohyun Lee, and Joseph J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. In *Robotics: Science and Systems*, 2023.

[31] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[32] Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, Jeff Braga, Dipam Chakraborty, Kinal Mehta, and João G.M. Araújo. Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *Journal of Machine Learning Research*, 23(274):1–18, 2022. URL http://jmlr.org/papers/v23/21-1342.html.

[33] Siyuan Huang, Liliang Chen, Pengfei Zhou, Shengcong Chen, Zhengkai Jiang, Yue Hu, Yue Liao, Peng Gao, Hongsheng Li, Maoqing Yao, et al. Enerverse: Envisioning embodied future space for robotics manipulation. *arXiv preprint arXiv:2501.01895*, 2025.

[34] Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, et al. $\pi_{\{0.6\}}$: A vla that learns from experience. *arXiv preprint arXiv:2511.14759*, 2025.

[35] Kyoungseok Jang, Kwang-Sung Jun, Ilja Kuzborskij, and Francesco Orabona. Tighter PAC-Bayes Bounds Through Coin-Betting. In *Proceedings of Thirty Sixth Conference on Learning Theory*, pages 2240–2264. PMLR, July 2023. URL https://proceedings.mlr.press/v195/jang23a.html.

[36] Heinrich Jiang. Uniform Convergence Rates for Kernel Density Estimation. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1694–1703. PMLR, July 2017. URL https://proceedings.mlr.press/v70/jiang17b.html.

[37] Leslie K John, George Loewenstein, and Drazen Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5):524–532, 2012. doi: 10.1177/0956797611430953.

[38] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X.

[39] Ninad Khargonkar, Sai Haneesh Allu, Yangxiao Lu, Balakrishnan Prabhakaran, Yu Xiang, et al. Scenereplica: Benchmarking real-world robot manipulation by creating replicable scenes. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8258–8264. IEEE, 2024.

[40] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[41] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246*, 2024. doi: 10.48550/arXiv.2406.09246.

[42] 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.

[43] Kristofer D. Kusano, John M. Scanlon, Yin-Hsiu Chen, Timothy L. McMurry, Tilia Gode, and Trent Victor. Comparison of Waymo Rider-Only crash rates by crash type to human benchmarks at 56.7 million miles. *Traffic Injury Prevention*, 26(sup1):S8–S20, October 2025. ISSN 1538-9588. doi: 10.1080/15389588.2025.2499887. URL https://www.tandfonline.com/doi/full/10.1080/15389588.2025.2499887.

[44] Tze Leung Lai. Optimal Stopping and Sequential Tests which Minimize the Maximum Expected Sample Size. *The Annals of Statistics*, 1(4):659–673, 1973. ISSN 0090-5364. URL https://www.jstor.org/stable/2958310. Publisher: Institute of Mathematical Statistics.

[45] Tze Leung Lai. Boundary Crossing Probabilities for Sample Sums and Confidence Sequences. *The Annals of*

Probability, 4(2):299–312, 1976. ISSN 0091-1798. URL https://www.jstor.org/stable/2959164. Publisher: Institute of Mathematical Statistics.

[46] Tze Leung Lai. First Exit Times from Moving Boundaries for Sums of Independent Random Variables. *The Annals of Probability*, 5(2):210–221, 1977. ISSN 0091-1798. URL https://www.jstor.org/stable/2242894. Publisher: Institute of Mathematical Statistics.

[47] Tze Leung Lai. Boundary Crossing Problems for Sample Means. *The Annals of Probability*, 16(1):375–396, 1988. ISSN 0091-1798. Publisher: Institute of Mathematical Statistics.

[48] Tze Leung Lai. Nearly Optimal Sequential Tests of Composite Hypotheses. *The Annals of Statistics*, 16(2): 856–886, 1988. ISSN 0090-5364. Publisher: Institute of Mathematical Statistics.

[49] Tze Leung Lai and Li Min Zhang. Nearly Optimal Generalized Sequential Likelihood Ratio Tests in Multivariate Exponential Families. *Lecture Notes-Monograph Series*, 24:331–346, 1994. ISSN 0749-2170. Publisher: Institute of Mathematical Statistics.

[50] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024.

[51] Yaxuan Li, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. Worldeval: World model as real-world robot policies evaluator. *arXiv preprint arXiv:2505.19017*, 2025.

[52] Michael Lindon and Nathan Kallus. Anytime-Valid A/B Testing of Counting Processes. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, pages 3529–3537. PMLR, April 2025. URL https://proceedings.mlr.press/v258/lindon25a.html.

[53] 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.

[54] Ziyuan Liu, Wei Liu, Yuzhe Qin, Fanbo Xiang, Minghao Gou, Songyan Xin, Maximo A Roa, Berk Calli, Hao Su, Yu Sun, et al. Ocrtoc: A cloud-based competition and benchmark for robotic grasping and manipulation. *IEEE Robotics and Automation Letters*, 7(1):486–493, 2021.

[55] Jianlan Luo, Charles Xu, Fangchen Liu, Liam Tan, Zipeng Lin, Jeffrey Wu, Pieter Abbeel, and Sergey Levine. Fmb: a functional manipulation benchmark for generalizable robotic learning. *The International Journal of Robotics Research*, 44(4):592–606, 2025.

[56] Rachel Luo, Heng Yang, Michael Watson, Apoorva Sharma, Sushant Veer, Edward Schmerling, and Marco Pavone. Leveraging correlation across test platforms for variance-reduced metric estimation. *arXiv preprint arXiv:2506.20553*, 2025.

[57] 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.

[58] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

[59] Jerzy Neyman, Egon Sharpe Pearson, and Karl Pearson. IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706): 289–337, January 1997. doi: 10.1098/rsta.1933.0009. Publisher: Royal Society.

[60] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An Open-Source Generalist Robot Policy. 2024. doi: 10.48550/arXiv.2405.12213.

[61] Francesco Orabona and Kwang-Sung Jun. Tight Concentrations and Confidence Sequences From the Regret of Universal Portfolio. *IEEE Transactions on Information Theory*, 70(1):436–455, January 2024. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2023.3330187. URL https://ieeexplore.ieee.org/document/10315047/.

[62] Emanuel Parzen. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962. ISSN 0003-4851. URL https://www.jstor.org/stable/2237880.

[63] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient Action Tokenization for Vision-Language-Action Models, January 2025. URL http://arxiv.org/abs/2501.09747. arXiv:2501.09747 [cs].

[64] 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.

[65] Julian Quevedo, Percy Liang, and Sherry Yang. Evaluating robot policies in a world model. *arXiv preprint arXiv:2506.00613*, 2025.

[66] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

[67] Glenn Shafer, Vladimir Vovk, and Cs Rhul Ac Uk. A Tutorial on Conformal Prediction.

[68] David Siegmund. *Sequential analysis: tests and confidence intervals*. Springer-Verlag, 1985. doi: 10.1007/978-1-4757-1862-1.

[69] David Snyder, Asher James Hancock, Apurva Badithela, Emma Dixon, Patrick Miller, Rares Andrei Ambrus, Anirudha Majumdar, Masha Itkina, and Haruki

Nishimura. Is your imitation learning policy better than mine? policy comparison with near-optimal stopping. *arXiv preprint arXiv:2503.10966*, 2025.

[70] 1X World Model Team. 1x world model: Evaluating bits, not atoms. Technical report, 1X, 2025.

[71] 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.

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

[73] Wei-Cheng Tseng, Jinwei Gu, Qinsheng Zhang, Hanzi Mao, Ming-Yu Liu, Florian Shkurti, and Lin Yen-Chen. Scalable policy evaluation with video world models. *arXiv preprint arXiv:2511.11520*, 2025.

[74] Rosanne J. Turner and Peter D. Grünwald. Exact anytime-valid confidence intervals for contingency tables and beyond. *Statistics & Probability Letters*, 198:109835, July 2023. ISSN 0167-7152. doi: 10.1016/j.spl.2023.109835. URL https://www.sciencedirect.com/science/article/pii/S0167715223000597.

[75] Pierre Van Moerbeke. Optimal Stopping and Free Boundary Problems. *The Rocky Mountain Journal of Mathematics*, 4(3):539–578, 1974. ISSN 0035-7596. Publisher: Rocky Mountain Mathematics Consortium.

[76] Jean Ville. *Etude Critique de la Notion de Collectif*. PhD thesis, Universite de Paris, 1939. Publisher:Gauthier-Villars, Paris.

[77] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer International Publishing, Cham, 2022. ISBN 978-3-031-06648-1 978-3-031-06649-8. doi: 10.1007/978-3-031-06649-8. URL https://link.springer.com/10.1007/978-3-031-06649-8.

[78] A. Wald. Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945. ISSN 0003-4851. Publisher: Institute of Mathematical Statistics.

[79] A. Wald and J. Wolfowitz. Optimum Character of the Sequential Probability Ratio Test. *The Annals of Mathematical Statistics*, 19(3):326–339, 1948. ISSN 0003-4851. Publisher: Institute of Mathematical Statistics.

[80] Larry Wasserman. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer, New York, NY, 2006. ISBN 978-0-387-25145-5. doi: 10.1007/0-387-30623-4. URL http://link.springer.com/10.1007/0-387-30623-4.

[81] 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.

[82] B. L. WELCH. The Generalization of 'Student's' Problem When Several Different Population Variances are Involved. *Biometrika*, 34(1-2):28–35, January 1947. ISSN 0006-3444. doi: 10.1093/biomet/34.1-2.28. URL https://doi.org/10.1093/biomet/34.1-2.28.

[83] Robert M West. Best practice in statistics: Use the Welch t-test when testing the difference between two groups. *Annals of Clinical Biochemistry*, 58(4):267–269, July 2021. ISSN 0004-5632. doi: 10.1177/0004563221992088. URL https://doi.org/10.1177/0004563221992088.

[84] Brian Yang, Dinesh Jayaraman, Jesse Zhang, and Sergey Levine. Replab: A reproducible low-cost arm benchmark for robotic learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8691–8697. IEEE, 2019.

[85] Lihan Zha, Apurva Badithela, Michael Zhang, Justin Lidard, Jeremy Bao, Emily Zhou, David Snyder, Allen Z. Ren, Dhruv Shah, and Anirudha Majumdar. Guiding Data Collection via Factored Scaling Curves, May 2025. URL http://arxiv.org/abs/2505.07728. arXiv:2505.07728 [cs].

[86] Gaoyue Zhou, Victoria Dean, Mohan Kumar Srirama, Aravind Rajeswaran, Jyothish Pari, Kyle Hatch, Aryan Jain, Tianhe Yu, Pieter Abbeel, Lerrel Pinto, et al. Train offline, test online: A real robot learning benchmark. *arXiv preprint arXiv:2306.00942*, 2023.

[87] 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.

[88] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: Learning interactive real-robot action simulators. *arXiv preprint arXiv:2406.14540*, 2024.