

Is Your Imitation Learning Policy Better than Mine? Policy Comparison with Near-Optimal Stopping

David Snyder^{1,2}, Asher James Hancock², Apurva Badithela², Emma Dixon¹, Patrick Miller¹,
Rares Andrei Ambrus¹, Anirudha Majumdar², Masha Itkina¹, and Haruki Nishimura¹

¹Toyota Research Institute (TRI), ²Princeton University
dasnyder@princeton.edu

Abstract—Imitation learning has enabled robots to perform complex, long-horizon tasks in challenging dexterous manipulation settings. As new methods are developed, they must be rigorously evaluated and compared against corresponding baselines through repeated evaluation trials. However, policy comparison is fundamentally constrained by a small feasible sample size (e.g., 10 or 50) due to significant human effort and limited inference throughput of policies. This paper proposes a novel statistical framework for rigorously comparing two policies in the small sample size regime. Prior work in statistical policy comparison relies on batch testing, which requires a fixed, pre-determined number of trials and lacks flexibility in adapting the sample size to the observed evaluation data. Furthermore, extending the test with additional trials risks inducing inadvertent p-hacking, undermining statistical assurances. In contrast, our proposed statistical test is *sequential*, allowing researchers to decide whether or not to run more trials based on intermediate results. This adaptively tailors the number of trials to the difficulty of the underlying comparison, saving significant time and effort without sacrificing probabilistic correctness. Extensive numerical simulation and real-world robot manipulation experiments show that our test achieves near-optimal stopping, letting researchers stop evaluation and make a decision in a near-minimal number of trials. Specifically, it reduces the number of evaluation trials by up to 32% as compared to state-of-the-art baselines, while preserving the probabilistic correctness and statistical power of the comparison. Moreover, our method is strongest in the most challenging comparison instances (requiring the most evaluation trials); in a multi-task comparison scenario, we save the evaluator more than 160 simulation rollouts.

I. INTRODUCTION

The importance of trustworthy and efficient robot policy evaluation protocols has become paramount in imitation learning as the scale of underlying deep learning models and the complexity of tasks continue to increase. This need is especially pronounced in dexterous manipulation where stochastic, contact-rich interactions between the robot and the environment introduce inherent randomness in outcomes.

A particularly important aspect of policy evaluation is *policy comparison*, where two policies are repeatedly deployed in an environment to assess their relative performance. Policy comparison forms the foundation of robot learning as “an empirical science” [28], enabling researchers to objectively measure the scientific progress of the field. Nevertheless, this setting introduces an additional source of stochasticity due to random outcomes of both the first and the second policies, making the reliability of the comparison more challenging to ensure than evaluation of single policies.

To motivate concretely, consider an example policy comparison scenario presented in Fig. 1 where the performance is quantified based on binary success/failure metrics, a common choice [7, 59, 41] as continuous-valued rewards are often difficult to define. This scenario naturally arises when researchers want to demonstrate the effectiveness of a particular intervention (e.g., a new policy architecture) by comparing the new policy π_1 against a baseline π_0 . Alternatively, π_1 and π_0 could represent the same policy evaluated under different environment distributions, providing insights on generalization. In either case, the evaluator faces two practical challenges. First, only a small number of trials (e.g., 10–60) [35, 18, 14, 40, 27, 5] can be performed per evaluation setting due to the large human effort needed to reset the environment between trials and the substantial wall-clock time imposed by limited inference throughput of large policies. While high-fidelity simulators can alleviate the human effort and still provide valuable insights into policy performance [32, 42], real-world evaluation remains indispensable for ensuring reliable deployment in downstream applications. Second, the evaluation results are revealed sequentially, possibly leading to fluctuating observations depending on when the evaluation is stopped. In the Fig. 1 example, the evaluator could observe more successes for π_0 after conducting additional trials, even though π_1 initially appeared superior after the first five.

Although recent work [53, 28] proposes statistical policy comparison approaches, it follows the conventional batch testing scheme, requiring complete results from a pre-determined number of trials before the statistical test can be performed. Furthermore, the test can be conducted only once for the corresponding results; even if the test fails to determine the relative performance due to closely matched results, the evaluator cannot append more evaluation trials to the existing results to run the test again, as doing so would constitute p-hacking that invalidates statistical assurances [43]. Unfortunately, this is a common but harmful research practice outside of robotics [23], which needs to be averted to ensure reproducible science.

To address these challenges, we propose a novel sequential testing framework named STEP (Sequential Testing for Efficient Policy Comparison) for rigorously comparing performance of imitation learning policies¹. Unlike batch testing,

¹Although this paper focuses on imitation learning, STEP is naturally applicable to evaluating any types of policies based on binary metrics, including reinforcement learning (RL) policies with sparse 0/1 reward.

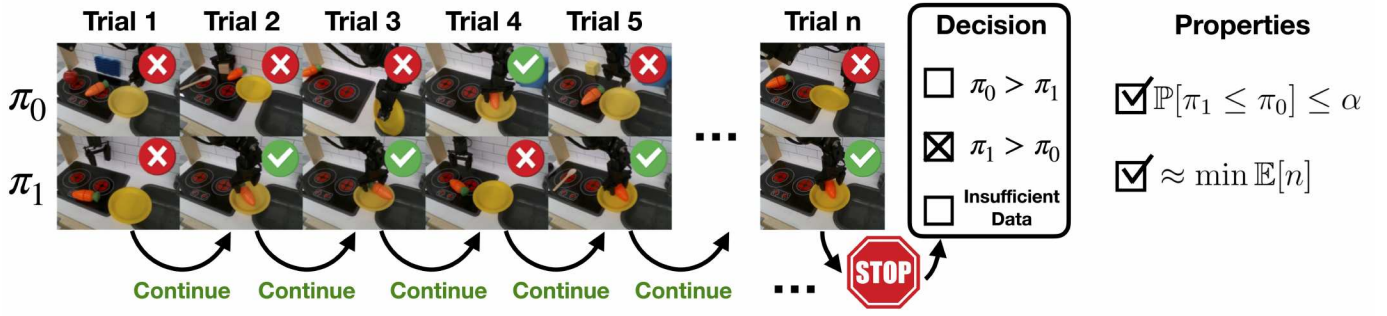


Fig. 1: Robot policy comparison problem under binary success/failure metrics. Novel policy π_1 is compared against baseline π_0 in a sequence of trials. Within a given evaluation budget, the evaluator seeks a statistically significant comparison in as few trials as possible. Due to the cost of hardware setup and calibration, as well as the limited inference speed of complex policies, these results generally arrive in sequence from a single (or few) hardware setups. Allowing the evaluator to adaptively and near-optimally tailor the number of trials based on the data observed so far — *without compromising statistical assurances of the comparison* — is a central contribution of this work.

our approach allows the number of trials to be varied within a given experimental budget. This critical feature offers two practical advantages. First, the evaluator can stop conducting trials early without sacrificing the probabilistic correctness of the comparison if enough statistical evidence is accumulated quickly in favor of π_1 (or π_0). Second, it reduces the epistemic risk of overconfident (and potentially incorrect) conclusions when π_0 and π_1 are closely matched, since the test will abstain from making a decision if statistical evidence remains low after all the planned trials have been executed. Alternatively, the evaluator may safely append additional trials to the original samples and re-conduct statistical analysis on all the results obtained thus far without inadvertent p-hacking. We demonstrate these advantages through simulation and real-world robot manipulation experiments. Furthermore, extensive numerical experiments show that STEP significantly outperforms state-of-the-art (SOTA) sequential methods, reducing the required number of trials by up to 32% without sacrificing probabilistic correctness. The specific contributions of this paper are as follows:

- We propose STEP, a novel sequential statistical framework for evaluating relative performance of two policies with tunable probabilistic correctness².
- Our sequential testing approach admits an adaptive number of evaluation trials tailored to the difficulty of the underlying comparison while achieving near-minimal sample complexity.
- We additionally present a straightforward extension of our framework to (1) multi-task and (2) multi-policy comparison settings via a reduction to a set of pair-wise statistical comparisons.

II. PRELIMINARIES

Consider a physical robot trained to complete a task in a variety of environment realizations. This setting is naturally

modeled as a partially observable Markov decision process (POMDP) [24], where the underlying state s represents the environment and robot states. The observation o is determined by the robot’s embodiment and sensing apparatus. The training pipeline is designed to synthesize a policy taking actions a that achieve a high reward $r(s, a) = \mathbb{1}[s \in \mathcal{S}_{\text{success}}]$ on a particular task, where $\mathcal{S}_{\text{success}}$ is a set of successful terminal states in which the episode will terminate. This reward encodes a binary success/failure criterion of the task. In imitation learning, a surrogate loss function is used to train a policy that matches the behavior of expert demonstrations [46] instead of directly solving the POMDP. At the end of a training process, a policy π_1 will be obtained. This policy has a true success rate (i.e., expected total episode reward) under a distribution $\mathcal{D}_{s_0, o_0}$ over the initial state and observation:

$$p_1 = \mathbb{E}_{\mathcal{D}_{s_0, o_0}, \pi_1} \left[\sum_{t=0}^T r(s_t, a_t) \right], \quad (1)$$

where dependence on the state transition and observation models are omitted for brevity. The true success rate is *unknown* and must be estimated via multiple evaluation trials.

Assumption 1 (Regularity): In each evaluation trial, the initial state s_0 and the observation o_0 are drawn in an independent and identically distributed (i.i.d.) fashion from the underlying distribution $\mathcal{D}_{s_0, o_0}$ ³. We assume access to samples from $\mathcal{D}_{s_0, o_0}$, but do not assume $\mathcal{D}_{s_0, o_0}$ itself to be known.

Under Assumption 1, the n^{th} evaluation trial involves making an i.i.d. draw of an environment from $\mathcal{D}_{s_0, o_0}$ and running the policy π_1 in this environment. This yields a binary success/failure outcome $z_{1,n}$ corresponding to an i.i.d. draw from a Bernoulli random variable with mean p_1 , which is the true performance (success rate) of the policy π_1 on the task:

$$z_{1,n} \sim \text{Ber}(p_1). \quad (2)$$

³Note that this is a standard assumption in statistical testing. A discussion of practical methods by which to approximately satisfy this condition in robotic evaluation is included in [28].

²Associated website and code can be found at: <https://tri-ml.github.io/step/>

Here, $z_{1,n} = 1$ indicates success and $z_{1,n} = 0$ failure. For any baseline policy π_0 , we similarly denote the outcome of its n^{th} trial as $z_{0,n} \sim \text{Ber}(p_0)$. For the sake of statistical analysis, we pair the outcomes of two policies by their indices in a vector $Z_n = (z_{0,n}, z_{1,n})$. The policy comparison problem can be formalized in the sense of Neyman-Pearson (N-P) statistical testing [37], where the null (skeptical) hypothesis is that the novel policy π_1 is no better than the baseline π_0 and the alternative is that the novel policy is indeed better:

$$\begin{aligned} \text{Null Hypothesis } \mathbb{H}_0 : p_1 &\leq p_0 \equiv (p_0, p_1) \in \mathcal{H}_0 \\ \text{Alt. Hypothesis } \mathbb{H}_1 : p_1 &> p_0 \equiv (p_0, p_1) \in \mathcal{H}_1. \end{aligned} \quad (3)$$

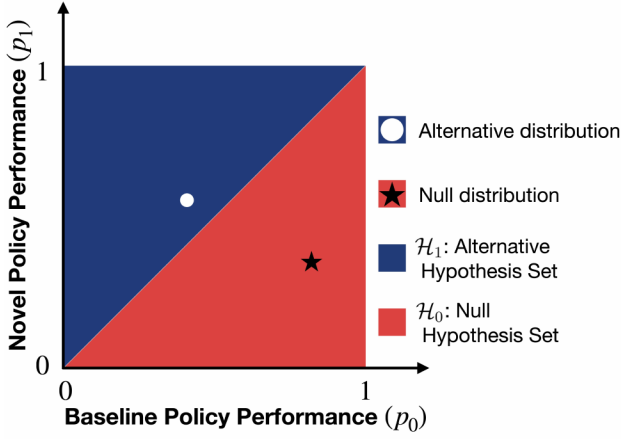


Fig. 2: The policy comparison problem as a composite-vs-composite statistical test. The null hypothesis set (red) corresponds to the novel policy being worse ($p_1 \leq p_0$), while the alternative hypothesis set (blue) corresponds to the novel policy being better ($p_1 > p_0$). The sets are each termed “composite” because they contain many elements. For any pair of policies, the truth corresponds to a single point; as examples, the white circle is a case where the alternative is true (baseline success 45%, novel policy success 55%), while the black star corresponds to the null being true (baseline success 82%, novel policy success 35%).

As illustrated in Fig. 2, this amounts to a composite vs-composite statistical test [4]. A hypothesis is termed “simple” if it is singleton, i.e., corresponds to a single data-generating distribution. For example, $(p_0, p_1) = (0.3, 0.7)$ would be a simple hypothesis. If the hypothesis corresponds to multiple data-generating distributions, it is termed “composite”.

The Type-I error rate (denoted α) of a statistical test corresponds to the probability of falsely rejecting the null *under the worst-case singleton element h_0 of the null hypothesis set $\mathcal{H}_0$* – that is, under the hardest-to-distinguish $(p_0, p_1) \in \mathcal{H}_0$. This type of error corresponds to falsely concluding that π_1 is better than π_0 when it is not. The *power* of a test (denoted $1 - \beta$) is the mixture probability of correctly rejecting the null when the alternative is true, under some measure on the set of

alternatives⁴. The Type-II error β is the associated (mixture) probability of failing to reject the null when the alternative is true. This represents failing to conclude that π_1 is better, when it is in fact better than π_0 .

Finally, while the preceding development has focused on one-sided testing, the comparison problem naturally admits a bidirectional version allowing decisions for membership in either the alternative set (**RejectNull**) or the null set (**AcceptNull**). Note that the decision **AcceptNull** formally amounts to rejecting the null of a “flipped” version of Eq. (3). In all subsequent discussion we will implicitly utilize the bidirectional test, which will allow for the decision **AcceptNull**.

III. PROBLEM FORMULATION

We assume that a robot evaluator is tasked with distinguishing two policies via successive evaluations, resulting in the testing paradigm described in Section II. We also assume that the evaluator has *pre-selected* the desired significance level α^* of the comparison and a maximum number of trials (for each policy) that they are willing or able to run: $N_{\max}$.

First, we note that the Type-I error must always be controlled in statistical testing, i.e., upper-bounded at the evaluator-specified rate $\alpha^* \in (0, 1)$. This represents a hard constraint (“validity”) that *will be enforced in all subsequent testing procedures*; a test is not feasible if it is not Type-I error controlling. The evaluator’s goal is to synthesize a decision rule that limits the Type-I Error to α^* while maximizing power and minimizing the expected number of evaluation trials.

We assume an underlying representation of the evaluation information collected *thus far* is available to the evaluator in a state $x_n = F(Z_1, Z_2, \dots, Z_n)$. Then, the evaluator’s procedure consists of deciding whether to **Continue** (gather another trial for each policy) or stop (and either **AcceptNull** or **RejectNull**). Given a decision set $\mathcal{U} = \{\text{AcceptNull}, \text{Continue}, \text{RejectNull}\}$, the problem is to find a state partition $\zeta = u(x)$ to optimally balance minimizing the expected sample size and maintaining high power, conditioned on the Type-I Error rate constraint, as shown in Eq. (4):

$$\begin{aligned} \min_{\zeta: \mathcal{X} \rightarrow \mathcal{U}} \quad & \mathbb{E}_{\mu(\mathcal{H}_1)}[n_{\text{stop}} + c\beta_{N_{\max}}] \\ \text{s.t.} \quad & \max_{h_0=(p_0, p_1) \in \mathcal{H}_0} \alpha(\zeta, h_0) \leq \alpha^* \\ & 0 \leq n_{\text{stop}} \leq N_{\max} \text{ w.p. } 1. \end{aligned} \quad (4)$$

This function is a multi-objective optimization which seeks to simultaneously minimize expected sample size and maximize power subject to the validity constraint. Informally, for any feasible terminal power $1 - \beta_{N_{\max}}^{\text{feasible}}$ (conditioned on $N_{\max}, \alpha^*$, and the true underlying distribution, which we emphasize is *unknown a priori*), there is a value of $c > 0$ that effects a decision rule in Eq. (4) approximating a test controlling

⁴For example, a minimax measure mimics the Type-I setting (e.g., the worst-case singleton $h_1 \in \mathcal{H}_1$). However, due to the finite termination condition, such a measure is vacuous because there will always exist a $(p_0, p_0 + \epsilon)$ for $\epsilon > 0$ sufficiently small such that the attainable power will only negligibly exceed random guessing. Therefore, in practice other measures must be used.

$\beta = \beta_{\text{feasible}}$. For example, this framework recovers the batch problem as $c \rightarrow \infty$ (demanding maximal power), and immediate termination as $c \rightarrow 0^{+5}$ (demanding minimal sample complexity). This objective will govern the methodology and analysis presented in the rest of the paper.

IV. METHODOLOGY

There are several important practical considerations to constructing a near-optimal solution to Eq. (4). We present the concrete challenges first, and then discuss the technical innovations that account for them. Throughout, we will leverage significant mathematical structure in the testing problem. Where insightful or intuitive, this will be explained *in situ*. See Section X-F in the Supplement for additional details.

A. State Representation

First, we express the information available to the evaluator at time n (i.e., after n evaluation trials have been performed for each policy) in a control-theoretic state representation; the evaluator's decision can then be understood as a state-feedback decision rule. In selecting this representation, the first instance of mathematical structure is the membership of Bernoulli distributions in the univariate exponential family; such distributions have known sufficient statistics which represent (informally) an optimal compression of the data for the purposes of testing and estimation [4]. Thus, a natural ("near-minimal") state representation is precisely the element-wise sufficient statistic for p_0 and p_1 respectively, augmented with a time state. For univariate exponential family distributions, the sufficient statistic is the sum of the observed data (i.e., respective number of successful trials under π_0 and π_1); this makes the state representation a first-order discrete integrator, as shown in Eq. (5):

$$\begin{aligned} \mathbf{x}_{n+1} &= \mathbf{x}_n + \mathbf{d}_n \\ \mathbf{d}_n &= (z_{0,n}, z_{1,n}, 1) \sim (\text{Ber}(p_0), \text{Ber}(p_1), 1) \\ \mathbf{x}_0 &= \mathbf{0}_3. \end{aligned} \quad (5)$$

Unlike in a typical control problem, we cannot actively guide the state trajectory; we are instead deciding when to stop based on the state trajectory. Concretely, the "control" involves partitioning the state space into stopping and continuation regions. In robotics, a similar set-theoretic notion arises in robust control and safe navigation, through the generation of invariant sets for dynamic systems (i.e., in reachability and barrier function-type methods [48, 2]), though those methods are primarily interested in nonstochastic uncertainty. Separately, many applications of asset pricing in mathematical finance utilize optimal stopping to set options prices under stochastic uncertainty [36].

⁵For the latter case: without looking at any data, draw a random number uniformly on $[0, 1]$. If it is less than α^* reject the null, otherwise fail to reject and terminate. This terminates at step 0 with probability 1, has power α^* , and is valid.

B. Decision Regions

In the sequential problem the state space partitions occur in sequence, yielding an effective representation of the decision rule $\zeta : \mathcal{X} \mapsto \mathcal{U}$ in the form of Eq. (6):

$$\zeta \equiv \left\{ \mathcal{X}_n^{\text{Reject Null}}, \mathcal{X}_n^{\text{Accept Null}}, \mathcal{X}_n^{\text{Continue}} \right\}_{n=1}^{N_{\max}}. \quad (6)$$

At each time step, the sets jointly encode control decisions for every state. Intuitively, the larger the size of the rejection region $\mathcal{X}_n^{\text{Reject Null}}$ (resp. $\mathcal{X}_n^{\text{Accept Null}}$), the smaller the number of expected trials needed to reject the null (resp. alternative), achieving a lower value of the stopping time component of Eq. (4). Focusing on $\mathcal{X}_n^{\text{Reject Null}}$ in the following development⁶, our auxiliary objective is then to maximize the size of the set $\mathcal{X}_n^{\text{Reject Null}}$ globally across all $n \in \{1, \dots, N_{\max}\}$. We take a probabilistic approach, allowing the states to reject with a probability less than 1. As we will see in Section IV-E, this yields an efficient optimization problem.

In addition to the maximization, there are two core challenges to synthesizing such a partition. First, the Type-I Error rate must be controlled. The decision to stop and reject the null hypothesis in a state $\mathbf{x}_n$ incurs risk due to the probability that $\mathbf{x}_n$ was reached under data generated from some $(p_0, p_1) \in \mathcal{H}_0$ – i.e., under the null hypothesis in Eq. (3). Thus, having a large $\mathcal{X}_n^{\text{Reject Null}}$ risks violating Type-I Error rate control. Second, the temporal rate at which the risk is accrued must be set appropriately to encourage early stopping on easy instances (i.e., π_1 significantly outperforms π_0) without harming performance too much on harder instances (i.e., π_1 and π_0 are closely competing, which requires many trials to distinguish). We will first address Type-I Error control in Section IV-C, and the temporal risk accumulation in Section IV-D. The resulting tractable optimization problem is presented in Section IV-E.

C. Type-I Error Control

The composite nature of the null hypothesis (which contains all pairs $\mathcal{H}_0 = \{(p_0, p_1) \in [0, 1]^2 \mid p_0 \geq p_1\}$) means that the decision-making problem can be thought of (informally) as analogous to distributionally robust control, where the uncertainty is over the particular worst-case $(p_0, p_1) \in \mathcal{H}_0$. Controlling the Type-I Error in policy comparison then amounts to controlling the Type-I Error uniformly (robustly) for all $h \in \mathcal{H}_0$. Suppose that some rejection region for the first $n-1$ steps has been obtained with accumulated risk α_{n-1} , and we are interested in bounding the Type-I Error for n by some $\alpha_n > \alpha_{n-1}$. Given the notion of stopping and continuation regions introduced in the previous section, this is expressed mathematically as:

$$\max_{h \in \mathcal{H}_0} \mathbb{P}_h \left(\mathbf{x}_n \in \mathcal{X}_n^{\text{Reject Null}} \mid \mathcal{X}_{n-1}^{\text{Reject Null}} \right) \leq \alpha_n. \quad (7)$$

The dependence on the (probabilistic) rejection region for $n-1$ is made explicit in Eq. (7), reflecting the internal dynamic structure. For example, if some state $\mathbf{x}_{n-1} = (a, b, n-1)$

⁶As discussed in Section II, the case of rejecting the alternative and accepting the null can be considered by flipping p_0 and p_1 .

is rejected at $n - 1$ with a non-zero probability, then the 1-step reachable states (e.g., $(a + 1, b, n)$) under the dynamics Eq. (5) are less likely to be feasible at n . The presence of many (infinite) elements h in the set of null hypotheses $\mathcal{H}_0$ makes verification of Type-I Error control challenging *a priori*, even for a single particular n . However: monotonicity, symmetry, and smoothness properties of this family of testing problems (noted in, for example, [13]) allow for efficient discretization procedures that preserve safety. Specifically, it suffices to consider a discrete set of the “worst-case” nulls $\hat{\mathcal{H}}_0 = \{(p^{(1)}, p^{(1)}), (p^{(2)}, p^{(2)}), \dots, (p^{(M)}, p^{(M)})\}$, where $0 \leq p^{(1)} \leq \dots \leq p^{(M)} \leq 1$. Further details of this discretization process is given in Section X-F2 in the Supplement. Thus, Eq. (7) reduces to a finite set of individual inequality constraints. We can equivalently represent this set of constraints as a linear inequality $P_n \mathbf{w}_n \leq \alpha_n \mathbb{1}$, where $\mathbf{w}_n$ is a vector representing the probability of rejecting the null in each state $\mathbf{x}_n \in \mathcal{X}_n = \{(0, 0, n), (0, 1, n), \dots, (N_{\max}, N_{\max}, n)\}$. P_n is a non-negative matrix of size $(M, |\mathcal{X}_n|)$ where each row represents the probability of reaching particular states under $h^{(i)} = (p^{(i)}, p^{(i)})$:

$$(P_n)_{ij} = \mathbb{P}_{h^{(i)}}(\mathbf{x}_n = \mathbf{x}_n^j \mid \mathcal{X}_{n-1}^{\text{Reject Null}}), \quad (8)$$

where $\mathbf{x}_n^j$ is the j^{th} (discrete) state in $\mathcal{X}_n$. Given the rejection region from the previous time step $n - 1$, we can accurately compute this probability Eq. (8) by forward-propagating the previous state occupancy distribution $(P_{n-1})_i$ according to the stochastic dynamics model Eq. (5) under $h^{(i)}$.

D. Power Adaptivity to Varying Difficulty: Risk Budgets

As discussed in Section IV-B, we must appropriately adjust the temporal rate of risk accumulation. To formally define this notion of risk, we introduce a non-negative scalar function $f(n)$ for $n \in \{1, \dots, N_{\max}\}$, which determines the maximum allowable Type-I Error under any null hypothesis at each step n . We impose a constraint $\sum_{n=0}^{N_{\max}} f(n) = \alpha^*$ to globally bound the Type-I error by α^* . This risk “budget” can be interpreted as encoding the evaluator’s competing objectives: to reject the null and stop the evaluation quickly in easier cases (front-loading the risk accumulation) against the desire to wait longer to achieve a significant decision in harder instances (delaying risk accumulation until more data is collected). Importantly, *any risk budget f that is nonnegative everywhere and sums to some $r \in [0, \alpha^*]$ maintains Type-I error control at level α^** . However, the shape of the risk budget will significantly influence the power of the resulting procedure, and thus represents an important component in solving (near-optimally) Eq. (4). With this said, in the following experiments we fix the risk budget to be uniform in order to focus on optimizing the decision regions (described in Section IV-E); to be explicit: this means the risk budget is $f(n) = \frac{\alpha^*}{N_{\max}}$ for each scenario. This selection is heuristically reasonable, but leaves potential for further improvements in future work.

Algorithm 1 STEP Decision-Rule Synthesis

Input: $N_{\max} > 0$, risk budget function $f(n)$, type-I error limit $\alpha^* \in (0, 1)$, number of approximation points M
Initialize: $\zeta_0 = \emptyset$, $(P_0)_{ij} = 1$ if $(i, j) = (0, 0)$ else 0.
for $n \in \{1, \dots, N_{\max}\}$ **do**
 $P_n \leftarrow \text{Propagate}(P_{n-1}, \zeta_{n-1}, M)$ {Eq. (8)}
 $\mathbf{w}_n \leftarrow \text{Opt}(P_n, f)$ {Eq. (10)}
 $\zeta_n \leftarrow \text{Compress}(\mathbf{w}_n)$
end for
return $\zeta = \{\zeta_1, \dots, \zeta_{N_{\max}}\}$ {STEP policy}

E. Tractable Optimization

Having specified the risk budget $f(n)$, it is straightforward to verify that the Type-I Error control is achieved if the following constraint is satisfied for all $n \in \{1, \dots, N_{\max}\}$:

$$P_n \mathbf{w}_n \leq \sum_{k=1}^n f(k) \mathbb{1} \quad (9)$$

Under this constraint, our objective is to maximize the size of the rejection region. We propose to solve the following series of optimization problems to tractably construct the rejection regions, one for each n :

$$\begin{aligned} \max_{\|\mathbf{w}_n\|_{\infty} \leq 1} \quad & \mathbb{1}^T \mathbf{w}_n \\ \text{s.t.} \quad & P_t \mathbf{w}_n \leq \sum_{k=1}^n f(k) \mathbb{1} \\ & 0 \leq \mathbf{w}_n \leq 1, \end{aligned} \quad (10)$$

This objective encourages the rejection from as many states as possible, maximizing the size of $\mathcal{X}_n^{\text{Reject Null}}$. Furthermore, it implicitly rejects from states less likely to occur under any null hypothesis, which are “cheaper” in terms of accruing risk. The first constraint ensures that the Type-I error is controlled up to time n , as discussed in Section IV-C and Section IV-D. The second constraint is to enforce boundedness of rejection probabilities in $[0, 1]$.

The optimization problem Eq. (10) is a simple linear program that can be efficiently solved by a standard optimization software; it is a maximization of $\|\mathbf{w}_n\|_1$ over the nonnegative orthant, subject to additional linear inequality constraints to control Type-I Error. Because each P_n depends on the corresponding ζ_{n-1} , the optimization needs to be performed sequentially for each n in increasing order. Nevertheless, all the computation can be performed offline prior to running the actual policy evaluation. See Algorithm 1 for the decision-rule synthesis procedure. The fact that the optimization problems are sequentially solvable is owed to the isolation of the risk accumulation rate $f(n)$ as a tunable parameter; otherwise, the rejection regions would be generally coupled across n and the optimization would be more challenging. Finally, we note in Algorithm 1 that the representation of the optimal policy can be compressed substantially from the vector $\mathbf{w}_n$ to a set of connected sets $\zeta_n = \left\{ \mathcal{X}_n^{\text{Reject Null}}, \mathcal{X}_n^{\text{Accept Null}}, \mathcal{X}_n^{\text{Continue}} \right\}$

defined by their boundaries, due primarily to the general monotonicity of the optimal regions and the unimodality of the distribution of the test statistic.

V. EXPERIMENTS

The experiments are designed to investigate the following key aspects of effective evaluation:

- A. How does sequential testing compare with batch testing in terms of statistical validity (Type-1 Error)?
- B. What are the unique advantages of STEP in sequential comparison problems of varying difficulty?
- C. How sample efficient is STEP in practical policy comparison settings?

The experiments address these questions through extensive numerical validation on simulated success/failure data, as well as practical validation on both simulated rollouts and tasks on physical hardware.

A. Baseline Procedures

The baselines constitute the SOTA sequential analysis methods described in Lai [30] and Lai and Zhang [31] (termed “Lai”) and the Safe, Anytime-Valid Inference (SAVI) method of Turner and Grünwald [50], which is specifically tailored to contingency table (i.e., policy comparison) problems (termed “SAVI”). Similar to STEP, Lai is a valid sequential method under a pre-determined $N_{\max}$ and is asymptotically optimal as $N_{\max}$ tends to infinity. The specific difference between these methods is that Lai makes the asymptotic assumption of normality of the empirical means. This allows for any distribution subject to the central limit theorem (CLT) to be analyzed tractably in the *large-data regime*, as the resulting optimal stopping PDE amounts to the solution of a reverse heat equation. STEP, by contrast, solves the *exact* PDE at the fundamental finite-sample resolution of the testing problem. As will be shown, the Lai procedure does best in difficult (i.e., requiring a lot of data) and symmetric (i.e., p_0, p_1 each near 0.5) test settings, as these best match the asymptotic approximation where the CLT approximation is well-matched to the exact distribution of the empirical means. SAVI, by contrast, does not require a fixed $N_{\max}$ *a priori* and maintains validity for an arbitrarily large $N_{\max}$. In essence, SAVI procedures construct a test statistic subject to stochastic dynamics (the sequential incorporation of data) which possesses a high-probability ‘stability certificate’ under any null data generating process. The ‘stability’ in this case grants time-uniformity (in perpetuity, i.e., not requiring any finite $N_{\max}$); thus, under any null hypothesis, the probability of *ever* violating the certificate is bounded (by design) to be less than α . In exchange for increased generality and robustness, however, SAVI methods inevitably sacrifice sample complexity in cases where a finite upper-bound to $N_{\max}$ is determined by practical time and resource constraints that researchers are subject to. We emphasize that each existing method is widely applicable to practical testing problems beyond robotics, as is STEP. Importantly, STEP is tailored specifically to the *small- $N_{\max}$ evaluation regime*.

Additionally, an Oracle Sequential Probability Ratio Test (SPRT) [55] is run using the true singleton alternative and associated worst-case singleton null; this method represents a near-optimal procedure for easier problems where feasible methods quickly approach a terminal power close to 1 and still serves as a reasonable benchmark in harder problems. *We emphasize that in practical cases, the Oracle method is infeasible to the evaluator*; it is included to give a conservative estimate of the optimality gap of each method. Additional information about each baseline is included in Section VI.

B. Numerical Simulation

For evaluation, we discretize a grid over the space of alternatives in 10-percentage point increments, offset from zero by five points. There are forty-five resulting alternatives in which $p_1 > p_0$: $(p_0, p_1) = \{(0.05, 0.15), (0.05, 0.25), \dots, (0.85, 0.95)\}$. For each of these alternatives, 5000 trajectories (each of length $N_{\max}$) are simulated from $(\text{Ber}(p_0), \text{Ber}(p_1))$. In addition, 400 additional trajectories are generated under the worst-case null for each of the 45 alternatives, in order to verify the Type-I Error control. This evaluation data is shared across each methodology.

The algorithms are first validated on multiple pairs of $(N_{\max}, \alpha^*)$. The ensuing numerical results will use the case $N_{\max} = 500, \alpha^* = 0.05$; similar figures for $N_{\max} = 100, \alpha^* = 0.05$ are included in Section X-B in the Supplement. For each method, we compute for each of the 45 alternatives the following characteristics: (1) Type-I Error – the fraction of associated worst-case null trajectories which have incorrectly rejected the null; (2) Terminal Power – the fraction of alternative trajectories which have correctly rejected the null by step $N_{\max}$; (3) Cumulative Power – a visualization of the fraction of alternative trajectories which have correctly rejected the null by step t for all $t \in \{0, 1, \dots, N_{\max}\}$. Note that another natural evaluation metric, the expected time-to-decision ($\mathbb{E}[n_{\text{stop}}]$), can be derived from the cumulative power as the area between the curve and the constant $y = 1$.

1) *How does sequential testing compare with batch testing in Type-I Error control?*: We show the Type-1 Error control of each sequential method in Fig. 3, as well as a widely-used batch method (Barnard’s Test [3]) run in sequence. Each sequential method maintains Type-1 Error control; however, STEP is the most efficient at using the full risk budget (lightest blue), while Lai is weaker (more conservative) in the lower-variance regime and SAVI is weaker in general (darker blue). Using a batch method like Barnard’s Test in sequence, conversely, violates Type-1 Error control (red). This highlights the aforementioned p-hacking issue of the batch testing scheme. A rigorous batch evaluator, having chosen N , cannot adapt to the data as it arrives lest they invalidate a resulting conclusion. On the other hand, STEP provides a safety margin for continued testing with the maximum allowable sample size $N_{\max} > N$.

2) *What are the unique advantages of STEP in comparison problems of varying difficulty?*:

a) *Terminal Power*: We begin a more fine-grained comparison of the efficiency of sequential procedures by presenting

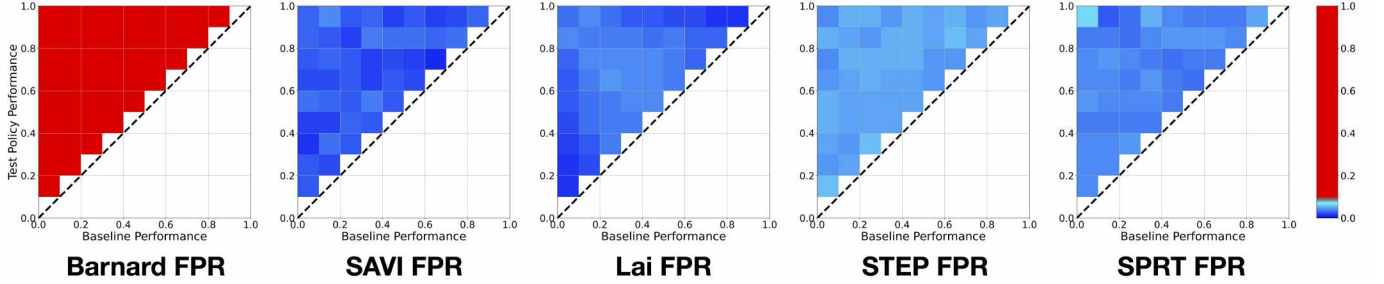


Fig. 3: False positive rate of four feasible methods (Barnard, SAVI, Lai, and Ours (STEP)) and the SPRT (Oracle method) for 400 simulated trajectories drawn from the respective worst-case null distributions for each of 45 alternatives (squares in color); $N_{\max} = 500$ and $\alpha^* = 0.05$. Note that naively utilizing a batch method in sequence leads to violation of Type-1 Error control (red). Additionally, note that SAVI and Lai struggle to utilize the full risk budget in finite $N_{\max}$ (darker blue regions).

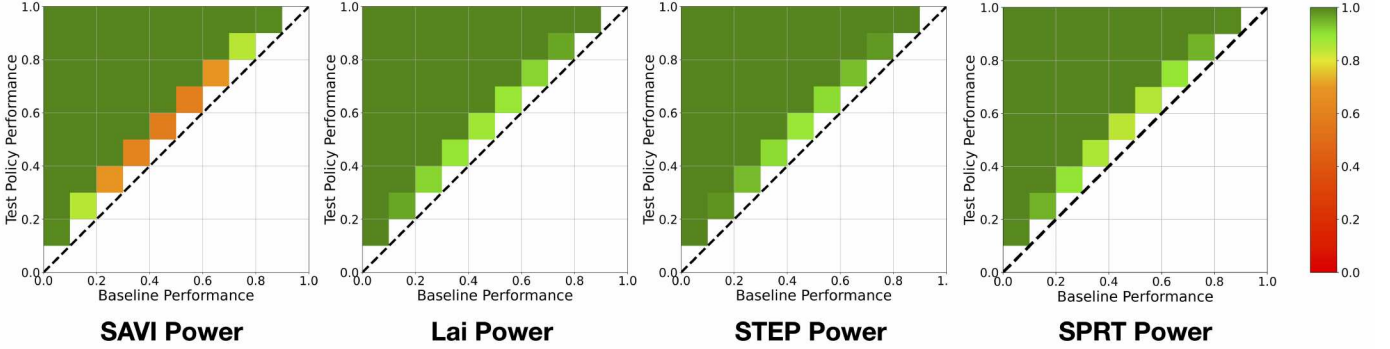


Fig. 4: Terminal power of three feasible methods (SAVI, Lai, and Ours (STEP)) and the SPRT (Oracle method) for 5000 simulated trajectories on each of 45 alternatives (squares in color); $N_{\max} = 500$ and $\alpha^* = 0.05$. Because $N_{\max}$ is large, the terminal power is high for all but the most difficult cases; however, note that SAVI has significantly worse power in these difficult instances where p_0 and p_1 are closely competing. This is due to its inherent validity for arbitrary $N_{\max}$, which is unnecessary in modern robotics evaluation contexts and renders the methods conservative in harder instances.

the terminal power (probability of deciding **RejectNull** by step $N_{\max}$) of all feasible methods and the Oracle SPRT in Fig. 4. This metric serves to illustrate a downside of SAVI: its inherent validity at arbitrarily large $N_{\max}$ imposes strong finite-time costs. In this case, the terminal power when the true gap in policy performance is 10 percentage points is significantly lower than the Lai baseline and our STEP, which closely approximate the SPRT Oracle.

b) Cumulative Power: We now demonstrate the downside of the Lai baseline procedure as compared to our method through the cumulative power of the procedure. In Fig. 5 (right), we illustrate the cumulative power on a hard evaluation case. This represents a second view of the observation presented in the terminal power analysis: the Lai procedure and our STEP each significantly outperform the SAVI procedure in the small-gap regime. Conversely, in Fig. 5 (center), we illustrate a difficult setting in which the Lai procedure struggles. In this setting the performance gap is again 10 percentage points, but the distribution is lower-variance and skewed as compared to that of Fig. 5 (right); as such, the Lai procedure cannot effectively adapt and our STEP significantly outperforms in terms of deciding more quickly for the alternative.

C. How sample efficient is STEP in practical policy comparison settings?

In addition to the numerical validation, we evaluate STEP through two sets of real-world robot evaluation experiments and a simulation experiment. In the first set of experiments, we compared policies with noticeable performance gaps to demonstrate the early-stopping capability of our approach. In the second set, we compared closely-competing policies to characterize necessary sample sizes for statistical validity when policy performance gaps become small. In the simulation experiment, we perform multi-task comparison of two SOTA imitation learning policies. See Section X-A in the Supplement for more details on the hardware experiments.

1) Hardware Evaluation in the Large/Medium-Gap Regime: In this experiment, we consider two manipulation tasks for a bimanual Franka Emika Panda robot: **FoldRedTowel** (Fig. 9a) and **CleanUpSpill** (Fig. 9b and Fig. 9c). We trained single-task diffusion policies [14] on each task, with 300 human demonstrations for **FoldRedTowel** and 150 for **CleanUpSpill**, respectively. In addition to the RGB images, the policy receives the proprioceptive states as additional observations.

In **FoldRedTowel**, we compare two policy checkpoints from a single training run. The baseline policy π_0 was trained for

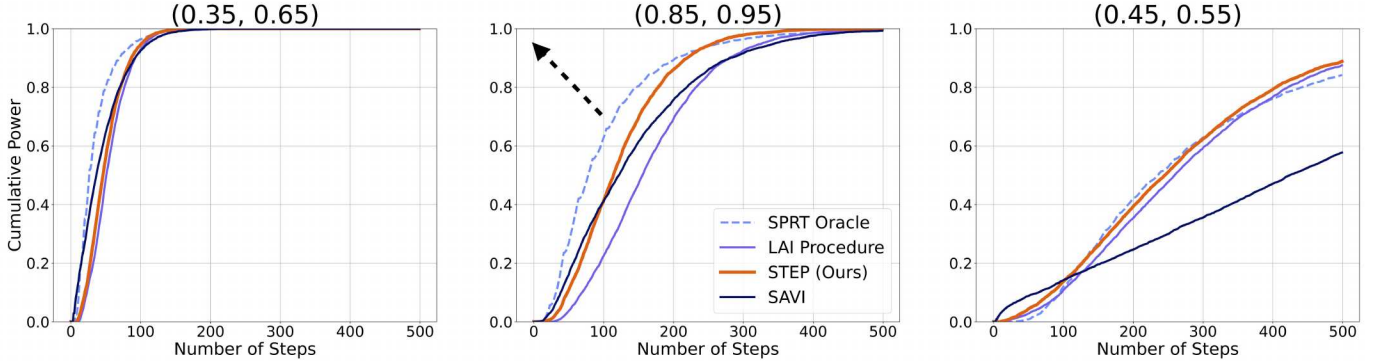


Fig. 5: Cumulative power of all feasible methods (Lai, SAVI, STEP (Ours)) and SPRT Oracle over 5000 trajectories in three evaluation settings of increasing difficulty; (p_0, p_1) for each setting title the respective figures. $N_{\max} = 500$ and $\alpha^* = 0.05$. The expected time-to-decision is the integral of the area *above* the cumulative power curve; therefore, curves higher and to the left are better (black arrow). (Left) For a gap of 30 percentage points, all methods demonstrate similar stopping times. (Center) For a gap of 10 percentage points in the low-variance regime (i.e., farther from 0.5), STEP significantly outperforms the Lai procedure and is better than SAVI. (Right) For a gap of 10 percentage points in the high-variance regime, SAVI struggles to maintain high power, and underperforms the other methods.

10000 gradient steps with an AdamW [34] optimizer and achieved the action mean-squared error (MSE) of $1.61\text{E-}3$ on the validation set. The other policy π_1 continued the training for 70000 additional steps, yielding the validation action MSE of $1.35\text{E-}3$. To evaluate each policy, five in-distribution (ID) initial conditions were chosen and repeated 10 times each, constituting 50 total trials. ID initial conditions visually resemble the ones that appear in the training dataset. As shown in Table I (rows 1-3), the empirical gap in success rates was 36 percentage points (56% to 92% success), suggesting that π_0 was under-trained. Each sequential method detected a significant difference at level $\alpha^* = 0.05$ in 17 – 23 trials (apiece). That is, had the algorithms been active during collection, the last 27 – 33 rollouts per policy were unnecessary for confirming the improvement of the later checkpoint over the former. Additionally, note that the Lai method and our STEP sequential procedures were each tuned for an $N_{\max}$ of 50 (row 1), 200 (row 2), and 500 (row 3) rollouts. In the latter two cases, additional rollouts could have been run up to 200 or 500 per policy if the gap was smaller *without compromising the validity of the decision*.

In **CleanUpSpill**, we compare the same policy on two different sets of initial conditions. The task is similar to the one originally presented by Xu et al. [60], which compares a set of ID initial conditions against the out-of-distribution (OOD) initial conditions. The ID set includes 10 initial conditions with a white towel and a short blue mug whereas the OOD set uses 10 with a checkered towel with a tall cyan mug (each initial condition is repeated five times). As shown in Table I (rows 4–6), the empirical gap of 52 percentage points was detected in 7–14 trials by all methods, though they were tuned (where applicable) to an $N_{\max}$ up to thirty to seventy times larger; this demonstrates the significant reduction in sensitivity (from an evaluator’s standpoint) arising from setting $N_{\max}$ versus

choosing a batch size N . Furthermore, STEP’s efficiency only minimally degrades when $N_{\max}$ is increased from 200 (row 5) to 500 (row 6). In this setting, any of the three sequential methods would have prevented the need for at least 70 of the 100 total rollouts (35 of the 50 batch trials per policy).

2) *Hardware Evaluation in the Small-Gap Regime:* We evaluate STEP on two open-source vision-language-action (VLA) models: Octo-Base [40], an action-chunking transformer-based diffusion policy, and OpenVLA [27], an autoregressive policy leveraging a pretrained large language model backbone. All experiments were conducted in a toy kitchen environment from the Bridge Data V2 dataset [57], which is included in both policies’ training data. We considered the task of placing a carrot on a plate (denoted **CarrotOnPlate**, see Fig. 9d and Fig. 9e), which is representative of evaluations investigated in [40, 27]. All policies were run on the Widow X 250S following the setup in [57].

For this task, the initial placement of the carrot is uniformly sampled from three possible locations (left, center, or right) on the counter, and the plate is placed next to the sink (see Figure 9d). At the start of each trial, the robot joint angles are initialized such that the gripper is roughly aligned with the carrot initial position. The environment follows a categorical distribution with two outcomes: no object distractors or two object distractors. We consider two environment distributions: **Env1** in which there are no distractors with probability 0.8 and **Env2** in which there are no distractors with probability 0.6. Object distractors are sampled uniformly (without replacement) from the following object categories: orange, apple, green and blue sponges, brown and yellow cubes, eggplant, spoon, and towel. The initial locations of distractors is also sampled uniformly without replacement from four possibilities: on the stove, left of the stove, above the stove, or next to the faucet. The distractors are physically placed according to a uniform (continuous) distribution within the selected region.

Task	Type	α^*	$N_{\max}$	N	$\hat{p}_0$	$\hat{p}_1$	SAVI	Lai	STEP (Ours)	SPRT***
FoldRedTowel	π_i	0.05	50	50	0.560	0.920	20	17	19	17
FoldRedTowel	π_i	0.05	200	50	0.560	0.920	20	21	21	17
FoldRedTowel	π_i	0.05	<u>500</u>	50	0.560	0.920	20	23	23	17
CleanUpSpill	$\mathcal{D}_{s_0, o_0}^i$	0.05	50	50	0.280	0.800	7	8	8	7
CleanUpSpill	$\mathcal{D}_{s_0, o_0}^i$	0.05	200	50	0.280	0.800	7	13	9	7
CleanUpSpill	$\mathcal{D}_{s_0, o_0}^i$	0.05	<u>500</u>	50	0.280	0.800	7	14	13	7
CarrotOnPlate	$\mathcal{D}_{s_0, o_0}^i$	0.05	100	100	0.590	0.680	—	—	—	—
CarrotOnPlate	π_i	0.05	100	100	0.680	0.760	—	—	—	—
SpoonOnTowel	π_i	0.01	500	500	0.084	0.386	33	36	36	26
EggplantInBasket	π_i	0.01	500	500	0.400	0.564	192	125	131	128
StackCube	π_i	0.01	500	500	0.000	0.030	329	417	225	135
Multitask	π_i	0.03	1500	1500	N/A	N/A	554	578	392	289

TABLE I: Empirical time-to-correct-decision for all **hardware (top)** and **simulation (bottom)** policy comparisons. The comparison type is described first; π_i is comparing two policies, $\mathcal{D}_{s_0, o_0}^i$ compares one policy under possible distribution shift. The utilized Type-I Error α^* and $N_{\max}$ describe the constraints applied *a priori* by the evaluator (we underline to emphasize the change in $N_{\max}$ for rows 1-3 and 4-6; observe that the sensitivity of the stopping times is very small). $N \leq N_{\max}$ represents the amount of data available for the statistical analysis. We report the terminal empirical success rates (after N trials) of each policy in each setting under $\hat{p}_i$ (this information is not available to any feasible algorithm). We do not have truth labels on this data; however, in all cases, every method arrived at the same decision, including the Oracle SPRT which has *a priori* access to privileged information $(\hat{p}_0, \hat{p}_1)_N$; this decision was Reject Null for all rows except the CarrotOnPlate tasks, which each returned Fail To Decide. We report the stopping times of all methods on the right of the table for every context; in all cases: lower is better. We put in **bold** any *feasible* method result that is near-optimal within ten trials (absolute) or 25% (relative) of the SPRT Oracle, which is *not implementable* by an evaluator. In the Multitask setting, we test $p_1 > p_0$ *uniformly across the preceding three tasks*. This stopping time is the sum by column of the stopping times for the three tasks. Our method saves the evaluator over 160 trials in uniform certification over these three tasks as compared to either feasible baseline.

We sample 100 environment configurations each for three settings: i) Octo under distribution **Env1**, ii) OpenVLA under distribution **Env1**, and iii) Octo under distribution **Env2**. We then make the following comparisons with STEP: i) Octo (**Env1**) and Octo (**Env2**), and ii) Octo (**Env1**) and OpenVLA (**Env1**). In the first comparison (Table I, row 7), we test the effect of distribution shift in the probability of distractors being present. Here, the $\hat{p}_0$ corresponds to the perturbed distribution Octo (**Env2**), and $\hat{p}_1$ to the nominal distribution Octo (**Env1**). We find that, while there is an empirical gap (59% vs 68%) at $N = 100$, no method returns a significant result at $\alpha^* = 0.05$. In the second comparison (Table I, row 8), $\hat{p}_0$ corresponds to Octo (**Env1**) and $\hat{p}_1$ corresponds to OpenVLA (**Env1**). We observe no significant gap despite the empirical gap of 8 percentage points in favor of OpenVLA. Importantly, insignificance of the test does not mean that the null hypothesis should be accepted [20]; there remains a possibility that OpenVLA actually outperforms Octo, or that Octo’s performance indeed degrades due to the presence of distractors. However, our budget of $N_{\max} = 100$ was likely not sufficient to accumulate enough evidence. In Section X-D in the Supplement, we further investigate this data insufficiency to show that, if the ground truth values were equal to the empirical success rates (59% vs. 68% and 68% vs. 76%), then we would require $N_{\max} = 500$ trials to confidently determine

$p_1 > p_0$. This number is an order of magnitude larger than the current norms, reflecting fundamental yet often overlooked challenges in trustworthy policy comparison.

3) *Multi-Task Evaluation in SimplerEnv Simulation*: Finally, we briefly consider the problem of multi-task and multi-policy extensions to this framework, and illustrate its potential via an example of policy evaluation in simulation (where costs are lower than on hardware, but still can be significant). Concretely, Octo-Small (π_1) and Octo-Base (π_0) [40] are compared in the SimplerEnv [32] simulation environment on three tasks of varying empirical difficulty. On the **EggplantInBasket** task, the policies each succeed at a near-50% rate, with a gap of 16.4 percentage points. For the **SpoonOnTowel** task, the gap is larger at 30 percentage points. For the hardest task, **StackCube**, the performance gap is 3 percentage points. The evaluation statistics are shown in Table I, rows 9-11. Note that the empirical success rates we observed are consistent with the findings of Li et al. [32] that Octo-Small is more performant on these tasks (see their Table V). We seek to evaluate the multitask comparison of the two policies. Letting $p_s^{[\tau]}$ denote the performance of Octo-Small and $p_b^{[\tau]}$ denote the performance of Octo-Base on task τ , we test:

$$\begin{aligned}
\mathbb{H}_0 : \exists \tau \in \{1, 2, 3\} \ p_s^{[\tau]} &\leq p_b^{[\tau]} \\
\mathbb{H}_1 : \forall \tau \in \{1, 2, 3\} \ p_s^{[\tau]} &> p_b^{[\tau]}.
\end{aligned} \tag{11}$$

Many established and sophisticated methods exist to efficiently run multi-hypothesis testing (in this case, we are essentially evaluating three separate hypotheses, one for each task⁷). As a simple illustration, we use the standard Bonferroni (union bound) correction [15] to evaluate the test: specifically, running each of the individual three tests at level $\alpha_{[\tau]}^* = 0.01$, we observe the stopping times shown in Table I, rows 9-11 (right-hand side). Via the Bonferroni correction, the combined decision (**RejectNull**, because every subtest decided **RejectNull**) expressed in Eq. (11) is then confirmed at $\alpha = 0.03 = \sum_{\tau} \alpha_{[\tau]}^*$. As illustrated in Table I, each sequential method saves a substantial number of simulation rollouts on the easiest subtest (**SpoonOnTowel**). As expected, SAVI begins to struggle when the tests become more challenging, such as in **EggplantInBasket**. Finally, we observe the weakness of the Lai procedure: in heavily skewed cases, it suffers substantially even compared with SAVI methods, as shown in **StackCube**. To summarize: naive multitask evaluation requires the aggregation of multiple batches of rollouts, here totaling 1500 per policy (500 per task per policy). Note in rows 9-11 how different the number of requisite trials can be, and therefore how hard it is to reliably run the evaluation using a fixed batch size. On the easiest task, even when tuned to $N_{\max} = 500$, the comparison was answered in fewer than 40 rollouts by all sequential methods, a savings of over 90%. On the progressively harder cases the number of required samples increased 5-10 times over the easiest, but our method (STEP) improved substantially over each of the other sequential procedures. In total, STEP would have saved the evaluator an additional 160 rollouts for each of Octo-small and Octo-base for the multitask comparison problem as compared to the current SOTA approaches.

VI. RELATED WORK

Thus far, our discussion has focused solely on the practical setting of comparative policy evaluation. However, mathematical statistical testing has a well-established near-century-long history, and sequential testing in particular has existed for nearly the same amount of time. This section provides an extensive review of the statistics literature to further highlight the significance of our approach.

A. Statistical Testing and Policy Evaluation

The Neyman-Pearson (N-P) statistical testing paradigm [37] forms the foundation of the last century of frequentist statistical decision theory. Applicable to questions spanning many fields, these methods have generally been applied in robotics for predictable policy *characterization*⁸ [53, 1], where the number of samples is fixed and a single decision or estimate is to be made. The Neyman-Pearson Lemma [37] and the Karlin-Rubin Theorem [25] give sufficient conditions for maximal

⁷Note that multi-hypothesis testing can naturally handle the case of **multi-policy** comparison as well, where we would reduce the test to a set of pairwise comparisons that are examined simultaneously.

⁸As a simple example, one can accurately predict *a priori* that for estimating $\text{Ber}(p)$ with $\hat{p} \in [0.25, 0.75]$ and $N \geq 36$, a 95% confidence interval for p will be approximately $\hat{p} \pm \frac{1}{\sqrt{N}}$.

power probability ratio tests (PRTs), which form the precursor to the SPRT Oracle used in this work. Specific methods have been developed for robot policy comparison-type problems in the context of 2x2 contingency tables. Of the tests by Fisher [17], Boschloo [6], and Barnard [3], the latter is a powerful batch procedure for comparison problems. However, while it has strong power in the batch setting, it does not provide a direct mechanism for choosing the appropriate size of the batch *a priori*.

B. Sequential Statistical Evaluation Methods

The difficulty in choosing the appropriate batch size motivates the sequential testing framework set out in Wald [55]. This is the setting adopted in this paper and described in Section III. Wald and Wolfowitz [56] showed that in the simple-vs-simple setting, the sequential probability ratio test (SPRT) minimizes the expected number of samples among all tests that control Type-I and Type-II error, extending the N-P result. Significant efforts have extended near-optimality results to the composite testing regime. Minimax results attempt to limit the worst-case expected sample size [26, 33, 16], while others minimize the expected sample size under a weighted mixture over the alternatives [47, 19]. Lai [30] reconciled this Bayesian interpretation with the frequentist developments of Chernoff [10, 11, 12].

1) *Optimal Stopping-Based Methods*: The direct approach to synthesizing near-optimal decision regions in the composite-vs-composite regime relies on developments in the theory of martingales and optional stopping [58]. Van Moerbeke [51] provides a clear reduction of the statistical decision making problem to that of optimal stopping and exposit previous developments in the area equating the optimal stopping procedure with the solution of a Stefan-type free-boundary partial differential equation (PDE) [8]. Unfortunately, while specification of the testing problem is usually feasible, its mapping to the PDE parameterization is generally difficult to specify under composite null hypotheses, rendering this method impractical. A parallel line of work considers utilizing asymptotic approximations of the solution to the free-boundary problem in order to construct near-optimal tests. The work of Lai [30] solves for a near-optimal procedure in the univariate composite-vs-composite setting and follow-on work [31, 9] extends this to the multivariate setting. While useful for proving asymptotic optimal rates, these methods suffer in the finite- $N_{\max}$ regime due to losing decision-making information.

2) *Safe, Anytime-Valid Inference (SAVI) Methods*: The difficulty in synthesizing optimal procedures in the multivariate setting (see, e.g., [31, 9]) motivates recent developments which fuse a distributionally robust safety invariance and the structure of the probability ratio test. This yields a family (SAVI) of methods which generalize to any $N_{\max}$ (i.e., arbitrarily small performance gaps) and a richer set of composite-hypothesis testing problems. Utilizing Ville's Inequality (a sequential generalization of Markov's Inequality) [52], SAVI methods construct a probability ratio test that enforces Type-I error control uniformly in time [22, 43]. In the small sample regime

it is the effect of the Power-1 nature of the test [44, 29, 43] that can most negatively affect performance.

C. Numerical Implementations

Numerical methods are of significant practical importance in statistical testing. Excellent implementations of batch procedures have been released in the SciPy [54] package; recently, the particular problem of optimal binomial confidence intervals was addressed in [53]. However, despite a significant history dating back to the 1980s [13], numerical methods for sequential analysis in evaluation are relatively limited. Further, in [13] as in more recent work, emphasis has remained on the simple-null or univariate composite setting [38, 39, 16] for non-SAVI procedures. However, to our knowledge, a numerical implementation of near-optimal policy comparison procedures in the full multidimensional composite-vs-composite setting has not yet been implemented.

VII. LIMITATIONS

Despite the strong performance of STEP as shown in Section V, it does have several practical and theoretical limitations. First, the development of STEP heavily relies on the mathematical structure of Bernoulli distributions, which essentially requires the policy evaluation results to be presented as binary success/failure metrics. Extending STEP to more complex discrete and continuous distributions would be a valuable future research direction for broader applicability. Second, in our current implementation a user needs to specify STEP’s risk accumulation rate $f(k)$. Although we empirically show that the uniform rate already achieves near-optimality, further research is needed to improve the performance.

Besides these limitations that are specific to STEP, researchers must be mindful of underlying assumptions and common misuse of statistical hypothesis testing [20]. Concretely, the foundational assumption is that evaluation data are i.i.d. as all the statistical assurances are built on top of it. To this end, we ensured that all hardware evaluations comported with the best practices of [28]. Finally, Goodhart’s Law [49] provides a useful warning about inadvertent p-hacking: having statistically rigorous evaluation is important, but if “significance at level α^* ” becomes a target and not a metric, it can induce undesirable research practices. We emphasize that p-values and significance levels are only as rigorous as the rigor of the research methods that utilize them. Their adoption has the potential to be enormously valuable for the empirical codification of robotics knowledge, but they are not a panacea.

VIII. FUTURE WORK

There are several concrete extensions of STEP which promise significant benefits to practical robotic evaluation. Most direct is incorporation of more principled multi-policy evaluation, building upon Section V-C3. Additionally, partial credit schema have found increasing importance in long-horizon evaluation [28]; generalizing STEP beyond binary

success/failure metrics to discrete (multinoulli) partial credit is direct in theory, although it presents additional challenges in implementation.

IX. CONCLUSION

We present STEP, a novel sequential statistical method to rigorously compare performance of imitation learning policies through a series of evaluation trials. STEP’s sequential evaluation scheme provides flexibility in adapting the number of necessary trials to the underlying difficulty of policy comparison. This leads to sample efficiency in cases where one policy clearly outperforms the other, while avoiding overconfident and potentially incorrect evaluation decisions when the policies are closely competing. We show that STEP near-optimally minimizes the expected number of trials required in the policy comparison problem. Furthermore, STEP matches or exceeds the performance of state-of-the-art baselines across a wide swath of practical evaluation scenarios in numerical and robotic simulation and on numerous physical hardware demonstrations. These results highlight the practical utility of STEP as a versatile statistical analysis tool for policy comparison, contributing to the foundation of robot learning as an empirical science.

ACKNOWLEDGMENTS

This work was primarily completed during an internship at Toyota Research Institute (TRI). The authors would like to thank Chen Xu for insightful feedback and discussions, as well as the robot teacher team at TRI headquarters for their data collection efforts, in particular Derick Seale and Donovan Jackson. Finally, we thank the anonymous reviewers and the area chair for incisive suggestions for improving the paper.

REFERENCES

- [1] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 29304–29320, 2021. doi: 10.48550/arXiv.2108.13264.
- [2] Aaron D. Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control Barrier Functions: Theory and Applications. In *European Control Conference (ECC)*, pages 3420–3431, June 2019. doi: 10.23919/ECC.2019.8796030.
- [3] G. A. Barnard. Significance Tests for 2x2 Tables. *Biometrika*, 34(1-2):123–138, January 1947. ISSN 0006-3444. doi: 10.1093/biomet/34.1-2.123.
- [4] 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.
- [5] 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. *arXiv preprint arXiv:2410.24164*, 2024. doi: 10.48550/arXiv.2410.24164.
- [6] 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.
- [7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. doi: 10.48550/arXiv.2307.15818.
- [8] 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.
- [9] Hock Peng Chan and Tze Leung Lai. Asymptotic Approximations for Error Probabilities of Sequential or Fixed Sample Size Tests in Exponential Families. *The Annals of Statistics*, 28(6):1638–1669, 2000. ISSN 0090-5364. Publisher: Institute of Mathematical Statistics.
- [10] 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.
- [11] Herman Chernoff. Sequential Test for the Mean of a Normal Distribution III (Small n). *The Annals of Mathematical Statistics*, 36(1):28–54, 1965. ISSN 0003-4851. Publisher: Institute of Mathematical Statistics.
- [12] 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.
- [13] 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.
- [14] 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*, 0, 2024. doi: 10.48550/arXiv.2303.04137.
- [15] Olive Jean Dunn. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. doi: 10.1080/01621459.1961.10482090.
- [16] 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.
- [17] 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>.
- [18] Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on Robot Learning (CoRL)*, pages 158–168. PMLR, 2022. doi: 10.48550/arXiv.2109.00137.
- [19] Robert Fortus. Approximations to Bayesian Sequential Tests of Composite Hypotheses. *The Annals of Statistics*, 7(3):579 – 591, 1979. doi: 10.1214/aos/1176344679.
- [20] Sander Greenland, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31(4):337–350, 2016.
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. doi: 10.48550/arXiv.1512.03385.
- [22] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, April 2021. ISSN 0090-5364, 2168-8966. doi: 10.1214/20-AOS1991. Publisher: Institute of Mathematical Statistics.
- [23] 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.

- [24] 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.
- [25] Samuel Karlin and Herman Rubin. The Theory of Decision Procedures for Distributions with Monotone Likelihood Ratio. *The Annals of Mathematical Statistics*, 27(2):272–299, June 1956. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177728259. Publisher: Institute of Mathematical Statistics.
- [26] J. Kiefer and Lionel Weiss. Some Properties of Generalized Sequential Probability Ratio Tests. *The Annals of Mathematical Statistics*, 28(1):57–74, 1957. ISSN 0003-4851. Publisher: Institute of Mathematical Statistics.
- [27] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Open-VLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246*, 2024. doi: 10.48550/arXiv.2406.09246.
- [28] 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. In *Robotics: Science and Systems*, 2024. doi: 10.48550/arXiv.2409.09491.
- [29] Tze Leung Lai. Power-One Tests Based on Sample Sums. *The Annals of Statistics*, 5(5):866 – 880, 1977. doi: 10.1214/aos/1176343943.
- [30] 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.
- [31] 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.
- [32] 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. doi: 10.48550/arXiv.2405.05941.
- [33] Gary Lorden. Nearly-Optimal Sequential Tests for Finitely Many Parameter Values. *The Annals of Statistics*, 5(1):1–21, 1977. ISSN 0090-5364. Publisher: Institute of Mathematical Statistics.
- [34] I Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. doi: 10.48550/arXiv.1711.05101.
- [35] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning (CoRL)*, volume 164, pages 1678–1690. PMLR, 2021. doi: 10.48550/arXiv.2108.03298.
- [36] Ravi Myneni. The Pricing of the American Option. *The Annals of Applied Probability*, 2(1):1–23, 1992. ISSN 1050-5164. Publisher: Institute of Mathematical Statistics.
- [37] 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.
- [38] Andrei Novikov and Fahil Farkhshatov. A computational approach to the Kiefer-Weiss problem for sampling from a Bernoulli population. *Sequential Analysis*, 41(2):198–219, 2022. doi: 10.1080/07474946.2022.2070212.
- [39] Andrey Novikov. A numerical approach to sequential multi-hypothesis testing for Bernoulli model. *Sequential Analysis*, 42(3):303–322, 2023. doi: 10.1080/07474946.2023.2215825.
- [40] 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. In *Robotics: Science and Systems (RSS)*, 2024. doi: 10.48550/arXiv.2405.12213.
- [41] Abby O’Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and RT-X models. *arXiv preprint arXiv:2310.08864*, 2023. doi: 10.48550/arXiv.2310.08864.
- [42] 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. doi: 10.48550/arXiv.2402.08191.
- [43] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-Theoretic Statistics and Safe Anytime-Valid Inference. *Statistical Science*, 38(4):576–601, November 2023. ISSN 0883-4237, 2168-8745. doi: 10.1214/23-STS894. Publisher: Institute of Mathematical Statistics.
- [44] H. Robbins and D. Siegmund. The Expected Sample Size of Some Tests of Power One. *The Annals of Statistics*, 2(3):415 – 436, 1974. doi: 10.1214/aos/1176342704.
- [45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

- [46] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011. doi: 10.48550/arXiv.1011.0686.
- [47] Gideon Schwarz. Asymptotic Shapes of Bayes Sequential Testing Regions. *The Annals of Mathematical Statistics*, 33(1):224 – 236, 1962. doi: 10.1214/aoms/1177704726.
- [48] Dušan M. Stipanović, Inseok Hwang, and Claire J. Tomlin. Computation of an over-approximation of the backward reachable set using subsystem level set functions. In *2003 European Control Conference (ECC)*, pages 300–305, September 2003. doi: 10.23919/ECC.2003.7084971.
- [49] Marilyn Strathern. ‘Improving ratings’: audit in the British University system. *European Review*, 5(3): 305–321, 1997. doi: 10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4.
- [50] 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>.
- [51] 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.
- [52] Jean Ville. *Etude Critique de la Notion de Collectif*. PhD thesis, Université de Paris, 1939. Publisher: Gauthier-Villars, Paris.
- [53] Joseph A. Vincent, Haruki Nishimura, Masha Itkina, Paarth Shah, Mac Schwager, and Thomas Kollar. How Generalizable is My Behavior Cloning Policy? A Statistical Approach to Trustworthy Performance Evaluation. *IEEE Robotics and Automation Letters*, 9(10):8619–8626, 2024. doi: 10.1109/LRA.2024.3445635.
- [54] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.
- [55] 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.
- [56] 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.
- [57] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023. doi: 10.48550/arXiv.2308.12952.
- [58] David Williams. *Probability with Martingales*. Cambridge University Press, 1991. doi: 10.1017/CBO9780511813658.
- [59] Xuan Xiao, Jiahang Liu, Zhipeng Wang, Yanmin Zhou, Yong Qi, Qian Cheng, Bin He, and Shuo Jiang. Robot learning in the era of foundation models: A survey. *arXiv preprint arXiv:2311.14379*, 2023. doi: 10.48550/arXiv.2311.14379.
- [60] Chen Xu, Tony Khuong Nguyen, Emma Dixon, Christopher Rodriguez, Patrick Miller, Robert Lee, Paarth Shah, Rares Ambrus, Haruki Nishimura, and Masha Itkina. Can We Detect Failures Without Failure Data? Uncertainty-Aware Runtime Failure Detection for Imitation Learning Policies. *arXiv preprint arXiv:2503.08558*, 2025.

X. SUPPLEMENT

A. Details of Real-World Robot Experiments

All of our real-world hardware tasks are visualized in Fig. 9. In **FoldRedTowel**, the robot first observes an unfolded red towel placed in random poses. The task is considered a success if the robot folds the towel twice and then moves the folded towel to a corner of the table. In **CleanUpSpill**, a mug is initially lying sideways on the table and a coffee spill exists near the mug. The task is successful if one arm puts the mug upright while the other arm picks up a white towel and wipes the spill. In both tasks, a total of four RGB cameras observe the Franka robot and the objects, where two monocular cameras are mounted on the table top and a stereo wrist camera on each of the arms. We trained single-task diffusion policies [14] on each task, with 300 human demonstrations for **FoldRedTowel** and 150 for **CleanUpSpill**, respectively. In addition to the RGB images, the policy receives the proprioceptive states as additional observations. Following [14], the image observations are passed to the ResNet-18 [21] encoder before fed into the U-Net [45] diffusion policy architecture. $T_o = 2$ observations are stacked and fed into the policy network to predict $T_p = 16$ steps of actions. The actions are re-planned after $T_a = 8$ actions are executed.

For the **CarrotOnPlate** task, an experiment is recorded as a success if the robot policy succeeds in placing the carrot on the plate within the max episode count without: i) pushing the carrot off the counter, ii) colliding with the back wall, iii) pushing the plate into the sink, and iv) accumulating a total of 3 cm of negative z commands when the end-effector is in contact with the table surface. For Octo evaluations, we use an action-chunking horizon of 2.

In all the experiments, we take the effort to mitigate distribution shift during trials, such as a change in lighting conditions. We also randomize the order of trials so that any distribution shift due to other factors (e.g., hardware degradation over the course of trials) is equally reflected in all the settings. Where applicable, we also separate the role of the evaluator from the demonstrator of the tasks for training. These practices are adopted from [28] to reduce unintended variability in environmental conditions during policy evaluation.

B. Additional Numerical Simulation Results

We plot results analogous to Fig. 4, Fig. 3, and Fig. 5 in Section V for $N_{\max} = 100$ and $\alpha^* = 0.05$ in Fig. 7, Fig. 6, and Fig. 8. We include for this case the power profile for a Barnard Test that is validly sequentialized (using Bonferroni’s correction); *this rectifies the Type-I Error violation in Fig. 6*. In so doing, it loses significant power and fails to meaningfully compete with the SOTA sequential procedures. In addition to inefficient computational properties, the Bonferroni-correct Barnard procedure becomes even weaker for larger $N_{\max}$.

A key point of emphasis in the $N_{\max} = 100$ regime is the low power of all tests for gaps of approximately 10 percentage points and smaller. Notably, no procedure has power over 50% in the hardest regimes (see the orange regions of every

method in Fig. 7). A small amount of this is due to the sequential procedure; however, a significant amount reflects fundamental uncertainty (variance in outcomes) present for small sample sizes in evaluation. The implication of this is the need for significant increases in evaluation trials in order to effect meaningful comparisons when the underlying gap is small. This will be considered further in the context of the **CarrotOnPlate** hardware experiments (Section V) in Section X-D below.

Finally, we note the presence of a small hint to the weakness of the Lai procedure in skewed settings. Note that in the bottom left and top right of the Lai panel of Fig. 7, the power significantly lags STEP and SPRT; in a similar vein, note the regions of darker blue in the Lai procedure panel of Fig. 6. These reflect an inherent inefficiency undergirding Lai method, which directly explain the significant gap on the highly-skewed **StackCube** task in Section V.

C. Empirical Results with Regenerated Sequences

To (approximately) evaluate the counterfactual noise in the data generation process for robotic evaluation, we randomly generate Bernoulli sequences using (as the true data-generating parameters) the empirical success rates of each task in Table I. This provides an estimate of the average sample complexity for each method *were the empirical success rates equal to the true rates ‘in the world’*, providing a Bernoulli counterfactual in Table IV; in that table, all entries present the empirical mean complexities (with standard deviation) over 400 regenerated sequences per task.

D. Further Analysis of **CarrotOnPlate** Experiments

We explore the results of the **CarrotOnPlate** hardware results in more detail. Fig. 10b and Fig. 10c illustrate how the running empirical success rates change as N grows. Note that the relative performance consistently fluctuates and even sometimes flips, which indicates the inherent difficulty of comparison when the two policies are closely competing. In order to estimate the minimum number of necessary trials for these challenging comparisons, we run the SPRT Oracle on multiple instances of $N_{\max}$. Namely, we assume that the true underlying distribution matches the terminal empirical success rates ($\mathbb{H}_1 : (p_0, p_1) = (0.59, 0.68)$ for **Octo (Env2)** vs **Octo (Env1)** and $\mathbb{H}_1 : (p_0, p_1) = (0.68, 0.76)$ for **Octo (Env1)** vs **OpenVLA (Env1)**). We determine the worst-case point null (corresponding $\mathbb{H}_0 : (p_0, p_1) = h_0^* \in \mathcal{H}_0$) for each case and run the SPRT Oracle on the associated simple-vs-simple test, where it is essentially optimal. We observe the following empirical power results (Table II), which can be understood as approximating the probability of rejecting the null (under the draw of the sequence of i.i.d. data) at each level of $N_{\max}$ when the true gap matches the empirical gap observed on 100 trials in hardware.

As shown Table II, nearly 400 trials are required before reaching an approximately 75% chance of rejection over the draw of observed sequences. We emphasize that this is computed via a method that is optimal with respect to the expected

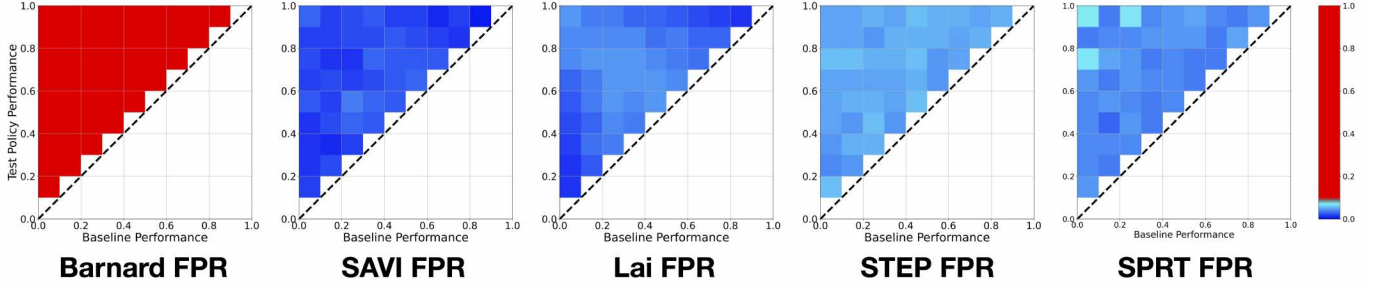


Fig. 6: False positive rate of four feasible methods (Barnard, SAVI, Lai, and Ours (STEP)) and the SPRT (Oracle method) for 1000 simulated trajectories on each of 45 alternatives (squares in color); $N_{\max} = 100$ and $\alpha^* = 0.05$. Note that naively utilizing a batch method in sequence leads to violation of Type-1 Error control (Barnard). Additionally, note that SAVI and Lai struggle to utilize the full risk budget in finite $N_{\max}$ (darker blue regions).

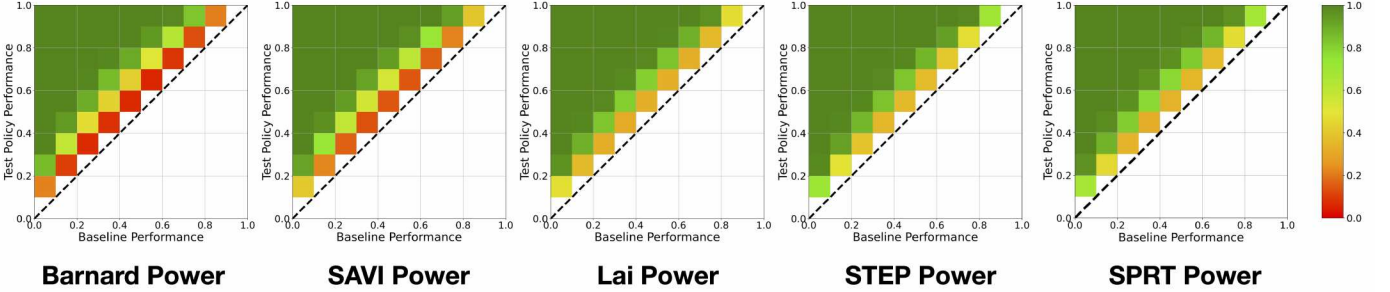


Fig. 7: Terminal power of four feasible methods (Barnard, SAVI, Lai, and Ours (STEP)) and the SPRT (Oracle method) for 5000 simulated trajectories on each of 45 alternatives (squares in color); $N_{\max} = 100$ and $\alpha^* = 0.05$. Because $N_{\max}$ is small, the terminal power is generally low for gaps less than 20 percentage points. Moving from left to right: sequentializing Barnard's Test is inefficient due to a loss of structure; SAVI methods also suffer when p_0 and p_1 are closely competing, due to the method inherently generalizing to arbitrary $N_{\max}$. The Lai procedure and our STEP are similar to the SPRT oracle; however, note that Lai struggles more at the extremes (bottom left and top right). This inefficiency in the skewed regime becomes more pronounced as N grows and the gaps shrink.

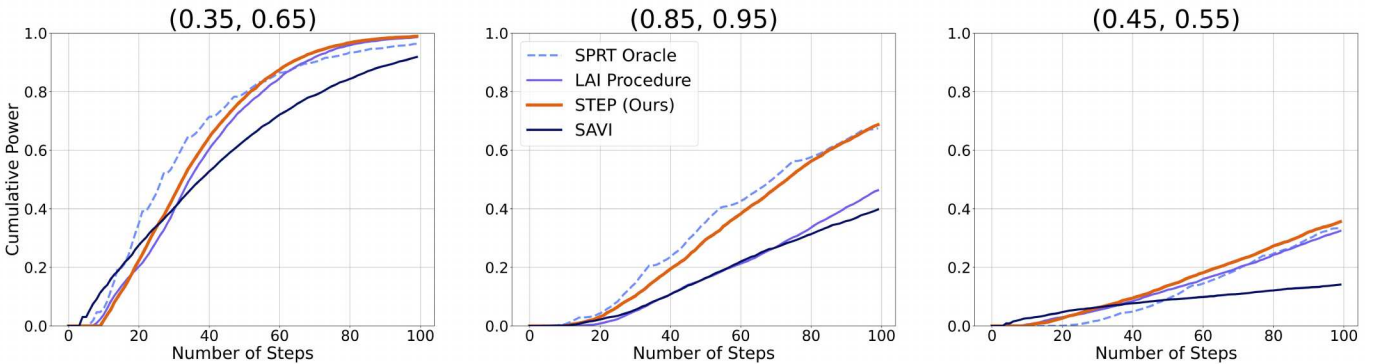


Fig. 8: Cumulative power of all feasible methods (Lai, SAVI, STEP (Ours)) and SPRT Oracle over 5000 trajectories in three evaluation settings of increasing difficulty; (p_0, p_1) for each setting title the respective figures. $N_{\max} = 100$ and $\alpha^* = 0.05$. The expected time-to-decision is the integral of the area *above* the cumulative power curve; therefore, curves higher and to the left are better. (Left) For a gap of 30 percentage points, all methods demonstrate similar stopping times. (Center) For a gap of 10 percentage points in the low-variance regime (i.e., farther from 0.5), STEP significantly outperforms the Lai and SAVI procedures. (Right) For a gap of 10 percentage points in the high-variance regime, STEP and Lai are similar but again SAVI struggles and underperforms the other methods.

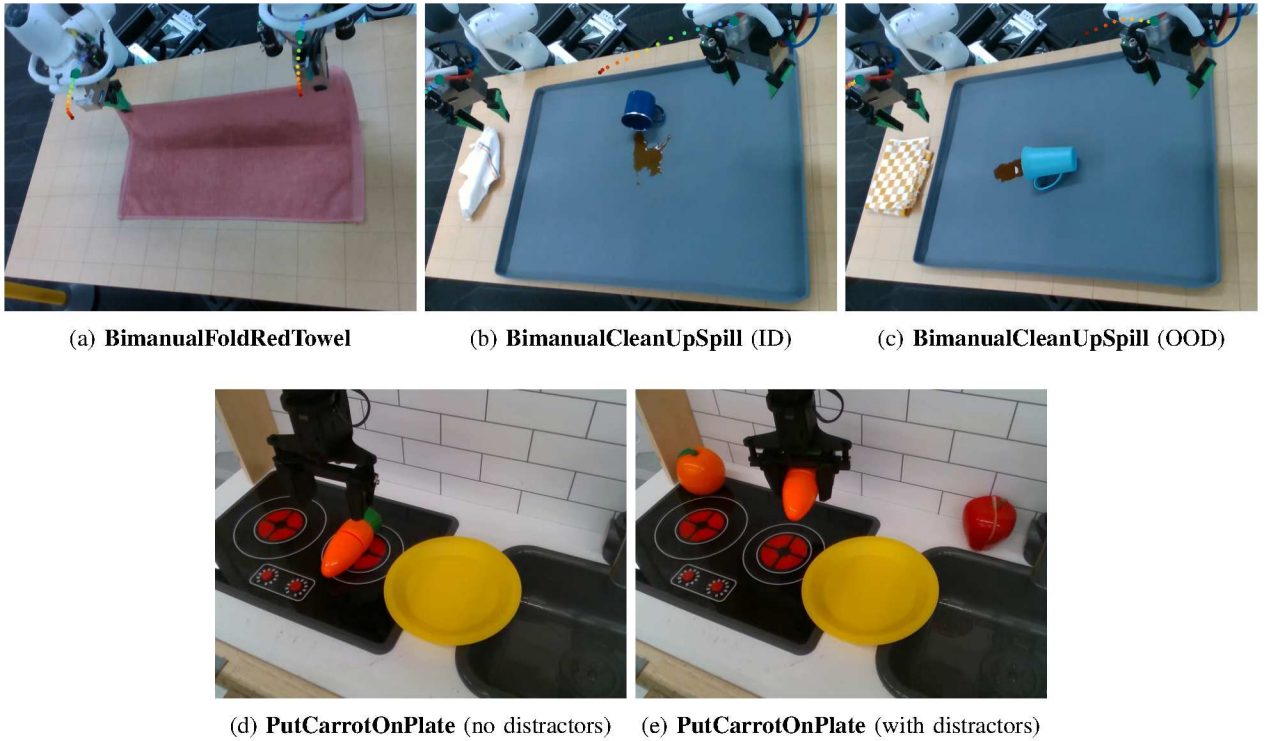


Fig. 9: Snapshots of robot policy evaluation tasks. (Top) Bimanual manipulation tasks with diffusion policy. Colored dots represent the camera projection of planned future end-effector positions. In **BimanualFoldRedTowel**, all the evaluations are done with in-distribution (ID) initial conditions and we compare two policy checkpoints from a single training run. In **BimanualCleanUpSpill**, we evaluate a single policy checkpoint in ID initial conditions with a white towel and out-of-distribution (OOD) initial conditions with a green towel to measure generalization performance. (Bottom) **PutCarrotOnPlate** task on the WidowX platform in a toy kitchen environment. The carrot is initially placed in one of three possible locations on the stove. The environment can either have no distractors or two distractors. We compare Octo and OpenVLA under the nominal environment distribution, and compare Octo performance in nominal environment distribution and under distribution shift. Detailed policy comparison metrics are given in Table I.

Case ($\downarrow$)	$N_{\max}$ ($\rightarrow$)	100	200	300	400	500
(0.59, 0.68)	SPRT Power ($\rightarrow$)	0.324	0.513	0.676	0.762	0.823
(0.68, 0.76)	SPRT Power ($\rightarrow$)	0.337	0.491	0.643	0.724	0.804

TABLE II: Empirical power of SPRT Oracle on distributions matching the empirical gaps observed in hardware trials of **CarrotOnPlate**. This suggests that at 200 trials per policy, there is only about a 50% chance of observing a sequence leading to rejection of the null; even for the oracle, 500 trials are required before this approximate probability reaches 80%.

sample size; as such, the evaluation requirements are primarily fundamental to the variance of Bernoulli random variables, and thus represent fundamental uncertainty and sample complexity for the policy comparison problem.

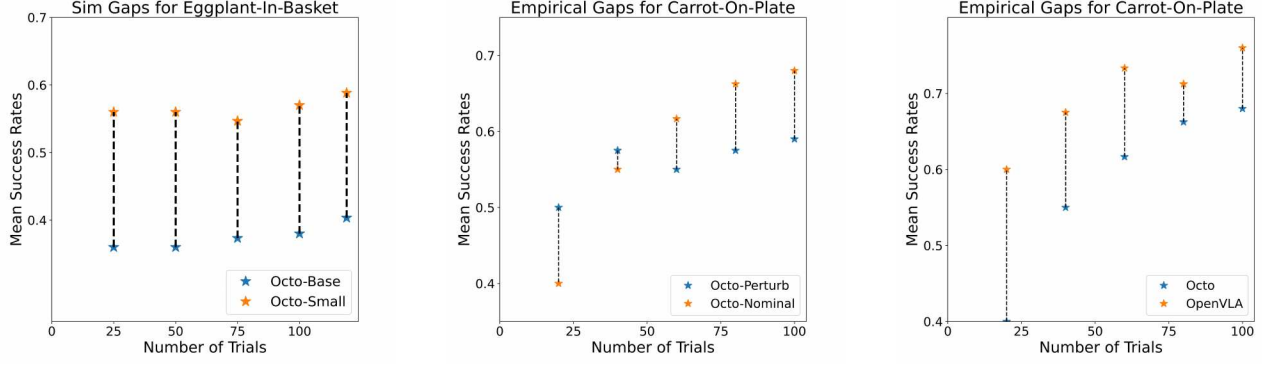
E. Additional **CarrotOnPlate** Experiments

A prior iteration of the **CarrotOnPlate** experiments (not reported in Section V) involved a hardware implementation error in which the end-effector rotation commands output by the policies were not correctly published. As a result, for the purpose of policy comparison, we label these policies **PolicyA**

(in place of Octo) and **PolicyB** (in place of OpenVLA). Note that the hardware implementation does not invalidate the policy comparison procedure itself. The environment distributions **Env1** and **Env2** are as described in Section V. We set $N_{\max} = 200$ as the evaluation budget for the following three settings: i) **PolicyA** (under **Env1**), ii) **PolicyB** (under **Env1**), and iii) **PolicyA** (under **Env2**). We compare **PolicyA** (**Env1**) with **PolicyB** (**Env1**), and **PolicyA** (**Env1**) with **PolicyA** (**Env2**). These results are listed in Table III. Observe that despite utilizing the full budget, no procedure yields conclusive results in the first comparison. However, in the second comparison, we ran the evaluation after collecting 150 trials per policy. Because every method terminated, we were able to stop early at 150 trials (i.e., saving ourselves 50 hardware evaluations).

F. Mathematical and Numerical Notes

1) *Worst-Case Null Hypotheses*: The worst-case null hypotheses are computed in this framework as the real number $p \in (0, 1)$ that maximizes the expected log-likelihood ratio. First, noting the monotonicity properties of the joint distribution, we claim that the worst-case null hypothesis must lie on the line $p_0 = p_1 = p \in (0, 1)$. Second, noting the optimal



(a) Octo-Base and Octo-Small in simulation (b) Octo-Base (**Env1**) and Octo-Base (**Env2**) (c) Octo-Base (**Env1**) and OpenVLA (**Env1**)
EggplantInBasket task in real-world **CarrotOnPlate** task in real-world **CarrotOnPlate** task

Fig. 10: Running empirical success rates of two policies as the number of trials increases. (a) In the **EggplantInBasket** task, there is a consistent gap in performance due to lower statistical uncertainty. This is reflected in Table I (row 8) where STEP terminates at $N = 119$. (b and c) On the other hand, in the **CarrotOnPlate** task, the relative performance consistently fluctuates and even sometimes flips due to high statistical uncertainty arising from the close competition between two policies. This leads to even SPRT oracle requiring more than 500 trials to confidently determine the relative performance (Table II).

N	$\hat{p}_0$	$\hat{p}_1$	SAVI	Lai	STEP	SPRT***
200	0.530	0.560	—	—	—	—
150	0.613	0.827	132	64	61	26

TABLE III: Additional **CarrotOnPlate** evaluations with $N_{\max} = 200$. Row 1: Comparing **PolicyA** under **Env1** (π_1) with **PolicyA** under **Env2** (π_0). Row 2: Comparing **PolicyA** under **Env1** (π_0) with **PolicyB** under **Env1** (π_1).

power properties of the SPRT for simple-vs-simple problems, we construct the log probability-ratio test maximization as:

$$\begin{aligned} & \arg \max_p \mathbb{E}_{x \sim (p, p)} \left[\left(\frac{p_1}{p} \right)^x \left(\frac{1-p_1}{1-p} \right)^{1-x} \left(\frac{p_0}{p} \right)^x \left(\frac{1-p_0}{p} \right)^{1-x} \right] \\ & \equiv \arg \max_p \mathbb{E} \left[x \log \frac{p_0 p_1}{p^2} + (1-x) \log \frac{(1-p_0)(1-p_1)}{(1-p)^2} \right] \end{aligned}$$

Differentiating, the solution is the interpolation in the natural parameter space of the Bernoulli distribution:

$$\begin{aligned} \log \frac{p^*}{1-p^*} &= \frac{\log \frac{p_0}{1-p_0} + \log \frac{p_1}{1-p_1}}{2} \\ \Rightarrow \eta^* &= \frac{\eta_0 + \eta_1}{2}; \end{aligned}$$

the reconstruction of p^* follows directly as

$$p^* = (1 + \exp \eta^*)^{-1}.$$

That is, the worst-case null in the sense of ‘falsely’ maximizing the probability ratio test under the null is precisely the interpolation in natural parameter space of (p_0, p_1) . In fact, the ‘true’ worst-case null is difficult to compute exactly; as we verify the Type-1 Error control against the additional methods of linear projection in the nominal parameter space

$$p^* = \frac{p_0 + p_1}{2}$$

and as the interpolation under the KL-divergence ‘pseudo-distance:’

$$p^* = \{p' \in (0, 1) \mid \text{KL}(p_0, p') = \text{KL}(p_1, p')\}.$$

In practice, assuming continuity corrections are applied to any case in which p_0 or p_1 belong to $\{0, 1\}$, these methods generally result in similar estimates of the worst-case null hypothesis, and form a small region in which the error control can be verified to greater numerical accuracy.

2) *Discretizing the Null Hypotheses:* In order to discretize the null hypotheses safely, it is necessary to ensure coverage over the set of possible worst-case nulls: $\{(p, p) : p \in [0, 1]\}$. First, we establish an interior bound $(\epsilon, 1-\epsilon)$ to the necessary values $p \in (0, 1)$. Specifically, for a fixed $N_{\max}$ one can derive a value of ϵ such that if $p \geq 1-\epsilon$ (or $p \leq \epsilon$), it holds w.p. $\geq 1-\alpha^*$ that $\hat{p}_{1,n} = 1$ (resp. 0) for all $n \in \{1, \dots, N_{\max}\}$. These extremal nulls pose no risk to the algorithm (because they cannot violate α^* Type-I error if we never **RejectNull** when $\hat{p}_1 \leq \hat{p}_0$). With this limitation in place we avoid problems arising from the rapid decay of the variance near 0 and 1 in the distribution set. Now, discretization in the range $(\epsilon, 1-\epsilon)$ can be undertaken to approximate all possible worst-case null hypotheses. In practice, we used approximately 100 points for $N_{\max}$ up to 500; this is significantly (3x) more than the default in the Scipy implementation of Barnard’s Test [54]. Note that formally, this discretization can be ensured to be safe numerically by using Pinsker’s inequality to relate the total variation distance (which upper bounds, for example, the event of a false rejection from a null hypothesis) to the KL-divergence; the implication of the inequality is that the false rejection rate error due to discretization is upper bounded by a monotonic function of the maximal KL divergence between any adjacent points in the discretization; for a sufficiently dense discretization, the error can be made arbitrarily small.

Task	Type	α^*	$N_{\max}$	N	$\hat{p}_0$	$\hat{p}_1$	SAVI	Lai	STEP (Ours)	SPRT***
FoldRedTowel	π_i	0.05	50	50	0.560	0.920	21.1 (0.63)	19.3 (0.49)	18.9 (0.44)	14.5 (0.53)
FoldRedTowel	π_i	0.05	200	50	0.560	0.920	21.1 (0.63)	28.6 (0.53)	24.4 (0.53)	14.5 (0.53)
FoldRedTowel	π_i	0.05	500	50	0.560	0.920	21.1 (0.63)	32.0 (0.56)	28.3 (0.56)	14.5 (0.53)
CleanUpSpill	$\mathcal{D}_{s_0, o_0}^i$	0.05	50	50	0.280	0.800	13.8 (0.44)	13.6 (0.33)	14.0 (0.30)	11.4 (0.41)
CleanUpSpill	$\mathcal{D}_{s_0, o_0}^i$	0.05	200	50	0.280	0.800	13.8 (0.44)	20.4 (0.38)	18.2 (0.38)	11.4 (0.41)
CleanUpSpill	$\mathcal{D}_{s_0, o_0}^i$	0.05	500	50	0.280	0.800	13.8 (0.44)	23.6 (0.44)	20.7 (0.43)	11.4 (0.41)
CarrotOnPlate	$\mathcal{D}_{s_0, o_0}^i$	0.05	100	100	0.590	0.680	–	–	–	–
CarrotOnPlate	π_i	0.05	100	100	0.680	0.760	–	–	–	–
SpoonOnTowel	π_i	0.01	500	500	0.084	0.386	44.6 (1.25)	59.8 (1.17)	49.6 (1.10)	34.5 (1.08)
EggplantInBasket	π_i	0.01	500	500	0.400	0.564	233.2 (6.7)	202.4 (5.0)	183.9 (4.9)	193.4 (5.3)
StackCube	π_i	0.01	500	500	0.000	0.030	328.2 (4.8)	409.8 (4.3)	265.5 (4.2)	70.5 (3.0)
Multitask	π_i	0.03	1500	1500	N/A	N/A	606.0	672.0	499.0	298.4

TABLE IV: Empirical **expected** time-to-correct-decision for all **hardware (top)** and **simulation (bottom)** policy comparisons. The contexts, parameters, and annotation are identical to those in Table I. Summary statistics are taken over 400 random trajectories generated according to a Bernoulli distribution with data-generating (i.e., ‘true’) parameters corresponding to the observed empirical success rates ($\hat{p}_0, \hat{p}_1$). We report the average stopping times of all methods on the right of the table for every context (standard deviation of the *empirical mean* in parentheses). In all cases: lower is better. In the Multitask setting, we test $p_1 > p_0$ *uniformly across the preceding three tasks*. This stopping time is the sum by column of the average stopping times for the three tasks. Our method saves the evaluator approximately 110 to 170 trials (in expectation) in uniform certification over these three tasks as compared to either feasible baseline. Note that for any single sequence of evaluations, the standard deviation for the stopping time on a given task can be approximately computed by multiplying the parenthetical standard deviation by 20 (the square root of the number of trials (400)). This correction confirms that the observed improvement of 160 evaluation trials saved in multitask simulation in Table I is consistent with the results observed here.

3) *Technical Setting*: The problem formulation in Section III was presented informally to avoid needless over-technical confusion. Slightly more precisely, we assume necessary measurability conditions on the random variables representing the success or failure of the policy in the environment. Given the probability space implicit in this assumption, the concatenation of observations constitutes the natural filtration on this space; that is, $\mathcal{F}_n = \{n, Z_1, Z_2, \dots, Z_n\}$. In practice, knowledge of the sufficient statistic for exponential families induces us to use the compressed filtration $\mathcal{F}_n^{[comp]} = \{n, \sum_{i=1}^n z_{0,i}, \sum_{i=1}^n z_{1,i}\}$. Interestingly, using the Neyman-Pearson lemma, one can show that the two-dimensional state represents a lossy compression as a three-dimensional state is needed to construct the optimal exact SPRT.

4) *Intuition for Tests*: We quickly summarize a few examples of extremal test procedures that can help provide scaffolding for the reader in terms of understanding the tradeoffs inherent between Type-I Error, Type-II Error, and expected sample size. First and foremost, safety is always possible in the Type-I sense: simply never reject the null (i.e., without looking at any data). Slightly more subtly, safety and small sample size is always feasible, as shown in Footnote 5 in Section III: decide without looking at any data, but first generate an independent random number uniformly on $[0, 1]$ and reject if the number is less than α^* , otherwise fail to reject. Power and small sample sizes can be obtained accordingly at the cost of violating Type-I Error (just reject instead of failing to reject). Power-1 tests finish out the last leg of the triangle – waiting an arbitrarily long time can allow for simultaneous

control of Type-I and Type-II error (the N-P Lemma only concludes that in the batch setting – where N is fixed and finite – there exist instances for which Type-I and Type-II Error cannot be simultaneously controlled).