How Generalizable Is My Behavior Cloning Policy? A Statistical Approach to Trustworthy Performance Evaluation

Joseph A. Vincent[1*], Haruki Nishimura[2], Masha Itkina[2], Paarth Shah[2], Mac Schwager[1], Thomas Kollar[2]

***Abstract*— With the rise of stochastic generative models in robot policy learning, end-to-end visuomotor policies are increasingly successful at solving complex tasks by learning from human demonstrations. Nevertheless, since real-world evaluation costs afford users only a small number of policy rollouts, it remains a challenge to accurately gauge the performance of such policies. This is exacerbated by distribution shifts causing unpredictable changes in performance during deployment. To rigorously evaluate behavior cloning policies, we present a framework that provides a tight lower-bound on robot performance in an arbitrary environment, using a minimal number of experimental policy rollouts. Notably, by applying the standard stochastic ordering to robot performance distributions, we provide a worst-case bound on the *entire distribution* of performance (via bounds on the cumulative distribution function) for a given task. We build upon established statistical results to ensure that the bounds hold with a user-specified confidence level and tightness, and are constructed from as few policy rollouts as possible. In experiments we evaluate policies for visuomotor manipulation in both simulation and hardware. Specifically, we (i) empirically validate the guarantees of the bounds in simulated manipulation settings, (ii) find the degree to which a learned policy deployed on hardware generalizes to new real-world environments, and (iii) rigorously compare two policies tested in out-of-distribution settings. Our experimental data, code, and implementation of confidence bounds are open-source.** [1]

I. Introduction

In this paper we focus on evaluating robot policies obtained through the *behavior cloning* (BC) framework of robot learning. BC policies, such as diffusion policies [1], have recently advanced the state-of-the-art (SOTA) in visuomotor policy synthesis, particularly in manipulation tasks [1], [2], [3], [4], [5]. BC is attractive because it avoids the significant challenges of the sim-to-real gap, which impede the transfer of reinforcement learning policies and other policy synthesis techniques to real-world robots [6]. In BC, humans give demonstrations, often directly on the robot hardware through teleoperation. Then, a policy is learned based on these demonstrations. This procedure removes the need for a simulation model, avoiding the sim-to-real gap.

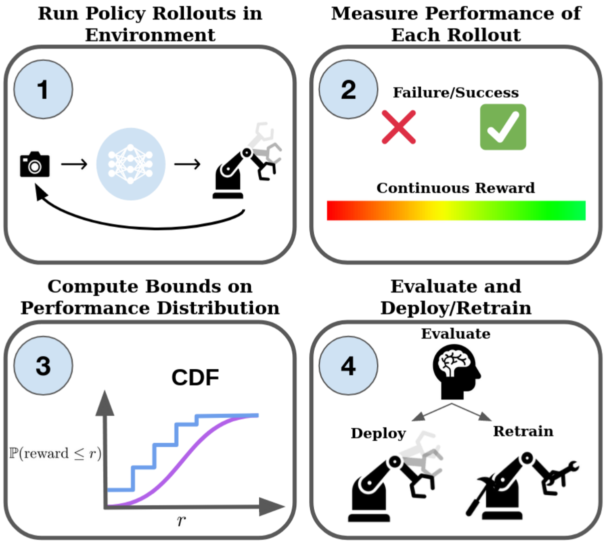
Run Policy Rollouts in Environment
1
Measure Performance of Each Rollout
2
Failure/Success
Continuous Reward
Compute Bounds on Performance Distribution
3
CDF
$\mathbb{P}(\text{reward} \leq r)$
r
Evaluate and Deploy/Retrain
4
Evaluate
Deploy
Retrain

Fig. 1: Our approach to evaluating BC policies. First, policy rollouts are collected in the environment of interest. Second, each rollout results in either a binary or continuous performance measurement. Third, statistical tools are used to compute an upper confidence bound on the CDF of performance. Finally, the user interprets the confidence bound and chooses to deploy or retrain the policy.

However, without an accurate simulation model, evaluation of BC policies relies on real-world tests. In robotics research it is common to evaluate these policies using fewer than 50 policy rollouts, recording the empirical success rate or average reward (see, e.g., [1], [2], [7]). However, with such small sample sizes, it can be difficult to interpret the significance of the recorded results. In addition, measuring average performance can be insufficient for applications with safety and reliability requirements.

To address this need, we propose statistical bounds to rigorously evaluate the performance of a BC policy. We quantify performance through a user-specified metric, either a binary success/failure or a continuous reward. Although many BC policies are trained and deployed in hardware, these metrics can also be given for simulated robotics settings. While specifying a suitable reward for policy training can be challenging in practice, we emphasize that we do not use the performance metric for training, only for evaluation. Simple performance metrics (e.g., task success/failure, distance to a goal location) are sufficient for our purposes.

Our proposed method is to compute worst-case bounds on the performance of policy based on the results of a small number of policy rollouts. Worst-case bounds are useful for determining a policy exceeds some performance threshold rather than determining it does not. Our approach can specify the fewest number of policy rollouts required to obtain a user-specified *confidence* and *tightness* for a bound on the

[*] Work primarily done during internship at Toyota Research Institute.

The NASA University Leadership initiative (grant #80NSSC20M0163) provided funds to assist the authors with their research, but this article solely reflects the opinions and conclusions of its authors and not any NASA entity. The first author was partially supported by the Dwight D. Eisenhower Transportation Fellowship.

This research was partially supported by ONR grant N00014-23-1-2354.

[1] Department of Aeronautics and Astronautics, Stanford University, Stanford, CA 94305, USA, {josephav, schwager}@stanford.edu

[2] Toyota Research Institute, Los Altos, CA 94022, USA, {haruki.nishimura, masha.itkina, paarth.shah, thomas.kollar}@tri.global

[1] https://github.com/TRI-ML/stochastic_verification
https://github.com/TRI-ML/binomial_CIs

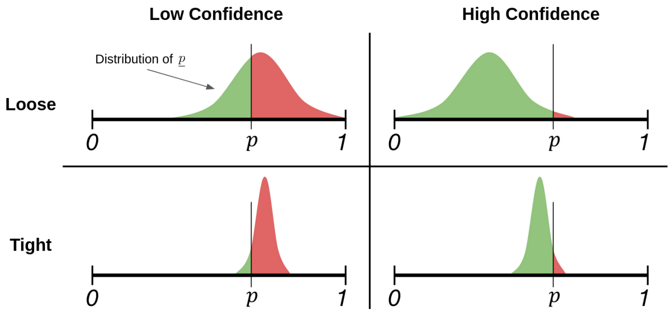

Fig. 2: Hypothetical distributions of a lower confidence bound $\underline{p}$ for an unknown probability of success p. High confidence levels give better chances that the $\underline{p}$ we obtain is lower than p (green shaded region), and tighter bounds give better chances that $\underline{p}$ is close to p.

performance metric. *Confidence* is the probability (fraction of random outcomes) for which the bound holds.[2] *Tightness* quantifies how close the bound tends to be to the true (but unknown) quantity of interest. A bound which is tighter provides a better estimate of the unknown quantity while a bound which is higher confidence provides a better chance that the estimate is conservative (rather than optimistic). In Fig. 2 we give a visual interpretation of how confidence and tightness relate to the distribution of a lower confidence bound for an unknown probability of success. Achieving higher confidence and tighter bounds on policy performance comes at the cost of more policy rollouts.

The bounds we use require no knowledge of the underlying probability distributions in the policy, training data, robot hardware, or environment physics. The confidence and tightness of these bounds do not depend on the dimensionality of the robot state or observation. We only require that the stochasticity of the policy, robot, and environment is stationary between rollouts so that the measured performance of the policy rollouts can be treated as independent and identically distributed (i.i.d.) random variables. This assumption is frequently implicit during BC training, and, depending on the task, can be effectively managed during evaluation.

In this paper we demonstrate how to construct worst-case bounds on performance when performance is measured by (i) a binary success/failure metric, and (ii) a continuous metric which measures the reward accumulated along a trajectory. For the binary metric, we seek a lower bound on the unknown probability of task success. We revisit a rarely-used bound from the statistics literature, which is a tighter version of the more common Clopper-Pearson bound [8]. Our bound exactly achieves a user-specified confidence level and is as tight as possible. This is enabled by a novel computational method based on mixed monotonic optimization [9]. For the continuous metric, the true distribution of reward is entirely unknown. We propose to use bounds for the cumulative distribution function (CDF) of reward. The CDF fully characterizes the reward distribution, allowing flexibility for the user to analyze performance, e.g., in the tails, at any quantile, or in the mean. We use a tighter version of the Dvoretzky–Kiefer–Wolfowitz (DKW) bound [10], [11] for this purpose. In both the binary and continuous cases, the user specifies the confidence level and the tightness of the bound. Since real-world evaluation is costly, the bounds we employ meet these specifications exactly using the minimum number of policy rollouts. This is in contrast to the more common Clopper-Pearson and DKW bounds.

Finally, in our experiments we are primarily interested in applying our framework to investigate the generalization of BC policies to out-of-distribution (OOD) settings. Because human intuition is often unreliable for predicting which OOD settings will be harmful to the policy, we believe this is an application where statistical evaluation methods shine, giving roboticists well-defined confidence levels for the results of their experiments. We demonstrate our bounds in simulation and real-world robotic manipulation settings. We empirically compare our bounds to more common bounds from the literature, showing improved tightness, particularly for few policy rollouts. We also show how our bounds can be used to compare the performance of two policies, and to help decide which policy is superior. Finally, in hardware, we bound the performance of a SOTA visuomotor diffusion policy [1] in generalizing to new environments. An early version of these results appeared in a workshop [12]. Here we substantially expand the analysis and simulation results, add hardware experiments, and add an example use-case for comparing policies. We summarize our contributions as follows:

- A method for bounding the performance of a BC robot policy in an arbitrary environment with minimal policy rollouts, by putting in practice the concept of worst-case bounds on the entire distribution of performance.
- A novel method for computing tightness for binomial confidence intervals, specifically *maximum expected shortage* (the worst-case value for the average degree of underestimation of the true success rate).
- An open-source implementation for computing the binomial bounds described in the paper, which optimally trade-off between confidence, tightness, and sample size: `https://github.com/TRI-ML/binomial_CIs`.

II. Related Work

In this section, we discuss related literature on rigorously evaluating learned robot policies. We first discuss approaches from formal methods, then discuss statistical approaches.

A. Formal Evaluation of Learned Policies

Formal methods applied to robots with learned components can be used to evaluate open-loop policy requirements (such as state-dependent action constraints [13]) and closed-loop system requirements (such as reach-avoid constraints [14], [15], [16], [17]). These methods typically have strict limitations on the structure and complexity of the robot policy and environment model. In addition, these methods are most commonly employed in settings with deterministic environments or with bounded (set-based) uncertainties. In settings where these requirements are met, formal methods can evaluate rich closed-loop behaviors by finding certificates such as invariant sets [15], [16], Lyapunov functions [18], [19], and contraction metrics [20], [21]. However, in this paper we do not assume

[2] Sometimes also referred to as the *coverage* of the bound.

access to an environment model or any particular policy structure, precluding the use of formal evaluation methods. When systems become too complex for formal evaluation, black-box validation methods are often used [22]. Statistical testing, discussed next, is one flavor of black-box validation that provides probabilistic guarantees.

B. Statistical Evaluation of Learned Policies

When constructing statistical bounds, there are trade-offs between confidence, tightness, and sample size. Related work tends to focus on the trade-off between confidence and number of samples, neglecting the important role that tightness plays in interpreting the bound. Placing statistical bounds on the performance of a learning-enabled robot from few policy rollouts is explored in [23], [24], [25], [26]. However, these works only bound particular scalar quantities such as the expected value, value-at-risk, and conditional value-at-risk of the performance distribution. We place bounds on the CDF that consider all aspects of the performance distribution. For example, policies with appealing expected performance may have unacceptable long tails of poor performance. Further, these papers do not quantify bound tightness and their bounds can require more policy rollouts than necessary to achieve the specified confidence.

Conformal prediction has been used for rigorous uncertainty quantification of robot perception components [27], [28], [29], [30] and policies [31], [32]. This research has enabled policies to attain a specified probability of task success, such as avoiding collisions or making language-based plans. However, like the other approaches in this section, (i) by bounding a quantile of the score distribution, conformal prediction bounds also ignore tails of distributions, (ii) tightness of bounds from conformal prediction is often not quantified, and (iii) conformal prediction bounds may require more samples than necessary to achieve the desired confidence level [33], [34]. In this paper we bound the entire distribution of performance, quantify bound tightness, and ensure that the number of samples used is minimal.

Confidence bounds for unknown success probabilities and CDFs are well studied with the Wald and Clopper-Pearson intervals being the most common for success probabilities [35], [8] and with the DKW inequality being the most common for CDFs [10], [11]. However, these bounds are not optimally tight and thus may require more policy rollouts than necessary. In Section IV and Section V, we discuss less common yet well-established confidence bounds which optimally trade-off between confidence, tightness, and number of samples. Lastly, in Section VI-C we compare the success rates of two policies using the bounds in this paper. Common approaches which test for differences in success rates include Fisher's and Barnard's tests [36]. However, while such tests determine which policy performs better, they do not return confidence intervals for the individual success rates (as we do), but rather provide confidence intervals for the odds ratio. Our view is that in practice it is not enough to know which policy performs better, but it is also important to know the performance of each policy individually.

III. Distributional Bounds

We want to obtain worst-case bounds on the *entire distribution* of performance (via bounds on the CDF), rather than conventional quantities such as expected value. This approach affords a rich understanding of performance, hedging against deploying a policy with some hidden drawbacks (e.g., high expected reward but a long tail of low reward).

Consider two CDFs $F_1(x) = \mathbb{P}(X_1 \leq x)$ and $F_2(x) = \mathbb{P}(X_2 \leq x)$, describing the distribution of performance under two different policies. If $F_2(x) \geq F_1(x)\ \forall x$, then $F_2(x)$ has more probability mass on *lower* performance values. Therefore $F_1(x)$ is preferable to $F_2(x)$, following the standard notion of stochastic ordering [37]. If the inequality holds for some, but not all, values of x, we do not claim one distribution is preferable to the other. Therefore we have a partial ordering of performance based on CDFs.

An upper confidence bound $\overline{F}(x)$ on a CDF with confidence level $1-\alpha$ satisfies

$$\mathbb{P}(F(x) \leq \overline{F}(x)\ \forall x) \geq 1-\alpha. \tag{1}$$

$\overline{F}(x)$ is constructed based on the observed performance of some number of policy rollouts. Since the outcome of policy rollouts is random, $\overline{F}(x)$ is also random, leading to the bound holding probabilistically as in Eq. (1). In light of the above discussion on partial ordering of CDFs, $\overline{F}(x)$ is a *worst-case* bound on the distribution of performance with confidence $1-\alpha$. Next, we give worst-case bounds for the both binary and continuous performance metrics. In the case of a binary performance metric, an upper bound on the CDF is simply a lower bound on the probability of success.

IV. Confidence Bounds - Binary Metric

We are interested in finding lower confidence bounds on the probability of success that require a minimal number of policy rollouts while achieving user-specified levels of confidence and tightness. To determine such bounds we use the approach in [37] to achieve a desired confidence level with minimal rollouts. We then introduce a computational method to achieve a desired tightness with minimal rollouts.

To find lower bounds on a policy's success rate in a new environment, we treat the result of each policy rollout as a Bernoulli random variable X, where $X=1$ denotes a success and $X=0$ denotes a failure. Then given the results of n i.i.d. policy rollouts, we construct a lower confidence bound $\underline{p}$. The number of observed successes follows a binomial distribution. Before describing the bound, we first introduce general optimality criteria for lower confidence bounds of an unknown parameter θ.

A. Optimality Criteria

The standard notion of optimality for lower confidence bounds is that of being *uniformly most accurate* (UMA).

Definition 1 (Uniformly Most Accurate, Eq. 3.22 of [37]): For test statistic T, a lower confidence bound $\underline{\theta}(T)$ satisfying

$$\mathbb{P}_\theta[\underline{\theta}(T) \leq \theta] \geq 1-\alpha \quad \forall\theta \tag{2}$$

and amongst all possible lower bounds $\underline{\theta}'(T)$ minimizes

$$\mathbb{P}_\theta[\underline{\theta}'(T) \leq \theta_0] \quad \forall \theta_0 < \theta \tag{3}$$

is a UMA lower confidence bound at confidence level $1-\alpha$. Intuitively, a UMA lower confidence bound underestimates the unknown parameter θ by as little as possible, no matter what the actual value of θ is. *Shortage* (i.e. excess width) is used to quantify the amount of underestimation,

$$\text{shortage} = \max\{\theta - \underline{\theta}, 0\}. \tag{4}$$

Using the Ghosh-Pratt identity [38], [39],

$$\text{ES}(\theta) = \mathbb{E}_\theta[\text{shortage}] = \int_{\theta_0 < \theta} \mathbb{P}_\theta[\underline{\theta} \leq \theta_0] d\theta_0. \tag{5}$$

Now, since we know that a UMA lower confidence bound minimizes the integrand of Eq. (5), we can conclude that a UMA lower confidence bound also minimizes expected shortage for all values of θ. Although expected shortage is a useful quantity, it depends on the value of θ, which is unknown. To avoid any assumptions on θ, we measure tightness according to *maximum expected shortage* (*MES*),

$$\text{MES} = \max_\theta \; \mathbb{E}_\theta[\text{shortage}]. \tag{6}$$

MES is the worst-case expected shortage over all possible parameter values θ. The notion of MES is not new [40], but it is not commonly used since the maximization in Eq. (6) can be challenging when expected shortage is not concave in θ. Next, we describe a UMA lower confidence bound for an unknown success probability and introduce the first tractable method for computing the associated MES.

B. Optimal Binomial Bounds

In this paper we let $\text{bin}(k, n, p)$ and $\text{Bin}(k, n, p)$ denote the binomial probability mass function (PMF) and CDF, with k successes, n trials, and success probability p. The floor function $\lfloor x \rfloor$ rounds the argument x down to the nearest integer. Now, we define a test statistic $T = U + \sum_{i=1}^n X_i$ where $U \sim \mathcal{U}[0,1]$ and X_i are the observed successes and failures. Then T is random with probability density

$$f_p(t) = \binom{n}{\lfloor t \rfloor} p^{\lfloor t \rfloor}(1-p)^{n-\lfloor t \rfloor}. \tag{7}$$

Then, we can compute the CDF as follows,

$$F_p(t) = \int_0^t \binom{n}{\lfloor x \rfloor} p^{\lfloor x \rfloor}(1-p)^{n-\lfloor x \rfloor} dx \tag{8a}$$

$$= \text{Bin}(\lfloor t \rfloor - 1, n, p) + (t - \lfloor t \rfloor) \cdot \text{bin}(\lfloor t \rfloor, n, p). \tag{8b}$$

$F_p(t)$ is continuous in both t and p. Then, by Corollary 3.5.1 of [37], we have the UMA lower confidence bound,

$$\mathbb{P}[\underline{p} \leq p] \geq 1 - \alpha \tag{9a}$$

$$\underline{p}(t) = \begin{cases} 0 & \text{if } t < 1-\alpha \\ 1 & \text{if } t > n+1-\alpha \\ p^* & \text{s.t. } F_{p^*}(t) = 1-\alpha \text{ otherwise.} \end{cases} \tag{9b}$$

The bisection method can be used to solve $F_{p^*}(t) = 1-\alpha$, as $F_p(t)$ is monotonically decreasing in p. Taking limits $p \to 0^+$ and $p \to 1^-$, one can show that $F_p(t) = 1-\alpha$ has no solution when $t < 1-\alpha$ and $t > n+1-\alpha$, we thus set $\underline{p}$ to either zero or one in these cases. This bound is well-established (see Example 3.5.2 of [37]). The Clopper-Pearson bound can be obtained by setting $U = 0$ in the test statistic.

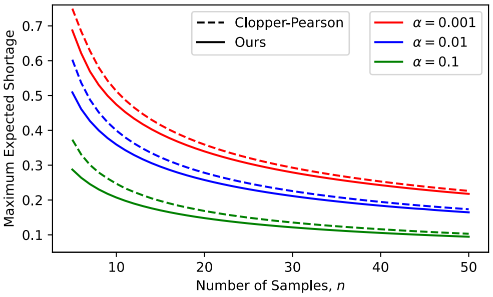

Fig. 3: Variation of MES, n, and α. Our method is always tighter than Clopper-Pearson, and appreciably so at small sample sizes.

Next, we compute expected shortage. A useful identity is

$$\mathbb{P}[\underline{p} \leq p_0] = \mathbb{P}[t \leq t^*(p_0)] \tag{10}$$

where $t^*(p_0)$ is the unique value that satisfies

$$F_{p_0}(t^*) = 1 - \alpha. \tag{11}$$

Then, we can compute expected shortage as

$$\text{ES}(p) = \int_0^p \mathbb{P}[\underline{p} \leq p_0] dp_0 = \int_0^p \mathbb{P}[t \leq t^*(p_0)] dp_0 \tag{12a}$$

$$= \int_0^p F_p(t^*(p_0)) dp_0. \tag{12b}$$

This integral can be evaluated with standard software.

Next, we find MES via global optimization. Specifically, consider the following reformulation of expected shortage,

$$\text{ES}(p_1, p_2) = \int_0^{p_1} F_{p_2}(t^*(p_0)) dp_0. \tag{13}$$

In this form, expected shortage is increasing in p_1 and decreasing in p_2; this property is known as *mixed monotonicity*. Note that for $p_1 = p_2$ we recover the original function. Mixed monotonic functions can be globally optimized using the branch and bound procedure given in [9]. This approach is the first tractable method that certifiably solves the optimization needed to compute MES. We provide an open-source implementation of this UMA bound at

https://github.com/TRI-ML/binomial_CIs.

Now, given two of the following: (i) confidence level $(1-\alpha)$, (ii) tightness (MES), (iii) number of policy rollouts (n), we can determine the third. In Fig. 3 we show how these quantities vary with one another for our bound and the Clopper-Pearson bound. Since the bound we use is UMA, it always has lower MES than Clopper-Pearson. MES computation for Clopper-Pearson is also novel. We derive the appropriate form of Eq. (12) and compute the corresponding MES using mixed monotonic programming. Lastly, one can compute a lower confidence bound on the *failure* probability using the same procedure, but constructing the test statistic T with failure counts rather than successes.

V. Confidence Bounds - Continuous Metric

When performance is measured with a continuous metric, the policy rollouts result in i.i.d. samples from some unknown distribution of reward. This distribution may be continuous, discrete, or mixed. In this section we describe how to obtain an upper confidence bound on the CDF of this distribution. As in Section IV, the CDF upper bound is constructed in such a way that given user-specified confidence level and tightness, the minimum number of policy rollouts is used to meet these specifications. The results used in this section are not novel (see [41]), but are uncommon in the statistics literature, especially as applied to robotics.

Consider an i.i.d. sample, $X_{1:n}$ from an unknown distribution with CDF $F(x)$. The empirical CDF is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \le x), \tag{14}$$

where $\mathbf{1}(x)$ evaluates to 1 if x is true and 0 otherwise. A one-sided Kolmogorov-Smirnov (KS) statistic is

$$D_n^- = \sup_x \; F(x) - F_n(x). \tag{15}$$

If $F(x)$ is continuous, then the distribution of D_n^- is

$$\mathbb{P}(D_n^- \le \epsilon) = 1 - \epsilon \sum_{k=0}^{\lfloor n(1-\epsilon) \rfloor} w_k \tag{16a}$$

$$w_k = \binom{n}{k} (1 - \epsilon - \frac{k}{n})^{n-k} (\epsilon + \frac{k}{n})^{k-1}. \tag{16b}$$

Note that there is no dependence on the true $F(x)$. Moreover, if $F(x)$ is not continuous, then Eq. (16a) holds with $\ge$ rather than equality. Next, note the equivalence

$$\sup_x \; F(x) - F_n(x) \le \epsilon \Leftrightarrow F(x) \le F_n(x) + \epsilon \; \forall x. \tag{17}$$

Then, given i.i.d. samples $X_{1:n}$ from some $F(x)$, we have the following upper confidence bound on the CDF,

$$\boxed{\begin{aligned} \mathbb{P}(F(x) \le \overline{F}(x) \; \forall x) &\ge 1 - \alpha & (18a) \\ \overline{F}(x) &= F_n(x) + \epsilon^* & (18b) \\ \text{with } \epsilon^* \text{ chosen s.t. } \mathbb{P}(D_n^- \le \epsilon^*) &= 1 - \alpha. & (18c) \end{aligned}}$$

The tightness of this bound is measured by ϵ^*, the offset from the empirical CDF. The ϵ^* chosen by Eq. (18c) is optimal, if made any smaller then Eq. (18a) would not hold. Practical computation of ϵ^* given n and α is easily done via a bisection search. Given α and ϵ^*, it is then simple to compute the minimum number of policy rollouts n needed to meet these specifications. One can also use the DKW inequality [10], [11] to obtain a similar bound, which takes the same form as Eq. (18b) except with $\epsilon^* = \sqrt{-\ln \alpha / (2n)}$. However, the DKW bound is always less tight as shown in Fig. 4.

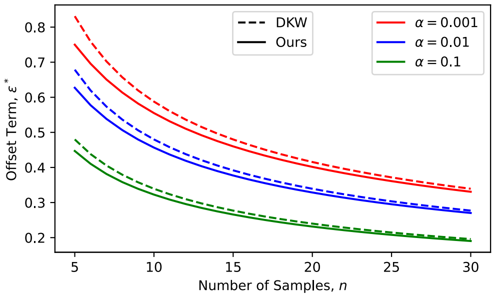

Fig. 4: Variation of ϵ^*, n, and α. Our method is always tighter than the DKW bound, and appreciably so at small sample sizes.

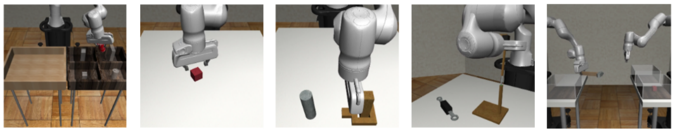

Fig. 5: Simulation tasks: can, lift, square, tool hang, transport [42].

VI. Experiments

We demonstrate our approach to evaluating BC policies in both simulation and real-world robotic manipulation settings. The performance of BC policies in OOD settings is hard to predict since it is difficult to characterize what aspects of the policy are invariant or sensitive to particular distribution shifts. This motivates the need for rigorous statistical evaluation. Thus, all our experiments evaluate policies in OOD settings. In simulation, we run many evaluations and show the empirical confidence and tightness of the bounds agree with the theory. In hardware, we (i) investigate the degree to which a learned policy generalizes to OOD settings, and (ii) compare the OOD generalization of two learned policies. Our approach requires i.i.d. performance scores; a sufficient (but not necessary) condition to ensure this is for the initial observations to be i.i.d. from some (potentially unknown) distribution. In simulation, this can be trivially satisfied. In hardware, we took efforts to minimize the impact of any time-varying conditions that may undermine this assumption. Because our bounds are strictly tighter than those of Clopper-Pearson and DKW, we report the results for our bounds only.

A. Simulation Results

We consider five diffusion policies from [1] for visuomotor robot manipulation that have shown recent SOTA performance, each trained for a separate task. Fig. 5 shows snapshots of each task in the training environments.

After training, we modified each environment as follows,

- **Can**: Changed the color of the can to lime green, and the color of the floor to beige.
- **Lift**: Changed the cube color to blue.
- **Square**: Changed the square color to tan.
- **Tool Hang**: Changed the wall color to floral white.
- **Transport**: Changed the color of the cube to lime green and the color of the lid handle to silver.

In Fig. 6 we show our simulation results. To obtain accurate estimates of the true in-distribution (ID) (orange) and OOD

	Confidence			**Expected Shortage**	
	Theory	Bin.	Cont.	Theory	Bin.
can	0.95	0.97	0.95	0.078	0.079
lift	0.95	0.94	0.97	0.115	0.114
square	0.95	0.90	0.91	0.085	0.077
tool hang	0.95	0.97	0.96	0.127	0.122
transport	0.95	0.96	0.96	0.115	0.105

TABLE I: Simulation results of the data in Fig. 6. *Theory* columns denote the theoretical values for confidence and expected shortage, *Bin.* columns denote the empirical estimates of these quantities for the binary metric bounds, the *Cont.* column denotes the empirical confidence level for the continuous metric bounds. Tightness of the continuous bounds is given by ϵ^*, which is not random and thus not listed. In general, we see good agreement between the theoretical values and their empirical estimates (100 trials). With more trials we expect these estimates to converge to their theoretical values.

(purple) performance distributions, we used 1000 policy rollouts from each environment. We use 4000 rollouts to construct confidence bounds (100 bounds generated from 40 samples each). We also show the Monte Carlo estimates of the OOD performance distribution (gray) one would obtain using 40 samples. We construct 100 bounds to empirically validate the guarantees of the bounds, but in practice one would only construct a single bound. The true ID and OOD performance distributions are typically unknown and are shown only for validating the bounds.

From the figure we see that the distribution shifts for the Can and Square environments were quite harmful to the performance of the policy, while those for the other tasks were benign. Our bounds tend to accurately reflect these trends. These results reflect the unintuitive nature about when learned policies generalize. For instance, it is not clear why changing the color of the manipulated object is harmful in some cases (Can and Square) but benign in other cases (Lift and Transport). Lastly, the Monte Carlo estimates of the reward distributions are optimistic about half the time, making them unfit to be interpreted as bounds. Table I shows good agreement between the empirical confidence and tightness estimates with their theoretical values.

B. Generalization of a Single Policy

Here we investigate the degree to which a learned policy generalizes to OOD environments. We test a diffusion policy [1] trained to pour ice from a red cup into a sink. During training there are no other objects on the tabletop. We test the policy in two OOD settings shown in Fig. 7. The first setting is hypothesized to be a harmful distribution shift (another red object in frame) and the second is hypothesized to be a benign distribution shift (no other red objects).

We performed 50 policy rollouts in each setting and found $4/50$ successes in the harmful setting and $38/50$ successes in the benign setting. Applying Eq. (9b) at a 95% confidence level we find $\underline{p}_{\text{harmful}} = 0.035$, and $\underline{p}_{\text{benign}} = 0.641$ with MES $= 0.118$. From this, we expect $\underline{p}_{\text{harmful}}$ and $\underline{p}_{\text{benign}}$ to be reasonably close to the true success rates. The low lower-bound for the harmful setting is attributed to frequent grasp failure despite its tendency to move towards the red mug.

The bounds we find support our hypotheses, and importantly also inform us as to the degree of performance we can assert at a 95% confidence level. If a user requires a 60% success rate, then we could assert with 95% confidence that the benign setting is suitable for deployment whereas we could not for the harmful setting. We may then choose to retrain the policy with data from this setting and re-evaluate.

C. Comparing the Generalization of Two Policies

In this experiment we compare the performance of two learned visuomotor manipulation policies. Specifically, we apply our statistical bounds to the recent results from [43], where the authors compare their RT-2 policy to a VC-1 policy in three settings designed to test emergent capabilities in symbol understanding, reasoning, and human recognition. We compute a lower confidence bound on the success rate of the RT-2 policy ($\underline{p}_{\text{RT-2}}$) and an upper confidence bound on the success rate of the VC-1 policy ($\overline{p}_{\text{VC-1}}$). Note, $\overline{p}_{\text{VC-1}} = 1 - \underline{q}_{\text{VC-1}}$, where $\underline{q}_{\text{VC-1}}$ is a lower confidence bound on the failure rate. We construct the bounds with confidence level 0.975, to ensure a joint confidence level of 0.95,

$$\mathbb{P}\Big[p_{\text{RT-2}} \in [\underline{p}_{\text{RT-2}}, 1] \ \cap \ p_{\text{VC-1}} \in [0, \overline{p}_{\text{VC-1}}]\Big] \geq 0.95. \tag{19}$$

Then, if $\overline{p}_{\text{VC-1}} < \underline{p}_{\text{RT-2}}$ (disjoint bounds), we conclude that the RT-2 policy outperforms the VC-1 policy. Furthermore, the chances we come to this conclusion incorrectly are $\leq 5\%$.

In Fig. 8, we show the bounds for each policy in each evaluation setting. We find in each case there is enough evidence to conclude (at a 95% confidence level) that the RT-2 policy outperforms the VC-1 policy. For some settings the gap between the bounds is larger, this may be due to the actual success rates being further apart, or an effect of sample size. With separate bounds for each policy, we can also conclude how well each policy performs individually, an important aspect that Fisher's exact test does not provide [36]. Finally, this approach easily extends to the case of continuous rewards. A lower bound on the CDF is computed similarly to an upper bound by noting that $D_n^+ = \sup_x \ F_n(x) - F(x)$ has the same distribution as D_n^- (Eq. (16)) [41].

VII. Conclusion

We build upon classical statistical techniques to rigorously evaluate BC policies. We use worst-case confidence bounds to measure policies' OOD generalization of to unseen environments. For binary and continuous performance metrics, we place confidence bounds on the entire distribution (i.e. CDF) of a policy's performance. These bounds hold with a user-specified confidence level, tightness, and are constructed from as few policy rollouts as necessary. For binary performance metrics, we introduce a solution for tractably evaluating the tightness via mixed-monotonic programming. Finally, in three experiments we show both the validity of our approach as well as how the approach can be used to make judgements about policy generalization in OOD settings.

In this paper, the number of policy rollouts is determined by a budget or by constraints on confidence and tightness. However, our ideas can be extended to use confidence sequences where the number of rollouts is not determined

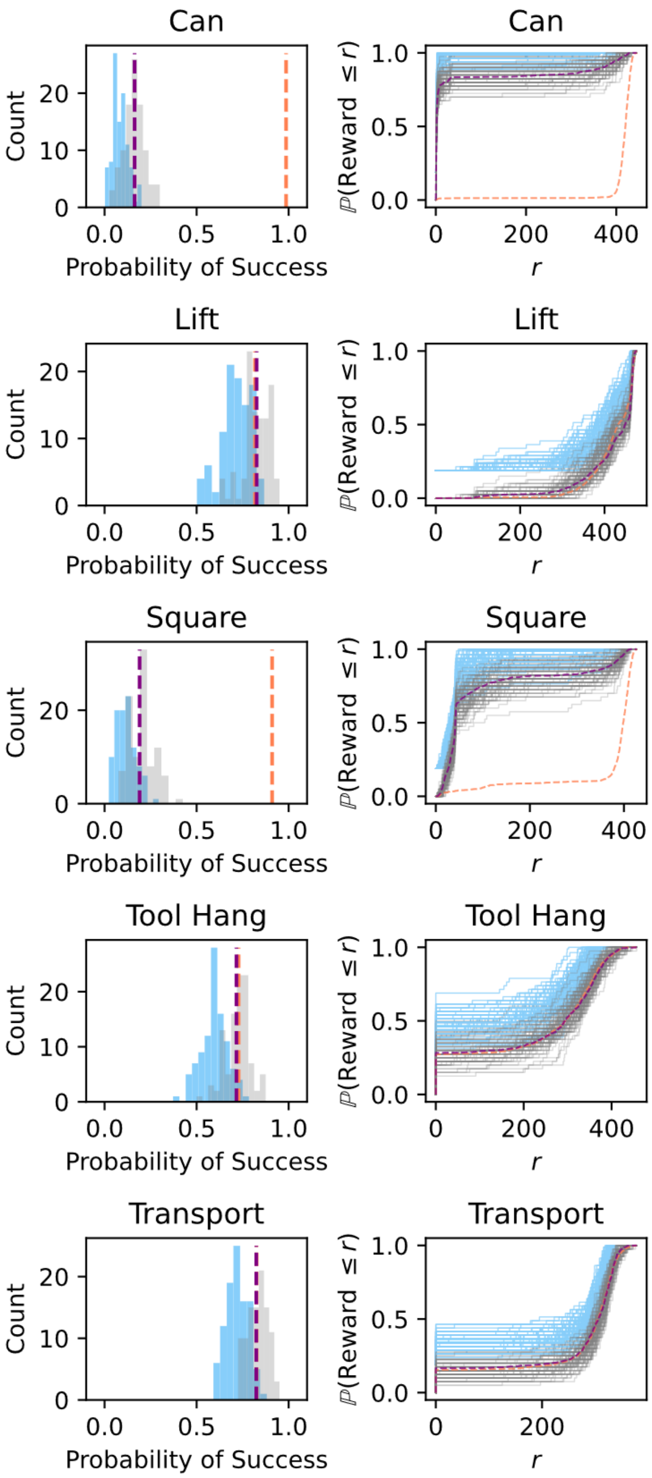

Fig. 6: Visualizations of the confidence bounds for each task. Each plot shows the OOD reward distribution (purple), Monte Carlo estimates of this distribution (gray), our confidence bounds for this distribution (blue), and the ID reward distribution (orange). We use 1000 rollouts to accurately estimate the OOD and ID reward distributions. We observe that the bounds appropriately bound the OOD reward distribution $\sim 95\%$ of the time, while a Monte Carlo estimate is overly optimistic $\sim 50\%$ of the time (since the Monte Carlo estimates are unbiased). From the ID reward distribution we can see that the distribution shifts in the Can and Square environments were harmful while the other shifts were benign.

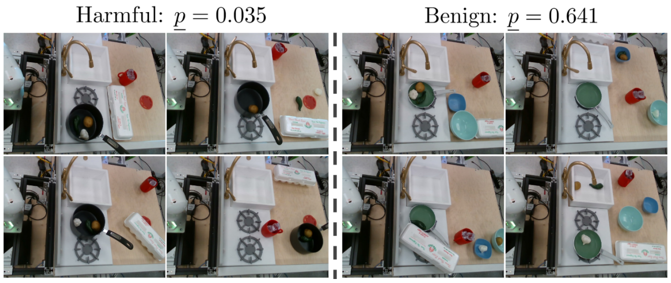
Harmful: $\underline{p} = 0.035$
Benign: $\underline{p} = 0.641$

Fig. 7: Snapshots of the environments used in Section VI-B. The robot is trained to pour the ice from the red cup into the sink. Four out of 50 initial conditions are shown for each OOD setting.

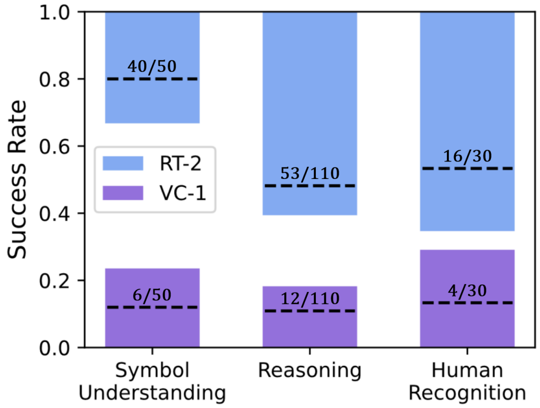

Fig. 8: Confidence intervals for success rates of RT-2 (blue) and VC-1 (purple) policies. For each setting, the intervals are disjoint and we conclude, with 95% confidence, that RT-2 outperforms VC-1. For each policy and setting we also show # successes / # trials and the corresponding empirical success rates (dashed black).

beforehand [44]. It may save on testing costs to stop running policy rollouts when enough evidence has been collected to conclude that the policy surpasses some performance threshold. Another direction is to evaluate performance during a single policy rollout, rather than multiple i.i.d. rollouts. As a final remark, the utility of any statistical analysis depends on the researchers being forthright in their data collection, assumptions/modeling, and interpretation of significance [45].

Acknowledgment

We thank Cheng Chi for guidance on diffusion policies.

References

[1] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," *arXiv preprint arXiv:2303.04137*, 2023.

[2] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 991–1002.

[3] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation," *arXiv preprint arXiv:2401.02117*, 2024.

[4] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[5] P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson, "Implicit Behavioral Cloning," in *Conference on Robot Learning*. PMLR, 2022, pp. 158–168.

[6] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey," in *2020 IEEE symposium series on computational intelligence (SSCI)*, 2020, pp. 737–744.

[7] S. Levine, C. Finn, T. Darrell, and P. Abbeel, " End-to-End Training of Deep Visuomotor Policies," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.

[8] C. J. Clopper and E. S. Pearson, "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

[9] B. Matthiesen, C. Hellings, E. A. Jorswieck, and W. Utschick, "Mixed Monotonic Programming for Fast Global Optimization," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2529–2544, 2020.

[10] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator," *The Annals of Mathematical Statistics*, pp. 642–669, 1956.

[11] P. Massart, "The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality," *The annals of Probability*, pp. 1269–1283, 1990.

[12] J. A. Vincent, H. Nishimura, M. Itkina, and M. Schwager, "Full-Distribution Generalization Bounds for Imitation Learning Policies," in *First Workshop on Out-of-Distribution Generalization in Robotics at CoRL 2023*, 2023.

[13] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks," in *Computer Aided Verification: 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I 30*. Springer, 2017, pp. 97–117.

[14] H.-D. Tran, X. Yang, D. Manzanas Lopez, P. Musau, L. V. Nguyen, W. Xiang, S. Bak, and T. T. Johnson, "NNV: The Neural Network Verification Tool for Deep Neural Networks and Learning-Enabled Cyber-Physical Systems," in *International Conference on Computer Aided Verification*. Springer, 2020, pp. 3–17.

[15] J. A. Vincent and M. Schwager, "Reachable Polyhedral Marching (RPM): An Exact Analysis Tool for Deep-Learned Control Systems," *arXiv preprint arXiv:2210.08339*, 2022.

[16] S. M. Katz, A. L. Corso, C. A. Strong, and M. J. Kochenderfer, "Verification of Image-Based Neural Network Controllers Using Generative Models," *Journal of Aerospace Information Systems*, vol. 19, no. 9, pp. 574–584, 2022.

[17] N. Rober, S. M. Katz, C. Sidrane, E. Yel, M. Everett, M. J. Kochenderfer, and J. P. How, "Backward Reachability Analysis of Neural Feedback Loops: Techniques for Linear and Nonlinear Systems," *IEEE Open Journal of Control Systems*, 2023.

[18] S. M. Richards, F. Berkenkamp, and A. Krause, "The Lyapunov Neural Network: Adaptive Stability Certification for Safe Learning of Dynamical Systems," in *Conference on Robot Learning*. PMLR, 2018, pp. 466–476.

[19] H. Dai, B. Landry, L. Yang, M. Pavone, and R. Tedrake, "Lyapunov-stable neural-network control," *arXiv preprint arXiv:2109.14152*, 2021.

[20] S. Singh, S. M. Richards, V. Sindhwani, J.-J. E. Slotine, and M. Pavone, "Learning stabilizable nonlinear dynamics with contraction-based regularization," *The International Journal of Robotics Research*, vol. 40, no. 10-11, pp. 1123–1150, 2021.

[21] D. Sun, S. Jha, and C. Fan, "Learning Certified Control Using Contraction Metric," in *Conference on Robot Learning*. PMLR, 2021, pp. 1519–1539.

[22] A. Corso, R. Moss, M. Koren, R. Lee, and M. Kochenderfer, " A Survey of Algorithms for Black-Box Safety Validation of Cyber-Physical Systems ," *Journal of Artificial Intelligence Research*, vol. 72, pp. 377–428, 2021.

[23] P. Akella, M. Ahmadi, and A. D. Ames, "A Scenario Approach to Risk-Aware Safety-Critical System Verification," *arXiv preprint arXiv:2203.02595*, 2022.

[24] P. Akella, A. Dixit, M. Ahmadi, J. W. Burdick, and A. D. Ames, "Sample-Based Bounds for Coherent Risk Measures: Applications to Policy Synthesis and Verification," *arXiv preprint arXiv:2204.09833*, 2022.

[25] M. Cleaveland, L. Lindemann, R. Ivanov, and G. J. Pappas, "Risk verification of stochastic systems with neural network controllers," *Artificial Intelligence*, vol. 313, p. 103782, 2022.

[26] J. A. Vincent, A. O. Feldman, and M. Schwager, "Guarantees on Robot System Performance Using Stochastic Simulation Rollouts," *arXiv preprint arXiv:2309.10874*, 2023.

[27] R. Luo, S. Zhao, J. Kuck, B. Ivanovic, S. Savarese, E. Schmerling, and M. Pavone, "Sample-Efficient Safety Assurances Using Conformal Prediction," in *International Workshop on the Algorithmic Foundations of Robotics*. Springer, 2022, pp. 149–169.

[28] L. Lindemann, M. Cleaveland, G. Shim, and G. J. Pappas, "Safe Planning in Dynamic Environments Using Conformal Prediction," *IEEE Robotics and Automation Letters*, 2023.

[29] A. Dixit, L. Lindemann, S. X. Wei, M. Cleaveland, G. J. Pappas, and J. W. Burdick, "Adaptive Conformal Prediction for Motion Planning among Dynamic Agents," in *Learning for Dynamics and Control Conference*. PMLR, 2023, pp. 300–314.

[30] A. Z. Ren, J. Clark, A. Dixit, M. Itkina, A. Majumdar, and D. Sadigh, "Explore until Confident: Efficient Exploration for Embodied Question Answering," *arXiv preprint arXiv:2403.15941*, 2024.

[31] A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley, *et al.*, "Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners," in *Conference on Robot Learning*. PMLR, 2023, pp. 661–682.

[32] B. Wu, B. D. Lee, B. Bucher, and N. Matni, "Uncertainty Aware Deployment of Pre-trained Task Conditioned Imitation Learning Policies," in *First Workshop on Out-of-Distribution Generalization in Robotics at CoRL 2023*, 2023.

[33] H. Papadopoulos, V. Vovk, and A. Gammerman, " Regression Conformal Prediction with Nearest Neighbours," *Journal of Artificial Intelligence Research*, vol. 40, pp. 815–840, 2011.

[34] A. N. Angelopoulos and S. Bates, "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification," *arXiv preprint arXiv:2107.07511*, 2021.

[35] P. G. Andersson, "The Wald Confidence Interval for a Binomial p as an Illuminating "Bad" Example," *The American Statistician*, pp. 1–6, 2023.

[36] M. P. Fay and S. A. Hunsberger, "Practical valid inferences for the two-sample binomial problem," *Statistics Surveys*, vol. 15, no. none, pp. 72 – 110, 2021. [Online]. Available: https://doi.org/10.1214/21-SS131

[37] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*. Springer, 2022, vol. 4.

[38] J. K. Ghosh, "On the Relation Among Shortest Confidence Intervals of Different Types," *Calcutta Statistical Association Bulletin*, vol. 10, no. 4, pp. 147–152, 1961.

[39] J. W. Pratt, "Length of Confidence Intervals," *Journal of the American Statistical Association*, vol. 56, no. 295, pp. 549–567, 1961.

[40] M. D. d. Edwardes, "THE EVALUATION OF CONFIDENCE SETS WITH APPLICATION TO BINOMIAL INTERVALS," *Statistica Sinica*, pp. 393–409, 1998.

[41] Z. Birnbaum and F. H. Tingey, "One-Sided Confidence Contours for Probability Distribution Functions," *The Annals of Mathematical Statistics*, pp. 592–596, 1951.

[42] Y. Zhu, J. Wong, A. Mandlekar, R. Martín-Martín, A. Joshi, S. Nasiriany, and Y. Zhu, "robosuite: A Modular Simulation Framework and Benchmark for Robot Learning," *arXiv preprint arXiv:2009.12293*, 2020.

[43] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, *et al.*, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," *arXiv preprint arXiv:2307.15818*, 2023.

[44] S. R. Howard and A. Ramdas, "Sequential estimation of quantiles with applications to A/B testing and best-arm identification," *Bernoulli*, vol. 28, no. 3, pp. 1704–1728, 2022.

[45] S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman, "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations," *European journal of epidemiology*, vol. 31, no. 4, pp. 337–350, 2016.