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

Learning What Matters: Adaptive Information-Theoretic Objectives for Robot Exploration

Youwei Yu[1], Jionghao Wang[2], Zhengming Yu[2], Wenping Wang[2] and Lantao Liu[1]
[1]Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA
Email: {youwyu, lantao}@iu.edu
[2]Department of Computer Science & Engineering, Texas A&M University, College Station, TX, USA
Email: {jionghao, yuzhengming, wenping}@tamu.edu

***Abstract*—Designing learnable information-theoretic objectives for robot exploration remains challenging. Such objectives aim to guide exploration toward data that reduces uncertainty in model parameters, yet it is often unclear what information the collected data can actually reveal. Although reinforcement learning (RL) can optimize a given objective, constructing objectives that reflect parametric learnability is difficult in high-dimensional robotic systems. Many parameter directions are weakly observable or unidentifiable, and even when identifiable directions are selected, omitted directions can still influence exploration and distort information measures. To address this challenge, we propose Quasi-Optimal Experimental Design (QOED), an adaptive information objective grounded in optimal experimental design. QOED (i) performs eigenspace analysis of the Fisher information matrix to identify an observable subspace and select identifiable parameter directions, and (ii) modifies the exploration objective to emphasize these directions while suppressing nuisance effects from non-critical parameters. Under bounded nuisance influence and limited coupling between critical and nuisance directions, QOED provides a constant-factor approximation to the ideal information objective that explores all parameters. We evaluate QOED on simulated and real-world navigation and manipulation tasks, where identifiable-direction selection and nuisance suppression yield performance improvements of $35.23\,\%$ and $21.98\,\%$, respectively. When integrated as an exploration objective in model-based policy optimization, QOED further improves policy performance over established RL baselines.**

I. Introduction

In robot exploration, information-theoretic objectives provide a principled mechanism for making effective use of limited interaction budgets by encouraging actions that collect data reducing uncertainty about unknown quantities. Bayesian optimal experimental design (BOED) [1] formalizes this idea by selecting actions that maximize expected information gain about a chosen set of parameters [1]. This perspective has been used across robotics, including active perception for scene representations [2–4], sim-to-real transfer via simulator calibration [5, 6], and manipulation via contact-rich interactions [7]. Due to its statistical rigor, BOED can substantially improve exploration efficiency by prioritizing informative data.

Website with additional information, videos, and code: https://www.youwei-yu.com/qoed.

However, the success of BOED-based exploration depends on a well-specified objective—in particular, choosing parameters that the collected data can reliably reveal (i.e., that are sufficiently observable and identifiable). In practice, learnability varies widely across settings: geometry can be more observable than appearance in active perception [2–4], and mass/motor properties can be easier to infer than aerodynamic effects in system identification [6]. In challenging scenarios, such as a quadruped traversing ice, it may be difficult to determine whether abnormal behavior is caused by the environment or by degraded actuators [8]. When the objective is well defined, BOED can also yield interpretable behaviors (e.g., rubbing to estimate friction) [7]. Nonetheless, specifying the "critical" parameters typically requires domain expertise and often assumes they can be pre-specified *a priori*. Moreover, naively ignoring unselected parameters can distort information objectives and lead exploration to pursue spurious or uninformative directions.

This work addresses the challenge of adaptively designing learnable information-theoretic objectives. We focus on identifying critical parameters online and reducing the influence of nuisance parameters on the exploration objective. To this end, we propose Quasi-Optimal Experimental Design (QOED). QOED first leverages eigenspace analysis of the Fisher information matrix to identify an observable subspace and select identifiable parameter coordinates. It then constructs an adaptive information objective that emphasizes the critical coordinates while suppressing nuisance directions.

Our main contributions include:

1) **Identifying Critical Parameters.** Many physical parameters in robotic systems are weakly observable or unidentifiable. We analyze the eigenstructure of the FIM to characterize an observable subspace and select a compact set of identifiable parameter directions.
2) **Adaptive Information Objective.** The proposed adaptive information-theoretic objective emphasizes critical directions of interest while suppressing the influence of nuisance parameters. Under bounded nuisance effects and limited coupling between critical and nuisance directions, this objective yields a constant-factor approximation to

the ideal objective that explores all parameters.

3) **QOED for Model-Based Policy Optimization.** Beyond pure exploration, we integrate QOED into model-based policy optimization (MBPO). We learn a physics-conditioned forward dynamics model that enables estimation of the QOED objective, and augment the task reward with this objective to guide robot exploration.

II. RELATED WORK

Exploration is a central challenge in robotics because real-world interaction is expensive. In this section, we review (i) exploration methods in RL, (ii) BOED for exploration objectives, and (iii) model-based policy optimization, highlighting the practical gaps that arise when bringing BOED into real-world robotic learning.

A. Learning to Explore

Exploration in robotics has moved from heuristics-driven [9] to learning-based methods. In RL, exploration is typically implemented through additional reward bonuses and is often grouped into two families: *uncertainty-driven* (information gathering) and *intrinsic-motivation* (curiosity).

Uncertainty-driven exploration. It encourages actions that reduce uncertainty. A common distinction is between: (i) uncertainty about learned parameters (e.g., value functions), and (ii) irreducible randomness in the environment (e.g., stochastic transitions). In discrete settings, the uncertainty can be propagated with Bayesian updates [10]. In continuous domains, it is often approximated using low-dimensional structure [11] or ensembles/bootstrapping [12, 13]. Exploration is then driven by reducing parameter uncertainty [10], controlling information regret [14], or reducing uncertainty in predicted returns [15]. Irreducible randomness is commonly modeled with distributional RL [16, 17]. Although theoretically distinct, these uncertainty sources often overlap in practice [15, 18].

Curiosity-driven exploration. Curiosity bonuses encourage agents to seek experiences that are either novel or hard to predict. For novelty, classic count-based methods in tabular RL [19] have been extended to continuous spaces using density estimation [20, 21]. Go-Explore [22] further biases exploration by returning to previously discovered states and expanding from them. For prediction-based curiosity, Random Network Distillation (RND) [23] uses the prediction error of a fixed random target network as an intrinsic reward. Related methods use forward dynamics prediction [24, 25] or inverse dynamics to emphasize controllable novelty [26]. Physics-based curiosity can also be defined through parameter estimation error [27], but it typically assumes a small and specified set of parameters.

B. Bayesian Optimal Experiment Design

BOED selects experiments (e.g., actions) to maximize expected information gain about unknown parameters. However, classical BOED typically relies on two assumptions: (i) a compact, expert-defined set of parameters, and (ii) an accurate probabilistic model for predicting information gain.

Choosing parameters of interest. BOED traditionally assumes key parameters are pre-specified by experts. With a good choice of parameters, BOED has achieved strong results in locomotion [6], underactuated control [28], and manipulation [7, 29]. However, success hinges on parameters that are both observable and identifiable (e.g., masses and damping in a cart–pendulum system [28]). This assumes that unselected parameters do not alter experiment utility. While active subspaces [30] provide tools for dimensionality reduction, they are typically applied to static experimental design. Standard Bayesian active learning can fail in the presence of nuisance parameters [31]. Nonetheless, identifying nuisance parameters and accounting for their influence remains a significant gap.

Statistical model. In policy learning, the statistical model usually represents forward dynamics [6]. Robotics applications often employ manually-designed physics models with Gaussian noise to ensure analytical tractability [5–7]. While effective in controlled settings, these are sensitive to model mismatch. Conversely, model-based reinforcement learning (MBRL) excels at learning dynamics [32] but often lacks the tractable likelihoods required for BOED. Consequently, a practical gap persists between predicting the next state and estimating the information gain it yields.

C. Model-Based Policy Optimization

Much recent progress in robotic RL relies on models that can generate large amounts of training experience with little additional data collection. These models are typically either (i) physics simulators or (ii) learned dynamics models.

Simulators. Physics simulators act as a library of validated physical rules and have enabled impressive sim-to-real transfer in agile locomotion and manipulation [33–37]. Ongoing work continues to improve simulator fidelity and breadth [38], but accurately modeling every aspect of the real world (e.g., fluids, contact micro-effects) remains difficult, and the sim-to-real gap can still limit performance.

Learned world models. MBRL learns a dynamics model directly from real experience and uses it to generate imagined rollouts in parallel [39, 40]. Training policies using imagined rollouts from a learned model is often referred to as model-based policy optimization [39]. Our proposed approach is based on this framework, but we emphasize an additional requirement that is critical for BOED: the model should support likelihood-based reasoning about unknown parameters.

In contrast to existing work, our model is designed to be both a forward dynamics model for imagined rollouts and a statistical model that supports BOED-driven exploration.

III. PRELIMINARY

We study a hidden-parameter Markov decision process (MDP) defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \Phi, p_0, P, p(\boldsymbol{\phi}), R, T, \gamma \rangle$. States are $\mathbf{s} \in \mathcal{S} \subseteq \mathbb{R}^{d_s}$, actions are $\boldsymbol{a} \in \mathcal{A} \subseteq \mathbb{R}^{d_a}$, and the hidden parameters are $\boldsymbol{\phi} \in \Phi \subseteq \mathbb{R}^m$ with prior $p(\boldsymbol{\phi})$. The initial state is $\mathbf{s}_0 \sim p_0(\mathbf{s}_0)$. Given $(\mathbf{s}_t, \boldsymbol{a}_t, \boldsymbol{\phi})$, the next state is sampled from the transition model $P(\cdot|\mathbf{s}_t, \boldsymbol{a}_t, \boldsymbol{\phi})$, i.e., $\mathbf{s}_{t+1} \sim$

$p(\mathbf{s}_{t+1}|\mathbf{s}_t, \boldsymbol{a}_t, \boldsymbol{\phi})$. The per-step reward is $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the horizon is $T \in \mathbb{N}^+$, and $\gamma \in (0, 1]$ is the discount factor. We write a trajectory prefix as $\boldsymbol{\tau}_t = [\mathbf{s}_0, \boldsymbol{a}_0, \mathbf{s}_1, \ldots, \boldsymbol{a}_{t-1}, \mathbf{s}_t]$. For a stochastic policy $\boldsymbol{\pi}(\boldsymbol{a}|\mathbf{s})$, the induced distribution over $\boldsymbol{\tau}_t$ is

$$p(\boldsymbol{\tau}_t|\boldsymbol{\phi}, \boldsymbol{\pi}) = p_0(\mathbf{s}_0) \prod_{k=0}^{t-1} \boldsymbol{\pi}(\boldsymbol{a}_k|\mathbf{s}_k)\, p(\mathbf{s}_{k+1}|\mathbf{s}_k, \boldsymbol{a}_k, \boldsymbol{\phi}). \quad (1)$$

Problem 1 (Policy learning with an exploration objective)**.** *We learn a stochastic policy $\boldsymbol{\pi}(\boldsymbol{a}|\mathbf{s})$ that maximizes reward while also collecting data that helps estimate the hidden parameters $\boldsymbol{\phi}$. Concretely, we solve*

$$\boldsymbol{\pi}^\star = \arg\max_{\boldsymbol{\pi} \in \Pi} \mathbb{E}_{p(\boldsymbol{\phi}),\, p(\boldsymbol{\tau}|\boldsymbol{\phi}, \boldsymbol{\pi})} \Bigg[\sum_{t=0}^{T-1} \gamma^t \Big(R(\mathbf{s}_t, \boldsymbol{a}_t) + \alpha\, \mathcal{B}(\boldsymbol{\phi}|\boldsymbol{\tau}_t) \Big) \Bigg]$$
$$\text{s.t.} \quad \mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \boldsymbol{a}_t, \boldsymbol{\phi}), \mathbf{s}_t \in \mathcal{S}, \quad \boldsymbol{a}_t \in \mathcal{A}, \quad \forall t. \quad (2)$$

where $\mathcal{B}(\boldsymbol{\phi}|\boldsymbol{\tau}_t)$ is an exploration objective computed from the trajectory prefix $\boldsymbol{\tau}_t$, and $\alpha \geq 0$ controls the trade-off between task reward and exploration.

Across the literature, $\boldsymbol{\phi}$ can represent different unknowns, ranging from explicit physical parameters [5] to weights of neural networks [41]. In this paper, $\boldsymbol{\phi}$ denotes physical parameters (e.g., friction). These parameters are interpretable and can be used as inputs to a forward dynamics model. We focus on physically meaningful parameters to avoid the degeneracy caused by arbitrary unit scalings.

A. Exploration Objective from Information Gain

We design the exploration objective $\mathcal{B}_t$ to encourage collecting data that is informative about $\boldsymbol{\phi}$. Intuitively, data is informative if changing $\boldsymbol{\phi}$ would noticeably change the likelihood of the observed transitions. We motivate this using Bayesian optimal experiment design (BOED) and the Fisher information matrix (FIM).

Definition 1 (BOED with Fisher Information)**.** *BOED chooses a design to maximize the expected information gain (EIG) about unknown parameters. In our setting, the "design" is the policy $\boldsymbol{\pi}$. We quantify the information content of a trajectory $\boldsymbol{\tau}_t$ with respect to $\boldsymbol{\phi}$ using FIM, which measures the sensitivity of the trajectory distribution to parameter perturbations. For a trajectory prefix $\boldsymbol{\tau}_t$, define the score $\mathbf{g}(\boldsymbol{\tau}_t, \boldsymbol{\phi}) := \nabla_{\boldsymbol{\phi}} \log p(\boldsymbol{\tau}_t|\boldsymbol{\phi}, \boldsymbol{\pi})$. The FIM is*

$$\boldsymbol{\mathcal{F}}_{\boldsymbol{\phi}} = \mathbb{E}_{\boldsymbol{\tau}_t \sim p(\boldsymbol{\tau}_t|\boldsymbol{\phi}, \boldsymbol{\pi})} \Big[\mathbf{g}(\boldsymbol{\tau}_t, \boldsymbol{\phi})\, \mathbf{g}(\boldsymbol{\tau}_t, \boldsymbol{\phi})^\top \Big], \quad (3)$$

assuming standard regularity conditions (Sect. 2.3.1 in [42]). Since the policy $\boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{s})$ and initial state distribution $p(\boldsymbol{s}_0)$ are independent of the parameters $\boldsymbol{\phi}$, their gradients vanish. The score depends solely on the transition dynamics:

$$\nabla_{\boldsymbol{\phi}} \log p(\boldsymbol{\tau}_t|\boldsymbol{\phi}, \boldsymbol{\pi}) = \sum_{k=0}^{t-1} \nabla_{\boldsymbol{\phi}} \log p(\mathbf{s}_{k+1}|\mathbf{s}_k, \boldsymbol{a}_k, \boldsymbol{\phi}). \quad (4)$$

The full derivation is given in Appendix B.

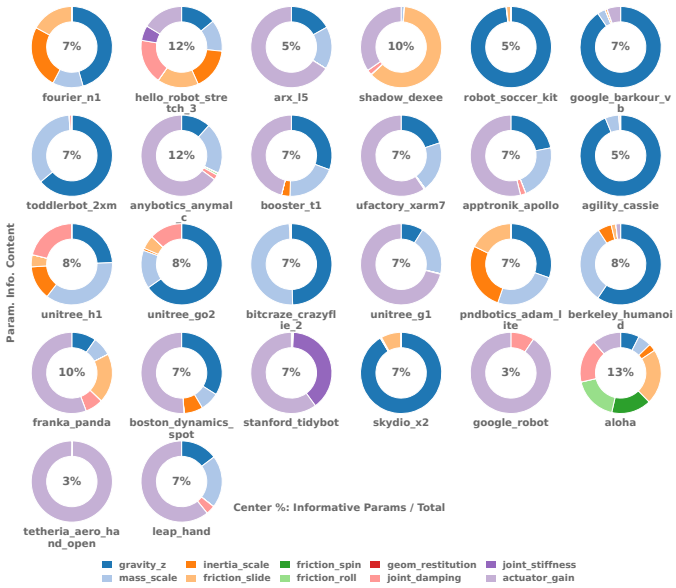

Fig. 1: Fisher information matrix values for physical parameters of 26 representative robots. Color denotes parameter type; size indicates information content. The large variation in information distribution indicates that information can be concentrated in only a few critical physical parameters.

To obtain a single scalar measure of informativeness, we summarize the FIM using its trace (i.e., T-optimality [43]):

$$\text{Ideal BOED Objective}: \quad \mathcal{B}_{\text{BOED}}(\boldsymbol{\phi}|\boldsymbol{\tau}_t) = \operatorname{tr}(\boldsymbol{\mathcal{F}}_{\boldsymbol{\phi}}). \quad (5)$$

B. Challenges to Address

The success of BOED-style exploration depends on ideally an expert-curated, low-dimensional set of critical parameters. This expert curation relies on the implicit assumption that other excluded parameters exert negligible influence on the policy exploration. Real robots often provide neither. This leads to two questions that we address in this paper:

- **Q1 Critical parameter subspace.** In high-dimensional systems, only a few "critical" parameters affect what the information can reveal [44, 45] (see Fig. 1 for an illustration). Can an agent identify these critical parameters automatically, without expert assumptions?
- **Q2 Subspace-optimal experimental design.** Simply optimizing for critical parameters is insufficient; discarded parameters inevitably affect the information objective. Consequently, how can we design an objective to compensate for these unmodeled effects, ensuring the policy remains near optimal to the full BOED—the ideal objective computed over the entire parameter space?

IV. METHOD

This section details our design of an exploration objective to focus on *critical* parameter directions that are learnable from data. Sect. IV-A identifies a compact, non-redundant set of critical parameters, and Sect. IV-B introduces QOED, which prioritizes these parameters while suppressing the influence of the remaining ones. Sect. IV-C integrates QOED into model-based policy optimization to improve policy performance.

A. Selecting Critical Parameters

To answer **Q1** presented in Sect. III-B, we select critical parameters in three steps: (i) estimate the current parameter values, (ii) find the well-observed subspace via eigenspace analysis of the FIM, and (iii) select a compact set of parameter coordinates within the observable subspace that are weakly correlated with each other, improving identifiability.

Step I – Parameter estimation. Because the FIM depends on the parameter value, we first estimate $\boldsymbol{\phi}$ from data. Given an observed trajectory prefix $\boldsymbol{\tau}_t^{\text{obs}}$ with action sequence $\mathbf{A}_t = [\boldsymbol{a}_0, \dots, \boldsymbol{a}_{t-1}]$, we estimate $\boldsymbol{\phi}$ by matching rollouts to the observed trajectory:

$$\hat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}} \mathbb{E}_{\boldsymbol{\tau}_t \sim p(\boldsymbol{\tau}_t|\boldsymbol{\phi}, \mathbf{A}_t)} \left[\|\boldsymbol{\tau}_t^{\text{obs}} - \boldsymbol{\tau}_t\|_2^2 \right]. \tag{6}$$

Here $p(\boldsymbol{\tau}_t|\boldsymbol{\phi}, \mathbf{A}_t)$ is the trajectory distribution with fixed actions $\mathbf{A}_t$, i.e., $p(\boldsymbol{\tau}_t|\boldsymbol{\phi}, \mathbf{A}_t) = p(\mathbf{s}_0) \prod_{k=0}^{t-1} p(\mathbf{s}_{k+1}|\mathbf{s}_k, \boldsymbol{a}_k, \boldsymbol{\phi})$. See Appendix B for the derivation. Although differentiability would allow gradient-based optimization, we use the cross-entropy method (CEM) [46], a derivative-free optimizer, which is effective in practice [5]. We maintain a Gaussian belief over parameters, $p(\boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\phi}}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}})$, and update uncertainty approximately as $\boldsymbol{\Sigma}_{\boldsymbol{\phi}} \leftarrow \left(\boldsymbol{\mathcal{F}}_{\hat{\boldsymbol{\phi}}} + \boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1}\right)^{-1}$, analogous to updates in extended Kalman filtering [47]. This Gaussian approximation may overestimate uncertainty, but it empirically facilitates parameter convergence in our setting [7, 47].

Step II – Identify the well-observed subspace. We use eigenvalues and eigenvectors of the FIM to identify the critical parts of the parameter space. Eigenvalues indicate how much information the trajectory provides along different directions in parameter space (larger is better). In high-dimensional settings, many FIM eigenvalues are often close to zero [44], meaning the data carries little information about many directions (as shown in Fig. 1). Eigenvectors indicate which physical parameters are contributing together in a particular informative direction (in some cases, individual parameters may not be separable: different physical parameters can produce nearly indistinguishable behavior [45], which prevents reliable identification of individual parameters). We start with the following assumption for FIM's eigen-decomposition.

Assumption 1 (Regularity)**.** *Assume $p(\boldsymbol{\tau}_t|\boldsymbol{\phi}, \boldsymbol{\pi})$ is continuously differentiable in $\boldsymbol{\phi}$ and its score has bounded norm:*

$$\left\|\nabla_{\boldsymbol{\phi}} \log p(\boldsymbol{\tau}_t|\boldsymbol{\phi}, \boldsymbol{\pi})\right\| \le \delta_{\text{reg}}, \qquad \forall \boldsymbol{\phi} \in \Phi,$$

for some constant $\delta_{\text{reg}} \ge 0$. This holds under standard smoothness and boundedness conditions; for example, if $\boldsymbol{\phi}$ is compact and the score is uniformly bounded.

Lemma 1. *For any $\boldsymbol{\phi} \in \Phi$, the FIM $\boldsymbol{\mathcal{F}}_{\boldsymbol{\phi}} \in \mathbb{R}^{m \times m}$ is symmetric positive semidefinite and admits an eigen-decomposition*

$$\boldsymbol{\mathcal{F}}_{\boldsymbol{\phi}} = \mathbf{W} \boldsymbol{\Lambda} \mathbf{W}^\top, \quad \boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_m),$$

where $\lambda_1 \ge \cdots \ge \lambda_m \ge 0$ and $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_m] \in \mathbb{R}^{m \times m}$ is orthonormal. The Fisher information along eigen-direction $\mathbf{w}_i$ equals λ_i:

$$\mathbb{E}_{\boldsymbol{\tau} \sim p(\boldsymbol{\tau}|\boldsymbol{\phi}, \boldsymbol{\pi})} \left[\left(\nabla_{\boldsymbol{\phi}} \log p(\boldsymbol{\tau}|\boldsymbol{\phi}, \boldsymbol{\pi})^\top \mathbf{w}_i \right)^2 \right] = \lambda_i.$$

Proof: Please refer to Appendix C. □

Lemma 1 shows that each eigenvalue λ_i measures information along eigen-direction $\mathbf{w}_i$. If $\lambda_i = 0$, then the score has zero projection onto $\mathbf{w}_i$, and trajectories provide no local information in that direction. The Cramér–Rao lower bound [48] further implies that small λ_i correspond to large lower bounds on estimation error.

Proposition 1 (Multivariate Cramér–Rao lower bound [48])**.** *Let $\hat{\boldsymbol{\phi}}$ be any unbiased estimator of the true parameters $\boldsymbol{\phi}$. Under standard regularity conditions,*

$$\text{Cov}(\hat{\boldsymbol{\phi}}) \succeq \boldsymbol{\mathcal{F}}_{\boldsymbol{\phi}}^{-1},$$

where $\boldsymbol{\mathcal{F}}_{\boldsymbol{\phi}}$ is the FIM. Consequently,

$$\mathbb{E}[\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}\|^2] = \text{tr}\big(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\phi}}}\big) \ \ge \ \text{tr}\big(\boldsymbol{\mathcal{F}}_{\boldsymbol{\phi}}^{-1}\big).$$

Let $\delta_{\text{eig}} > 0$ be an eigenvalue threshold and define the index set of well-observed directions $\mathbf{o} = \{i \mid \lambda_i \ge \delta_{\text{eig}}\}$, with $|\mathbf{o}| = n$. We partition the eigen-decomposition as

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{\mathbf{o}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_{\bar{\mathbf{o}}} \end{bmatrix}, \ \mathbf{W} = \begin{bmatrix} \mathbf{W}_{\mathbf{o}} & \mathbf{W}_{\bar{\mathbf{o}}} \end{bmatrix}, \ \mathbf{o} = \{i | \lambda_i \ge \delta_{\text{eig}}\}, \tag{7}$$

where $\mathbf{W}_{\mathbf{o}} = \mathbf{W}[:, \mathbf{o}] \in \mathbb{R}^{m \times n}$ spans the well-observed subspace and $\mathbf{W}_{\bar{\mathbf{o}}} \in \mathbb{R}^{m \times (m-n)}$ spans the weakly observed (or unobserved) subspace. Importantly, individual physical parameters may not be observable by themselves; the observable direction is defined as $\mathbf{W}_{\mathbf{o}}^\top \boldsymbol{\phi}$ in the literature [30], which represents linear combinations of physical parameters.

Step III – Select identifiable parameter coordinates. Even if a direction is well observed, it can correspond to a linear combination of multiple physical parameters. To obtain a compact and interpretable set of coordinates, we select a subset of parameter indices $\mathbf{k} \subseteq \{1, \dots, m\}$ that (i) have strong components in the well-observed subspace and (ii) are not redundant with each other. Let $\mathbf{r}_j^\top$ be the j-th row of $\mathbf{W}_{\mathbf{o}}$. Intuitively, $\mathbf{r}_j$ describes how parameter j loads onto the well-observed directions. Selecting indices $\mathbf{k}$ gives the row-submatrix $\mathbf{W}_{\mathbf{ko}} = \mathbf{W}[\mathbf{k}, \mathbf{o}] \in \mathbb{R}^{|\mathbf{k}| \times n}$. With a budget $|\mathbf{k}| \le n$,

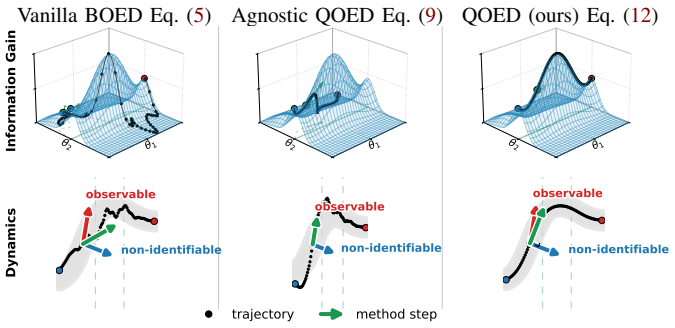

Fig. 2: **Information gain trajectories.** *Top:* Fisher information landscape over mass θ_1 and friction θ_2 with induced trajectories for the box pushing task. *Bottom:* local geometry of the dynamics. Optimizing information in both parameters can drift off the identifiable ridge. Restricting the objective to selected coordinates can stay closer to the ridge but may still excite non-identifiable directions. Our QOED method (Sect. IV-B) adds constraints/regularization to suppress these directions and keep trajectories near the ridge.

we choose

$$\begin{aligned} \mathbf{k} = & \underset{\mathbf{k}\subseteq\{1,\dots,m\}:\,|\mathbf{k}|\le n}{\arg\max} \log\det\Big(\mathbf{W}_{\mathbf{ko}}\,\mathbf{W}_{\mathbf{ko}}^{\top}\Big) \\ \text{s.t.} \quad & \big|\cos(\mathbf{r}_i,\mathbf{r}_j)\big| \le \delta_{\cos}, \quad \forall\, i\neq j,\ i,j\in\mathbf{k}, \end{aligned} \tag{8}$$

where $\cos(\mathbf{r}_i,\mathbf{r}_j) = \mathbf{r}_i^{\top}\mathbf{r}_j/\|\mathbf{r}_i\|\|\mathbf{r}_j\|$. The log-determinant term favors a diverse set of rows that spans the well-observed subspace, and the cosine constraint prevents selecting highly correlated parameters. We approximately solve Eq. (8) with a lazy-greedy procedure [49] that adds one index at a time and rejects candidates that violate the cosine constraint. In the rare case where two parameters induce the same observable direction, this procedure keeps only one representative.

Solution to Q1. Steps I–III output the index set $\mathbf{k}$. We refer to the selected coordinates $\boldsymbol{\phi}_{\mathbf{k}} = \boldsymbol{\phi}[\mathbf{k}]$ as the *critical (identifiable) parameters*. A straightforward exploration objective is the FIM restricted to these coordinates. Let $\boldsymbol{\mathcal{F}}_{\mathbf{kk}} = \boldsymbol{\mathcal{F}}_{\boldsymbol{\phi}}[\mathbf{k},\mathbf{k}]$ be the principal submatrix of the FIM. We define

$$\text{Agnostic QOED Objective}: \quad \mathcal{B}_{\text{Agnostic}}(\boldsymbol{\phi}|\boldsymbol{\tau}_t) = \operatorname{tr}(\boldsymbol{\mathcal{F}}_{\mathbf{kk}}), \tag{9}$$

where the term "Agnostic" indicates that this objective ignores discarded (i.e., non-critical) parameters.

B. Adaptive Information-Theoretic Objective

We now present our *Quasi-Optimal Experimental Design* (QOED). To answer **Q2** presented in Sect. III-B, QOED prioritizes information gain in *critical* parameter coordinates $\boldsymbol{\phi}_{\mathbf{k}}$ while *suppressing* the influence of the remaining ones. This operation is important because *non-critical or weakly identifiable parameters can still influence the dynamics and induce non-negligible Fisher information*. In such settings, simply dropping parameters as in Eq. (9) can distort the information objective [30]. Formally, denote the score as $\mathbf{g} = [\mathbf{g}_{\mathbf{k}}; \mathbf{g}_{\bar{\mathbf{k}}}]$ so that $\boldsymbol{\mathcal{F}}_{\boldsymbol{\phi}} = \mathbb{E}[\mathbf{g}\mathbf{g}^{\top}]$ implies the block form in Eq. (10). The agnostic objective $\mathcal{B}_{\text{Agnostic}} = \operatorname{tr}(\boldsymbol{\mathcal{F}}_{\mathbf{kk}})$ maximizes the retained score energy $\mathbb{E}\|\mathbf{g}_{\mathbf{k}}\|_2^2$ while treating $\mathbf{g}_{\bar{\mathbf{k}}}$ as irrelevant. However, the discarded energy $\mathbb{E}\|\mathbf{g}_{\bar{\mathbf{k}}}\|_2^2 = \operatorname{tr}(\boldsymbol{\mathcal{F}}_{\bar{\mathbf{k}}\bar{\mathbf{k}}})$ can be non-negligible, and $\mathbf{g}_{\mathbf{k}}$ can be correlated with $\mathbf{g}_{\bar{\mathbf{k}}}$ (via $\boldsymbol{\mathcal{F}}_{\mathbf{k}\bar{\mathbf{k}}}$). As a result, maximizing $\operatorname{tr}(\boldsymbol{\mathcal{F}}_{\mathbf{kk}})$ can overestimate how much the data truly informs $\boldsymbol{\phi}_{\mathbf{k}}$: the apparent information may arise mainly from correlations with nuisance directions (via $\boldsymbol{\mathcal{F}}_{\mathbf{k}\bar{\mathbf{k}}}$).

Solution to Q2. Let $\mathbf{k}$ denote the *critical parameter indices* returned by Sect. IV-A, and let $\bar{\mathbf{k}} = \{1,2,\dots,m\}\setminus\mathbf{k}$ be the complementary indices. We block-partition the FIM as

$$\boldsymbol{\mathcal{F}}_{\boldsymbol{\phi}} = \begin{bmatrix} \boldsymbol{\mathcal{F}}_{\mathbf{kk}} & \boldsymbol{\mathcal{F}}_{\mathbf{k}\bar{\mathbf{k}}} \\ \boldsymbol{\mathcal{F}}_{\bar{\mathbf{k}}\mathbf{k}} & \boldsymbol{\mathcal{F}}_{\bar{\mathbf{k}}\bar{\mathbf{k}}} \end{bmatrix}. \tag{10}$$

QOED uses the *nuisance-adjusted* Fisher information for $\boldsymbol{\phi}_{\mathbf{k}}$, given by the Schur complement[1]

$$\boldsymbol{\mathcal{I}}_{\mathbf{k}|\bar{\mathbf{k}}} = \boldsymbol{\mathcal{F}}_{\mathbf{kk}} - \boldsymbol{\mathcal{F}}_{\mathbf{k}\bar{\mathbf{k}}}\boldsymbol{\mathcal{F}}_{\bar{\mathbf{k}}\bar{\mathbf{k}}}^{-1}\boldsymbol{\mathcal{F}}_{\bar{\mathbf{k}}\mathbf{k}}. \tag{11}$$

We define the QOED information objective using the trace:

$$\text{QOED Objective}: \qquad \mathcal{B}_{\text{QOED}}(\boldsymbol{\phi}|\boldsymbol{\tau}_t) = \operatorname{tr}\Big(\boldsymbol{\mathcal{I}}_{\mathbf{k}|\bar{\mathbf{k}}}\Big). \tag{12}$$

Algorithm 1: QOED Information Objective

Input : Trajectories $\{\boldsymbol{\tau}_{t,n}\}_{n=1}^{N}$, parameter $\boldsymbol{\phi}$, policy $\boldsymbol{\pi}$
Default: Thresholds of eigenvalue $\delta_{\text{eig}} = 0.1$, $\alpha_{\text{eig}} = 0.01$, and similarity $\delta_{\cos} = 0.95$

/* Fisher Information Matrix Eq. (3) */
$\boldsymbol{\mathcal{F}}_{\boldsymbol{\phi}}(\boldsymbol{\pi}) \approx \frac{1}{N}\sum_{n=1}^{N}\mathbf{g}_n\mathbf{g}_n^{\top}, \quad \mathbf{g}_n = \nabla_{\boldsymbol{\phi}}\log p(\boldsymbol{\tau}_{t,n}|\boldsymbol{\phi},\boldsymbol{\pi})$
/* eigen-decomposition */
$\mathbf{W}\boldsymbol{\Lambda}\mathbf{W}^{\top} \leftarrow \boldsymbol{\mathcal{F}}_{\boldsymbol{\phi}}(\boldsymbol{\pi})$
/* Observable Subspace Eq. (7) */
$\delta_{\text{eig}} \leftarrow \max(\delta_{\text{eig}},\ \alpha_{\text{eig}}\cdot\max\operatorname{diag}(\boldsymbol{\Lambda}))$
Split $(\boldsymbol{\Lambda},\mathbf{W})$ into $(\boldsymbol{\Lambda}_{\mathbf{o}},\mathbf{W}_{\mathbf{o}})$, $(\boldsymbol{\Lambda}_{\bar{\mathbf{o}}},\mathbf{W}_{\bar{\mathbf{o}}})$ with δ_{eig}
/* Identifiable Parameters Eq. (8) */
Greedy select $\mathbf{k}$ from $\mathbf{W}_{\mathbf{o}}$ with $\delta_{\cos}$
$\bar{\mathbf{k}} \leftarrow \{1,2,\dots,|\boldsymbol{\phi}|\}\setminus\mathbf{k}$

Return: $\operatorname{tr}\left(\boldsymbol{\mathcal{F}}_{\mathbf{kk}}\right) - \operatorname{tr}\Big(\boldsymbol{\mathcal{F}}_{\mathbf{k}\bar{\mathbf{k}}}\boldsymbol{\mathcal{F}}_{\bar{\mathbf{k}}\bar{\mathbf{k}}}^{-1}\boldsymbol{\mathcal{F}}_{\bar{\mathbf{k}}\mathbf{k}}\Big)$

Alg. 1 summarizes the computation of the QOED information objective. We use a relative threshold $\alpha_{\text{eig}}\cdot\max\operatorname{diag}(\boldsymbol{\Lambda})$ to make the eigenvalue split robust to global scalings. Here, we also give an interpretation of the Schur complement. Partition the score as $\mathbf{g} = [\mathbf{g}_{\mathbf{k}}; \mathbf{g}_{\bar{\mathbf{k}}}]$. Then the Schur complement $\boldsymbol{\mathcal{I}}_{\mathbf{k}|\bar{\mathbf{k}}}$ equals the covariance of the residual score after removing the best linear prediction from $\mathbf{g}_{\bar{\mathbf{k}}}$: $\operatorname{tr}(\boldsymbol{\mathcal{I}}_{\mathbf{k}|\bar{\mathbf{k}}}) = \min_{\mathbf{A}}\ \mathbb{E}\big[\|\mathbf{g}_{\mathbf{k}} - \mathbf{A}\mathbf{g}_{\bar{\mathbf{k}}}\|_2^2\big]$. Thus QOED rewards information about the critical parameters that cannot be *linearly predicted* from the non-critical ones. As shown in Fig. 2, QOED stays near the identifiable ridge while optimizing the full trace drifts off the ridge. On the other hand, from an eigen-subspace viewpoint [30], projecting the score onto a selected FIM eigenspace discards an expected squared score magnitude equal to the sum of the omitted eigenvalues (Appendix F). The following theorem shows that optimizing QOED is quasi-optimal, in the sense that it achieves a constant-factor approximation to the full BOED objective.

[1] We use the standard stabilization $\boldsymbol{\mathcal{F}}_{\bar{\mathbf{k}}\bar{\mathbf{k}}} \leftarrow \boldsymbol{\mathcal{F}}_{\bar{\mathbf{k}}\bar{\mathbf{k}}} + \epsilon\mathbf{I}$ before inversion.

Theorem 1 (Quasi-optimality w.r.t. full BOED)**.** *Make the policy dependence explicit by writing* $\mathcal{F}^{\pi} := \mathcal{F}_{\phi}$ *in Eq. (3). Fix a set of critical indices* $\mathbf{k}$ *(with complement* $\overline{\mathbf{k}}$*) and let* $\mathcal{I}^{\pi}_{\mathbf{k}|\overline{\mathbf{k}}} := \mathcal{F}^{\pi}_{\mathbf{k}\mathbf{k}} - \mathcal{F}^{\pi}_{\mathbf{k}\overline{\mathbf{k}}} \left(\mathcal{F}^{\pi}_{\overline{\mathbf{k}}\overline{\mathbf{k}}}\right)^{-1} \mathcal{F}^{\pi}_{\overline{\mathbf{k}}\mathbf{k}}$*. Define the full BOED and QOED objectives*

$$\mathcal{B}_{BOED}(\boldsymbol{\pi}) := \operatorname{tr}\left(\mathcal{F}^{\pi}\right), \qquad \mathcal{B}_{QOED}(\boldsymbol{\pi}) := \operatorname{tr}\left(\mathcal{I}^{\pi}_{\mathbf{k}|\overline{\mathbf{k}}}\right). \quad (13)$$

Let $\boldsymbol{\pi}^{\star} \in \arg\max_{\boldsymbol{\pi}\in\Pi} \mathcal{B}_{BOED}(\boldsymbol{\pi})$ *and* $\widehat{\boldsymbol{\pi}} \in \arg\max_{\boldsymbol{\pi}\in\Pi} \mathcal{B}_{QOED}(\boldsymbol{\pi})$*. Assume* $\mathcal{F}^{\pi}_{\mathbf{k}\mathbf{k}} \succ \mathbf{0}$ *and* $\mathcal{F}^{\pi}_{\overline{\mathbf{k}}\overline{\mathbf{k}}} \succ \mathbf{0}$ *for all* $\boldsymbol{\pi} \in \Pi$ *and define*

$$\begin{aligned} \eta &= \sup_{\boldsymbol{\pi}\in\Pi} \frac{\operatorname{tr}(\mathcal{F}^{\pi}_{\overline{\mathbf{k}}\overline{\mathbf{k}}})}{\operatorname{tr}(\mathcal{F}^{\pi}_{\mathbf{k}\mathbf{k}})}, \\ \beta &= \sup_{\boldsymbol{\pi}\in\Pi} \left\| \left(\mathcal{F}^{\pi}_{\mathbf{k}\mathbf{k}}\right)^{-1/2} \mathcal{F}^{\pi}_{\mathbf{k}\overline{\mathbf{k}}} \left(\mathcal{F}^{\pi}_{\overline{\mathbf{k}}\overline{\mathbf{k}}}\right)^{-1/2} \right\|_2^2. \end{aligned} \quad (14)$$

If $\eta < \infty$ *and* $\beta < 1$*, then the QOED-optimal policy is a constant-factor approximation to the full BOED-optimal policy:*

$$\operatorname{tr}\left(\mathcal{F}^{\widehat{\pi}}\right) \geq \frac{1-\beta}{1+\eta} \operatorname{tr}\left(\mathcal{F}^{\pi^{\star}}\right). \quad (15)$$

Proof: Please refer to Appendix E for a justification of η and β, an extension to our time-varying $\mathbf{k}$, and a proof that the agnostic objective is not quasi-optimal in general. □

C. Policy Learning with QOED

We consider Problem 1 in a model-based policy optimization (MBPO) setting [39]. Because analytical physics models can be mismatched to real dynamics, we learn a differentiable dynamics model

$$q_{\boldsymbol{\theta}}(\mathbf{s}_{t+1}|\mathbf{s}_t, \boldsymbol{a}_t, \boldsymbol{\phi}), \quad (16)$$

parameterized by $\boldsymbol{\theta}$. Together with a policy $\boldsymbol{\pi}(\boldsymbol{a}|\mathbf{s})$, this induces a trajectory likelihood $q_{\boldsymbol{\theta}}(\boldsymbol{\tau}_t|\boldsymbol{\phi}, \boldsymbol{\pi})$ that serves as a surrogate for the real likelihood $p(\boldsymbol{\tau}_t|\boldsymbol{\phi}, \boldsymbol{\pi})$. We instantiate $q_{\boldsymbol{\theta}}$ with a shortcut model [50]: we sample latent noise $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and generate a one-step state increment via a transport map $T_{\boldsymbol{\theta}}$,

$$\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \qquad \mathbf{s}_{t+1} = \mathbf{s}_t + T_{\boldsymbol{\theta}}(\boldsymbol{\delta}, \mathbf{s}_t, \boldsymbol{a}_t, \boldsymbol{\phi}). \quad (17)$$

We train $T_{\boldsymbol{\theta}}$ using the shortcut-model objective [50] (Appendix A). If $T_{\boldsymbol{\theta}}$ is invertible in $\boldsymbol{\delta}$, then the conditional log-density follows by change of variables [51]:

$$\log q_{\boldsymbol{\theta}}(\mathbf{s}_{t+1}|\mathbf{s}_t, \boldsymbol{a}_t, \boldsymbol{\phi}) = \log p_0(\boldsymbol{\delta}) - \log \left|\det \nabla_{\boldsymbol{\delta}} T_{\boldsymbol{\theta}}(\boldsymbol{\delta}, \mathbf{s}_t, \boldsymbol{a}_t, \boldsymbol{\phi})\right|, \quad (18)$$

where p_0 is the standard normal density and $\boldsymbol{\delta}$ is the latent satisfying $\mathbf{s}_{t+1} - \mathbf{s}_t = T_{\boldsymbol{\theta}}(\boldsymbol{\delta}, \mathbf{s}_t, \boldsymbol{a}_t, \boldsymbol{\phi})$. This yields a tractable estimate of the FIM; the trajectory-level derivation is in Appendix B, following Eq. (4). We parameterize $T_{\boldsymbol{\theta}}$ with a Transformer [52]: state, action, and parameter inputs are embedded by separate MLPs, processed by a six-block Transformer with Adaptive Layer Normalization (AdaLN), and mapped to the output by a final AdaLN modulation layer. Architecture details are in Appendix G.

To learn the policy, we follow the MBPO paradigm [40] using a learned dynamics model $q_{\boldsymbol{\theta}}$ to train a PPO [53] policy $\boldsymbol{\pi}$. Specifically, we follow the Robotic World Model [40] pipeline. To prevent catastrophic failure and initialize the dynamics model with physics knowledge, we pretrain both the policy $\boldsymbol{\pi}$ and the dynamics model $q_{\boldsymbol{\theta}}$ in simulation with domain randomization of physics parameters [54]. Upon deployment, we transition to online learning to bridge the reality gap. At each iteration:

1) Collect a transition $(\mathbf{s}_t, \boldsymbol{a}_t, \hat{\boldsymbol{\phi}}, \mathbf{s}_{t+1})$ with the current policy $\boldsymbol{\pi}$, where $\hat{\boldsymbol{\phi}}$ is estimated using Eq. (6), and update $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \boldsymbol{a}_t, \hat{\boldsymbol{\phi}}, \mathbf{s}_{t+1})\}$.
2) Update dynamics $q_{\boldsymbol{\theta}}$ with the shortcut model objective [50] using data sampled from dataset $\mathcal{D}$.
3) Roll out imagined trajectories with domain randomization branched from states in $\mathcal{D}$ under $q_{\boldsymbol{\theta}}$, $\boldsymbol{\pi}$, and $p(\boldsymbol{\phi})$. Update $\boldsymbol{\pi}$ with rewards augmented by Alg. 1.

Exploration terminates when the trace of the parameter posterior covariance falls below δ_{var} and the dynamics prediction error falls below δ_{dyn}. The prediction error is measured as the horizon-H ℓ_2 deviation between real states and model rollouts.

V. Simulation Experiment

We evaluate the two questions raised in the preliminary section. **(Q1)** Does the QOED strategy collect information that improves identification of unknown parameters? **(Q2)** Does QOED improve policy learning compared with state-of-the-art methods? What is the contribution of each component? Finally, we compare the learned dynamics model against the purely analytical physics model.

Environments. We evaluate on seven environments in MuJoCo [55], covering locomotion and manipulation platforms: Unitree Quadruped `Go1-Flat` and `Go1-Rough` (54), Humanoid `G1-Flat` and `G1-Rough` (119), Clearpath Vehicle `Jackal-Flat` and `Jackal-Rough` (31), and the dexterous Inspire-FTP `Hand-Rotate` (104). The numbers in parentheses denote the dimensionality of the physical parameters. `Flat` and `Rough` indicate flat ground and uneven terrain, respectively. All policies are pretrained in mjlab [56] with 100 episodes, using the mjlab default configuration, and then run in MuJoCo with 20 episodes, on a single NVIDIA 4090 GPU. In MuJoCo, we use a single robot as a surrogate for real-world challenges, and we randomize physics coefficients by sampling from mjlab's default domain randomization distribution. We additionally test multiple noise levels, with the first level set to $\sigma = 0.025$. Full settings are provided in Appendix G.

A. Does QOED yield informative data?

In the MuJoCo evaluation, we analyze **Q1** using the RMSE of (i) parameter estimates and (ii) dynamics predictions. We compare QOED against two ablations. **(1) QOED-AGNOSTIC** only considers identifiable parameters [5, 6], defined in Eq. (9) as $\mathcal{B}_{\text{Agnostic}}(\boldsymbol{\phi}|\boldsymbol{\tau}_t) = \operatorname{tr}\left(\mathcal{F}_{\mathbf{k}\mathbf{k}}\right)$, which is agnostic to the discarded parameters. **(2) BOED** is the canonical and ideal formulation that uses all parameters, defined in Eq. (5) as

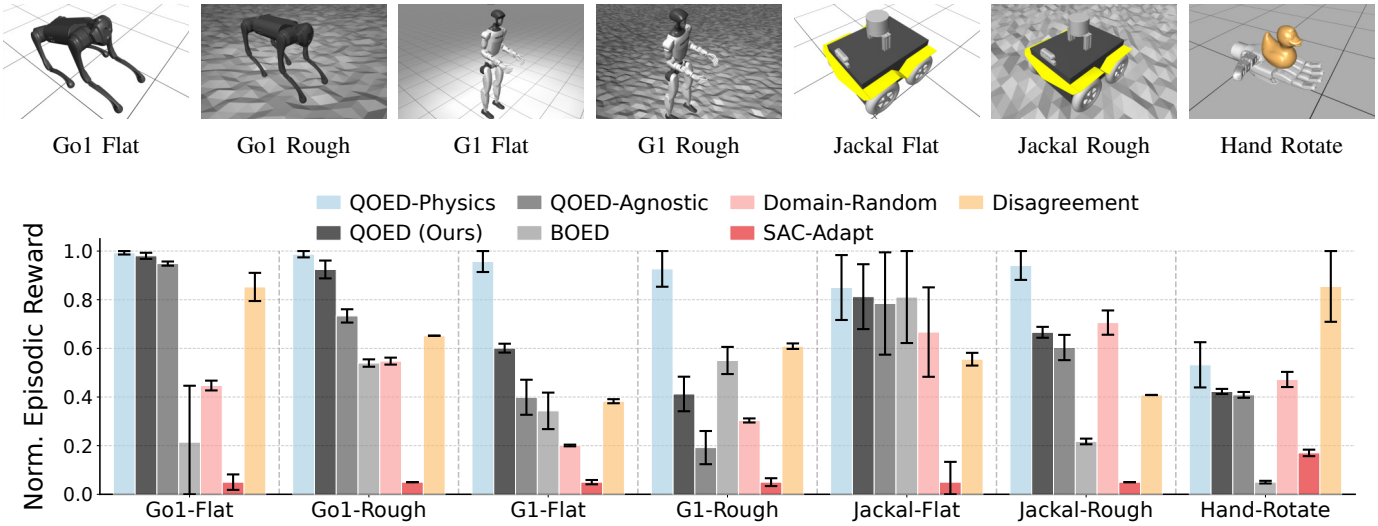

Fig. 3: Policy performance across diverse robot environments. QOED-PHYSICS with ground-truth physics consistently outperforms baselines, validating our adaptive information objective. QOED with learned dynamics also performs well, highlighting the promise of learned models for exploration.

$\mathcal{B}_{\text{BOED}}(\boldsymbol{\phi}|\boldsymbol{\tau}_t) = \text{tr}\,(\boldsymbol{\mathcal{F}}_{\boldsymbol{\phi}})$. We do not include an "observable-subspace" BOED variant because the observable directions $\mathbf{W}_{\text{o}}^{\top}\boldsymbol{\phi}$ need not map uniquely to physical parameters. All methods use the same CEM optimizer and learned dynamics model, running 5 optimization steps with 2048 samples per step. Following [57], we use $\alpha = 1$, $H = 2\,\text{s}$, $\delta_{\text{var}} = 0.05^2$, and $\delta_{\text{dyn}} = 1$ as defaults.

Results. Table I reports (i) the RMSE of parameter estimates and (ii) the RMSE of dynamics prediction when using the within-episode parameter estimate, averaged over 25 random seeds per scenario. Across all scenarios and noise levels, QOED achieves the strongest dynamics-prediction RMSE and competitive parameter-estimation RMSE. QOED improves dynamics prediction by $\mathbf{21.98}\%$ over QOED-AGNOSTIC and by $\mathbf{35.23}\%$ over BOED. BOED often produces poor estimates because it does not explicitly account for observability and can allocate exploration effort to directions that are hard to learn. QOED-AGNOSTIC restricts exploration to the identifiable subspace, but it still underperforms QOED because it treats the discarded parameters as irrelevant; in practice, these dropped directions can remain coupled to the critical ones and degrade estimation. Although the parameter estimation RMSE is similar across methods, the prediction RMSE differs because not all parameters contribute equally to the dynamics. QOED gathers data that is informative about the critical parameters while reducing the influence of the dropped ones, enabling CEM to recover the parameters that matter most for accurate prediction. Our CEM runtime averages 6.3 ms and FIM estimation runtime averages 28.3 ms. Overall, the results highlight the importance of identifying critical parameters and suppressing the influence of discarded directions.

Why dynamics prediction improves despite similar parameter errors? To explain why dynamics-prediction RMSE can improve even when aggregate parameter RMSE changes less, we perform a post-hoc attribution analysis. Specifically, by fixing all non-critical parameters to their ground-truth

Env.	Noise	QOED (Ours)		QOED-Agnostic		BOED	
		Param. Est.↓	Dyn. Pred.↓	Param. Est.↓	Dyn. Pred.↓	Param. Est.↓	Dyn. Pred.↓
Go1	1σ	848.29 ± 26.95	$\mathbf{28.57 \pm 4.56}$	$\mathbf{847.95 \pm 27.0}$	59.64 ± 18.08	848.15 ± 26.92	119.04 ± 94.74
	2σ	$\mathbf{847.89 \pm 27.0}$	$\mathbf{32.22 \pm 4.97}$	848.08 ± 26.99	73.16 ± 35.84	847.91 ± 27.0	59.36 ± 22.39
	3σ	847.98 ± 26.99	$\mathbf{37.91 \pm 5.17}$	833.94 ± 25.42	47.19 ± 10.54	$\mathbf{825.69 \pm 27.69}$	101.88 ± 9.51
G1	1σ	$\mathbf{963.02 \pm 33.73}$	$\mathbf{37.65 \pm 1.24}$	1059.73 ± 30.58	39.52 ± 1.75	1014.05 ± 56.02	44.93 ± 2.43
	2σ	$\mathbf{1051.11 \pm 35.28}$	$\mathbf{35.61 \pm 0.88}$	1063.02 ± 33.95	37.18 ± 1.02	1062.99 ± 33.74	40.73 ± 1.23
	3σ	1063.17 ± 33.72	$\mathbf{35.49 \pm 0.70}$	1063.11 ± 33.75	43.46 ± 1.27	$\mathbf{1063.05 \pm 33.73}$	44.52 ± 1.45
Jackal	1σ	$\mathbf{291.81 \pm 75.29}$	$\mathbf{1.00 \pm 0.13}$	369.74 ± 102.23	1.02 ± 0.11	370.74 ± 101.87	1.01 ± 0.08
	2σ	$\mathbf{296.90 \pm 66.54}$	$\mathbf{1.64 \pm 0.11}$	348.36 ± 91.93	1.73 ± 0.16	363.12 ± 99.15	6.57 ± 3.34
	3σ	$\mathbf{293.43 \pm 63.38}$	$\mathbf{2.64 \pm 0.25}$	360.77 ± 87.60	4.92 ± 2.40	375.36 ± 95.16	5.29 ± 2.38
Hand	1σ	$\mathbf{50.98 \pm 3.89}$	$\mathbf{3.55 \pm 0.57}$	52.68 ± 2.72	4.66 ± 0.68	52.68 ± 2.84	4.75 ± 0.69
	2σ	$\mathbf{51.74 \pm 3.7}$	$\mathbf{3.86 \pm 0.55}$	52.59 ± 2.99	4.57 ± 0.63	52.67 ± 2.72	4.72 ± 0.67
	3σ	$\mathbf{52.59 \pm 2.99}$	$\mathbf{4.8 \pm 0.61}$	53.18 ± 2.91	5.16 ± 0.69	52.68 ± 2.72	5.04 ± 0.7

TABLE I: Parameter-estimation and dynamics-prediction RMSE ($\times 100$), averaged over 25 seeds and (when available) over `Flat` and `Rough` environments. Lower is better; best results are in **bold**, and least favorable results are in gray.

values, we measure how much of the dynamics-prediction error is explained by the identified critical parameter set. The critical parameters account for 65.9% (Go1), 53.1% (G1), 51.9% (Jackal), and 47.2% (Hand) of the total dynamics-prediction error. For G1, for example, QOED identifies only 16 parameters (13.4% of the full parameter space), indicating that a compact identifiable subset can explain a disproportionate share of dynamics prediction.

Robustness to Hyperparameters. We assess QOED robustness to δ_{eig}, α_{eig}, and $\delta_{\cos}$ using a grid search. We sweep $\delta_{\text{eig}} \in [0.05, 0.5]$ (step 0.05), $\alpha_{\text{eig}} \in [0.005, 0.05]$ (step 0.005), and $\delta_{\cos} \in [0.9, 0.99]$ (step 0.01). In `Jackal` environment, QOED achieves an average dynamics prediction RMSE of 2.42 ± 0.11, while QOED-AGNOSTIC yields 4.77 ± 2.51. These results indicate that QOED is robust to the choice of hyperparameters over a wide range. While better settings may exist for specific tasks, we use our defaults because they are simple and perform well.

B. Does QOED help policy learning?

Baselines and Ablations. We compare against the following strong exploration baselines. **(1) SAC-ADAPT** (SAC [58]) adaptively tunes the weights of task and exploration rewards [59] to balance exploration and performance. **(2) DISAGREEMENT** (SAC) follows an explore-then-exploit schedule: it maximizes an exploration reward for the first 25% of interactions, then switches to maximizing task reward [60–62]. We use ensemble disagreement as the exploration reward, computed from an ensemble of five learned dynamics. **(3) DOMAIN-RANDOM** trains PPO with domain randomization (DR) [54] using mjlab settings. We also evaluate QOED ablations: **(4) QOED-PHYSICS** (finite differences with ground-truth simulator physics), **(5) QOED-AGNOSTIC** (matching prior settings [5, 6]), and **(6) BOED** [1]. PPO and SAC follow Stable-Baselines3 [63] (with adjustments in Appendix G). For fairness, all methods except QOED-PHYSICS use our learned dynamics for policy optimization.

Results. Fig. 3 answers **Q2** by reporting episodic *task* reward (excluding exploration reward), averaged over ten seeds. Results are normalized to $[0, 1]$ per environment for comparability. QOED-PHYSICS performs best across all environments since it has access to ground-truth physics, confirming that when the dynamics model is accurate, QOED can achieve strong policy performance. In contrast, BOED is greedy for information, leading to weaker policies. QOED-AGNOSTIC focuses on critical parameters but underperforms QOED because it ignores the influence of dropped parameters; as in the previous experiment, poorer dynamics estimates yield miscalibrated physics during policy optimization. A similar issue appears in DOMAIN-RANDOM, where broad domain-randomization priors do not efficiently improve online performance. DISAGREEMENT is a strong baseline by targeting regions of high dynamics-model uncertainty, an approach that is effective in model-based RL (e.g., PETS [64]) compared to SAC baselines. Nonetheless, in our robotics scenarios, conditioning the dynamics model on physics parameters yields further gains; see Scaffolder [65] for a broader discussion of incorporating privileged information in dynamics models. QOED performance drops in `G1-Rough` because the learned dynamics model does not reliably capture the critical parameters during early learning. Appendix H provides further result analysis of the learned dynamics model.

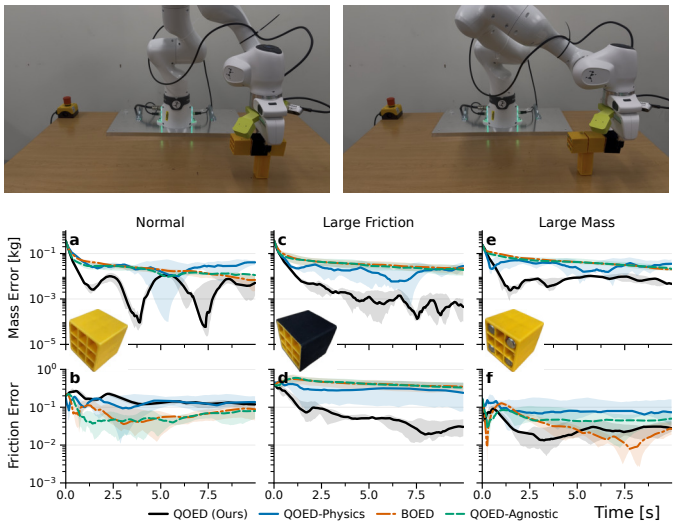

Fig. 4: Rod balancing demonstration and parameter-estimation error bars. Our QOED identifies the parameters quickly and accurately.

VI. REAL-WORLD EXPERIMENT

We evaluate **Q1** and **Q2** in real-world experiments on two platforms: a Franka Emika Panda arm (manipulation) and a Clearpath Jackal mobile vehicle (navigation).

A. Rod balance

This section primarily answers **Q1**, since the Franka setup provides high-accuracy reference measurements for comparison. We use the same ablations as in Sect. V-B. Note that QOED-PHYSICS updates the real-to-sim residual distribution using an unscented Kalman filter [66].

Environment. We use a Franka Emika Panda equipped with a RealSense D435i camera ($30\,\text{Hz}$) and an NVIDIA 4090 GPU. The robot must identify the mass, inertia, and friction of three cubes, then pick the cube with the highest friction and balance the stacked cubes (Fig. 4). We create three task variants by randomly stacking the cubes. For each variant and each method, we run six trials.

Results. Despite the apparent simplicity of the task, the ablations often fail. QOED achieves an $\mathbf{89}\%$ success rate, whereas QOED-PHYSICS, BOED, and QOED-AGNOSTIC achieve $\mathbf{42}\%$, $\mathbf{8}\%$, and $\mathbf{17}\%$, respectively. Failures are typically caused by center-of-mass misalignment: an error of only $2\,\text{cm}$ is sufficient to topple the stack, contributing to the low success rate of QOED-PHYSICS. During exploration, QOED adopts a simple objective-adaptation strategy—first identifying mass and inertia, then estimating friction—leading to faster parameter convergence (Fig. 4). The parameter estimation curve is not smoothed, as we opt to present the raw results rather

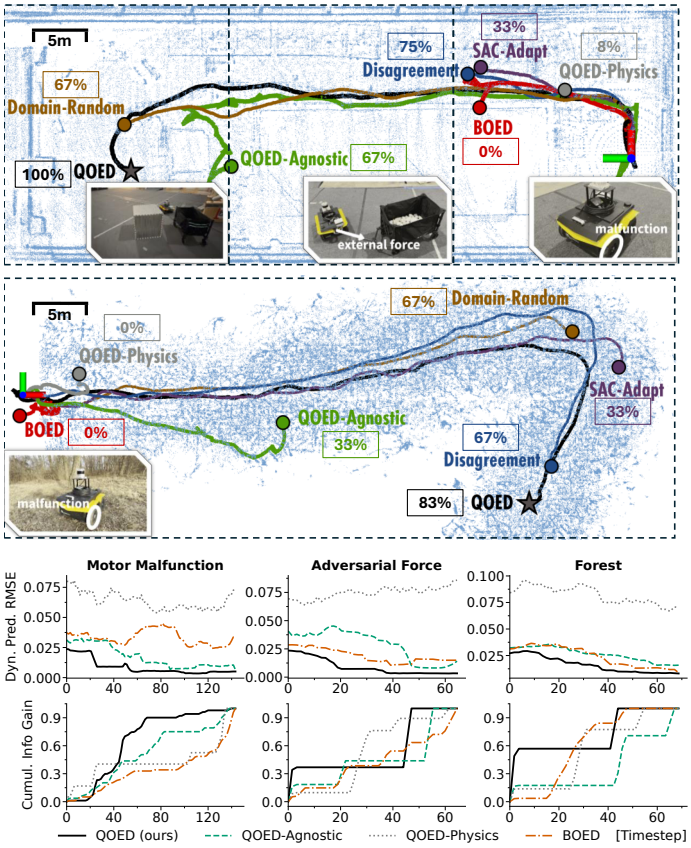

Fig. 5: Real-world snapshots with success rates shown in the text boxes. QOED achieves the highest success rate and the lowest dynamics prediction RMSE. By explicitly suppressing nuisance directions, it attains the highest cumulative information gain across environments.

than apply the filters. Although QOED-AGNOSTIC targets a similar objective-adaptation strategy, it does not account for the influence of friction when estimating mass, making the resulting exploration behavior (e.g., hefting the cube) less reliable. Even with the Franka arm's accurate sensing, jointly identifying mass/inertia and friction is challenging, which helps explain the poor performance of QOED-AGNOSTIC and BOED.

B. Wild Wheeled Navigation

This section further evaluates **Q1** and **Q2** in real-world navigation. Since ground-truth parameters are unavailable, we use the dynamics prediction RMSE as the primary indicator. We use the same baselines and ablations as in Sect. V-B.

Environment. We use a Clearpath Jackal equipped with an OS1-64 LiDAR for LiDAR–inertial odometry [67, 68] at 100 Hz and a RealSense D435i camera for navigation at 30 Hz; all computation runs on a Jetson Orin SoC. We evaluate three scenarios and a transition: `Motor Malfunction` (front-left wheel axle broken), `Adversarial Force` (towing a cart loaded with rocks), and `Forest` under `Motor Malfunction`. The robot must reach a goal while avoiding sparse obstacles. After `Adversarial Force`, we unload the trailer and move it to the `Forest` environment. For each scenario, we run six trials.

Results. Fig. 5 reports real-world trajectories together with success rate, dynamics prediction RMSE, and normalized cumulative information gain. QOED reaches the goal by quickly identifying key parameters, whereas baselines often crash due to miscalibrated physical coefficients. It also achieves the highest cumulative information gain across scenarios. QOED first identifies mass and friction, then diagnoses wheel-related effects and longitudinal force while suppressing other coupled directions, yielding faster and more stable estimation than QOED-AGNOSTIC; BOED estimation often fails to converge. In the recorded data, QOED spends $52.5\,\%$ of exploration time on mass and friction, whereas QOED-AGNOSTIC and BOED spend $87.5\,\%$ and $89.5\,\%$, respectively, which slows exploration and degrades estimation of motor and adversarial-force effects. In the real world, the learned dynamics model further improves performance over QOED-PHYSICS, benefiting from the data efficiency of MBRL. DISAGREEMENT is also strong due to its ensemble-based uncertainty signal, whereas SAC-ADAPT exhibits wobbling behaviors that are uninformative for policy learning and can lead to obstacle collisions. See Appendix H for additional discussions.

C. Discussions and Limitations.

Although the SAC-based baselines perform poorly in our scenarios, they are usually more general than our MBPO setup because we require physics parameterization to define the exploration objective. One way to relax this requirement is to learn a latent parameterization (e.g., with a variational autoencoder) and perform exploration in the latent space. In the real world, limited training time also leads to less refined behavior than in simulation: BOED-style objectives can be difficult to learn with RL, as observed in prior work [5, 6]. This creates a tension between objective complexity and exploration efficiency, which remains an important direction for future work.

VII. CONCLUSION AND FUTURE WORK

We presented *Quasi-Optimal Experimental Design* (QOED), an adaptive information objective for robot exploration. Grounded in optimal experimental design, QOED identifies critical parameters online and prioritizes them while suppressing nuisance directions. Extensive simulation and real-world experiments show that QOED collects informative data and efficiently reduces model error. We also demonstrate benefits in model-based policy optimization with a learned dynamics model, allowing QOED to outperform purely physics-based analytical models. Finally, we plan to extend QOED to broader robot learning problems, such as meta-learning RL update rules by treating them as parameters of interest, or optimizing the hyperparameters of differentiable algorithms, including differentiable MPC.

References

[1] Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

[2] Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher information. *arXiv preprint arXiv:2311.17874*, 2023.

[3] Yuhan Xie, Yixi Cai, Yinqiang Zhang, Lei Yang, and Jia Pan. GauSS-MI: Gaussian Splatting Shannon Mutual Information for Active 3D Reconstruction. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025.

[4] Matthew Strong, Boshu Lei, Aiden Swann, Wen Jiang, Kostas Daniilidis, and Monroe Kennedy. Next Best Sense: Guiding Vision and Touch with FisherRF for 3D Gaussian Splatting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3204–3210, 2025.

[5] Marius Memmel, Andrew Wagenmaker, Chuning Zhu, Dieter Fox, and Abhishek Gupta. ASID: Active Exploration for System Identification in Robotic Manipulation. In *The Twelfth International Conference on Learning Representations*, 2024.

[6] Nikhil Sobanbabu, Guanqi He, Tairan He, Yuxiang Yang, and Guanya Shi. Sampling-based system identification with active exploration for legged robot sim2real learning. *arXiv preprint arXiv:2505.14266*, 2025.

[7] Hrishikesh Sathyanarayan and Ian Abraham. Behavior Synthesis via Contact-Aware Fisher Information Maximization. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025.

[8] Parker Ewen, Hao Chen, Yuzhen Chen, Anran Li, Anup Bagali, Gitesh Gunjal, and Ram Vasudevan. You've Got to Feel It To Believe It: Multi-Modal Bayesian Inference for Semantic and Property Prediction. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024.

[9] C. Cao, H. Zhu, Z. Ren, H. Choset, and J. Zhang. Representation granularity enables time-efficient autonomous exploration in large, complex worlds. *Science Robotics*, 8(80):eadf0970, 2023.

[10] Richard Dearden, Nir Friedman, Stuart Russell, et al. Bayesian Q-learning. *Aaai/iaai*, 1998:761–768, 1998.

[11] Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through bayesian deep q-networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

[12] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and Exploration via Randomized Value Functions. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2377–2386, New York, New York, USA, 20–22 Jun 2016. PMLR.

[13] Ian Osband, Benjamin Van Roy, Daniel J. Russo, and Zheng Wen. Deep Exploration via Randomized Value Functions. *Journal of Machine Learning Research*, 20 (124):1–62, 2019.

[14] Johannes Kirschner and Andreas Krause. Information Directed Sampling and Bandits with Heteroscedastic Noise. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 358–384. PMLR, 06–09 Jul 2018.

[15] Thomas M Moerland, Joost Broekens, and Catholijn M Jonker. Efficient exploration with double uncertain value networks. *arXiv preprint arXiv:1711.10789*, 2017.

[16] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR, 2017.

[17] Borislav Mavrin, Hengshuai Yao, Linglong Kong, Kaiwen Wu, and Yaoliang Yu. Distributional reinforcement learning for efficient exploration. In *International conference on machine learning*, pages 4424–4434. PMLR, 2019.

[18] Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. Information-Directed Exploration for Deep Reinforcement Learning. In *International Conference on Learning Representations*, 2019.

[19] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. #exploration: a study of count-based exploration for deep reinforcement learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[20] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying Count-Based Exploration and Intrinsic Motivation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[21] Leshem Choshen, Lior Fox, and Yonatan Loewenstein. Dora the explorer: Directed outreaching reinforcement action-selection. *arXiv preprint arXiv:1804.04012*, 2018.

[22] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. First return, then explore. *Nature*, 590(7847):580–586, 2021.

[23] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

[24] Bradly C Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015.

[25] Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V.

Hafner. Intrinsic Motivation Systems for Autonomous Mental Development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286, 2007.

[26] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven Exploration by Self-supervised Prediction. In *ICML*, 2017.

[27] Misha Denil, Pulkit Agrawal, Tejas D Kulkarni, Tom Erez, Peter Battaglia, and Nando De Freitas. Learning to perform physics experiments via deep reinforcement learning. *arXiv preprint arXiv:1611.01843*, 2016.

[28] Andrew D. Wilson, Jarvis A. Schultz, and Todd D. Murphey. Trajectory Synthesis for Fisher Information Maximization. *IEEE Transactions on Robotics*, 30(6): 1358–1370, 2014.

[29] Rafael Oliveira, Dino Sejdinovic, David Howard, and Edwin V. Bonilla. Bayesian Adaptive Calibration and Optimal Design. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[30] Paul G. Constantine. *Active Subspaces*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2015.

[31] Sabina J. Sloman, Ayush Bharti, Julien Martinelli, and Samuel Kaski. Bayesian Active Learning in the Presence of Nuisance Parameters. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

[32] Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, Robust World Models for Continuous Control. In *The Twelfth International Conference on Learning Representations*, 2024.

[33] David Hoeller, Nikita Rudin, Dhionis Sako, and Marco Hutter. ANYmal parkour: Learning agile navigation for quadrupedal robots. *Science Robotics*, 9(88):eadi7566, 2024.

[34] Ziwen Zhuang, Shenzhe Yao, and Hang Zhao. Humanoid Parkour Learning. In *8th Annual Conference on Robot Learning*, 2024.

[35] Fabian Jenelten, Junzhe He, Farbod Farshidian, and Marco Hutter. DTC: Deep Tracking Control. *Science Robotics*, 9(86):eadh5401, 2024.

[36] Ankur Handa, Arthur Allshire, Viktor Makoviychuk, Aleksei Petrenko, Ritvik Singh, Jingzhou Liu, Denys Makoviichuk, Karl Van Wyk, Alexander Zhurkevich, Balakumar Sundaralingam, and Yashraj Narang. DeXtreme: Transfer of Agile In-hand Manipulation from Simulation to Reality. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5977–5984, 2023.

[37] Marcel Torne Villasevil, Anthony Simeonov, Zechu Li, April Chan, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. Reconciling Reality through Simulation: A Real-To-Sim-to-Real Approach for Robust Manipulation. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024.

[38] Newton Contributors. Newton: GPU-accelerated physics simulation for robotics, and simulation research., 2025.

[39] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to Trust Your Model: Model-Based Policy Optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[40] Chenhao Li, Andreas Krause, and Marco Hutter. Robotic world model: A neural network simulator for robust policy optimization in robotics. *arXiv preprint arXiv:2501.10100*, 2025.

[41] Runa Eschenhagen, Alexander Immer, Richard E Turner, Frank Schneider, and Philipp Hennig. Kronecker-Factored Approximate Curvature for Modern Neural Network Architectures. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[42] Mark J Schervish. *Theory of statistics*. Springer Science & Business Media, 2012.

[43] Friedrich Pukelsheim. *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006.

[44] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1032–1041. PMLR, 2019.

[45] Franz-Georg Wieland, Adrian L. Hauber, Marcus Rosenblatt, Christian Tönsing, and Jens Timmer. On structural and practical identifiability. *Current Opinion in Systems Biology*, 25:60–69, 2021. ISSN 2452-3100.

[46] Reuven Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and computing in applied probability*, 1(2):127–190, 1999.

[47] Tanner Schmidt, Richard Newcombe, and Dieter Fox. DART: Dense Articulated Real-Time Tracking. In *Proceedings of Robotics: Science and Systems*, Berkeley, USA, July 2014.

[48] C. Radhakrishna Rao. Minimum variance and the estimation of several parameters. *Mathematical Proceedings of the Cambridge Philosophical Society*, 43(2):280–283, 1947.

[49] Michel Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In J. Stoer, editor, *Optimization Techniques*, pages 234–243, Berlin, Heidelberg, 1978. Springer Berlin Heidelberg. ISBN 978-3-540-35890-9.

[50] Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One Step Diffusion via Shortcut Models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[51] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[53] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[54] Xiaoyu Chen, Jiachen Hu, Chi Jin, Lihong Li, and Liwei Wang. Understanding Domain Randomization for Sim-to-real Transfer. In *International Conference on Learning Representations*, 2022.

[55] Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[56] Kevin Zakka, Qiayuan Liao, Brent Yi, Louis Le Lay, Koushil Sreenath, and Pieter Abbeel. mjlab: A Lightweight Framework for GPU-Accelerated Robot Learning. 2026.

[57] Alejandro Murillo-González and Lantao Liu. Action Flow Matching for Continual Robot Learning. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025.

[58] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

[59] Eric Chen, Hong Zhang-Wei, Joni Pajarinen, and Pulkit Agrawal. Redeeming intrinsic rewards via constrained optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[60] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-Supervised Exploration via Disagreement. In *ICML*, 2019.

[61] Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to Explore via Self-Supervised World Models. In *ICML*, 2020.

[62] Russell Mendonca, Oleh Rybkin, Kostas Daniilidis, Danijar Hafner, and Deepak Pathak. Discovering and Achieving Goals via World Models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24379–24391. Curran Associates, Inc., 2021.

[63] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

[64] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[65] Edward S. Hu, James Springer, Oleh Rybkin, and Dinesh Jayaraman. Privileged Sensing Scaffolds Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[66] Alexander Schperberg, Yusuke Tanaka, Feng Xu, Marcel Menner, and Dennis Hong. Real-to-Sim: Predicting Residual Errors of Robotic Systems with Sparse Data using a Learning-Based Unscented Kalman Filter. In *2023 20th International Conference on Ubiquitous Robots (UR)*, pages 27–34, 2023.

[67] Chunge Bai, Tao Xiao, Yajie Chen, Haoqian Wang, Fang Zhang, and Xiang Gao. Faster-LIO: Lightweight Tightly Coupled Lidar-Inertial Odometry Using Parallel Sparse Incremental Voxels. *IEEE Robotics and Automation Letters*, 7(2):4861–4868, 2022.

[68] Youwei Yu, Yanqing Liu, Fengjie Fu, Sihan He, Dongchen Zhu, Lei Wang, Xiaolin Zhang, and Jiamao Li. Fast Extrinsic Calibration for Multiple Inertial Measurement Units in Visual-Inertial System. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 01–07, 2023.

[69] Youwei Yu, Junhong Xu, and Lantao Liu. Adaptive Diffusion Terrain Generator for Autonomous Uneven Terrain Navigation. In *8th Annual Conference on Robot Learning*, 2024.