Behavior-agnostic Task Inference for Robust Offline In-context Reinforcement Learning

Long Ma [1 2] **Fangwei Zhong** [3 2] **Yizhou Wang** [4 5 6]

Abstract

The ability to adapt to new environments with noisy dynamics and unseen objectives is crucial for AI agents. In-context reinforcement learning (ICRL) has emerged as a paradigm to build adaptive policies, employing a **context** trajectory of the test-time interactions to infer the true task and the corresponding optimal policy efficiently without gradient updates. However, ICRL policies heavily rely on context trajectories, making them vulnerable to distribution shifts from training to testing and degrading performance, particularly in offline settings where the training data is static. In this paper, we highlight that most existing offline ICRL methods are trained for approximate Bayesian inference based on the training distribution, rendering them vulnerable to distribution shifts at test time and resulting in poor generalization. To address this, we introduce Behavior-agnostic Task Inference (BATI) for ICRL, a model-based maximum-likelihood solution to infer the task representation robustly. In contrast to previous methods that rely on a learned encoder as the approximate posterior, BATI focuses purely on dynamics, thus insulating itself against the behavior of the context collection policy. Experiments on MuJoCo environments demonstrate that BATI effectively interprets out-of-distribution contexts and outperforms other methods, even in the presence of significant environmental noise.

[1]Center for Data Science, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China [2]Beijing Institute for General Artificial Intelligence, Beijing, China [3]School of Artificial Intelligence, Beijing Normal University, Beijing, China [4]School of Computer Science, Peking University, Beijing, China [5]Institute for Artificial Intelligence, Peking University, Beijing, China [6]State Key Laboratory of General Artificial Intelligence, Beijing, China. Correspondence to: Fangwei Zhong <fangweizhong@bnu.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

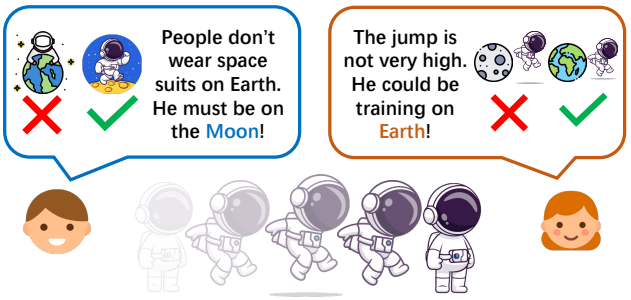


Figure 1. Illustration of the core idea behind BATI. The boy (left) infers the astronaut's location using **behavioral** cues (e.g., wearing space suits), while the girl (right) relies on physical **dynamics** (e.g., jump height). Although the former behavioral correlation is easier to spot, the latter is more robust to distribution shifts and leads to the correct answer in this example. BATI mirrors the latter strategy, prioritizing dynamics over behavior.

1. Introduction

The ability of AI agents to adapt to new environments with noisy dynamics and unknown objectives is becoming increasingly important as we push the boundaries of artificial intelligence applications. Meta-reinforcement learning (Finn et al., 2017; Duan et al., 2016; Beck et al., 2023) has emerged as a promising paradigm for developing adaptive policies. This approach leverages the concept of *learning to learn*, enabling agents to generalize from previous experiences and effectively tackle novel tasks quickly. Recently, the marriage of in-context learning (Brown et al., 2020; Min et al., 2022; Hendel et al., 2023) and meta-RL has attracted attention from the community, referred to as in-context reinforcement learning (ICRL) (Laskin et al., 2023; Grigsby et al., 2024; Ma et al., 2024). It leverages a **context** trajectory of interactions during testing to infer the true task and determine the corresponding optimal policy without any gradient update, showing significant potential for effective generalization across diverse and unknown environments. However, such context-conditioned policies can be expensive and time-consuming to train with online interactions due to the sample inefficiency of RL algorithms and the costs associated with online data collection (Yu, 2018; Gu et al., 2024). People thus turn to offline ICRL to extract in-context policies from **offline data** without online interactions (Laskin et al., 2023; Li et al., 2024).

Despite its promise, ICRL faces substantial challenges, as the performance of ICRL agents can be sensitive to the behavior shown in the context. During testing, any shifts in context distribution can lead to a marked decrease in performance (Gao et al., 2024). This issue is exacerbated when learning from offline data, where the distribution of training data is likely to differ from the conditions encountered during deployment. As a result, the robustness and generalization capabilities of offline ICRL methods are often compromised by **context shifts**.

We identify a critical limitation in existing offline ICRL methods (Li et al., 2021b; Yuan & Lu, 2022; Gao et al., 2024; Li et al., 2024): they are typically trained for approximate Bayesian inference based on the training distribution. This approach makes them vulnerable to distribution shifts at test time, undermining their ability to generalize to new, unseen contexts. Addressing this limitation is crucial for advancing the practical applicability of ICRL in real-world scenarios. Figure 1 shows an example of the effect of context shift when inferring the location of the jumping astronaut. In this case, the commonsense inference using behavioral characteristics leads to a wrong answer under distribution shifts, while the correct answer is given by an analysis of environmental dynamics.

In this paper, we propose a new **Behavior-agnostic Task Inference**[1] (BATI) framework to enhance the robustness of ICRL. BATI is a model-based maximum-likelihood approach that infers task representations without being influenced by the behavior of the context collection policy. Unlike previous methods that depend on a learned encoder to approximate the posterior, BATI focuses exclusively on the dynamics of the environment. This focus allows BATI to remain insulated from the variability introduced by different behavior policies, thereby improving its resilience to distribution shifts. We conduct extensive experiments in several MuJoCo environments to evaluate the effectiveness of BATI. Our results demonstrate that BATI not only interprets out-of-distribution contexts more effectively than existing methods but also outperforms them in environments with significant noise. These findings highlight the potential of BATI to enhance the adaptability and reliability of AI agents operating in complex and dynamic settings.

Our contributions are three-fold: 1) We identify and address a key vulnerability in existing ICRL methods related to context distribution shifts; 2) We introduce BATI, a robust task inference framework that enhances generalization by focusing on environmental dynamics; and 3) We validate the generalization of BATI through comprehensive experimental evaluations, setting a new benchmark for ICRL in noisy and unpredictable environments.

[1]Project page: https://sites.google.com/view/bati-icrl

2. Related Works

In-Context / Meta-Reinforcement Learning. In-context reinforcement learning (ICRL) aims to train agents that can generalize to solve new tasks using test-time interactions and reward signals, or **contexts** (Duan et al., 2016; Laskin et al., 2023; Grigsby et al., 2024; Gao et al., 2024; Li et al., 2024). ICRL falls into the broader category of meta-reinforcement learning (Beck et al., 2023), which encompasses both gradient-based methods (Finn et al., 2017; Song et al., 2020; Yoon et al., 2018) and context-based methods (where ICRL belongs) for learning new skills at test time. Critically, to improve the test-time efficiency(Ma et al., 2024), ICRL policies need to acquire new capabilities without any gradient updates, resembling the in-context learning phenomenon of large language models (Brown et al., 2020; Min et al., 2022; Hendel et al., 2023). In this paper, we focus on the ICRL problem to obtain agents that can efficiently adapt to new tasks without parameter updates. While ICRL has many appealing properties, an in-context policy can be very expensive or time-consuming to train due to the sample inefficiency of online RL algorithms and various costs of collecting online interactions in real-world scenarios (Grigsby et al., 2024; Yu, 2018; Gu et al., 2024). Offline ICRL (Li et al., 2024; Gao et al., 2024) has emerged to harness the advantages of both offline RL (Kostrikov et al., 2022; Wang et al., 2024; Zhong et al., 2025) and ICRL. Our work seeks to address the context shift problem in this offline setting.

Task Inference. Task inference methods (Humplik et al., 2019; Liu et al., 2021; Rakelly et al., 2019; Zintgraf et al., 2020; 2021) cast ICRL as a two-stage problem, where a latent representation of the true task is first inferred from the context and an in-context policy is conditioned on this latent to execute the corresponding optimal behavior. Previous works use supervision (Humplik et al., 2019), contrastive-like objectives (Li et al., 2021b), or RL losses (Rakelly et al., 2019) to guide the task inference. However, the context distribution may shift between training and test time, posing a great challenge to the generalization capabilities of task inference methods (Lin et al., 2020; Yuan & Lu, 2022; Li et al., 2024; Gao et al., 2024; Xu et al., 2024), especially in the offline setting. To address this **context shift**, CSRO (Gao et al., 2024) proposes to minimize the mutual information between the task latent and the context collection policy to promote the true correlation between the latent and the task. Furthermore, UNICORN (Li et al., 2024) provides a unified information-theoretic framework for understanding task inference methods and proposes a tighter approximation to the true objective. In this paper, we analyze flaws in previous works and propose a maximum-likelihood-based robust task inference method.

3. Preliminary

3.1. Problem Formulation

We define the ICRL problem on a distribution of Markov decision process (MDP) $M = (\mathcal{S}, \mathcal{A}, P, R, \rho_0, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition function, $R : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathbb{R}}$ is the reward function, $\rho_0 \in \Delta_{\mathcal{S}}$ is the initial state distribution, and γ is the discount factor. Hereafter we refer to each MDP as a **task**. We assume that all the tasks in support of the task distribution share the same $\mathcal{S}, \mathcal{A}, \rho_0$, and γ so that we can identify a task with P_M, R_M. Define **context** $X = \{(s_i, a_i, r_i, s'_i)\}_{i=0}^{C-1}$ as a set of C transition tuples in a task M; further denote its components as $X^b = \{(s_i, a_i)\}_i, X^t = \{(r_i, s'_i)\}_i$. Such a context could be generated by rolling out a (plain) policy $\mu : \mathcal{S} \to \Delta_{\mathcal{A}}$ in M, so $s_0 \sim \rho_0, a_i \sim \mu(s_i), r_i \sim R_M(s_i, a_i), s'_i = s_{i+1} \sim P_M(s_i, a_i) \forall i$, in which case we write the context random variables as $\mathbf{X}_{M,\mu}, \mathbf{X}^b_{M,\mu}, \mathbf{X}^t_{M,\mu}$.

Our objective is to learn an **in-context policy**, or meta-policy $\pi_\theta : \mathcal{S} \times \mathcal{X} \to \Delta_{\mathcal{A}}$ parameterized by θ to optimize the following discounted objective:

$$\max_\theta \mathbb{E}_{M,\mu \sim p(M,\mu), X_{M,\mu} \sim \mathbf{X}_{M,\mu}} \left[\sum_{t \geq 0} \gamma^t r_t \right] \tag{1}$$

where r_t is generated by rolling out $\pi_\theta(\cdot \mid \cdot, X_{M,\mu})$ in M, $\mathcal{X} = (\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S})^*$ is the context space. As the task M is sampled from a distribution and not directly revealed to π_θ, π_θ may only learn about the properties of the current task it needs to solve via the sampled context $X_{M,\mu}$.

Finally, to characterize the context shift problem, note that the above objective is an expectation over the joint distribution of task M and context collection policy μ. Denote the training distribution as $p_{\text{train}}(M, \mu)$ and the testing distribution as $p_{\text{test}}(M, \mu)$. For the same task M, the training context $X_{M,\mu_{\text{train}}}$ induced by $\mu_{\text{train}} \sim p_{\text{train}}(\mu \mid M)$ may have a very different distribution from the testing context $X_{M,\mu_{\text{test}}}, \mu_{\text{test}} \sim p_{\text{test}}(\mu \mid M)$, forming a **context shift**.

3.2. Background & Analysis

In recent years, as interest in ICRL and meta-RL increased, many methods have been developed for training effective in-context policies. Among the existing works, task inference approaches consider ICRL to be the problem of inferring a latent task representation from a given test-time interaction context. Once the test-time task is known, the optimal policy can be determined and executed to achieve a good performance efficiently without time-consuming and unstable test-time gradient updates. Early works thus focus on effectively extracting task information from the context using a **context encoder**, with the main differences being the learning objectives used (Humplik et al., 2019; Ren et al., 2022; Sohn et al., 2020; Zhang & Kan, 2022; Lee et al., 2019; Peng et al., 2021; Kamienny et al., 2020). For example, Humplik et al. (2019) assumes that supervision of task parameters is available during training, while FOCAL (Li et al., 2021b) takes a distance metric learning approach.

However, in an offline learning setting, where the ICRL policy needs to be extracted from a fixed dataset without any online interactions, **context shift** becomes a major concern. The context distribution may dramatically change between the offline training and the online testing, causing the context encoder to go out of distribution and produce incorrect task encodings, which leads to unsatisfactory policy performance. Among works that have sought to address the context shift problem (Lin et al., 2020; Li et al., 2024; Gao et al., 2024; Yuan & Lu, 2022; Li et al., 2021b;a), recently, UNICORN (Li et al., 2024) proposed an information-theoretic framework for offline ICRL, arguing that the task representation learning should be done by optimizing the following mutual information objective:

$$\max_\phi I(\mathbf{Z}; \mathbf{M}) \tag{2}$$

where $Z := f_\phi(X)$ is the context encoding, f_ϕ is the context encoder parameterized by ϕ. Intuitively, $\mathbf{Z}$ should capture the part of $\mathbf{X}$ that describes the true task $\mathbf{M}$. From this perspective, UNICORN proved that the objectives of several prior works can be streamlined as approximations or bounds of the mutual information objective. It proposed an alternative objective to achieve a tighter and more robust approximation. A classification-like loss can also be used to directly optimize Eq. 2 using task indices.

However, we observe that this framework fails to formally characterize the change of $I(\mathbf{Z}; \mathbf{M})$ under context shift. In particular, its modeling of the ICRL problem does not take account of the distribution of the context-collection policy μ. Expanding Eq. 2 under our formulation in Sec. 3.1 yields

$$\begin{aligned} & I(\mathbf{Z}; \mathbf{M}) \\ = \; & H(\mathbf{M}) - H(\mathbf{M} \mid \mathbf{Z}) \\ = \; & H(\mathbf{M}) + \mathbb{E}_{M,Z}\, p(M \mid Z) \\ = \; & H(\mathbf{M}) + \mathbb{E}_{M,\mu}\, \mathbb{E}_{X_{M,\mu}}\, p(M \mid f_\phi(X_{M,\mu})) \end{aligned}$$

This derivation reveals that the mutual information objective depends on the joint distribution $p(M, \mu)$, which may shift between training and testing time. As a result, **even if we were able to optimize the true objective on the training data, the testing MI is still not guaranteed to be large**. Furthermore, the posterior estimate $p(M \mid f_\phi(X_{M,\mu}))$ is also dependent on $p(M, \mu)$ and may give incorrect estimates about M in an out-of-distribution scenario, hurting the policy performance as shown in our experiments.

To address the discrepancy, CSRO (Gao et al., 2024) observed that context shifts are introduced when the testing

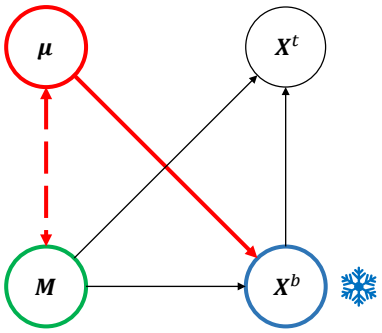


Figure 2. The graphical model of our proposed formulation. The dotted red arrow indicates the joint distribution of the task $\mathbf{M}$ and the context collection policy μ, which may change between training and testing, causing the context $\mathbf{X}^b, \mathbf{X}^t$ to shift. However, after conditioning on $\mathbf{X}^b$ (blue node), the effect of μ is blocked (solid red arrow), so we can safely infer $\mathbf{M}$ (green node) via $\mathbf{X}^t$.

contexts are collected by a different policy from the training one, echoing our formulation. Subsequently, CSRO proposed to maximize the **true correlation** $I(\mathbf{Z};\mathbf{M})$ while minimizing the **spurious correlation** $I(\mathbf{Z};\mathbf{X}^b)$ to decorrelate $\mathbf{Z}$ with the behavior of the context collection policy contained in the state-action pair X^b. However, as UNICORN noted, this mixed objective poses a trade-off because the two components are sometimes contradictory. For example, consider a bandit-like environment where every task corresponds to a specific action and training contexts always take that action. In this case, the mutual information between the task $\mathbf{M}$ and the behavior $\mathbf{X}^b$ is already high, and it's not possible to find a suitable $\mathbf{Z}$ with high $I(\mathbf{Z};\mathbf{M})$ yet low $I(\mathbf{Z};\mathbf{X}^b)$. More formally, we have

Theorem 3.1. *For arbitrary task representation Z,*

$$I(\mathbf{Z};\mathbf{M}) - I(\mathbf{Z};\mathbf{X}^b) \leq H(\mathbf{M}) - I(\mathbf{M};\mathbf{X}^b) \tag{3}$$

See App. A for the proof. Consequently, when the training context collection policy is highly correlated with the task, the competing objectives of CSRO cannot be achieved simultaneously, and a good encoder (in terms of the CSRO objective) does not exist.

4. Method

4.1. Behavior-agnostic Task Inference

In the analysis of previous works, we note that all of them fit a context encoder f_ϕ to estimate the posterior of the task variable $p(M \mid X)$ using different objectives and regularizers. We argue that this direct Bayesian inference approach is inherently flawed, since

$$p(M \mid X) \propto \int_\mu p(M,\mu)p(X \mid M,\mu)d\mu \tag{4}$$

which inevitably depends on the joint distribution $p(M,\mu)$. This is illustrated in the graphical model of our formulation (Fig. 2), where the inference of M is disrupted by the shifting distribution of μ, which subsequently contaminates the distributions of X^b and X^t.

To circumvent this failure mode, we make a further observation from the graphical model that we can remove the influence of μ by blocking $\mathbf{X}^b$. Once conditioned on $\mathbf{X}^b$ and the variable to be inferred $\mathbf{M}$, all paths going from μ to $\mathbf{X}^t$ are blocked while μ and $\mathbf{X}^t$ become independent, so we can correctly infer $\mathbf{M}$ using $\mathbf{X}^t$. Taking advantage of this observation, we propose **behavior-agnostic task inference** (BATI), a maximum-likelihood-based solution to replace the Bayesian posterior inference of $p(M \mid X)$:

$$\arg\max_M \log p(X^t \mid X^b, M) \tag{5}$$

This term, corresponding to the environment dynamics of M and irrelevant of μ, can now be safely estimated from offline data and transferred to online inference without worrying about the shifted distribution of μ.

We can give another interpretation to this solution by viewing it as a *robust* version of the full likelihood $p(X \mid M)$:

$$\begin{aligned} & p(X \mid M) \\ = \; & \textstyle\int_\mu p(\mu \mid M)p(X \mid M,\mu)d\mu \\ = \; & \textstyle\int_\mu p(\mu \mid M)p(X^b \mid M,\mu)p(X^t \mid X^b, M,\mu)d\mu \\ = \; & p(X^t \mid X^b, M)\textstyle\int_\mu p(\mu \mid M)p(X^b \mid M,\mu)d\mu \\ \approx \; & p(X^t \mid X^b, M) \end{aligned}$$

The third equality stems from the properties of MDP, where the reward and the next state depend only on the state-action pair in any given MDP and are independent of the overall policy. In this expansion, only the integral over μ is affected by the distribution shift. To estimate it, we would need information about the test-time joint distribution $p(M,\mu)$ and the behaviors of every μ in every M, a tall order to fulfill. We thus assume that the integral is approximately the same over different tasks and ignore this term. Empirically, we find this approximation to have satisfactory performance with extensive experiments in Sec. 5.

Given our analysis above, as direct Bayesian inference is problematic, why do existing methods still work to some extent? Expanding Eq. 4 in a similar manner as above yields

$$\begin{aligned} & p(M \mid X) \\ \propto \; & \textstyle\int_\mu p(M,\mu)p(X \mid M,\mu)d\mu \\ = \; & \textstyle\int_\mu p(M,\mu)p(X^b \mid M,\mu)p(X^t \mid M,\mu,X^b)d\mu \\ = \; & p(X^t \mid M,X^b)\textstyle\int_\mu p(M,\mu)p(X^b \mid M,\mu)d\mu \end{aligned}$$

which also includes the robust term $p(X^t \mid M, X^b)$. Furthermore, we observe that the evaluation environments of previous works have **deterministic dynamics**, which means that $p(X^t \mid M, X^b)$ is close to a delta function. Consequently, the true correlation may nevertheless dominate the Bayesian

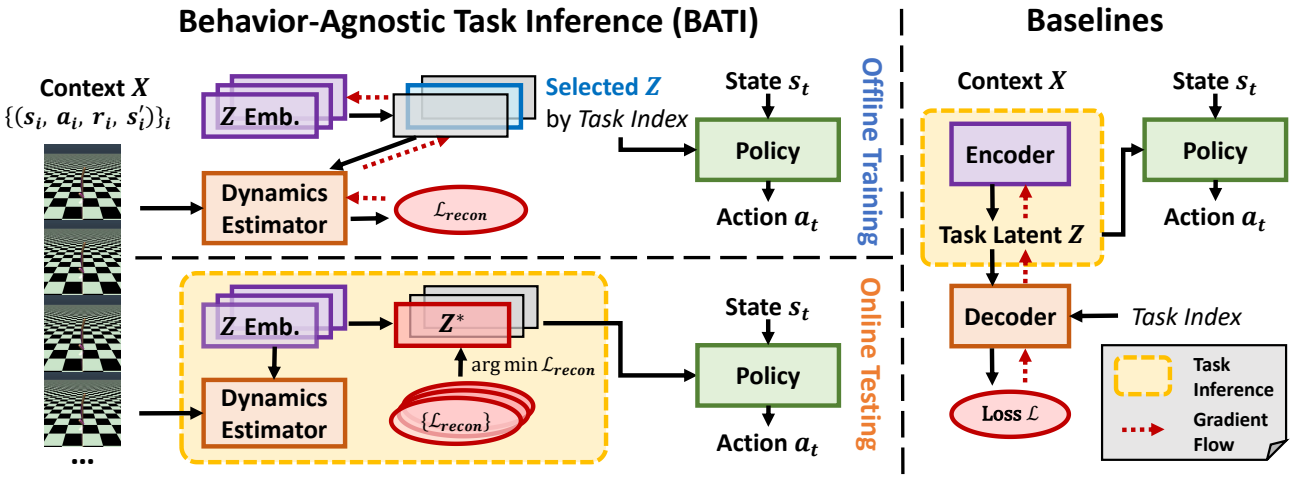


Figure 3. Illustrations of Behavior-Agnostic Task Inference (BATI, left) and baselines (right). Baselines use the same encoder to encode the context X into a task representation Z for both offline training and online testing, in effect performing Bayesian inference over the training distribution. In contrast, BATI uses the task index during training and searches for an optimal task latent during online evaluation, avoiding the influence of context shifts.

inference and overwhelm the shifted integral. However, when the dynamics are **noisy**, as is common in real-world scenarios, the discriminative power of $p(X^t \mid M, X^b)$ might be weakened, leading the fitted posterior to rely on the spurious correlation $p(X^b \mid M, \mu)$ as a shortcut (Geirhos et al., 2020). We demonstrate this sensitivity to noise with our ablations in Sec. 5.4.

4.2. Offline Training Pipeline

We now describe the instantiation of BATI and our offline training pipeline, as shown in Fig. 3. To estimate $p(X^t \mid X^b, M)$ without ground-truth task parameterizations, we build a table of task embedding distributions $\{\mathbf{Z}_\phi^M\}_M$ (Fig. 3, stacked purple rectangles) containing an entry for each training task. We parameterize the Gaussian distributions with ϕ, replacing the learned encoder f_ϕ in previous works. Both the task embeddings and the dynamics estimator (orange rectangle) are supervised by the following:

$$\min_{\phi,\psi} \mathbb{E}_{M\sim p_{\text{train}}(M), X\sim\mathcal{D}_M, Z\sim\mathbf{Z}_\phi^M} \mathcal{L}_{\text{recon}}^{X,Z}(\phi,\psi) \tag{6}$$

where

$$\mathcal{L}_{\text{recon}}^{X,Z}(\phi,\psi) := \frac{(X^t - g_\psi(X^b, Z))^2}{\exp h_\psi(X^b, Z)} + h_\psi(X^b, Z) \tag{7}$$

is the **dynamics reconstruction loss** as the negative log-likelihood of $\mathcal{N}\left(g_\psi(X^b, Z); \exp h_\psi(X^b, Z)\right)$, $p_{\text{train}}(M)$ is the training task distribution, $\mathcal{D}_M$ is the offline training dataset for task M, and g_ψ, h_ψ represent the dynamics estimator.

To train the in-context policy π_θ, we use IQL (Kostrikov et al., 2022) as the offline RL algorithm. During training, π_θ and the value functions receive the same task embedding as the dynamics estimator, and all modules are trained simultaneously. However, gradients from the policy and value functions are detached from the task embeddings, which are supervised exclusively with $\mathcal{L}_{\text{recon}}$. See Alg. 1 for the pseudocode of the full training pipeline. Note that both the dynamics estimator and the policy are trained with purely offline data without any online interactions.

4.3. Online Evaluation Procedure

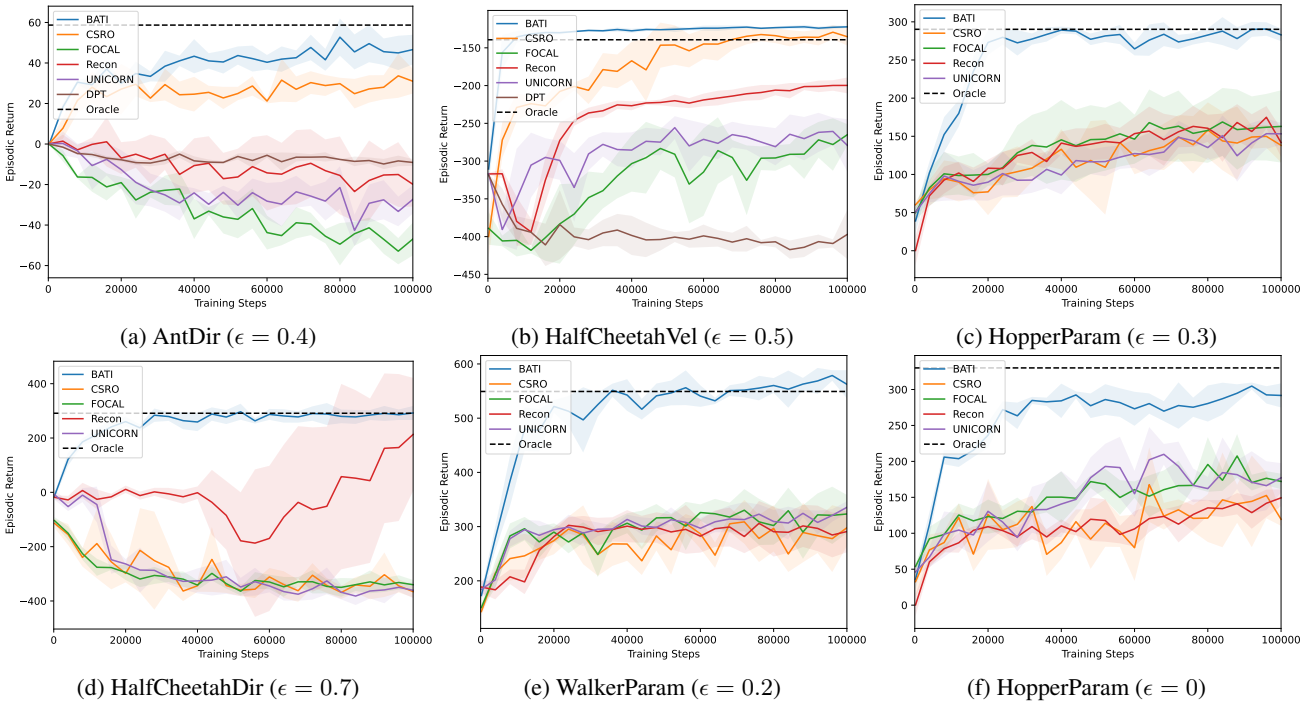
During the online evaluation, given contexts X collected by an unknown policy for an unknown task sampled from $p_{\text{test}}(M, \mu)$, we perform the proposed maximum-likelihood optimization in Eq. 5. Specifically, we compute

$$\arg\min_{Z^*} \mathcal{L}_{\text{recon}}^{X,Z^*} \tag{8}$$

with Z^* sampled from $\{\mathbf{Z}_\phi^M\}_M$ for a fixed number of times N. The in-context policy π_θ takes Z^* as input and interacts with the test task to evaluate its performance. See Alg. 2 for the online evaluation procedure.

5. Experiments

We empirically validate BATI's performance in several MuJoCo-based environments commonly used for offline ICRL. With the experiments in this section, we aim to answer the following questions: 1) Can BATI achieve robust performance in the presence of significant context shifts? 2) How does BATI's inference procedure compare with a learned encoder? 3) How does the noise level impact the performance of BATI and baselines? 4) Can BATI outperform baselines in noiseless environments? 5) How does



(a) AntDir ($\epsilon = 0.4$) (b) HalfCheetahVel ($\epsilon = 0.5$) (c) HopperParam ($\epsilon = 0.3$)

(d) HalfCheetahDir ($\epsilon = 0.7$) (e) WalkerParam ($\epsilon = 0.2$) (f) HopperParam ($\epsilon = 0$)

Figure 4. Online evaluation episodic return curves of BATI and baselines on contexts from $p_{\text{test}}(M, \mu)$ during training in our evaluation environments. BATI consistently achieves the best performance in all environments and settings, converging faster and more stably.

BATI scale with more context data? 6) Can BATI handle OOD contexts and tasks simultaneously?

5.1. Setup

We choose five representative robot locomotion environments based on the MuJoCo simulator (Todorov et al., 2012) with varying properties and levels of difficulty. Among the environments, **AntDir**, **HalfCheetahVel**, and **HalfCheetahDir** have different reward functions R_M, while **HopperParam** and **WalkerParam** have parameterized dynamics P_M. To simulate real-world scenarios more closely, which are often **non-deterministic and noisy**, we perturb the dynamics of the environments with various noises of scale ϵ. As noted in Sec. 4.1, dynamics noise weakens the true task correlation and adds to the challenge of task inference. The details of the environments are in App. C.1. We also conduct an ablative study in Sec. 5.4 demonstrating the effect of noise on baselines, compared to BATI which remains unscathed in all scenarios.

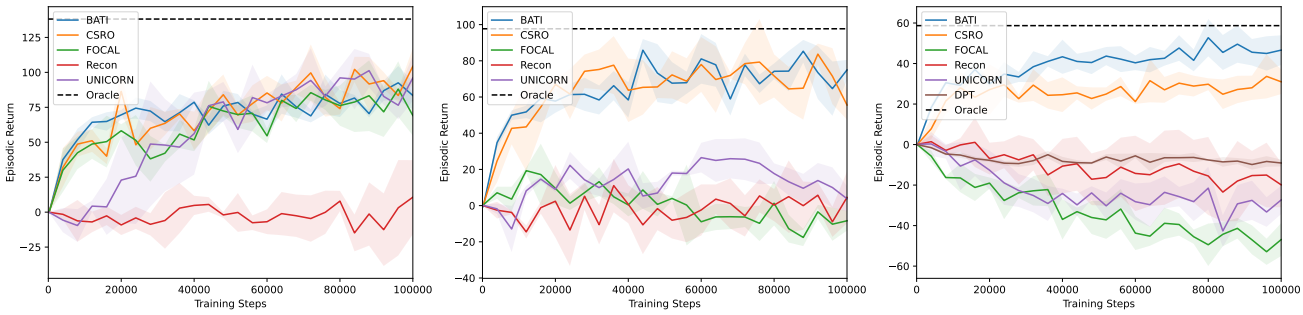
In each evaluation environment, we randomly sample 20 tasks as $p_{\text{train}}(M)$ and another 20 tasks as $p_{\text{test}}(M)$ according to its task parameterization, e.g. target directions in AntDir or physics parameters in HopperParam (see App. C.1 for details). During the online evaluation, the true task indices or parameters are **not directly provided** to the in-context policy and must be **inferred** from the context. To construct contexts with a large shift, we create a policy μ_M as the primary context collection policy associated with each task M and set

$$p_{\text{train}}(M, \mu) = p_{\text{train}}(M) \cdot \begin{cases} 0.9, & \mu = \mu_M \\ \frac{0.1}{|p_{\text{train}}(M)|-1}, & \mu \neq \mu_M \end{cases} \tag{9}$$

$$p_{\text{test}}(M, \mu) = p_{\text{test}}(M) \cdot \mathbb{I}[\mu = \mu_{\bar{M}}] \tag{10}$$

where $\bar{M}$ is a task **most different** from M. This creates an extremely challenging testing distribution, as the in-context policy will be misguided by the context to execute behavior that is good for an adversarial task but quite bad for the true task. Unless otherwise stated, all evaluation results in this section are on $p_{\text{test}}(M, \mu)$. See App. C.2 for details on dataset construction.

We compare BATI with representative baselines from the field of offline ICRL: a) **UNICORN** (Li et al., 2024), a recent state-of-the-art method using an information-theoretic objective, we use the UNICORN-SS variant reported to have better performance; b) **CSRO** (Gao et al., 2024), a strong baseline that promotes robustness by minimizing mutual information between the task representation and the context behavior; c) **FOCAL** (Li et al., 2021b), a classic method that uses distance metric learning for self-supervised task inference; d) **Recon**, a model-based method that uses the same reconstruction loss as BATI but with a learned encoder for inference. This is also called UNICORN-SS-0 or GENTLE (Zhou et al., 2024) and serves as an ablative baseline to isolate the effects of our task inference procedure. In AntDir



(a) $\epsilon = 0$ (b) $\epsilon = 0.2$ (c) $\epsilon = 0.4$ (Default)

Figure 5. Online evaluation episodic return curves during training for different noise levels on AntDir. With the noise level rising, baselines increasingly rely on the behavior of the context collection policy μ to determine the true task, causing their OOD performance to become flat (b) or even decrease throughout training (c). At the same time, BATI consistently improves during training and performs similarly to or better than the baselines for all noise levels.

and HalfCheetahVel, we also compare with **DPT** (Lee et al., 2024), a recent transformer-based method for ICRL. We implement BATI and the baselines on the same codebase with IQL (Kostrikov et al., 2022) as the base offline RL algorithm. See App. C.3 for implementation details.

5.2. Can BATI achieve robust performance in the presence of significant context shifts?

In this section, we report the online evaluation results on $p_{\text{test}}(M, \mu)$ as described in the previous section. Each online evaluation rollouts the in-context policy in all 20 testing environments with their respective contexts (Alg. 2) and returns the average episodic return. Fig. 4 shows the episodic return curves during training, while Tab. 4 contains the numerical results at convergence, computed as the average of the final 5 evaluations of each training run. Standard deviations are reported over 5 training seeds. Across all environments, BATI outperforms the baselines, usually by a large margin. In AntDir and HalfCheetahDir, we observe that the evaluation performances of several baselines decrease over time, indicating the gradual learning of spurious correlation. As a comparatively strong baseline, CSRO performs somewhat similarly to BATI in AntDir and HalfCheetahVel but converges more slowly and is less stable. It is dramatically outperformed in the more difficult environments. The reason for this underperformance and slow learning is that the competing objectives of CSRO cannot be fulfilled simultaneously when the context collection policy is highly correlated with the true task (see Sec. 3.2).

5.3. How does BATI's inference procedure compare with a learned encoder?

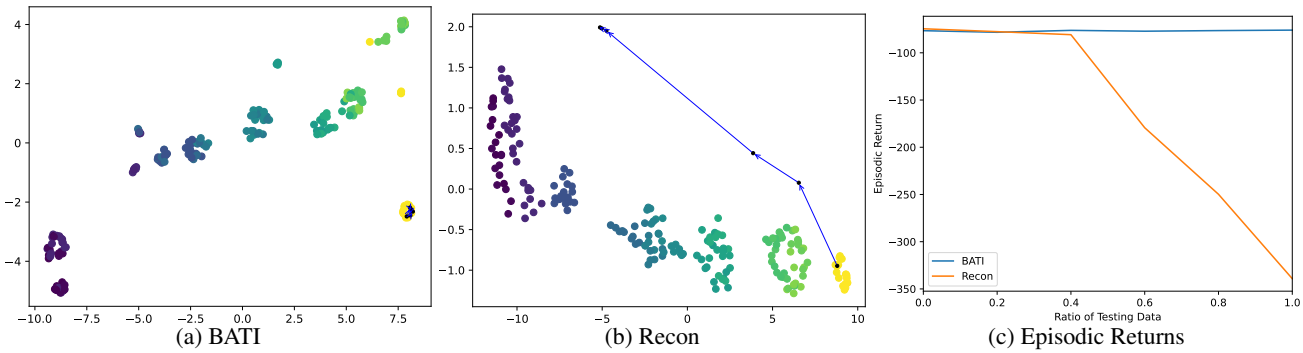
To demonstrate the effect of BATI's maximum-likelihood inference procedure, we compare BATI with the Recon baseline (also known as GENTLE and UNICORN-SS-0) which shares the same reconstruction objective $\mathcal{L}_{\text{recon}}$ with BATI. The key difference is that Recon encodes the full trajectory X to produce Z, which is then used with X^b to decode X^t. In contrast, BATI uses an embedding during training and maximum-likelihood inference during testing for Z. As shown in Tab. 4 and Fig. 4, Recon underperforms BATI in every environment and can have very high variances in certain cases (e.g. HalfCheetahDir). In Fig. 4a, the performance of Recon decreases over time, indicating the capture of spurious correlations. This result reinforces our argument in Sec. 3.2 that a learned encoder performs approximate Bayesian posterior inference and cannot handle context shifts well.

5.4. How does the noise level impact the performance of BATI and baselines?

We now show the effect of dynamics noise on BATI and baselines in AntDir and HalfCheetahDir, two representative environments with continuous (AntDir) and discrete (HalfCheetahDir) task parameterizations. We rerun the experiments with lower or no dynamics noise. As shown in Figures 5a to 5c above and Figures 7a to 7c in App. D, increasing dynamics noise progressively destabilizes baselines, while BATI maintains robust performance. The phenomenon of spurious correlation capture for baselines is especially evident in AntDir. While most baselines can achieve a decent performance in the noiseless setting (Fig. 5a), they fail to learn with a medium level of noise (Fig. 5b) and even grow steadily worse as learning progresses in the default high-noise setting (Fig. 5c). This finding supports our analysis in Sec. 4.1 that noisy dynamics weaken the true correlation, forcing the learned Bayesian posterior to rely more on the spurious correlation of the context collection policy as a shortcut.

5.5. Can BATI outperform baselines in noiseless environments?

With a deeper understanding of the role of noise, we now further show that baselines may fail even without any noise in



(a) BATI (b) Recon (c) Episodic Returns

Figure 6. t-SNE visualizations of inferred task latents of BATI (a) and Recon (b) in HalfCheetahVel. The dots denote the latents of all tasks on p_{train}, and colors indicate the target velocities. Black dots linked by blue arrows are those inferred from interpolated contexts of a specific task. The corresponding online evaluation episodic returns are shown in (c). As the ratio of out-of-distribution p_{test} contexts rises, the latents inferred by Recon drift off in the direction of the spurious task (from lower-right to upper-left of (b)), At the same time, BATI consistently produces correct and in-support latent representations (lower-right of (a)), resulting in better performance.

Table 1. Online evaluation episodic returns of the final checkpoints of BATI and baselines on HalfCheetahDir under various context lengths. All policies are trained with $C = 50$ (*) and $\epsilon = 0.9$.

	$C = 50$ (*)	$C = 200$	$C = 400$
BATI	-4.7 ± 1.2	1.8 ± 0.5	4.7 ± 1.5
CSRO	-23.8 ± 1.0	-24.7 ± 0.7	-25.3 ± 1.1
FOCAL	-25.2 ± 0.2	-25.6 ± 0.3	-26.6 ± 0.2
Recon	-11.4 ± 0.2	-12.2 ± 0.3	-12.0 ± 0.2
UNICORN	-22.8 ± 2.3	-24.0 ± 1.9	-24.8 ± 2.4

hard scenarios. Fig. 4f shows the performances of BATI and baselines in HopperParam with no dynamics noise. Compared with the noisy results in Fig. 4c, baselines still underperform BATI with only modest gains for the UNICORN baseline and little change for the others. As explained in Sec. 4.1, baselines fail when the discriminative power of the true correlation $p(X^t \mid M, X^b)$ is weakened. The dynamics of HopperParam are sufficiently complicated and nonlinear that $p(X^t \mid M, X^b)$ is already hard to estimate even under a noiseless setting. Baselines thus opt to learn the shortcut of $p(X^b \mid M, \mu)$ instead and fail when the context collection policy changes.

5.6. How does BATI scale with more context data?

A crucial desideratum for ICRL is the ability to keep improving with more context data. In this section, we demonstrate this property for BATI in an extremely challenging setting. We choose HalfCheetahDir, an environment with two discrete task variants (going left or right), and conduct an experiment with very high noise ($\epsilon = 0.9$) and short training context length ($C = 50$). Under this setting, the reward for each time step is computed in the true direction with probability 0.55 and in the other direction with 0.45, so the ICRL policy may only obtain an exceedingly weak signal of the true direction with each time step. Coupled with a short time horizon, this setting presents an enormous challenge to ICRL algorithms. Tab. 1 contains the performances of BATI and baselines trained in the setting described above and tested with different context lengths. BATI outperforms all baselines with the shortest context length $C = 50$ and improves continuously when provided with longer contexts. As the context grows, BATI can estimate log-likelihoods with lower variance, leading to improved quality of the inferred task latent and better policy performance. However, the baselines do not exhibit such a pattern and fluctuate around the initial performance, even getting slightly worse.

5.7. Can BATI handle OOD contexts and tasks simultaneously?

In addition to generalizing to OOD contexts, we demonstrate BATI's OOD task generalization capability through experiments in AntDir. OOD task generalization coupled with context generalization is even more challenging, since the policy is forced to adapt to an unfamiliar task that differs dramatically from those seen during training and requires very different behaviors to solve. We sample target directions from $[0, \pi)$ during training while using directions from $[\pi, 2\pi)$ for testing. Results are presented in Tab. 2. While the performance of BATI decreases compared with the in-distribution case, it still outperforms baselines by a large margin and is the only method to achieve positive returns. This further demonstrates the generalization capability of BATI to both OOD contexts and tasks.

5.8. t-SNE Visualization with Interpolated Contexts

In this section, we show t-SNE visualizations of task latents inferred by BATI and Recon in HalfCheetahVel to demonstrate the impact of context shifts more clearly. We construct interpolated contexts of various ratios r consisting of $N \cdot r$ steps from p_{test} and $N \cdot (1 - r)$ steps from p_{train}. Visualizations and episodic returns are shown in Fig. 6. While

Table 2. Final online evaluation episodic returns of BATI and baselines on $p_{\text{test}}(M, \mu)$ with OOD tasks in AntDir. BATI outperforms baselines and is the only method to achieve positive returns, further demonstrating its generalization potential.

Method	BATI	CSRO	FOCAL	Recon	UNICORN
Episodic Return	$\mathbf{12.1 \pm 0.5}$	-45.8 ± 14.5	-74.7 ± 6.1	-51.8 ± 37.0	-104.6 ± 9.1

both methods produce reasonable latent patterns on p_{train}, the latents inferred by Recon drift off when the ratio of p_{test} rises (Fig. 6b), causing its performance to drop (Fig. 6c). In contrast, the latents inferred by BATI consistently stay in the correct region (Fig. 6a) and yield performant policies.

6. Conclusion

Intelligent agents in the real world must adapt to unseen tasks and environments in noisy conditions based on their experience. As in-context reinforcement learning emerges to train context-adaptive policies, the **context shift** problem presents a great challenge to the task inference process and generalization capabilities. In this paper, we introduce Behavior-Agnostic Task Inference (BATI), a maximum-likelihood-based approach for robust task inference in ICRL. In contrast to the previous works, which predominantly use a learned encoder to perform direct Bayesian inference, BATI carefully analyzes the problem and infers task latents using dynamics estimators only, thereby freeing itself from the influence of context shifts. Extensive experiments in various benchmark environments validate the performance of BATI over strong baseline methods.

BATI still has certain limitations. We conducted experiments in several state-based MuJoCo environments commonly used in the field of offline ICRL. Going forward, we hope to scale BATI to vision-based embodied scenarios (Wu et al., 2022; Wang et al., 2023; Chen et al., 2023; Zhong et al., 2024) to enhance its applicability in the real world. The counterfactual estimation of $p(X^b \mid M, \mu)$ may further improve the performance of BATI. We may also scale BATI to multi-agent scenarios (Wang et al., 2022; Pan et al., 2022; Ci et al., 2023; Long et al., 2024) in the future, where the tasks involve interactions with or are defined by other agents.

Acknowledgements

This work was supported by the National Science and Technology Major Project (2022ZD0114904), NSFC-6247070125, NSFC-62406010, the State Key Lab of General Artificial Intelligence at Peking University, the Fundamental Research Funds for the Central Universities, Qualcomm University Research Grant.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Beck, J., Vuorio, R., Liu, E. Z., Xiong, Z., Zintgraf, L., Finn, C., and Whiteson, S. A survey of meta-reinforcement learning. *arXiv preprint arXiv:2301.08028*, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901, 2020.

Chen, Y., Geng, Y., Zhong, F., Ji, J., Jiang, J., Lu, Z., Dong, H., and Yang, Y. Bi-dexhands: Towards human-level bimanual dexterous manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2804–2818, 2023.

Ci, H., Liu, M., Pan, X., Zhong, F., and Wang, Y. Proactive multi-camera collaboration for 3d human pose estimation. *arXiv preprint arXiv:2303.03767*, 2023.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., and Abbeel, P. Rl ^{2}: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Gao, Y., Zhang, R., Guo, J., Wu, F., Yi, Q., Peng, S., Lan, S., Chen, R., Du, Z., Hu, X., et al. Context shift reduction for offline meta-reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Grigsby, J., Fan, L., and Zhu, Y. AMAGO: Scalable in-context reinforcement learning for adaptive agents. In *The*

Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=M6XWoEdmwf.

Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., and Knoll, A. A review of safe reinforcement learning: Methods, theories and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870. PMLR, 2018.

Hendel, R., Geva, M., and Globerson, A. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, 2023.

Humplik, J., Galashov, A., Hasenclever, L., Ortega, P. A., Teh, Y. W., and Heess, N. Meta reinforcement learning as task inference. *arXiv preprint arXiv:1905.06424*, 2019.

Kamienny, P.-A., Pirotta, M., Lazaric, A., Lavril, T., Usunier, N., and Denoyer, L. Learning adaptive exploration strategies in dynamic environments through informed policy regularization. *arXiv preprint arXiv:2005.02934*, 2020.

Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8.

Laskin, M., Wang, L., Oh, J., Parisotto, E., Spencer, S., Steigerwald, R., Strouse, D., Hansen, S. S., Filos, A., Brooks, E., maxime gazeau, Sahni, H., Singh, S., and Mnih, V. In-context reinforcement learning with algorithm distillation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=hy0a5MMPUv.

Lee, G., Hou, B., Mandalika, A., Lee, J., and Srinivasa, S. S. Bayesian policy optimization for model uncertainty. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJGvns0qK7.

Lee, J., Xie, A., Pacchiano, A., Chandak, Y., Finn, C., Nachum, O., and Brunskill, E. Supervised pretraining can learn in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Li, L., Huang, Y., Chen, M., Luo, S., Luo, D., and Huang, J. Provably improved context-based offline meta-rl with attention and contrastive learning. *arXiv preprint arXiv:2102.10774*, 2021a.

Li, L., Yang, R., and Luo, D. FOCAL: Efficient fully-offline meta-reinforcement learning via distance metric learning and behavior regularization. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=8cpHIfgY4Dj.

Li, L., Zhang, H., Zhang, X., Zhu, S., YU, Y., Zhao, J., and Heng, P.-A. Towards an information theoretic framework of context-based offline meta-reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=QFUsZvw9mx.

Lin, Z., Thomas, G., Yang, G., and Ma, T. Model-based Adversarial Meta-Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10161–10173. Curran Associates, Inc., 2020.

Liu, E. Z., Raghunathan, A., Liang, P., and Finn, C. Decoupling exploration and exploitation for meta-reinforcement learning without sacrifices. In *International Conference on Machine Learning*, pp. 6925–6935. PMLR, 2021.

Long, Q., Zhong, F., Wu, M., Wang, Y., and Zhu, S.-C. Socialgfs: Learning social gradient fields for multi-agent reinforcement learning. *arXiv preprint arXiv:2405.01839*, 2024.

Ma, L., Wang, Y., Zhong, F., Zhu, S.-C., and Wang, Y. Fast peer adaptation with context-aware exploration. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 33963–33982. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ma24n.html.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022.

Pan, X., Liu, M., Zhong, F., Yang, Y., Zhu, S.-C., and Wang, Y. Mate: Benchmarking multi-agent reinforcement learning in distributed target coverage control. *Advances in Neural Information Processing Systems*, 35:27862–27879, 2022.

Peng, M., Zhu, B., and Jiao, J. Linear representation meta-reinforcement learning for instant adaptation. *arXiv preprint arXiv:2101.04750*, 2021.

Rakelly, K., Zhou, A., Finn, C., Levine, S., and Quillen, D. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International Conference on Machine Learning*, pp. 5331–5340. PMLR, 2019.

Ren, Z., Liu, A., Liang, Y., Peng, J., and Ma, J. Efficient meta reinforcement learning for preference-based fast adaptation. *Advances in Neural Information Processing Systems*, 35:15502–15515, 2022.

Sohn, S., Woo, H., Choi, J., and Lee, H. Meta reinforcement learning with autonomous inference of subtask dependencies. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkgsWxrtPB.

Song, X., Gao, W., Yang, Y., Choromanski, K., Pacchiano, A., and Tang, Y. Es-maml: Simple hessian-free meta learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1exA2NtDB.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012.

Wang, D., Zhong, F., Li, M., Wen, M., Peng, Y., Li, T., and Yang, A. Romat: Role-based multi-agent transformer for generalizable heterogeneous cooperation. *Neural Networks*, 174:106129, 2024.

Wang, H., Wang, Y., Zhong, F., Wu, M., Zhang, J., Wang, Y., and Dong, H. Learning semantic-agnostic and spatial-aware representation for generalizable visual-audio navigation. *IEEE Robotics and Automation Letters*, 8(6): 3900–3907, 2023.

Wang, Y., Zhong, F., Xu, J., and Wang, Y. Tom2c: Target-oriented multi-agent communication and cooperation with theory of mind. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=M3tw78MH1Bk.

Wu, T., Zhong, F., Geng, Y., Wang, H., Zhu, Y., Wang, Y., and Dong, H. Grasparl: Dynamic grasping via adversarial reinforcement learning. *arXiv preprint arXiv:2203.02119*, 2022.

Wu, Y., Tucker, G., and Nachum, O. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Xu, T., Li, Z., and Ren, Q. Meta-reinforcement learning robust to distributional shift via performing lifelong in-context learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 55112–55125. PMLR, July 2024.

Yoon, J., Kim, T., Dia, O., Kim, S., Bengio, Y., and Ahn, S. Bayesian model-agnostic meta-learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Yu, Y. Towards sample efficient reinforcement learning. In *IJCAI*, pp. 5739–5743, 2018.

Yuan, H. and Lu, Z. Robust Task Representations for Offline Meta-Reinforcement Learning via Contrastive Learning. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 25747–25759. PMLR, June 2022. URL https://proceedings.mlr.press/v162/yuan22a.html. ISSN: 2640-3498.

Zhang, H. and Kan, Z. Temporal logic guided meta q-learning of multiple tasks. *IEEE Robotics and Automation Letters*, 7(3):8194–8201, 2022.

Zhong, F., Wu, K., Wang, C., Chen, H., Ci, H., Li, Z., and Wang, Y. Unrealzoo: Enriching photo-realistic virtual worlds for embodied ai. *arXiv preprint arXiv:2412.20977*, 2024.

Zhong, F., Wu, K., Ci, H., Wang, C., and Chen, H. Empowering embodied visual tracking with visual foundation models and offline rl. In *European Conference on Computer Vision*, pp. 139–155. Springer, 2025.

Zhou, R., Gao, C.-X., Zhang, Z., and Yu, Y. Generalizable task representation learning for offline meta-reinforcement learning with data limitations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17132–17140, 2024.

Zintgraf, L., Shiarlis, K., Igl, M., Schulze, S., Gal, Y., Hofmann, K., and Whiteson, S. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hkl9JlBYvr.

Zintgraf, L., Devlin, S., Ciosek, K., Whiteson, S., and Hofmann, K. Deep interactive bayesian reinforcement learning via meta-learning. *arXiv preprint arXiv:2101.03864*, 2021.

A. Proof of Theorem 3.1

To prove this inequality, we use the following simple lemma:

Lemma A.1. *For arbitrary random variables* $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$*,*

$$H(\mathbf{X} \mid \mathbf{Y}) + H(\mathbf{Y} \mid \mathbf{Z}) \geq H(\mathbf{X} \mid \mathbf{Z}). \tag{11}$$

Proof.

$$\begin{aligned} & H(\mathbf{X} \mid \mathbf{Y}) + H(\mathbf{Y} \mid \mathbf{Z}) \\ \geq\ & H(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) + H(\mathbf{Y} \mid \mathbf{Z}) \\ =\ & H(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) \\ \geq\ & H(\mathbf{X} \mid \mathbf{Z}) \end{aligned}$$

□

Now, for the main proof:

Proof. Applying Lemma A.1 and rearranging, we have

$$H(\mathbf{Z} \mid \mathbf{X}^b) - H(\mathbf{Z} \mid \mathbf{M}) \leq H(\mathbf{M} \mid \mathbf{X}^b)$$

Now,

$$\begin{aligned} & I(\mathbf{Z}; \mathbf{M}) - I(\mathbf{Z}; \mathbf{X}^b) \\ =\ & H(\mathbf{Z}) - H(\mathbf{Z} \mid \mathbf{M}) - (H(\mathbf{Z}) - H(\mathbf{Z} \mid \mathbf{X}^b)) \\ =\ & H(\mathbf{Z} \mid \mathbf{X}^b) - H(\mathbf{Z} \mid \mathbf{M}) \\ \leq\ & H(\mathbf{M} \mid \mathbf{X}^b) \\ =\ & H(\mathbf{M}) - (H(\mathbf{M}) - H(\mathbf{M} \mid \mathbf{X}^b)) \\ =\ & H(\mathbf{M}) - I(\mathbf{M}; \mathbf{X}^b) \end{aligned}$$

We can now see that the CSRO objective (at $\lambda = 1$) is upper bounded by $H(\mathbf{M}) - I(\mathbf{M}; \mathbf{X}^b)$. Consequently, when $I(\mathbf{M}; \mathbf{X}^b)$ is high, e.g. $\mathbf{M}$ can be determined fully from $\mathbf{X}^b$, we have $I(\mathbf{M}; \mathbf{X}^b) = H(\mathbf{M})$ and $I(\mathbf{Z}; \mathbf{M}) \leq I(\mathbf{Z}; \mathbf{X}^b)$. As a result, we cannot find a $\mathbf{Z}$ that simultaneously captures $\mathbf{M}$ and is not correlated with $\mathbf{X}^b$. □

B. Pseudocode of the Offline Training and Online Testing Procedure

Algorithm 1 Offline Training Pipeline of BATI

Require: Training task distribution $p_{\text{train}}(M)$ and associated offline datasets $\{\mathcal{D}_M\}$

1: Randomly initialize θ, ϕ, ψ
2: **while** Maximum training step not reached **do**
3: Sample a batch of tasks $\{M_i\} \sim p_{\text{train}}(M)$ and corresponding training contexts $\{X_{M_i}\} \sim \{\mathcal{D}_{M_i}\}$
4: Sample $\{Z_{M_i}\} \sim \{\mathbf{Z}_\phi^{M_i}\}$ with reparameterization
5: Update ϕ, ψ with $\{X_{M_i}\}$ and $\{Z_{M_i}\}$ using the dynamics reconstruction loss Eq. 6
6: Update θ with $\{X_{M_i}\}$ and (detached) $\{Z_{M_i}\}$ using IQL losses
7: **end while**
8: Return final θ, ϕ, ψ

C. Experiment Details

C.1. Environments

We conduct our experiments in several commonly used environments in the field of offline ICRL. Based on MuJoCo (Todorov et al., 2012), a popular physics simulator, the environments feature several robot locomotion tasks where the robots must move to accomplish a certain task:

Algorithm 2 Online Evaluation Procedure of BATI

Require: Testing task distribution $p_{\text{test}}(M)$ and associated context datasets $\{\mathcal{C}_M\}$, evaluated parameters θ, ϕ, ψ, number of latent samples per inference N

$R \leftarrow 0$
for $M \in p_{\text{test}}(M)$ **do**
  Compute Z^* with $X \sim \mathcal{C}_M$ using the dynamics reconstruction loss Eq. 8
  Sample initial state $s \sim \rho_0$ of M
  while Episode not done **do**
    Sample action $a \sim \pi_\theta(s, Z^*)$
    Execute action a in task M and get the next state s, reward r, done flag d
    $R \leftarrow R + r$
  end while
end for
Return $R/|p_{\text{test}}(M)|$

- **AntDir**: A four-legged ant-like robot needs to go along the direction specified by the task. The task is a goal direction uniformly sampled from $[0, 2\pi)$, and the reward is the inner product between the position delta and the goal direction. At every time step t, we set the goal direction M_t to the sum of the true task M^* and a Gaussian noise of mean 0 and standard deviation ϵ, such that the reward function is stochastic with respect to the true task.

- **HalfCheetahVel**: A bipedal cheetah-like robot needs to run with the velocity specified by the task. The task is a target velocity uniformly sampled from $[0, 2]$, and the reward is the negative absolute difference between the current and the goal velocity minus a control cost. We add a similar Gaussian noise of mean 0 and standard deviation ϵ to the target velocity at every time step.

- **HalfCheetahDir**: A bipedal cheetah-like robot needs to run along the direction specified by the task. The task is a target direction uniformly sampled from $\{\text{left}, \text{right}\}$, and the reward is the velocity in the target direction minus a control cost. We use the true target direction to compute the reward with probability $1 - \epsilon$, and a uniformly random direction with probability ϵ.

- **HopperParam, WalkerParam**: A single-leg hopper or a bipedal walker needs to move forward. In these two environments, tasks differ by the transition (physics) P_M of the MDP, instead of the reward function R_M in the environments above. The task is a log friction coefficient uniformly sampled from $[-3, 3]$, and the reward is the velocity in the forward direction minus a control cost. Furthermore, the friction coefficients are also used to perform an affine projection to the action to ensure that different tasks have different optimal behaviors. We add a Gaussian noise of mean 0 and standard deviation ϵ to the log friction coefficient at every time step.

C.2. Datasets Construction

To construct the datasets used for training and testing, we train an expert policy μ_M using SAC (Haarnoja et al., 2018) for each training task M. For the environments differing in rewards, we use $\epsilon = 0$ when training the context-collection policies. For the testing task M, we use the expert policy for the most similar training task as its primary context collection policy μ_M, and that for the most different training task as the testing context collection policy $\mu_{\bar{M}}$. Furthermore, to ensure that p_{test} remains supported by the training data p_{train}, we mix 90% data from $\mu_{\bar{M}}$ with 10% data from μ_M to construct p_{test} for HopperParam and WalkerParam. Distance between tasks is defined as follows for each environment:

- **AntDir**: The absolute difference of the goal direction, measured in radians modulo 2π: $\min(\max(M_1, M_2) - \min(M_1, M_2), \min(M_1, M_2) + 2\pi - \max(M_1, M_2))$.

- **HalfCheetahVel**: The absolute difference in target velocities.

- **HalfCheetahDir**: The target direction is the same (0) or different (1).

- **HopperParam, WalkerParam**: The absolute difference in log friction coefficients.

Table 3. Hyperparameters used in each of our evaluation environments.

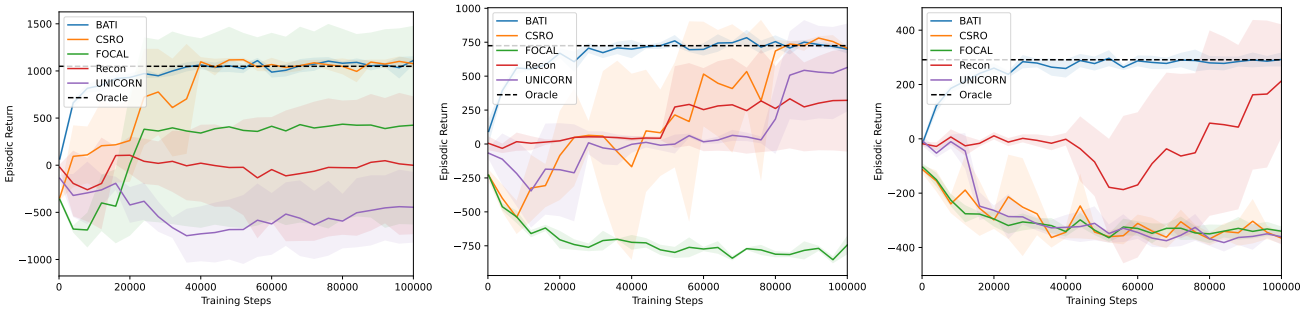
Parameter Name	Environments				
	AntDir	HalfCheetahVel	HalfCheetahDir	HopperParam	WalkerParam
Learning Rate	$3 * 10^{-4}$	$3 * 10^{-4}$	$3 * 10^{-4}$	$3 * 10^{-4}$	$3 * 10^{-4}$
Batch Size	4096	4096	4096	4096	4096
Task Contrastive Batch Size	16	16	16	16	16
IQL τ	0.8	0.8	0.8	0.8	0.8
IQL β	0.05	0.05	0.05	0.05	0.05
IQL Exp. Adv. Clip	100	100	100	100	100
# Gradient Steps	10^5	10^5	10^5	10^5	10^5
Episode Length	200	200	200	200	200
Dataset Size	10^5	$2 * 10^5$	$2 * 10^5$	$3 * 10^5$	$3 * 10^5$
Task Latent Dim	5	5	5	40	40
BATI # Latent Samples N	40	40	40	40	40
UNICORN Weight	0.15	0.15	0.15	1.5	1.5
CSRO CLUB Weight	5.0	1.0	1.0	2.5	2.5
CLUB Encoder Hidden Dims	[200] * 3	[200] * 3	[200] * 3	[200] * 3	[200] * 3
Encoder Hidden Dims	[64, 64]	[64, 64]	[64, 64]	[128, 128]	[128, 128]
Decoder Hidden Dims	[64, 64]	[64, 64]	[64, 64]	[128, 128]	[128, 128]
RL Hidden Dims	[256, 256]	[256, 256]	[256, 256]	[256, 256]	[256, 256]

Table 4. Final online evaluation episodic returns of BATI and baselines on $p_{\text{test}}(M, \mu)$. BATI achieves superior performance in all of the benchmark environments, dramatically outperforming baselines. IQL provides additional optimizations on the basis of oracle data and may lead to methods with even better performance.

	AntDir	HalfCheetahVel	HalfCheetahDir	HopperParam	WalkerParam
Oracle (Pre-IQL)	58.7	-139.3	291.2	290.2	549.2
BATI	$\mathbf{46.4 \pm 2.9}$	$\mathbf{-122.8 \pm 1.6}$	$\mathbf{286.1 \pm 13.8}$	$\mathbf{285.9 \pm 5.5}$	$\mathbf{565.2 \pm 7.9}$
CSRO	28.9 ± 3.0	-134.8 ± 8.9	-340.0 ± 17.2	144.4 ± 21.3	279.5 ± 34.0
FOCAL	-46.5 ± 2.2	-279.4 ± 9.4	-336.4 ± 8.5	162.6 ± 44.5	317.0 ± 37.0
Recon	-18.3 ± 13.7	-201.8 ± 7.4	127.1 ± 258.5	157.8 ± 16.7	292.6 ± 38.4
UNICORN	-32.0 ± 4.2	-267.3 ± 20.5	-362.9 ± 6.8	144.7 ± 21.4	319.0 ± 11.5

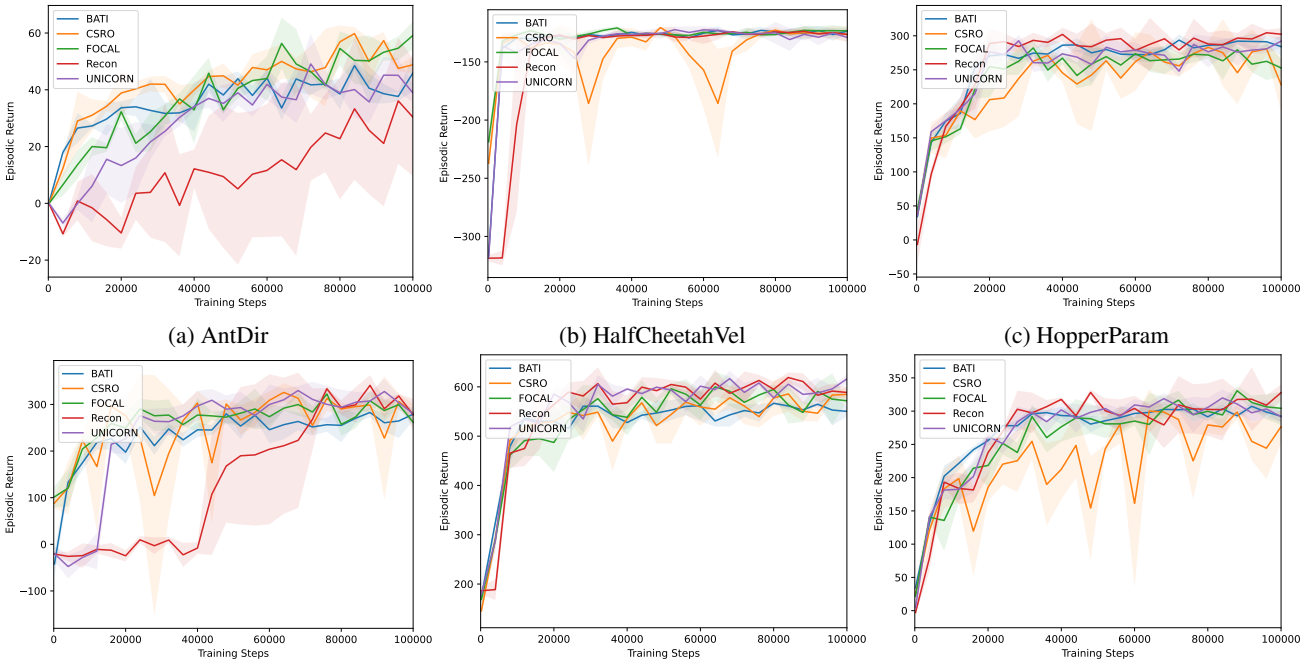
C.3. Implementation Details

Building on the codebase of the official implementation of UNICORN (Li et al., 2024), we implement BATI and all our baselines. All methods share the same offline RL algorithm and relevant architectures. The networks are implemented as MLPs with ReLU activations, with hidden dimensions specified in Tab. 3. The CLUB model used in the CSRO baseline is ported from the official implementation of CSRO (Gao et al., 2024). We found BRAC (Wu et al., 2019), the base offline RL algorithm used in several prior works, to behave unstably in some cases, and switched to IQL (Kostrikov et al., 2022) instead for all methods, which yields consistent and satisfactory performance. See Tab. 3 for hyperparameters of the main experiments. Note that some hyperparameters apply only to the methods that require them, e.g. Task Contrastive Batch Size applies only to baselines with FOCAL-like losses (Li et al., 2021b), and Encoder Hidden Dims apply only to baselines and not BATI (which does not have an encoder network). The task latent embedding distributions of BATI $\{\mathbf{Z}_\phi^M\}_M$ are parameterized as Gaussian distributions with learnable mean and fixed log variance of -4. An optional KL divergence between the latent distributions and $\mathcal{N}(\mathbf{0}, \mathbf{I})$ can be used to regularize the task latent distributions. We empirically find the unregularized version to have good performance and do not use this regularization for all methods, echoing UNICORN.



(a) $\epsilon = 0$ (b) $\epsilon = 0.3$ (c) $\epsilon = 0.7$ (Default)

Figure 7. Online evaluation episodic return curves during training for different noise levels on HalfCheetahDir. The performance of baselines mostly decreases relative to BATI as the noise level increases.



(a) AntDir (b) HalfCheetahVel (c) HopperParam

(d) HalfCheetahDir (e) WalkerParam (f) HopperParam($\epsilon = 0$)

Figure 8. Online evaluation episodic return curves of BATI and baselines on contexts from $p_{\text{train}}(M, \mu)$ during training in our evaluation environments.

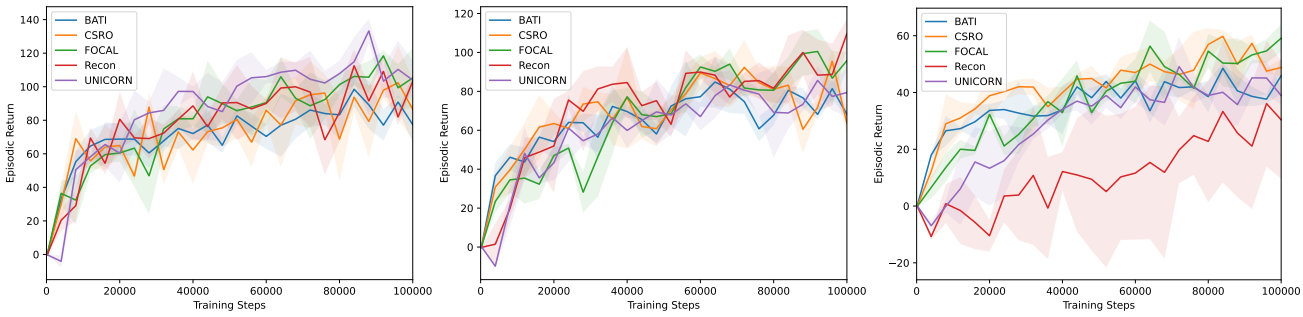
D. Additional Figures and Experiments

D.1. Additional Figures and Tables Referred by the Main Text

Due to space constraints, we moved some figures and tables used in the main text here. Tab. 4 contains the final online evaluation episodic returns of the main experiment, while Fig. 7 contains noise level ablations in HalfCheetahDir.

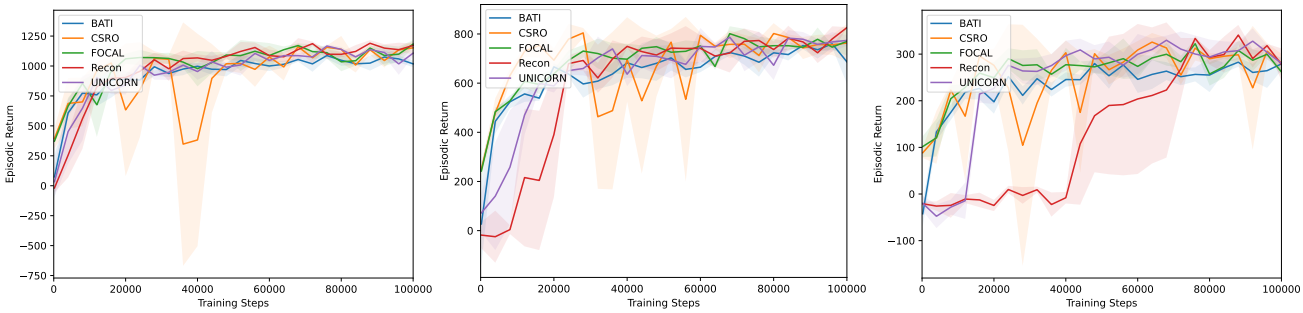
D.2. In-distribution Performances

We show the online evaluation episodic return curves on p_{train} here. In Fig. 8, most methods can achieve a similar performance when evaluated on the training p_{train} contexts, echoing our argument that learned encoders can be trained to perform in-distribution Bayesian inference. The performance of CSRO is somewhat unstable, potentially due to the conflicts among its objectives. Similar phenomena can be observed in the in-distribution performances of the noise level ablations in AntDir (Fig. 9 and HalfCheetahDir (Fig. 10).



(a) $\epsilon = 0$ (b) $\epsilon = 0.2$ (c) $\epsilon = 0.4$ (Default)

Figure 9. Online evaluation episodic return curves on $p_{\text{train}}(M, \mu)$ during training for different noise levels on AntDir.



(a) $\epsilon = 0$ (b) $\epsilon = 0.3$ (c) $\epsilon = 0.7$ (Default)

Figure 10. Online evaluation episodic return curves on contexts from $p_{\text{train}}(M, \mu)$ during training for different noise levels on HalfCheetahDir. All methods perform similarly while CSRO and Recon are somewhat unstable.

D.3. Adaptation to Different Numbers of Training Tasks

We demonstrate the robustness of BATI with respect to the task embedding table size (corresponding to the number of training tasks). We split the 40 evaluation tasks in AntDir into different training and testing splits and reran the experiments. As shown in Tab. 5, across all splits, BATI achieves the best performance uniformly and is highly stable.

D.4. Adaptation to Multi-agent Scenarios

We conduct preliminary experiments to showcase the general applicability of BATI to multi-agent domains. We choose Kuhn Poker, a two-player card game with discrete state and action spaces, differing from the continuous MuJoCo environments used in our paper. We generate different player-2 (opponent) policies as "tasks" and learn an adaptive policy for player-1 over 10 episodes (20 steps) of contexts. As shown in Tab. 6, BATI maintains superior performance over all baselines, further showcasing its capabilities and generalization.

Table 5. Final online evaluation episodic returns of BATI and baselines on $p_{\text{test}}(M, \mu)$ with different train/test splits in AntDir. BATI outperforms baselines and has consistently good performance across different numbers of training tasks.

Train/Test Split	BATI	CSRO	FOCAL	Recon	UNICORN
10/30	$\mathbf{45.0 \pm 6.0}$	4.7 ± 4.2	-33.2 ± 3.4	-6.7 ± 2.8	-22.2 ± 2.5
20/20 (Main)	$\mathbf{46.4 \pm 2.9}$	28.9 ± 3.0	-46.5 ± 2.2	-18.3 ± 13.7	-32.0 ± 4.2
30/10	$\mathbf{49.9 \pm 4.6}$	46.2 ± 4.8	-33.9 ± 1.4	-4.3 ± 5.0	-9.4 ± 7.6

Table 6. Final online evaluation episodic returns of BATI and baselines on $p_{\text{test}}(M, \mu)$ in KuhnPoker.

Method	Oracle	BATI	CSRO	FOCAL	Recon	UNICORN
Episodic Return	0.0734	$\mathbf{-0.049 \pm 0.025}$	-0.180 ± 0.032	-0.191 ± 0.020	-0.185 ± 0.052	-0.243 ± 0.042