BILE: An Effective Behavior-based Latent Exploration Scheme for Deep Reinforcement Learning

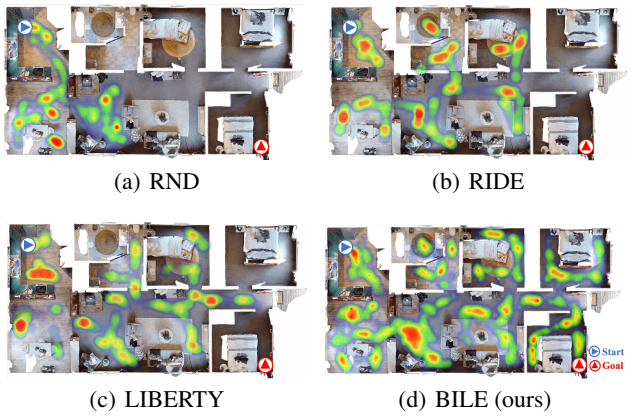
Yiming Wang[1], **Kaiyan Zhao**[1,2], **Yan Li**[3] and **Leong Hou U**[1*]
[1]State Key Laboratory of Internet of Things for Smart City, University of Macau, Macao SAR, China
[2]School of Computer Science, Wuhan University, Wuhan, China
[3]School of Artificial Intelligence, Shenzhen Polytechnic University, Shenzhen, China
{wang.yiming,yb57411}@connect.um.edu.mo, zhao.kaiyan@whu.edu.cn, ryanlhu@um.edu.mo

Abstract

Efficient exploration of state spaces is critical for the success of deep reinforcement learning (RL). While many methods leverage exploration bonuses to encourage exploration instead of relying solely on extrinsic rewards, these bonus-based approaches often face challenges with learning efficiency and scalability, especially in environments with high-dimensional state spaces. To address these issues, we propose BehavIoral metric-based Latent Exploration (BILE). The core idea is to learn a compact representation within the behavioral metric space that preserves value differences between states. By introducing additional rewards to encourage exploration in this latent space, BILE drives the agent to visit states with higher value diversity and exhibit more *behaviorally distinct* actions, leading to more effective exploration of the state space. Additionally, we present a novel behavioral metric for efficient and robust training of the state encoder, backed by theoretical guarantees. Extensive experiments on high-dimensional environments, including realistic indoor scenarios in Habitat, robotic tasks in Robosuite, and challenging discrete Minigrid benchmarks, demonstrate the superiority and scalability of our method over other approaches.

1 Introduction

Striking an appropriate balance between *exploration* and *exploitation* remains a long-standing challenge in reinforcement learning (RL) [Sutton and Barto, 2018]. To address this problem, classical exploration strategies such as ϵ-greedy or Boltzmann exploration randomly choose from all possible actions with a non-zero probability [Mnih *et al.*, 2015]. However, these methods alone cannot perform deep exploration, which requires the agent to discover rewards far from the initial states. Recent approaches leverage exploration bonus to encourage deep exploration. These exploration bonuses estimate the novelty of states and incentivize the policy to visit new states. For instance, [Bellemare *et al.*, 2016a] models the bonus as inversely proportional to the pseudo-count of visited states, encouraging the agent to visit infrequently-seen states. Other curiosity-driven approaches [Burda *et al.*, 2018b; Pathak *et al.*, 2017] aim to learn the dynamics of the environment and use prediction error as the intrinsic reward. Another line of research [Wang *et al.*, 2023; Raileanu *et al.*, 2020; Zhang *et al.*, 2021; Wang *et al.*, 2024] utilizes state differences as exploration bonuses. Despite the excellent performance of exploration bonus on some tasks, the *scalability* is a significant limitation especially in high-dimensional spaces where state differences is subtle. Furthermore, existing work [Burda *et al.*, 2018a; Badia *et al.*, 2020b; Henaff *et al.*, 2022] reports that the performance of bonuses varies significantly across different tasks and learning stages, especially in environments with high-dimensional state spaces where the novelty between states is rooted in the stochasticity in the environment's dynamics and have little to do with the agent's exploration capabilities (e.g., the "noisy TV" problem [Pathak *et al.*, 2017]), which signifi-



(a) RND (b) RIDE (c) LIBERTY (d) BILE (ours)

Figure 1: Visualization of heatmap of trajectories from policies trained with different exploration algorithms in the indoor environment with high-dimensional state spaces. The agent needs to navigate from the start position to reach the goal position. Other methods fails to reach the goal due to insufficient exploration, since the bonus can distract the agent, leading it to take repetitive actions and accumulate numerous rewards without reaching the goal. Our method BILE explores the whole room space including the goal position by encouraging the agent to produce diverse behavior while maintaining sufficient exploration over the high-dimensional state space.

*Corresponding author.

cantly limits the widespread adoption of exploration bonuses as a default exploration strategy in the realm of deep RL. For example, as shown in Figure 1, in a navigation task within a realistic indoor environment, recent advanced bonus-based exploration methods, such as those curiosity-driven [Burda *et al.*, 2018b], latent-based [Raileanu *et al.*, 2020], and state-difference approaches [Wang *et al.*, 2023], struggle to achieve comprehensive exploration, with the agent typically exploring only up to four rooms. The limitation arises because these methods encourage repetitive actions aimed at maximizing the exploration bonus, often disregarding the extrinsic goal, leading to what can be described as *meaningless exploration*. In contrast, our method encourages more *diverse behavior*, enabling the agent to fully explore all the rooms.

In this paper, we aim to address the aforementioned challenges by providing an exploration strategy that is both efficient and effective across various domains. To achieve this, we train a state encoder by projecting the high-dimensional states into the *behavioral* metric space, where the value difference between states is upper-bounded by the distance in the latent space. By learning to span the metric space, the agent is encouraged to explore states with higher value diversity, which leads to taking more diverse behavior and, in turn, promotes more effective exploration. Moreover, by exploring a more compact latent space through the proposed behavioral metric, the agent becomes better suited to handle complex, high-dimensional state spaces (e.g., images), which ensures that the agent can effectively navigate and learn in environments with high-dimensional state representations.

Our main contribution can be summarized as follows. Firstly, we propose a novel exploration strategy that captures diverse behaviors by maximally exploring the compact latent behavioral metric space. Secondly, we introduce a novel behavioral metric that enables efficient and robust training of the state encoder, supported by theoretical guarantees. Thirdly, we introduce a plug-in exploration bonus that promotes more effective exploration and is scalable across various environments. Lastly, we conduct extensive experiments on challenging tasks within different environments. We also evaluate our algorithm in the real-life indoor environment Habitat. The results demonstrate that our algorithm significantly enhances exploration while maintaining broad scalability.

2 Preliminaries

Reinforcement Learning. We assume the underlying environment is a Markov Decision Process (MDP), defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s' \mid s, a)$ is state transition function from state $s \in \mathcal{S}$ to state $s' \in \mathcal{S}$, r is the reward function, and $\gamma \in [0, 1)$ is the discount factor. Generally, the policy of an agent in an MDP is a mapping $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$. An agent chooses actions $a \in \mathcal{A}$ according to a policy function $a \sim \pi(s)$, which updates the system state $s' \sim P(s, a)$ yielding a reward $r(s, a)$. The goal of the agent is to learn a policy π that maximizes expected return $\mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$ in a trajectory $(s_0, a_0, s_1, \ldots,)$ by learning a value function (or value network) V_π from the interaction that approximates $V^\pi(s_0) \approx \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$.

Behavioral Metric. Behavioral metrics are designed to quantify the dissimilarity between states by considering differences in immediate reward signals and the divergence of their next-state distributions. These methods establish approximate metrics within the representation space, ensuring that behavioral similarities among states are preserved. Furthermore, behavioral metrics have been shown to provide an upper bound on state-value discrepancies between corresponding states. A widely used behavioral metric is the bisimulation metric [Ferns *et al.*, 2011], which defines a pseudometric space $(\mathcal{S}, d)$, where a distance function $d : \mathcal{S} \times \mathcal{S} \mapsto \mathbb{R}_{\geq 0}$ measures the *behavioral similarity* between two states. Recently, the π-bisimulation metric [Castro, 2020] has been introduced to address scalability challenges, making it applicable to continuous tasks. The metric consists of two key components: `Reward Difference` term and `Distribution Divergence` term:

Definition 1 (π-bisimulation metric)**.** *Given a fixed policy π, the following π-bisimulation metric exists and is unique:*

$$d^\pi(s_i, s_j) = \underbrace{|r^\pi(s_i) - r^\pi(s_j)|}_{\texttt{Reward Difference}} + \underbrace{\gamma W_1(d^\pi)(P^\pi(\cdot|s_i), P^\pi(\cdot|s_j))}_{\texttt{Distribution Divergence}} \tag{1}$$

where $r^\pi(s_i) = \mathbb{E}_{a\sim\pi(\cdot|s_i)} r(s_i, a_i)$, $P^\pi(\cdot \mid s_i) = \mathbb{E}_{a\sim\pi(\cdot|s_i)} P(\cdot \mid s_i, a)$, and W_1 is the 1-Wasserstein distance.

Exploration Bonus. Addressing the sparse reward challenge prevalent in many environments necessitates augmenting the original external reward function, r, with an intrinsic reward bonus, b, resulting in a combined reward: $r' = r + b$. As such, the agent is encouraged to visit novel states with the guidance of the extra exploration bonus. A number of bonuses have been proposed, based on pseudo-counts [Bellemare *et al.*, 2016a], prediction errors between dynamic models [Burda *et al.*, 2018b], and state differences in the latent space [Wang *et al.*, 2023], among others. Table 1 highlights the key differences between our approach and recent baselines. Notably, our method is the first to exhibit both the state-value difference bound and robust representation properties, while demonstrating effectiveness in sparse reward tasks and high-dimensional environments. A detailed discussion of these results is provided in the following section.

3 Methodology

Our goal is to develop an effective exploration scheme that enhances *deep exploration* across various environments.

Challenge of High-dimensional Exploration. A key challenge for exploration bonus-based methods is scaling to realistic environments with high-dimensional state spaces. While recent bonus-based methods, such as curiosity-driven exploration [Burda *et al.*, 2018b], perform well in toy environments like grid games, they often struggle in high-dimensional spaces. These methods typically prioritize maximizing the intrinsic bonus without considering the extrinsic task or goal,

Algorithms	Value Diff. Bound	Robust Representation	Learn Dyn.
ICM [Pathak *et al.*, 2017]	✗	✓	✓
RND [Burda *et al.*, 2018b]	✗	✓	✗
RIDE [Raileanu *et al.*, 2020]	✗	✓	✓
EME [Wang *et al.*, 2024]	✓	✗	✗
RLE [Mahankali *et al.*, 2024]	✗	✗	✗
LIBERTY [Wang *et al.*, 2023]	✓	✗	✓
BILE (ours)	✓	✓	✓

Table 1: Comparison of different algorithms. "Value Diff. Bound" means that the state distance computed in the latent space is upper-bounds the state value difference. "Robust Representation" indicates that the representation learning process remains robust, avoiding representation collapse as described in Theorem 1. "Learn Dyn." means that whether the algorithm learns the transition dynamics.

which leads to repetitive actions or meaningless exploration. For instance, as shown in Figure 1(a), the RND agent explores solely based on the intrinsic bonus, neglecting the extrinsic goal, which impairs both exploration and task performance. Recent state-difference-based methods [Wang *et al.*, 2023; Wang *et al.*, 2024] have shown strong performance across a variety of environments. However, subtle state differences complicate the evaluation of state novelty in high-dimensional environments. Additionally, we prove that these methods exhibit representation collapse (c.f. Theorem 1) in sparse reward scenarios. For example, LIBERTY [Wang *et al.*, 2023], as shown in Figure 1(c), struggles to reach the goal in realistic indoor environments. More recent approaches attempt to model novelty in low-dimensional latent spaces, but they still face difficulties in complex and high-dimensional settings. As illustrated in Figure 1(b), the latent-based exploration method RIDE [Raileanu *et al.*, 2020] fails to adequately explore the goal space, with only a limited portion of the environment explored, which highlights the reduced exploration capability of latent exploration methods in high-dimensional spaces.

Solution: Ensuring Behavioral Diversity of Bonuses. The key to improve the overall performance and scalability of bonus-based methods lies in their ability to generate diverse behaviors in environments with high-dimensional state spaces. Previous methods fail to promote the diversity of action behaviors in large state spaces, which can lead to repetitive or meaningless actions. Specifically, exploration bonuses can misguide the agent into repeating actions without progressing toward the task. Inspired by the behavioral meric-based representation learning methods [Zhang *et al.*, 2020; Castro, 2020], we integrate behavioral differences into the exploration bonus by training a behavioral metric-based state encoder. This approach not only improves learning efficiency in high-dimensional environments but also enhances the diversity of actions and behaviors. By connecting the value of states with the extrinsic reward, which is tied to task completion, our method encourages the agent to take actions that are both diverse and goal-directed. The remaining questions are:

Q1: What behavioral metric should we choose to construct the state encoder (latent space)?

Q2: How can we integrate behavioral diversity with bonus in an easy-to-plug and computationally efficient way?

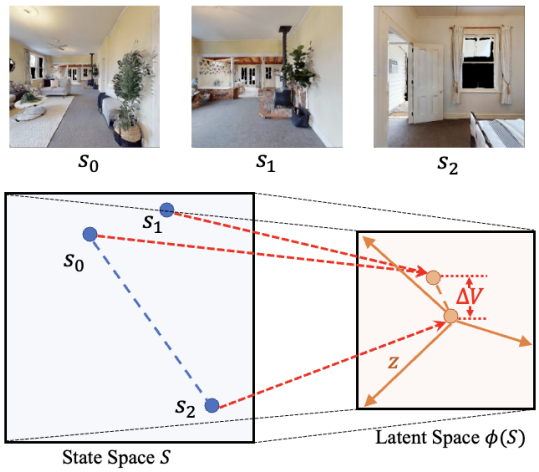


Figure 2: Behavioral metric-based latent exploration. The states are projected into a compact latent space where the the state-value diversity are incorporated into the novelty of states. By randomly sampling latent vector z, the agent is encouraged to explore the state space with diverse behavior.

3.1 Behavioral Metric-based State Encoder

We propose BILE (BehavIoral metric-based Latent Exploration), a data-efficient approach to promote effective exploration from unstructured high-dimensional states across different environments. We begin by training a representation function $d_\phi : \mathcal{S} \to \mathcal{Z}$. Regarding the aforementioned Q1 and Q2, our representation function ϕ has two key desiderata. First, d_ϕ should map ***behaviorally equivalent***, high-dimensional states into a compact low-dimensional representation. Second, ϕ should encode the distance between states in terms of their value differences to ensure state-value diversity in novel states.

Limitation of Previous Methods. Previous approaches, such as LIBERTY [Wang *et al.*, 2023] and DBC [Zhang *et al.*, 2020], set bisimulation-based metric as the behavioral metric (1) for the training of state encoders. These methods can be generalized by minimize the following objective:

$$\mathcal{L}(\phi) = \frac{1}{2}\mathbb{E}[(d^\pi_\phi(s_i, s_j) - \underbrace{|r^\pi(s_i) - r^\pi(s_j)|}_{\texttt{Reward Difference}} - \underbrace{\gamma W_2(P^\pi(\cdot \mid s_i), P^\pi(\cdot \mid s_j))}_{\texttt{Distribution Divergence}} - \text{Const.}]^2 \tag{2}$$

where $d^\pi_\phi(s_i, s_j) = ||\phi(s_i) - \phi(s_j)||_2$, 2-Wasserstein metric W_2 is used instead of W_1 to estimate the distribution divergence term due to its convenient closed form for Gaussian distributions, and the inverse dynamic output introduced in [Wang *et al.*, 2023] can be scalarized into a constant.

Theorem 1 (Representation Collapse under Spare Rewards)**.** *Let ξ denote the distribution over pairs of states (s_i, s_j) sampled from ξ. Assume deterministic transitions and the existence of a stationary distribution over states. Given a bisim-*

ulation metric d^π of the form (1), we have:

$$\mathbb{E}_{(s_i,s_j)\sim\xi}\left[d^\pi_\phi(s_i,s_j)\right] = \frac{1}{1-\gamma}\mathbb{E}_{(s_i,s_j)\sim\xi}\left[|r^\pi_{s_i} - r^\pi_{s_j}|\right] \tag{3}$$

Under sparse rewards, the right-hand side (RHS) of Equation (3), i.e., $\mathbb{E}_{(s_i,s_j)\sim\xi}[|r^\pi_{s_i} - r^\pi_{s_j}|] \approx 0$ for most cases. Consequently, the learned embedding collapses to a constant value: $\hat{d_\phi}(s_i) = d_{\phi_c}$.

Proof in Appendix A. Based on Theorem 1, in environments with sparse rewards, the leaned embedding under the loss (2) collapses to a constant value, effectively discarding all state information. Such a collapse significantly impairs exploration performance, as the embedding fails to distinguish between states. The issue is still prevalent even in the latest bisimulation-based approaches [Wang *et al.*, 2024; Zang *et al.*, 2023], which highlights a fundamental limitation in their ability to handle sparse reward scenarios.

Behavioral Metric of BILE. Our goal is to learn a robust representations $\mathcal{Z}$ using the behavioral metric-based encoder d_ϕ, even in sparse reward environments and then use these representations to improve exploration. To address the representation collapse issue highlighted in Theorem 1, we introduce the BILE operator for the behavioral metric:

Definition 2. *Given a policy π, the probabilistic transition dynamics model $P_\eta(\cdot|s) = \mathcal{N}(\mu_\eta, \sigma_\eta)$, BILE operator is updated as :*

$$\begin{aligned}\mathcal{F}(d^{BILE}_\phi, \pi) = &\underbrace{|r^\pi_{s_i} - r^\pi_{s_j}| + \frac{\alpha}{2}\sum_{s_i,s_j}|r^\pi_{PE}(s)|}_{\texttt{Reward Difference}} \\ &+ \underbrace{\gamma\mathbb{E}_{s'\sim P_\eta(\cdot|s)} d^{BILE}_\phi(s'_i, s'_j)}_{\texttt{Distribution Divergence}}\end{aligned} \tag{4}$$

where $r^\pi_{PE}(s) = ||\hat{P}_\eta(s,a) - s'||_2$, $\hat{P}_\eta(s,a)$ represents the predicted next state from the learned probabilistic transition dynamics model, and α is a scaling hyperparameter.

For the `Reward Difference` term, to address the sparse reward issue faced by previous methods, we incorporate the prediction error of the probabilistic transition dynamics model, $r^\pi_{\text{PE}}(s)$, as an extra reward into the state encoder learning objective. This ensures the reward signal remains informative during encoder training, preventing it from vanishing. Since the transition dynamics model is already trained for the encoder loss, the additional computational cost of this approach is negligible. Over time, as the agent explores and becomes better at predicting its environment, the prediction error diminishes, naturally reducing its impact on the metric learning process.

Furthermore, for the `Distribution Divergence` term, unlike previous bisimulation-based approaches that rely on the Wasserstein distance which is computationally intractable and requires relaxation, we replace it with a sample-based next-state distribution divergence. Our approach requires only sampling, making it significantly more computationally efficient without loss of theoretical integrity.

Theoretical Guarantee. The BILE metric operator preserves the following theoretical properties as a behavioral metric:

Theorem 2 (Convergence Guarantee)**.** *Given a policy π, with the convergence of the latent transition dynamic model, the BILE operator $\mathcal{F}(d^{BILE}_\phi, \pi)$ has a fixed-point.*

Theorem 3 (State-Value Difference Upper-bound)**.** *For any two states s_i and s_j, a given policy π and the BILE metric encoder d^{BILE}_ϕ, we have:*

$$|V^\pi(s_i) - V^\pi(s_j)| \le d^{BILE}_\phi(s_i, s_j) \tag{5}$$

Proof in Appendix A. Theorem 2 establishes the convergence guarantee for the BILE metric-based encoder. Meanwhile, Theorem 3 shows that the value difference between states is upper-bounded by their distance in the BILE behavioral metric space, similar to bisimulation-based approaches. As depicted in Figure 2, states with higher novelty retain greater value diversity, encouraging the agent to explore novel states with larger value differences, which drives the agent to exhibit more diverse behaviors and facilitates more effective exploration. Additionally, as the bonus increases, it results in larger temporal difference (TD) errors for adjacent state pairs during learning, which will significantly enhance learning efficiency, as the BILE metric promotes the prioritization of states that improve both novelty and learning efficacy.

3.2 Latent Exploration via BILE

Improvements over Previous Approaches. Previous latent exploration methods define the exploration bonus using prediction error [Pathak *et al.*, 2017], randomized rewards [Mahankali *et al.*, 2024], and state differences [Wang *et al.*, 2023]. However, these methods often struggle in environments with high-dimensional state spaces. As the state differences become subtle and the exploration space grows exponentially, their performance degrades significantly in realistic environments with high-dimensional observation spaces. A detailed comparison is provided in Table 1. To address these limitations and build upon the conclusions of prior work, we identify key principles to ensure behavioral diversity and improve exploration performance:

- P1: First, the exploration bonus require correlating with states; otherwise, they would appear as white noises, which does not help exploration as shown by [Plappert *et al.*, 2017].
- P2: Second, to ensure the bonus is fully observable and actionable by the policy, the bonus function must allow the policy to condition on it (see Figure 10).
- P3: Last, the exploration bonus should incorporate value diversity between states to encourage diverse and purposeful *behavior* even in sparse reward setting.

To fulfill these principles, as illustrated in Figure 2, we assert that encouraging the agent to effectively traverse the compact latent space will facilitate exploration of the true state space. We implement the bonus function (denoted as b) as

$$b(s,z) = f(s,s') \cdot z \tag{6}$$

where $f(s,s') = d^{\text{BILE}}_\phi(s,s') =< \phi(s) - \phi(s') >$ and $z \in \mathcal{Z}$ is a random sampled vector from the latent space (the analysis on z will be elaborated in Section 4 and Appendix B.

As Equation (6) satisfies **P1**, we address **P2** by augmenting the input to the policy with the latent vector z, enabling the latent-conditioned policy $\pi(s, z)$ to be aware of both the state and the random latent variable that factorizes the exploration bonus function. For **P3**, recalling Theorem 3, we set $d_\phi^{\text{BILE}}(s)$ as the state encoder, which encodes the value difference between states, encouraging the agent to visit novel states with higher value diversity even in sparse reward setting.

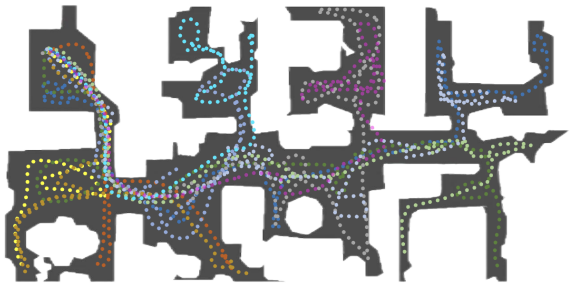


Figure 3: Visualization of trajectories from a BILE agent midway through training in an indoor environment, where each color denotes a distinct trajectory under different z.

As shown in Figure 3 , changing the latent vector z in BILE leads to diverse trajectories across all rooms. Since the latent-conditioned policy $\pi(s, z)$ must maximize the reward in Equation (6) for all randomly sampled vectors z, the agent is able to explore the entire latent space. Consequently, the agent is encouraged to produce diverse behavior while maintaining sufficient exploration over the state space.

Based on Equation (4), we denote the state encoder by $d_\phi^{\text{BILE}} : \mathcal{S} \to \mathbb{R}^n$, the probability transition dynamic model is parameterized by η, and draw batches of state pairs, and minimize the mean square error:

$$\mathcal{L}(\phi) = \mathbb{E}[(d_\phi^{\text{BILE}}(s_i, s_j) - |r^\pi_{s_i} - r^\pi_{s_j}| - \frac{\alpha}{2} \sum_{s_i, s_j} |r^\pi_{\text{PE}}(s)| - \gamma \mathbb{E}_{s'_i \sim P^\pi_\eta(\cdot|s_i), s'_j \sim P^\pi_\eta(\cdot|s_j)} (d_{\hat{\phi}}^{\text{BILE}}(s'_i, s'_j))^2] \tag{7}$$

where $\hat{\phi}$ is a copy of parameters for the target network. We outline the full training procedure in Algorithm 1.

Algorithm 1 BILE

Initialize policy π_θ, probability transition dynamic model $P_\eta(\cdot|s)$ and behavioral metric-based encoder d_ϕ^{BILE}
while *not converged* **do**
  Sample the latent vector z from distribution $P(\mathcal{Z})$
  for $t = 1$ to MAX_STEP_PER_EPISODE **do**
    Sample action $a_t \sim \pi_\theta(\cdot \mid s_t, z)$ and reach s_{t+1}
    Record transition in the buffer $\mathcal{D}$
    Compute bonus: $r'_t = r_t + b_t(s, z)$
    Train policy $\pi_\theta(s, z)$ via policy gradient
    Sample and permute mini batch B from the buffer
    Update metric encoder $\mathbb{E}_B[\mathcal{L}(\phi)]$ ▷Equation (7)
    Update dynamic models by minimizing the prediction error $r^\pi_{\text{PE}}(s)$
  end for
end while

4 Experiment

To evaluate the performance of BILE, we conduct comprehensive experiments on various settings of different environments to assess the effectiveness of our algorithm. The project site is https://sites.google.com/view/ijcai25bile.

Baselines. We compare BILE with the following baselines: **(1) ICM** [Pathak *et al.*, 2017]: The widely adopted curiosity-driven exploration method. **(2) RND** [Burda *et al.*, 2018b]: A representative baseline from the family of bonus-based exploration methods. **(3) RIDE** [Raileanu *et al.*, 2020]: An exploration method that constructs the exploration bonus in the latent space, modeled as the state difference within inverse and forward dynamics representation spaces. **(5) EME** [Wang *et al.*, 2024]: A competitive dynamic bonus-based method designed for high-dimensional environments, excelling in challenging exploration tasks. **(6) RLE** [Mahankali *et al.*, 2024]: A recent exploration method that leverages randomized rewards in the latent space. **(7) LIBERTY** [Wang *et al.*, 2023]: An approach utilizing a bonus based on state differences evaluated via the bisimulation metric. **(8) PPO** [Schulman *et al.*, 2017]: the standard RL benchmark algorithm.

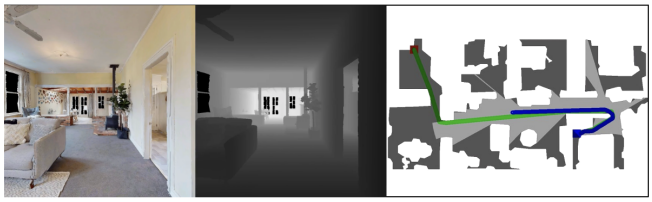


Figure 4: Snapshot of the real indoor Habitat environment

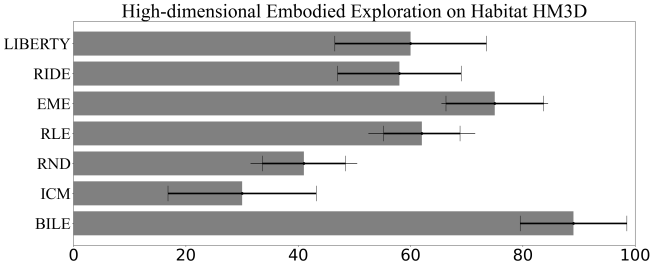


Figure 5: Averaged explored portion of map (%)

4.1 Real Indoor Habitat Environment

Habitat is a platform for embodied AI research which provides an interface for agents to navigate and act in photorealistic simulations of real indoor environments. As shown in Figure 4, at each episode, the agent finds itself in a different indoor space. Full details can be found in Appendix C.

High-dimensional Exploration. We assess the exploration capability of different algorithms by measuring how much of the environment is revealed by the agent's line of sight over

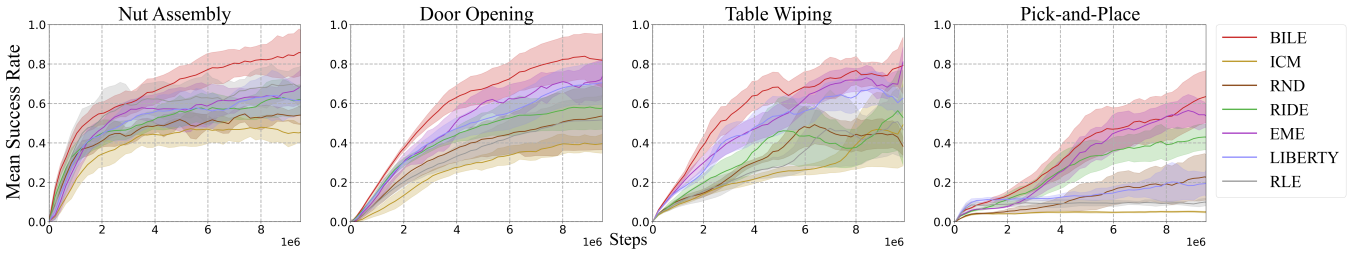


Figure 6: Comparison results for various exploration tasks from Robosuite. The solid lines and shaded areas in the plots represent the mean values and the standard errors, respectively, over five different seeds.

the course of an episode. Quantitative results are presented in Figure 5, where BILE reveals significantly more of the maps than any other method. The heatmaps of trajectories for all these methods can be found in Appendix B, which clearly demonstrates that BILE explores a larger portion of the space compared to other methods. These results provide strong evidence of BILE's scalability to high-dimensional observations and reinforce its practical applicability.

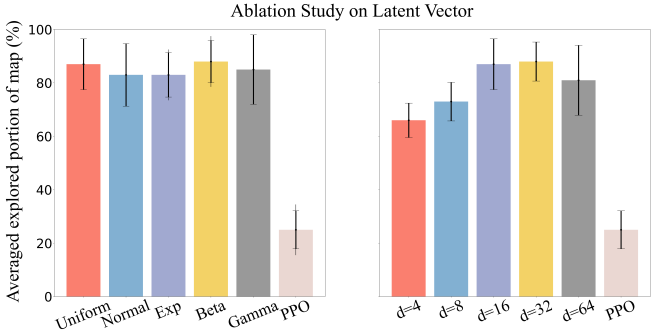


Figure 7: Ablation experiments on impact of latent vector

The Impact of Latent Vector. To evaluate the impact of z on the induced policy's behaviors. we sample different latent vector z and rollout trajectories with the policy conditioning on those latent vectors, plotting each trajectory in a different color in Figure 3, indicating that altering the latent vector z can produce diverse trajectories. We also investigate the impact of different latent vector distributions on BILE's performance and how sensitive BILE is to the dimension of the latent vector. Our study involves training BILE with various distributions, including Uniform, Normal, Exp (Exponential), Beta, and Gamma distributions. The dimension of the representation, or the latent vector, is an important hyperparameter. We test BILE with $d \in 4, 8, 16, 32, 64$, where $d = 16$ is the default setting in Habitat environments. As shown in Figure 7, BILE outperforms the baseline PPO across different latent vector distributions, suggesting that BILE's performance is robust to the choice of distributions. However, the performance of $d = 4$ and $d = 8$ lags behind, while $d = 16$ and $d = 32$ yield comparable results. When the dimension increases further, performance declines. Overall, BILE's performance with different dimensions remains superior to the baseline method PPO, indicating that BILE's performance is not highly sensitive to the choice of dimension.

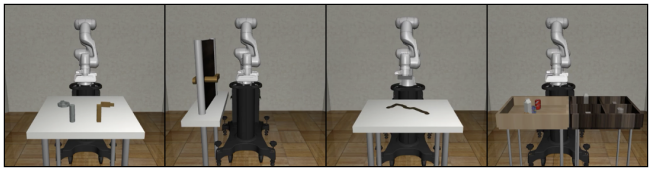


Figure 8: Continuous control tasks from Robosuite: Nut Assembly, Door Opening, Table Wiping and Pick-and-Place.

4.2 Robotic Continuous Control

Overall Performance of Continuous Control. In our continuous control experiments conducted on the realistic robotic platform [Zhu *et al.*, 2020], we evaluate agents across a variety of challenging tasks, including Nut Assembly, Door Opening, Table Wiping, and Pick-and-Place, as illustrated in Figure 8. The overall results, shown in Figure 6, demonstrate that our method consistently outperforms the baselines, highlighting its superiority in handling continuous control tasks. Among the baselines, EME achieves the second-best performance across all four tasks, underscoring the advantages of using a dynamic bonus. In contrast, episodic count-based methods such as RIDE and randomized bonus methods like RLE perform less effectively, as episodic counts and randomized bonuses lose their effectiveness in high-dimensional environments. Similarly, for the bisimulation metric-based method LIBERTY, performance declines in high-dimensional environments due to the increasingly subtle state differences. Curiosity-driven approaches such as ICM and RND struggle with insufficient exploration, leading to bad performance.

The Importance of Latent Vector Conditioning. As stated in P2 of Section 3.1, we emphasized the necessity for the policy to be conditioned on the latent vector. In order to verify the statement, we compare BILE with and without latent-conditioned policy network. As depicted in Figure 10, the performance of BILE exhibits a decline in the Pick-and-Place task (full results in Appendix B). The absence of latent vector conditioning results in limited behavioral variability in the policy network, which leads to failures in more challenging exploration tasks that necessitate a diverse behavior.

4.3 Sparse Reward Minigrid Environment

To evaluate BILE in scenarios with sparse rewards and discrete actions, we test our method on hard exploration tasks in the MiniGrid [Chevalier-Boisvert *et al.*, 2024] benchmark.

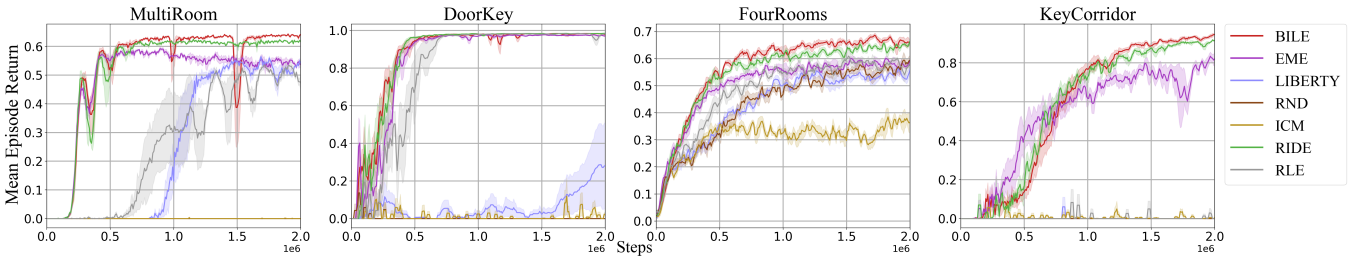


Figure 9: Performance on sparse reward Minigrid environments. The solid line and shaded regions represent the mean and standard deviation respectively, across three runs.

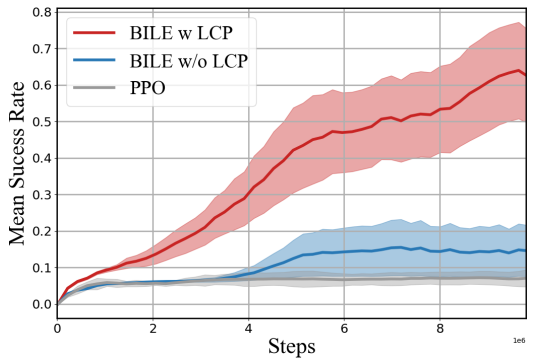


Figure 10: Learning curves of BILE with (w LCP) and without (w/o LCP) the Latent vector Conditioned Policy (LCP).

This benchmark includes grid-world exploration games with varying room layouts, interactive objects, and goals. In these environments, the agent must learn a specific sequence of actions to reach the final goal while operating within a limited view size. Valid actions include picking up a key, unlocking a door, unpacking a box, and moving objects, with no extrinsic reward provided until the goal is reached. We select four representative settings from MiniGrid: FourRooms, MultiRoom-N7, DoorKey-16x16, and KeyCorridorS6R3. Details on the specific environmental settings are provided in Appendix C. BILE successfully solves all hard-level exploration environments, achieving the best performance across all settings.

Noisy TV Problem. In addition to the standard tasks, we assess the model's robustness to environmental stochasticity by incorporating a manually-created Noisy TV setting introduced in [Raileanu *et al.*, 2020]. We also test the performance of our method as the room size and number of rooms increase. Results are detailed in Appendix B, showing that BILE consistently outperforms other baselines, further demonstrating the effectiveness and scalability of our approach.

5 Related Work

Exploration remains a long-standing problem in RL. Common approaches include ϵ-greedy [Sutton and Barto, 2018], count-based exploration [Bellemare *et al.*, 2016b; Ostrovski *et al.*, 2017a; Zhao *et al.*, 2024], and curiosity-based exploration [Stanton and Clune, 2016; Stanton and Clune, 2018; Burda *et al.*, 2018a]. Several exploration strategies use a dynamics model to provide intrinsic rewards [Pathak *et al.*, 2017; Burda *et al.*, 2018b; Houthooft *et al.*, 2016; Pathak *et al.*, 2019]. Latent variable dynamics have also been studied for exploration [Bai *et al.*, 2021; Tao *et al.*, 2020; Seo *et al.*, 2021; Raileanu *et al.*, 2020; Yang *et al.*, 2024]. Maximum entropy in the state representation has been used for exploration, through random encoders, in RE3 [Seo *et al.*, 2021], and prototypical representations [Yarats *et al.*, 2021]. Attention-based intrinsic reward is used to improve communication in traffic flow control [Yang *et al.*, 2023b; Yang *et al.*, 2023a]. Alternative approaches to modelling the environment's dynamics are based on pseudo-counts [Bellemare *et al.*, 2016a; Ostrovski *et al.*, 2017b; Tang *et al.*, 2017], which use density estimations techniques to explore less seen areas of the environment. Some methods combine model-based intrinsic motivation with pseudo-counts, such as RIDE [Raileanu *et al.*, 2020], which rewards the agent with for transitions that have an impact on the state representation, and NGU [Badia *et al.*, 2020b] and agent57 [Badia *et al.*, 2020a], which modulates a pseudo-count bonus with the intrinsic rewards provided by RND. Another line of research try to model the difference between states as bonuses to incentive exploration [Wang *et al.*, 2023; Zhang *et al.*, 2021; Henaff *et al.*, 2022]. Other unsupervised learning approaches [Eysenbach *et al.*, 2018; Park *et al.*, 2023; Park *et al.*, 2022] propose unsupervised frameworks to induce diverse skills or long-horizon behaviors. Recent breakthroughs concerning exploration in RL have also focused on using the learned environment dynamics to plan to explore [Shyam *et al.*, 2019; Ratzlaff *et al.*, 2020; Hafner *et al.*, 2019], where they use imaginary rollouts from their dynamics models to plan exploratory behaviors.

6 Conclusion

In this work, we identify the limitations of previous exploration methods, particularly in environments with high-dimensional state spaces. To address these challenges, we propose a novel exploration strategy that captures diverse behaviors by maximally exploring the compact latent behavioral metric space. We introduce a new behavioral metric that facilitates efficient and robust training of the state encoder, supported by theoretical guarantees. Additionally, we encourage agents to explore the latent space by providing exploration bonuses based on randomized latent vectors. Extensive experiments across various challenging tasks, including continuous control, minigrid games, and realistic environments, demonstrate the effectiveness and scalability of our method.

Acknowledgments

This work was supported by the Science and Technology Development Fund Macau SAR (0003/2023/RIC, 0052/2023/RIA1, 0031/2022/A, 001/2024/SKL for SKL-IOTSC) , Shenzhen-Hong Kong-Macau Science and Technology Program Category C (SGDX20230821095159012), NSF of China 62402325 and the Research Foundation of Shenzhen Polytechnic University under Grant 6022310014K and 6022312054K. This work was performed in part at SICC which is supported by SKL-IOTSC, University of Macau.

References

[Badia *et al.*, 2020a] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. Agent57: Outperforming the atari human benchmark. In *International conference on machine learning*, pages 507–517. PMLR, 2020.

[Badia *et al.*, 2020b] Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andew Bolt, et al. Never give up: Learning directed exploration strategies. *arXiv preprint arXiv:2002.06038*, 2020.

[Bai *et al.*, 2021] Chenjia Bai, Peng Liu, Kaiyu Liu, Lingxiao Wang, Yingnan Zhao, Lei Han, and Zhaoran Wang. Variational dynamic for self-supervised exploration in deep reinforcement learning. *IEEE Transactions on neural networks and learning systems*, 34(8):4776–4790, 2021.

[Bellemare *et al.*, 2016a] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

[Bellemare *et al.*, 2016b] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

[Burda *et al.*, 2018a] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

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

[Castro, 2020] Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10069–10076, 2020.

[Chevalier-Boisvert *et al.*, 2024] Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

[Eysenbach *et al.*, 2018] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

[Ferns *et al.*, 2011] Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011.

[Hafner *et al.*, 2019] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[Henaff *et al.*, 2022] Mikael Henaff, Roberta Raileanu, Minqi Jiang, and Tim Rocktäschel. Exploration via elliptical episodic bonuses. *Advances in Neural Information Processing Systems*, 35:37631–37646, 2022.

[Houthooft *et al.*, 2016] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

[Mahankali *et al.*, 2024] Srinath Mahankali, Zhang-Wei Hong, Ayush Sekhari, Alexander Rakhlin, and Pulkit Agrawal. Random latent exploration for deep reinforcement learning. *arXiv preprint arXiv:2407.13755*, 2024.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[Ostrovski *et al.*, 2017a] Georg Ostrovski, Marc G Bellemare, Aäron Oord, and Rémi Munos. Count-based exploration with neural density models. In *International conference on machine learning*, pages 2721–2730. PMLR, 2017.

[Ostrovski *et al.*, 2017b] Georg Ostrovski, Marc G Bellemare, Aäron Oord, and Rémi Munos. Count-based exploration with neural density models. In *International conference on machine learning*, pages 2721–2730. PMLR, 2017.

[Park *et al.*, 2022] Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, and Gunhee Kim. Lipschitz-constrained unsupervised skill discovery. In *International Conference on Learning Representations*, 2022.

[Park *et al.*, 2023] Seohong Park, Oleh Rybkin, and Sergey Levine. Metra: Scalable unsupervised rl with metric-aware abstraction. *arXiv preprint arXiv:2310.08887*, 2023.

[Pathak *et al.*, 2017] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven

exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

[Pathak *et al.*, 2019] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International conference on machine learning*, pages 5062–5071. PMLR, 2019.

[Plappert *et al.*, 2017] Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. Parameter space noise for exploration. *arXiv preprint arXiv:1706.01905*, 2017.

[Raileanu *et al.*, 2020] Roberta Raileanu, Tim Rocktäschel, et al. Ride: Rewarding impact-driven exploration for procedurally-generated environments. *arXiv preprint arXiv:2002.12292*, 2020.

[Ratzlaff *et al.*, 2020] Neale Ratzlaff, Qinxun Bai, Li Fuxin, and Wei Xu. Implicit generative modeling for efficient exploration. In *International Conference on Machine Learning*, pages 7985–7995. PMLR, 2020.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017.

[Seo *et al.*, 2021] Younggyo Seo, Lili Chen, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. State entropy maximization with random encoders for efficient exploration. In *International Conference on Machine Learning*, pages 9443–9454. PMLR, 2021.

[Shyam *et al.*, 2019] Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. In *International conference on machine learning*, pages 5779–5788. PMLR, 2019.

[Stanton and Clune, 2016] Christopher Stanton and Jeff Clune. Curiosity search: producing generalists by encouraging individuals to continually explore and acquire skills throughout their lifetime. *PloS one*, 11(9):e0162235, 2016.

[Stanton and Clune, 2018] Christopher Stanton and Jeff Clune. Deep curiosity search: Intra-life exploration improves performance on challenging deep reinforcement learning problems. *arXiv preprint arXiv:1806.00553*, 2018.

[Sutton and Barto, 2018] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[Tang *et al.*, 2017] 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. *Advances in neural information processing systems*, 30, 2017.

[Tao *et al.*, 2020] Ruo Yu Tao, Vincent François-Lavet, and Joelle Pineau. Novelty search in representational space for sample efficient exploration. *Advances in Neural Information Processing Systems*, 33:8114–8126, 2020.

[Wang *et al.*, 2023] Yiming Wang, Ming Yang, Renzhi Dong, Binbin Sun, Furui Liu, et al. Efficient potential-based exploration in reinforcement learning using inverse dynamic bisimulation metric. *Advances in Neural Information Processing Systems*, 36:38786–38797, 2023.

[Wang *et al.*, 2024] Yiming Wang, Kaiyan Zhao, Furui Liu, et al. Rethinking exploration in reinforcement learning with effective metric-based exploration bonus. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Yang *et al.*, 2023a] Ming Yang, Renzhi Dong, Yiming Wang, Furui Liu, Yali Du, Mingliang Zhou, and Leong Hou U. Tiecomm: Learning a hierarchical communication topology based on tie theory. In *International Conference on Database Systems for Advanced Applications*, pages 604–613. Springer, 2023.

[Yang *et al.*, 2023b] Ming Yang, Yiming Wang, Yang Yu, Mingliang Zhou, et al. Mixlight: Mixed-agent cooperative reinforcement learning for traffic light control. *IEEE Transactions on Industrial Informatics*, 20(2):2653–2661, 2023.

[Yang *et al.*, 2024] Ming Yang, Kaiyan Zhao, Yiming Wang, Renzhi Dong, Yali Du, Furui Liu, Mingliang Zhou, and Leong Hou U. Team-wise effective communication in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 38(2):36, 2024.

[Yarats *et al.*, 2021] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, pages 11920–11931. PMLR, 2021.

[Zang *et al.*, 2023] Hongyu Zang, Xin Li, Leiji Zhang, Yang Liu, Baigui Sun, Riashat Islam, Remi Tachet des Combes, and Romain Laroche. Understanding and addressing the pitfalls of bisimulation-based representations in offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36:28311–28340, 2023.

[Zhang *et al.*, 2020] Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *ArXiv*, abs/2006.10742, 2020.

[Zhang *et al.*, 2021] Tianjun Zhang, Huazhe Xu, Xiaolong Wang, Yi Wu, Kurt Keutzer, Joseph E Gonzalez, and Yuandong Tian. Noveld: A simple yet effective exploration criterion. *Advances in Neural Information Processing Systems*, 34:25217–25230, 2021.

[Zhao *et al.*, 2024] Kaiyan Zhao, Yiming Wang, Yuyang Chen, Yan Li, Xiaoguang Niu, et al. Efficient diversity-based experience replay for deep reinforcement learning. *arXiv preprint arXiv:2410.20487*, 2024.

[Zhu *et al.*, 2020] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.